MOVING BEYOND NON-INFORMATIVE PRIOR DISTRIBUTIONS: ACHIEVING THE FULL POTENTIAL OF BAYESIAN METHODS FOR PSYCHOLOGICAL RESEARCH

EDITED BY: Christoph Koenig, Sarah Depaoli, Haiyan Liu
and Rens Van De Schoot

PUBLISHED IN: Frontiers in Psychology

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88974-214-1
DOI 10.3389/978-2-88974-214-1

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

MOVING BEYOND NON-INFORMATIVE PRIOR DISTRIBUTIONS: ACHIEVING THE FULL POTENTIAL OF BAYESIAN METHODS FOR PSYCHOLOGICAL RESEARCH

Topic Editors:
Christoph Koenig, Goethe University Frankfurt, Germany
Sarah Depaoli, University of California, United States
Haiyan Liu, University of California, United States
Rens Van De Schoot, Utrecht University, Netherlands

Citation: Koenig, C., Depaoli, S., Liu, H., Van De Schoot, R., eds. (2022). Moving Beyond Non-Informative Prior Distributions: Achieving the Full Potential of Bayesian Methods for Psychological Research. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88974-214-1

Table of Contents

frontiers in Psychology

EDITORIAL
published: 09 December 2021
doi: 10.3389/fpsyg.2021.809719

Editorial: Moving Beyond Non-informative Prior Distributions: Achieving the Full Potential of Bayesian Methods for Psychological Research

Christoph Koenig[1], Sarah Depaoli[2], Haiyan Liu[2] and Rens van de Schoot[3]*

[1] Department of Educational Psychology, Goethe University Frankfurt am Main, Frankfurt, Germany, [2] Department of Psychological Sciences, University of California, Merced, Merced, CA, United States, [3] Department of Methodology and Statistics, Utrecht University, Utrecht, Netherlands

Keywords: Bayesian statistics, Bayesian modeling and inference, prior distributions, informative prior distributions, cumulative science, psychological research

Editorial on the Research Topic

Moving Beyond Non-informative Prior Distributions: Achieving the Full Potential of Bayesian Methods for Psychological Research

OPEN ACCESS

Edited and reviewed by:
Alexander Robitzsch,
IPN - Leibniz Institute for Science and Mathematics Education, Germany

****Correspondence:***
Christoph Koenig
koenig@psych.uni-frankfurt.de

Specialty section:
This article was submitted to Quantitative Psychology and Measurement,
a section of the journal Frontiers in Psychology

Received: *05 November 2021*
Accepted: *22 November 2021*
Published: *09 December 2021*

Citation:
Koenig C, Depaoli S, Liu H and van de Schoot R (2021) Editorial: Moving Beyond Non-informative Prior Distributions: Achieving the Full Potential of Bayesian Methods for Psychological Research. Front. Psychol. 12:809719. doi: 10.3389/fpsyg.2021.809719

Over the last two decades, Bayesian statistics have been established as an alternative to the well-known frequentist approaches primarily based on maximum likelihood (ML) estimation (van de Schoot et al., 2017; Koenig and van de Schoot, 2018). With the possibility of incorporating background knowledge into new analyses, Bayesian methods can potentially transform psychological research into a truly cumulative scientific discipline. However, the primary tool to achieve this, namely informative prior distributions, remains a seemingly elusive concept, especially for novice users. Reasons include, but are not limited to, the frequent criticism regarding their alleged subjective nature and a lack of knowledge about methods to formalize background knowledge (Goldstein, 2006; Vanpaemel, 2011). These two aspects are the primary point of departure for the twelve articles in this Special Issue.

The first set of two articles provides interested readers with means to **comprehend** the nature and potential impact of prior distributions in general. As Depaoli et al. (p. 3) state, "Understanding the impact of priors, and then making subsequent decisions about these priors, is perhaps the trickiest element of implementing Bayesian methods." Consequently, their tutorial paper presents an interactive Shiny app that enables novice and experienced users of Bayesian statistics to investigate and determine the impact of their specified prior distributions on model results. Arts et al. examine the impact of different prior distribution specifications for the variance parameter in a Bayesian approximate measurement invariance with alignment optimization (e.g., van de Schoot et al., 2013). The authors illustrate visually how the prior specification for the variance parameter affects the rank ordering of 30 countries in a large-scale assessment of the latent construct" willingness to sacrifice the environment." Visualizing different outcomes aids in understanding the effect of various prior specifications on model results.

The second set of articles aims to **convince** interested readers of the benefits and advantages of weakly and fully informative prior distributions compared to their non-informative counterparts and frequentist ML estimation. The five articles illustrate these benefits across a wide range of statistical models, with a particular focus on small-sample situations. Tong and Ke show the benefits and advantages of using weakly and fully informative priors for the precision parameter in

Bayesian non-parametric growth curve models. Their simulation demonstrates that using weakly or fully informative priors aids model convergence and the accuracy of the precision parameter of the Dirichlet process. This conclusion is essential, as previous research showed that the precision parameter is crucial for obtaining good results.

Similarly, Zyphur et al. show that using weakly or fully informative priors also aids model convergence and parameter accuracy for cross-lagged panel models. They concluded that using such priors increases model parsimony, estimate stability, and thus the general trustworthiness of results, compared to results obtained with ML estimation. When dealing with small samples, the role of Bayesian prior distributions becomes even more crucial for model convergence and parameter accuracy. Smid and Winter present a tutorial discussing the dangers and pitfalls of using default priors implemented in software for Bayesian structural equation models. They introduce an interactive Shiny app, where users can investigate the impact of various priors on model estimates, depending on sample size. Lüdtke et al. examine the stability of estimates across different Bayesian estimators in small-sample confirmatory factor analysis. The results show that estimates based on the posterior mean (EAP) produced more accurate estimates. Parameter estimates can be further stabilized using the four-parameter beta distribution for loadings and factor correlations (e.g., Merkle and Rosseel, 2018). The benefits of using this prior distribution in the weakly informative specification are present even when prior distributions are mildly misspecified. Another specification of weakly informative priors is illustrated in Zitzmann et al. In the context of multilevel latent variable models, they describe two strategies (direct and indirect; Zitzmann et al., 2015) to specify weakly informative priors for the group-level slope parameter. Their simulation results show that introducing additional information via these priors stabilizes the model and provides more accurate parameter estimates in small-sample situations.

Finally, the third set of articles focuses on different approaches to formalize background knowledge objectively. The ultimate aim is to build **confidence** in the specification and use of informative prior distributions. In this regard, Veen et al. focus on expert knowledge for specifying informative prior distributions. In their paper, they illustrate how the five-step method (Veen et al., 2017) is used for prior elicitation for the parameters of a latent growth curve model. They show how to aggregate expert knowledge and specify appropriate densities to be used in a Bayesian analysis. Moreover, they compare the prior densities with posterior densities from traditionally collected data and guide how to set up procedures for appropriate expert elicitation.

Van de Schoot et al. provide another example of eliciting expert knowledge and using it to specify informative prior distributions. They also use lesser-known Bayesian methods, such as tests for prior-data conflicts (Box, 1980), a scoring algorithm to incentivize truthful responses (John et al., 2012), and Bayes factors for replication success (Verhagen and Wagenmakers, 2014), to investigate the prevalence of questionable research practices among Dutch and Belgian early career researchers. These articles are complemented by three illustrations focusing on more quantitative ways to formalize background knowledge. In this regard, Tran et al. focus on formalizing background knowledge with systematic parameter reviews. These reviews consist of a systematic literature search for studies containing estimates of relevant model parameters and necessary transformations to make the parameter comparable across studies. They illustrate how to specify informative prior distributions based on these synthesized parameter estimates in the context of the Diffusion Decision Model (DDM; Ratcliff and McKoon, 2008). The two remaining studies extend this approach and illustrate ways to consider the similarity of the available background knowledge and demonstrate how to apply the necessary weighting of the contributions of the individual studies to the informative prior distribution. In this regard, Schulz et al. implement a distribution-based approach. In the context of mother-adolescent interaction behavior, they illustrate three methods for pooling results from previously conducted studies to specify informative prior distributions. Moreover, they show how to use expert knowledge to weigh the contribution of each previously conducted study and how to use these weights in a power prior approach (Carvalho and Ibrahim, 2021).

Lastly, Koenig illustrates how to specify informative prior distributions using random-effects meta-analytic models. In the context of Bayesian multiple regression models, they present a novel method based on propensity-score and mixed-effects meta-analytic approaches (Tipton, 2014; Cheung, 2015) for quantifying the similarity of background knowledge. Moreover, they illustrate how to use this similarity measure to specify similarity-weighted informative prior distributions, an evidence-based informative prior also based on the power prior concept (Kaplan and Depaoli, 2013; Ibrahim et al., 2015).

To enhance reproducibility, crucial for Bayesian papers with informative priors (van de Schoot et al., 2021), each article in this Special Issue is accompanied by comprehensive supplementary material, including annotated code, which provides researchers with the means to apply the models and methods directly to their Bayesian analyses. In conclusion, we hope that this Special Issue enables novice and more experienced Bayesian researchers to move beyond non-informative prior distributions and unlock the full potential of Bayesian methods for psychological research.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

REFERENCES

Box, G. E. (1980). Sampling and Bayes' inference in scientific modelling and robustness. *J. R. Statist. Soc. A* 143, 383–430. doi: 10.2307/2982063

Carvalho, L. M., and Ibrahim, J. G. (2021). On the normalized power prior. *Stat. Med.* 40, 5251–5275. doi: 10.1002/sim.9124

Cheung, M. W.-L. (2015). metaSEM: an R package for meta-analysis using structural equation modeling. *Front. Psychol.* 5, 1521. doi: 10.3389/fpsyg.2014.01521

Goldstein, M. (2006). Subjective Bayesian analysis: principles and practice. *Bayesian Anal.* 1, 403–420. doi: 10.1214/06-BA116

Ibrahim, J., Chen, M.-H., Gwon, Y., and Chen, F. (2015). The power prior: theory and applications. *Stat. Med.* 34, 3724–3749. doi: 10.1002/sim.6728

John, L. K., Loewenstein, G., and Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychol. Sci.* 23, 524–532. doi: 10.1177/0956797611430953

Kaplan, D., and Depaoli, S. (2013). "Bayesian statistical methods," in *The Oxford Handbook of Quantitative Methods, Vol. 1: Foundations*, ed T. D. Little (Oxford, UK: Oxford University Press), 407–437. doi: 10.1093/oxfordhb/9780199934874.013.0020

Koenig, C., and van de Schoot, R. (2018). Bayesian statistics in educational research: a look at the current state of affairs. *Educ. Rev.* 70, 486–509. doi: 10.1080/00131911.2017.1350636

Merkle, E. C., and Rosseel, Y. (2018). blavaan: Bayesian structural equation models via parameter expansion. *J. Stat. Softw.* 85, 1–30. doi: 10.18637/jss.v085.i04

Ratcliff, R., and McKoon, G. (2008). The diffusion decision model: theory and data for two-choice decision tasks. *Neural Comput.* 20, 873–922. doi: 10.1162/neco.2008.12-06-420

Tipton, E. (2014). How generalizable is your experiment? An index for comparing experimental samples and populations. *J. Educ. Behav. Stat.* 39, 478–501. doi: 10.3102/1076998614558486

van de Schoot, R., Depaoli, S., King, R., Kramer, B., Märtens, K., Tadesse, M. G., et al. (2021). Bayesian statistics and modelling. *Nat. Rev. Methods Prim.* 1, 1. doi: 10.1038/s43586-020-00001-2

van de Schoot, R., Kluytmans, A., Tummers, L., Lugtig, P., Hox, J., and Muthe'n, B. (2013). Facing off with Scylla and Charybdis: a comparison of scalar, partial, and the novel possibility of approximate measurement invariance. *Front. Psychol.* 4, 770. doi: 10.3389/fpsyg.2013.00770

van de Schoot, R., Winter, S. D., Ryan, O., Zondervan-Zwijnenburg, M., and Depaoli, S. (2017). A systematic review of Bayesian articles in psychology: the last 25 years. *Psychol. Methods* 22, 217–239. doi: 10.1037/met0000100

Vanpaemel, W. (2011). Constructing informative model priors using hierarchical methods. *J. Math. Psychol.* 55, 106–117. doi: 10.1016/j.jmp.2010.08.005

Veen, D., Stoel, D., Zondervan-Zwijnenburg, M., and van de Schoot, R. (2017). Proposal for a five-step method to elicit expert judgement. *Front. Psychol.* 8, 2110. doi: 10.3389/fpsyg.2017.02110

Verhagen, J., and Wagenmakers, E.-J. (2014). Bayesian tests to quantify the result of a replication attempt. *J. Exp. Psychol. Gen.* 143, 1457. doi: 10.1037/a0036731

Zitzmann, S., Lüdtke, O., and Robitzsch, A. (2015). A Bayesian approach to more stable estimates of group-level effects in contextual studies. *Multivariate Behav. Res.* 50, 688–705. doi: 10.1080/00273171.2015.1090899

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Expert Elicitation for Latent Growth Curve Models: The Case of Posttraumatic Stress Symptoms Development in Children With Burn Injuries

Duco Veen[1], Marthe R. Egberts[2], Nancy E. E. van Loey[2,3] and Rens van de Schoot[1,4]*

[1] Department of Methodology and Statistics, Utrecht University, Utrecht, Netherlands, [2] Department of Clinical Psychology, Utrecht University, Utrecht, Netherlands, [3] Association of Dutch Burn Centres, Beverwijk, Netherlands, [4] Optentia Research Program, North-West University, Potchefstroom, South Africa

OPEN ACCESS

Edited by:
*Stéphane Bouchard,
University of Quebec in Outaouais,
Canada*

Reviewed by:
*Philip Hans Franses,
Erasmus University Rotterdam,
Netherlands
Richard S. John,
University of Southern California,
Los Angeles, United States*

****Correspondence:***
*Duco Veen
ducoveen@gmail.com*

Specialty section:
*This article was submitted to
Quantitative Psychology
and Measurement,
a section of the journal
Frontiers in Psychology*

Received: *06 February 2020*
Accepted: *07 May 2020*
Published: *18 June 2020*

Citation:
*Veen D, Egberts MR,
van Loey NEE and van de Schoot R
(2020) Expert Elicitation for Latent
Growth Curve Models: The Case
of Posttraumatic Stress Symptoms
Development in Children With Burn
Injuries. Front. Psychol. 11:1197.
doi: 10.3389/fpsyg.2020.01197*

Experts provide an alternative source of information to classical data collection methods such as surveys. They can provide additional insight into problems, supplement existing data, or provide insights when classical data collection is troublesome. In this paper, we explore the (dis)similarities between expert judgments and data collected by traditional data collection methods regarding the development of posttraumatic stress symptoms (PTSSs) in children with burn injuries. By means of an elicitation procedure, the experts' domain expertise is formalized and represented in the form of probability distributions. The method is used to obtain beliefs from 14 experts, including nurses and psychologists. Those beliefs are contrasted with questionnaire data collected on the same issue. The individual and aggregated expert judgments are contrasted with the questionnaire data by means of Kullback–Leibler divergences. The aggregated judgments of the group that mainly includes psychologists resemble the questionnaire data more than almost all of the individual expert judgments.

Keywords: Bayesian statistics, elicitation, expert judgment, expert knowledge, Latent Growth Curve Model, prior, prior-data (dis)agreement

INTRODUCTION

Expert elicitation entails the extraction of information from experts and the translation of this information into a probabilistic representation. There are many reasons to elicit expert knowledge. In some cases, it is done to supplement existing data using priors that are informed by expert knowledge (van de Schoot et al., 2018). Alternatively, expert judgments allow for filling information gaps of certain data (Fischer et al., 2013; Dodd et al., 2017) or they can serve as a quality control for obtained data (Veen et al., 2018). Elicitation can also be used for forecasting purposes (Murphy and Winkler, 1974, 1984) or when there are no data available at all (Ho and Smith, 1997; Hald et al., 2016).

The use of expert knowledge is widespread across many disciplines. To give some examples, Dodd et al. (2017) elicited expert-based estimates for case-fatality ratios in HIV-positive children with tuberculosis who did not receive treatment; Barons et al. (2018) describe the use of expert judgments to create decision support systems with an example in food security; and Dewispelare et al. (1995) describe expert elicitation in relation to the long-term behavior of high-level nuclear waste repositories. For numerous other examples on elicitation practices, see for instance Chapter 10 of O'Hagan et al. (2006), listing applications in sales, medicine, nuclear industry, veterinary science, and many more. Other examples using a specific elicitation tool are given in Gosling (2018), while Cooke and Goossens (2008) describe a database of over 67,000 elicited judgments.

Recently, there is a growing interest in the use of expert elicitation in the social sciences. Where van de Schoot et al. (2017) only found two cases that reported the use of expert opinions to inform priors in 25 years of Bayesian statistics in psychology, this trend might slowly be changing. For instance, in their example related to a replication study in the field of psychology, Gronau et al. (2019) elicited expert judgments on effect sizes such that these could be used in informed Bayesian *t*-tests; Lek and van de Schoot (2018) elicited prior distribution from teachers concerning the math abilities of their students; and Zondervan-Zwijnenburg et al. (2017) elicited expert judgments on the correlation between cognitive potential and academic performance. Moreover, methods are being developed to facilitate expert elicitation in a flexible manner such that experts are guided in the elicitation process (Veen et al., 2017).

Whatever the reasons of the elicitation, the goal is to get an accurate representation of the experts' beliefs and associated (un)certainty, which enables the representation of the experts' domain knowledge in terms of a probability distribution. Overconfidence of experts is one of the crucial issues in expert elicitation (O'Hagan et al., 2006), resulting in elicited probability distributions with little uncertainty. In the seminal work of O'Hagan et al. (2006), feedback is named as the most natural way to improve the accuracy of elicited beliefs, with interactive software being almost essential for the effective use of feedback. This is corroborated by Goldstein and Rothschild (2014) who found that visual feedback can increase even laypeople's intuitions about probability distributions. Over a decade has passed since the advice by O'Hagan et al. (2006), and many have followed it. Elicitation software can be split into more general and more customized variations. Some more general frameworks are, for instance, ElicitN, which was developed by Fisher et al. (2012) for the elicitation of count data. Truong et al. (2013) made a web-based tool for the elicitation of variogram estimates which describe a degree of spatial dependence. The elicitator was developed for indirect elicitation, creating a scenario-based elicitation (James et al., 2010; Low-Choy et al., 2012). Morris et al. (2014) developed MATCH which is based on the R package SHELF (Oakley, 2019) and which is a very general elicitation tool that allows multiple elicitation methods to be used interactively to elicit single parameters. Garthwaite et al. (2013) developed an elicitation procedure for generalized linear and piecewise-linear models. Runge et al. (2013) developed one for seismic-hazard analysis and Elfadaly and Garthwaite (2017) for eliciting Dirichlet and Gaussian copula prior distributions. Sometimes, more customized software is developed for specific elicitation settings (e.g., Bojke et al., 2010; Haakma et al., 2014; Hampson et al., 2014, 2015). To sum up, the use of software, customized or not, to increase the accuracy of the elicited beliefs is now common practice.

In this paper, we present an elicitation methodology especially designed for eliciting parameters of a Latent Growth Curve Model (LGM) regarding the development of posttraumatic stress symptoms (PTSSs) in children with burn injuries. LGMs are commonly used to analyze longitudinal data, especially in the social sciences (e.g., Buist et al., 2002; Catts et al., 2008; Orth et al., 2012). These models include repeated measurements of observed variables and allow researchers to examine change or development over time in the construct of interest. For extensive explanations of LGMs, see Duncan and Duncan (2004), Little (2013), and Little et al. (2006). Because in Western high-income countries, the incidence of severe burn injuries in school-aged children and adolescents is relatively low and obtaining a large enough sample to estimate LGMs is challenging. Nevertheless, to gain knowledge on the development of PTSSs in this group of children, these types of models are favored over simpler models. Expert elicitation might provide an alternative to data collection for cases like our motivating example where traditional data are sparse or they might supplement such data.

The main aim of this paper is to compare domain expertise expressed by experts in an elicitation setting to data on the same topic collected by means of traditional data collection methods (Egberts et al., 2018). Comparing experts' domain knowledge to traditional data collection methods can provide unique insights into the topic of interest and the perception thereof. In the remainder of this paper, we first describe the methodology that is used to elicit the expert judgments. The methodology is an extension of the Five-Step Method (Veen et al., 2017) adapted to elicit multiple parameters. We elicit expert judgments from 14 experts, including nurses and psychologists working in the burn centers where data on PTSS in children were collected. Thereafter, we compare individual expert judgments to aggregated group-level expert judgments and data collected by means of traditional methods, followed by a reflection on the elicitation procedure. We conclude the paper with a *Discussion* section including recommendations for future research. All related materials for this study, including code and data, can be found on the Open Science Framework (OSF) website for this project at https://osf.io/y5evf/.

METHODS

In the first section, we describe the motivating example for this study. In the next section, we elaborate on the elicitation procedure and on software that has been developed. Finally, we describe the sample of experts ($N = 14$) participating in the elicitation study. The study received ethical approval from our internal ethics committee of the Faculty of Social and Behavioral Sciences of Utrecht University. The letter of approval can be found in the data archive on the OSF website for this project.

Motivating Example

The motivating example for this paper is the development of PTSS in children after a burn event. In a prospective study on child and parent adjustment after pediatric burns, data on these symptoms were collected in three Dutch and four Belgian burn centers. Children aged 8–18 years were eligible to participate in the study if they had been hospitalized for more than 24 h and if the percentage of total body surface area (TBSA) burned was at least 1%. A more detailed description of the overall study and sample can be found in Egberts et al. (2018). This sample consists of 100 children who reported on their symptoms of traumatic stress within the first month after the burn event (T1) and subsequently at 3 (T2) months post-burn. For the purpose of the current study, we also included the measurements obtained at 12 months (T3) post-burn. Children filled out the Children's Responses to Trauma Inventory (CRTI, revised version; Alisic et al., 2006). This measure assesses four symptom clusters of posttraumatic stress, including intrusion (e.g., repetitive, intrusive recollections of the trauma), avoidance (e.g., avoiding conversations of the event), arousal (e.g., difficulty concentrating), and other child-specific responses (e.g., feelings of guilt). Further details on this measure can be found in Alisic et al. (2011).

As the current study includes three measurements of PTSS at different time points, a straightforward model to analyze the development of PTSS is an LGM. **Figure 1** provides a visual representation of an LGM for this motivating example. The model is parameterized such that the latent intercept provides an estimate for PTSS in the first month after the burn event. The latent slope describes the change in PTSS at 1 year post-burn. Parameterizing the slope by year instead of per month is done to ease the reasoning in the elicitation procedure. Furthermore, the scale of the PTSS scores has been standardized for the data of the prospective study and for the elicitation study. The scores can fall between 0 and 100. A zero score means that none of the symptoms of any of the clusters of posttraumatic stress is present. A score of 100 means that all symptoms from all clusters are present to their maximum extent. A standardized cutoff value of 42 was used to indicate clinical relevance of symptoms and corresponds to the cutoff value provided in the CRTI manual. *Via* the OSF website for this project, supplementary materials can be found that describe the LGM analysis for these data, including assessment of the extent to which the LGM fits the data over the three time points.

Expert Elicitation

To optimally prepare the experts within the limited time that was allocated for each elicitation, a short introduction was presented by the researchers conducting the elicitation (DV and ME), hereafter named the facilitators. The facilitators presented the experts with a brief overview of what expert elicitation is, what it can be used for, and how to interpret the probability distributions that are used to represent their beliefs. Thereafter, to familiarize the experts with the elicitation procedure itself, an example elicitation for an unrelated topic was presented to the experts using the same elicitation tool. After the example elicitation, the facilitators introduced the specifics related to the motivating example and the actual elicitation. Experts were instructed to think of the same reference population as used in the questionnaire study (i.e., children hospitalized for at least 24 h in one of the three Dutch or four Belgian burn centers with a minimum of 1% TBSA burned). Moreover, the CRTI symptom clusters were introduced, including specific examples of symptoms assessed with this measure. In addition, the measurement scale and research question were introduced, and experts were invited to ask questions to clarify any part of the procedure. Once the experts stated that they were ready to continue with the elicitation, they were requested to sign the informed consent letter, which they received prior to the elicitation. If they agreed, they also agreed to the recording of the elicitation procedure. The experts were requested to reason aloud during the elicitation. The recordings were transcribed to provide additional insights into the elicitation procedure and to track possible differences between experts. The experts carried out the elicitation procedure using the software that is described next.

The software and procedure in this study were based on the Five-Step Method developed by Veen et al. (2017), with a slight adaptation to elicit multiple parameters instead of a single parameter. The Five-Step Method decomposes the elicitation process in multiple smaller steps, providing visual feedback at each stage of the elicitation procedure. By decomposing the elicitation task and providing visual feedback, the procedure aims to reduce bias, for instance from overconfidence. The software has been developed in the form of a Shiny web application (Chang et al., 2019). Using Shiny to develop elicitation tools is not uncommon, see, for instance, Hampson et al. (2014), Hampson et al. (2015), and the original Five-Step Method by Veen et al. (2017). In what follows, we describe the Five-Step Method as implemented for this specific study for each expert. Note that steps 3 and 4 were repeated for each parameter.

Step 1. Ten fictive individual PTSS trajectories were elicited for an LGM. These individual trajectories should be representative for the population. From these individual trajectories, we could deduce information on the point estimates for the average intercept and average slope parameters. This first step is called indirect elicitation because no statement is required directly concerning the parameters of interest. **Figure 2** provides a visual representation of step 1.

Step 2. Feedback was provided on the average trajectory that was based upon the 10 individual trajectories that the expert provided. The expert could accept this as the average trajectory and thereby accept point estimates for the average intercept and slope, or the expert could adjust his or her input in step 1. **Figure 3** provides a visual representation of step 2.

Step 3. The experts provided a reasonable lowerbound and upperbound for the point estimates of the group mean intercept and the group mean slope that were obtained using steps 1 and 2. The lowerbound and upperbound were used to determine the scale and shape of the probability distribution that was used to represent the experts' beliefs. This is called direct elicitation because the experts provided information directly related to the parameters of interest.

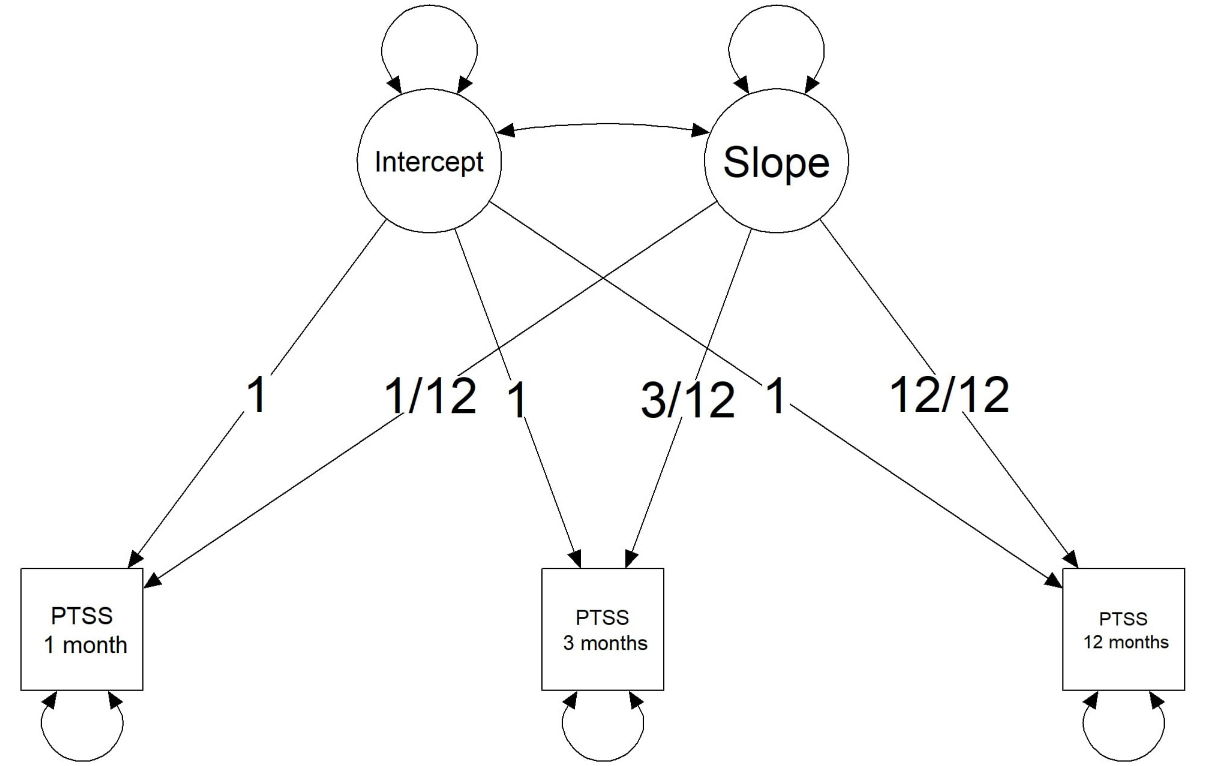
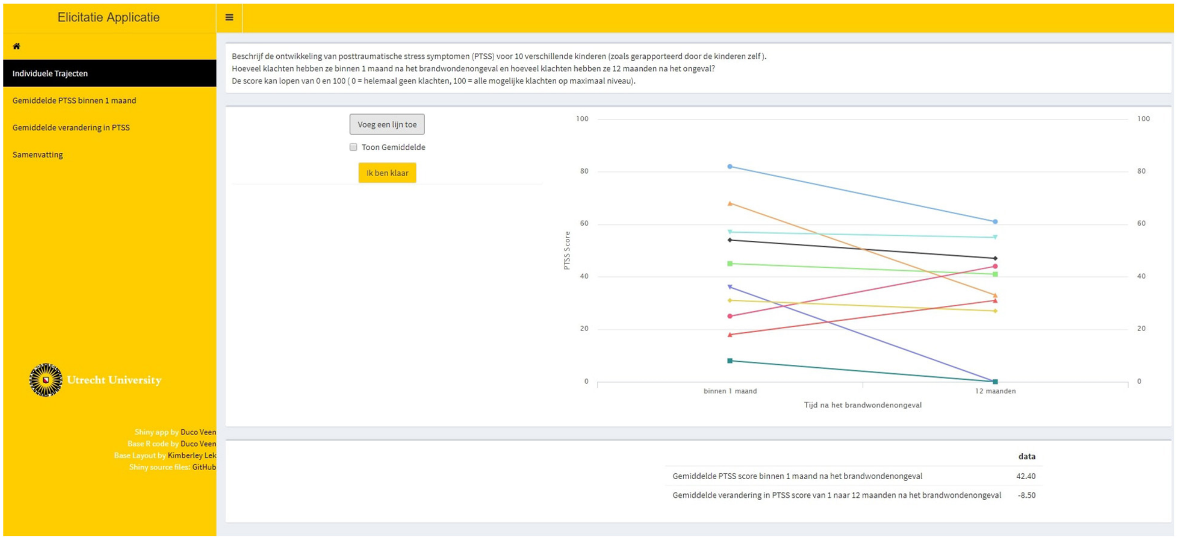


FIGURE 1 | Visual representation of a Latent Growth Curve Model with three observed time points for posttraumatic stress symptoms (PTSSs).



Beschrijf de ontwikkeling van posttraumatische stress symptomen (PTSS) voor 10 verschillende kinderen (zoals gerapporteerd door de kinderen zelf).
Hoeveel klachten hebben ze binnen 1 maand na het brandwondenongeval en hoeveel klachten hebben ze 12 maanden na het ongeval?
De score kan lopen van 0 en 100 (0 = helemaal geen klachten, 100 = alle mogelijke klachten op maximaal niveau).

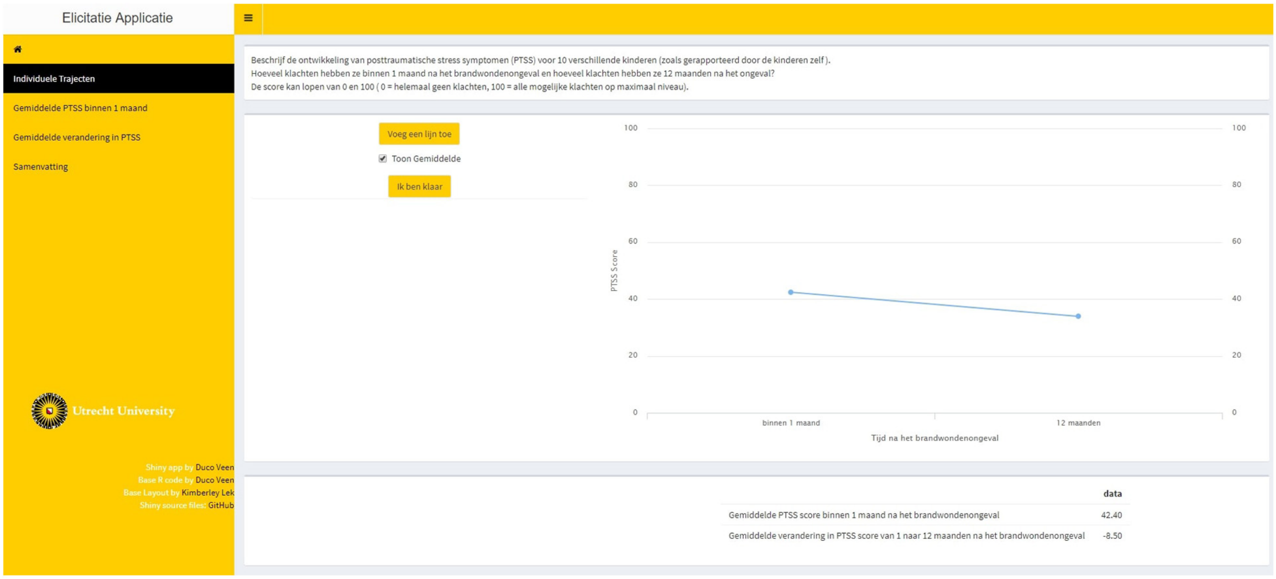
	data
Gemiddelde PTSS score binnen 1 maand na het brandwondenongeval	42.40
Gemiddelde verandering in PTSS score van 1 naar 12 maanden na het brandwondenongeval	-8.50

Shiny app by Duco Veen
Base R code by Duco Veen
Base Layout by Kimberley Lek
Shiny source files: GitHub

FIGURE 2 | Step 1 of the elicitation procedure. Trajectories of posttraumatic stress symptom (PTSS) development were elicited for 10 individuals that are representative for the population. From these trajectories, point estimates for the average intercept and the average slope were obtained.

Step 4. Feedback was provided on the probability distribution that was used to represent the experts' beliefs. **Figure 4** provides a visual representation of steps 3 and 4 with respect to the average intercept, top panel, and the average slope, bottom panel. Single-parameter feedback was provided in the form of a prior density plot, as well as the effect on the implied average trajectory. The experts could accept and confirm the representation of their beliefs or adjust their input in step 3.

Step 5. The experts were shown a summary page on the elicitation, see **Figure 5**. If the experts accepted the representation



Beschrijf de ontwikkeling van posttraumatische stress symptomen (PTSS) voor 10 verschillende kinderen (zoals gerapporteerd door de kinderen zelf).
Hoeveel klachten hebben ze binnen 1 maand na het brandwondenongeval en hoeveel klachten hebben ze 12 maanden na het ongeval?
De score kan lopen van 0 en 100 (0 = helemaal geen klachten, 100 = alle mogelijke klachten op maximaal niveau).

Shiny app by Duco Veen
Base R code by Duco Veen
Base Layout by Kimberley Lek
Shiny source files: GitHub

	data
Gemiddelde PTSS score binnen 1 maand na het brandwondenongeval	42.40
Gemiddelde verandering in PTSS score van 1 naar 12 maanden na het brandwondenongeval	-8.50

FIGURE 3 | Step 2 in the elicitation procedure, providing visual feedback on the extracted average trajectory based upon the experts' provided individual trajectories.

of their beliefs, the probability distributions were now ready to be saved and used in the analyses.

Sample of Experts

Fourteen experts from all three Dutch burn centers participated in the elicitation study. These experts had different professions, including (child) psychologists, pediatric nurses, specialized nurses for burn injuries, and nurses with an additional master's degree [master of science (MSc)]. During the process of obtaining this degree, these nurses worked closely with psychologists and observed their work. Though they are employed at the same burn centers, the tasks and expertise of nurses and psychologists differ: nurses are assumed to have a broader clinical view, taking into account physical and psychological aspects of adjustment, but not necessarily PTSS. Psychologists have a more focused clinical view and have specific expertise on PTSS after traumatic events. Because reporting the individual expert professions would remove almost all anonymity, we ensured that no elicited probability distributions can be associated with individual experts and therefore categorized the experts into two groups. The first group consisted of experts who have obtained an MSc degree ($N = 7$), and the second group consisted of experts who have not ($N = 7$). As the first group consisted mostly of psychologists or experts with at least some education in psychology, we shall refer to this group as the psychologists. The second group consisted mostly of nurses with a variety of additional specializations, and we shall refer to this group as the nurses. The two groups are considered large enough for elicitation studies. Cooke and Goossens (1999) recommend to use the largest possible number of experts, stating that four is the minimum. We were able to include seven experts in both groups of experts.

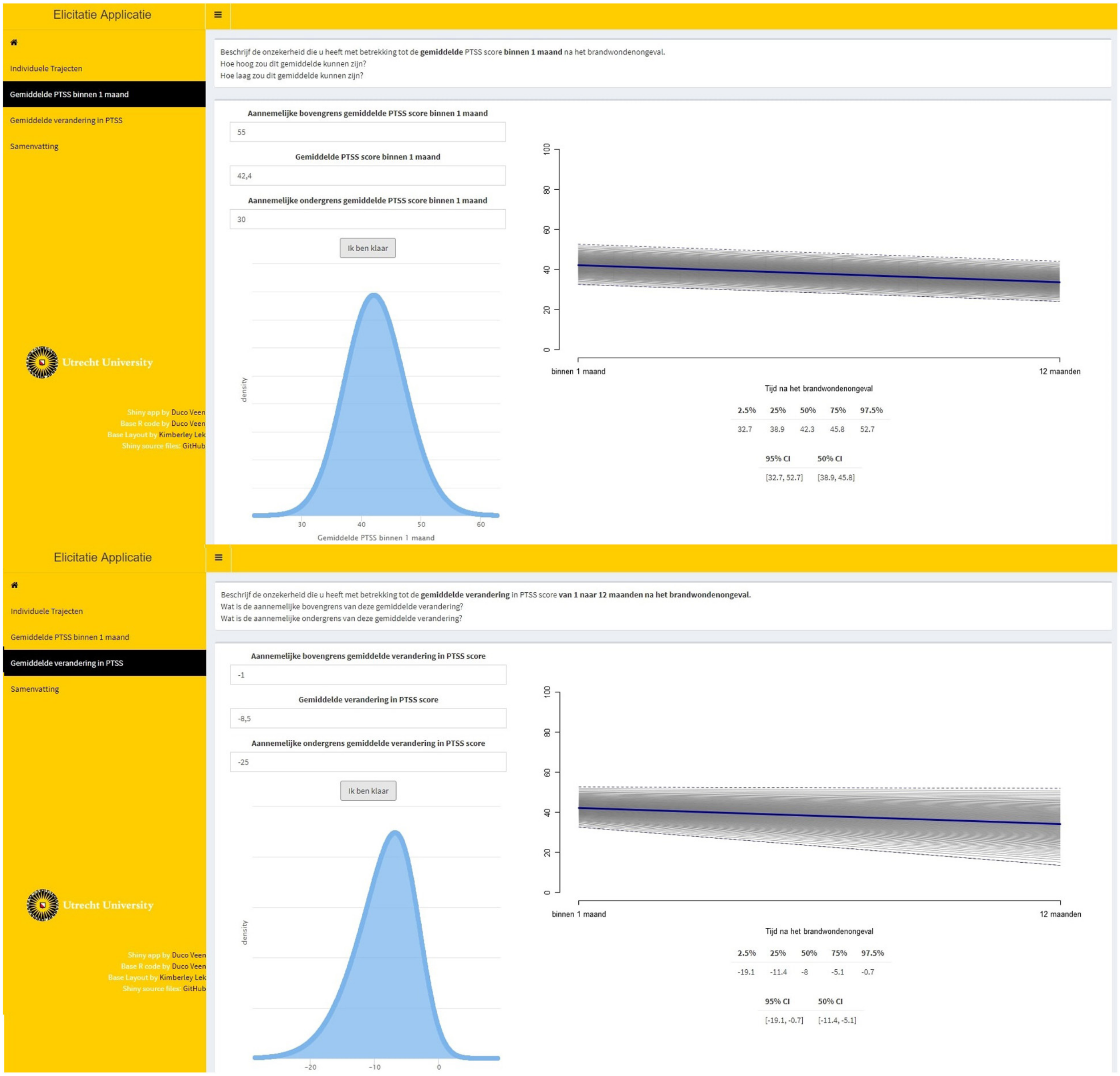
RESULTS

This section first covers a descriptive part on the expert judgments. We report the priors that the experts provided and the mixture priors that can be made from these expert judgments on an aggregated and group distinct level. Thereafter, we report prior-data (dis)agreement measures for all individual expert judgments and the mixture distributions. These prior-data (dis)agreements are based upon the data that were collected in the prospective study by Egberts et al. (2018). Finally, we report notable results from the audio recordings. Note that the quantitative results, analyses, and an overview of individual expert judgments can be found *via* the OSF website for this project at https://osf.io/y5evf/. The transcripts of the audio recordings include many identifying characteristics with respect to both the experts and patients they described during the elicitation and to preserve privacy, so these are not available. This is in accordance with the ethical approval agreement.

Individual and Group Expert Judgments

All 14 expert judgments had been elicited, allowing them to specify a skewed normal distribution parameterized according to Burkner (2019). In **Figure 6**, all the elicited individual expert prior densities can be found as well as the mixture density for all experts, the psychologists' group and the nurses' group regarding both the mean intercept and the mean slope of PTSS development[1]. **Figure 6** shows that the expert judgments differed

[1] Note that the mixtures are based on normal approximations of the elicited skewed normal distributions due to computational instability of the mixture distributions when skewed normal expert priors were used. All experts are weighted equally in the mixture for all experts. The mixture distributions of the nurses and psychologists can be seen as a special case of weighting in which half of the experts receive a weight of 0 and the other half are equally weighted.


Elicitatie Applicatie
≡
Individuele Trajecten
Gemiddelde PTSS binnen 1 maand
Gemiddelde verandering in PTSS
Samenvatting
Utrecht University
Shiny app by Duco Veen
Base R code by Duco Veen
Base Layout by Kimberley Lek
Shiny source files: GitHub
density
30
40
50
60
Gemiddelde PTSS binnen 1 maand
0
20
40
60
80
100
binnen 1 maand
12 maanden
Tijd na het brandwondenongeval

Beschrijf de onzekerheid die u heeft met betrekking tot de **gemiddelde** PTSS score **binnen 1 maand** na het brandwondenongeval.
Hoe hoog zou dit gemiddelde kunnen zijn?
Hoe laag zou dit gemiddelde kunnen zijn?

Aannemelijke bovengrens gemiddelde PTSS score binnen 1 maand
55

Gemiddelde PTSS score binnen 1 maand
42,4

Aannemelijke ondergrens gemiddelde PTSS score binnen 1 maand
30

Ik ben klaar

2.5%	25%	50%	75%	97.5%
32.7	38.9	42.3	45.8	52.7

95% CI	50% CI
[32.7, 52.7]	[38.9, 45.8]

Beschrijf de onzekerheid die u heeft met betrekking tot de **gemiddelde verandering** in PTSS score **van 1 naar 12 maanden na het brandwondenongeval.**
Wat is de aannemelijke bovengrens van deze gemiddelde verandering?
Wat is de aannemelijke ondergrens van deze gemiddelde verandering?

Aannemelijke bovengrens gemiddelde verandering in PTSS score
-1

Gemiddelde verandering in PTSS score
-8,5

Aannemelijke ondergrens gemiddelde verandering in PTSS score
-25

Ik ben klaar

2.5%	25%	50%	75%	97.5%
-19.1	-11.4	-8	-5.1	-0.7

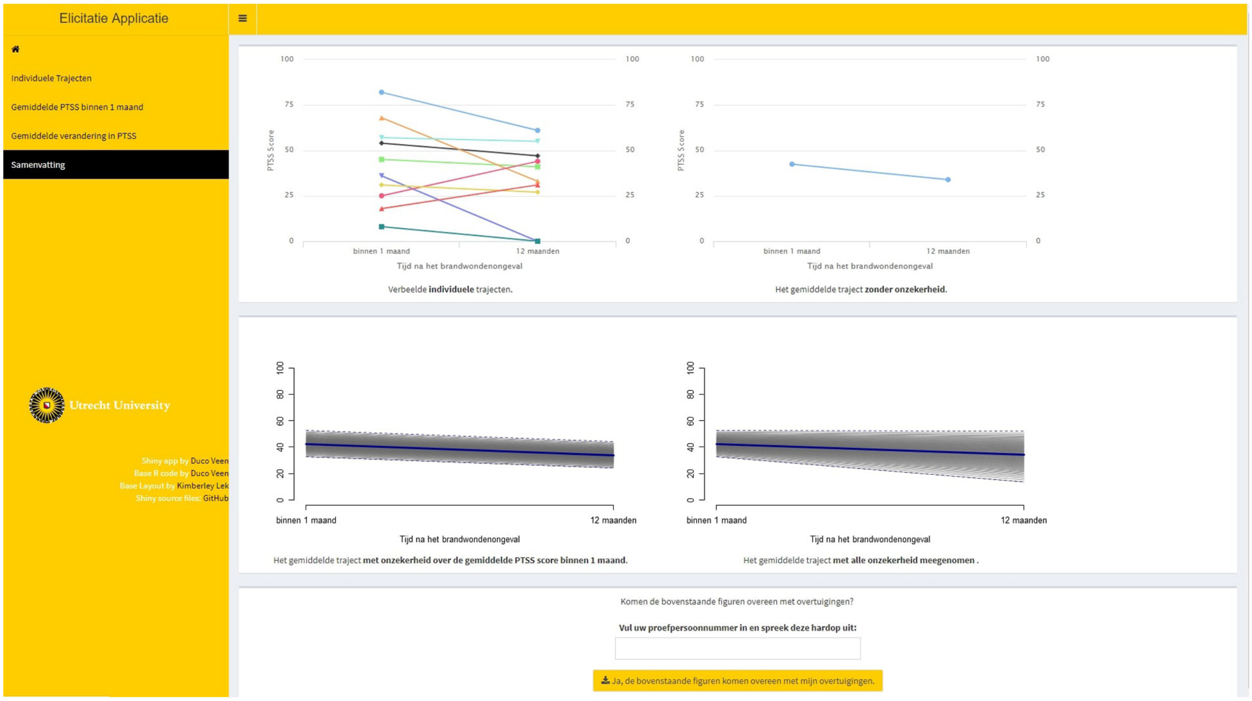
95% CI	50% CI
[-19.1, -0.7]	[-11.4, -5.1]

FIGURE 4 | Steps 3 and 4 of elicitation procedure for the average intercept, top panel, and the average slope, bottom panel. The input that was required for step 3 was provided in the fields on the top left of the tab in the elicitation software. The single-parameter feedback was provided on the bottom left of the tab, displaying the fitted prior density with respect to that parameter. The effect on the implied average trajectory was displayed on the right-hand side of the tab. The average trajectory that was accepted in step 2 is displayed, and a gray band has been added around this average trajectory that represents the 95% credible interval (CI) for the average trajectory. In the top panel, only the uncertainty with respect to the intercept was added to the average trajectory. In the bottom panel, the uncertainty with respect to both the intercept and the slope was added.

quite substantially. Especially concerning the development of PTSS as expressed by the slope parameter, we can see that experts disagreed on the direction of the effect and with a lot of confidence. When we look at the groups of experts, an interesting pattern emerges. If we combine the expert judgments of the psychologists and the nurses into their respective group, the nurses turn out to have a substantially different view from the psychologists. Not only did the nurses' judgments express on average a higher initial amount of PTSS in the population, their combined view also expressed that these initial PTSS scores are quite likely to increase on average over time. The psychologists in contrast assigned almost no probability to an increase in the average PTSS score over the time period of a year; see **Figure 7** for a closer look.

Prior-Data (Dis)Agreement

To assess the (dis)agreement of experts' judgments with the data from the prospective study by Egberts et al. (2018), we used


Elicitatie Applicatie
≡
Individuele Trajecten
Gemiddelde PTSS binnen 1 maand
Gemiddelde verandering in PTSS
Samenvatting
Utrecht University
Shiny app by Duco Veen
Base R code by Duco Veen
Base Layout by Kimberley Lek
Shiny source files: GitHub
PTSS Score
100
75
50
25
0
binnen 1 maand
12 maanden
Tijd na het brandwondenongeval
0
20
40
60
80
100

Verbeelde **individuele** trajecten.

Het gemiddelde traject **zonder onzekerheid.**

Het gemiddelde traject **met onzekerheid over de gemiddelde PTSS score binnen 1 maand.**

Het gemiddelde traject **met alle onzekerheid meegenomen .**

Komen de bovenstaande figuren overeen met overtuigingen?

Vul uw proefpersoonnummer in en spreek deze hardop uit:

Ja, de bovenstaande figuren komen overeen met mijn overtuigingen.

FIGURE 5 | Summary page of the elicitation procedure. The top left plot within the page displays all individual trajectories that the expert specified. The top right plot displays the average trajectory that was obtained based on those individual trajectories. The bottom left plot displays the average trajectory with uncertainty (95% CI) concerning the intercept value taken into account. The bottom right plot displays the average trajectory with uncertainty (95% CI) concerning both the intercept value and the slope value taken into account.

Kullback–Leibler (KL) divergences (Kullback and Leibler, 1951) between the posterior distribution that is based upon the data and an uninformative benchmark prior as well as the individual and aggregated expert judgments. Using information theoretical distance measures to asses prior-data (dis)agreement in this manner has previously been discussed by, for instance, Bousquet (2008), Lek and van de Schoot (2019), and Veen et al. (2018). KL divergences provide us with an indication of how much information is lost as we approximate distribution π_1 by another distribution π_2. A higher divergence indicates a higher loss of information. In this case, π_1 will be the posterior distribution based upon the data and an uninformative benchmark prior, to which we refer as the reference posterior. We approximate the reference posterior with the elicited prior distributions and report the loss of information. For an overview of the priors that are used to compute the reference posterior, see **Figure 8**. **Figure 9** visualizes the reference posteriors for the group mean latent intercept and slope. We used the uninformative benchmark 2 priors that are described in the next paragraph. The differences are negligible with the use of benchmark 1 priors, as can be seen in the supplementary materials that describe the LGM analysis. This demonstrates the principle of stable estimation; the priors are overwhelmed by the data.

In addition to comparing the expert priors to the benchmark posterior, we added two other comparisons to create a frame of reference. Two benchmark situations are added, and their loss of information is calculated. In the situation of benchmark 1, we would take some information regarding the measurement instrument into account. The scale of the measurement instrument was standardized such that values are between 0 and 100; therefore, a $U(0, 100)$ prior on the group mean intercept would cover all possible parameter values. With the parameterization such that the final time measurement implies a change of 1 times the individual latent slope parameter, taking the standardized scale into account, a $U(-100, 100)$ prior on the latent slope covers all possible parameter values and declares them equally possible. For benchmark 2, we take two $N(0, 10^8)$ priors on the latent group mean intercept and slope. It is still common practice, when using Bayesian statistics, to rely on default or uninformative priors when calculating posterior distributions. For instance, in Mplus, the default priors for these specific parameters are $N\,(0, \infty)$ (Asparouhov and Muthén, 2010, Appendix A), which are used in, for instance, McNeish (2016), and van de Schoot et al. (2015). Lynch (2007, chapter 9), using precision instead of variance, specifies $N\,(0,\ 0.0001)$ priors for these parameters. Benchmark 2 reflects this practice.

The KL divergences are reported in **Table 1** and are the numerical representation of the loss of information that occurs by approximating the reference posterior densities from **Figure 9** by the densities that can be seen in **Figure 6** for the experts' priors. It seems that most experts are in disagreement with the collected data from Egberts et al. (2018). There are some individual exceptions, notably experts 9 and 13, who have a view that is very similar to the collected data, while some experts provide a similar

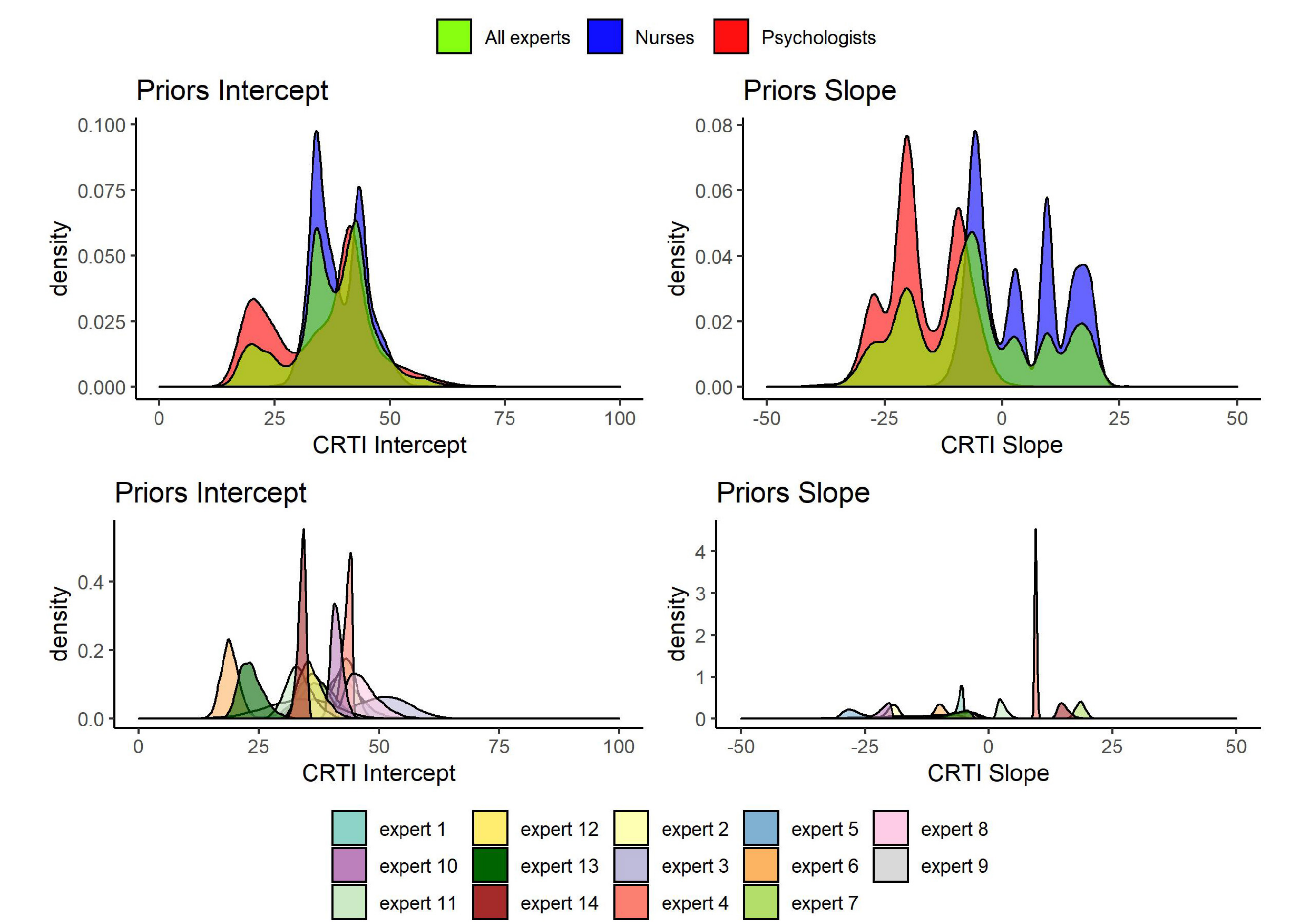


FIGURE 6 | Elicited prior densities from all experts and the associated mixture priors for all experts, the psychologists' group, and the nurses' group regarding both the mean intercept and the mean slope of posttraumatic stress symptom (PTSS) development.

view with respect to one of the two parameters, e.g., experts 3 and 6. It is notable that the group of psychologists in particular and the group of experts as a whole show less loss of information with respect to the data than most experts on both parameters. Finally, what is noteworthy is that benchmark 1, which has no preference for any part of the parameter space covered by the measurement instrument, resembles the data more than most expert judgments and more than the nurses' judgments as a group.

Audio Recordings

The following observations were noteworthy in the transcripts of the audio recordings. All psychologists referred specifically to the concept of PTSS during the elicitation procedure. The group of nurses mentioned stress a lot, but only two nurses actually referred to PTSS specifically. Three psychologists reflected on the linearity assumption of the model and noted that non-linear trajectories often occur. Five of the nurses expressed sentiments that the more severe cases came to mind more easily and therefore might be overrepresented in their beliefs. Only one psychologist expressed a similar statement. Three experts, one psychologist and two nurses, actively reflected on the visual feedback and adjusted their input in the elicitation tool based on this. One expert, a nurse, stated that although he or she was sure about the direction of the trajectory, he or she felt unsure about the associated numerical representation. Finally, one expert, a nurse, repeatedly mentioned that he or she found the task hard to do.

DISCUSSION

We were able to elicit expert judgments with respect to the development of PTSS in young burn victims from 14 experts and contrasted this with data collected in a traditional way by means of a questionnaire. Our study demonstrates differences in views between experts. On an individual basis, the experts were particularly in disagreement with regard to the change of PTSS at 1 year post-burn. There is little overlap in expert beliefs when we look at the elicited prior densities for the slope parameter. The expert judgments not only differed from one individual to the next, but there also seems to be a relationship between the experts' role in the post-burn treatment process and their view

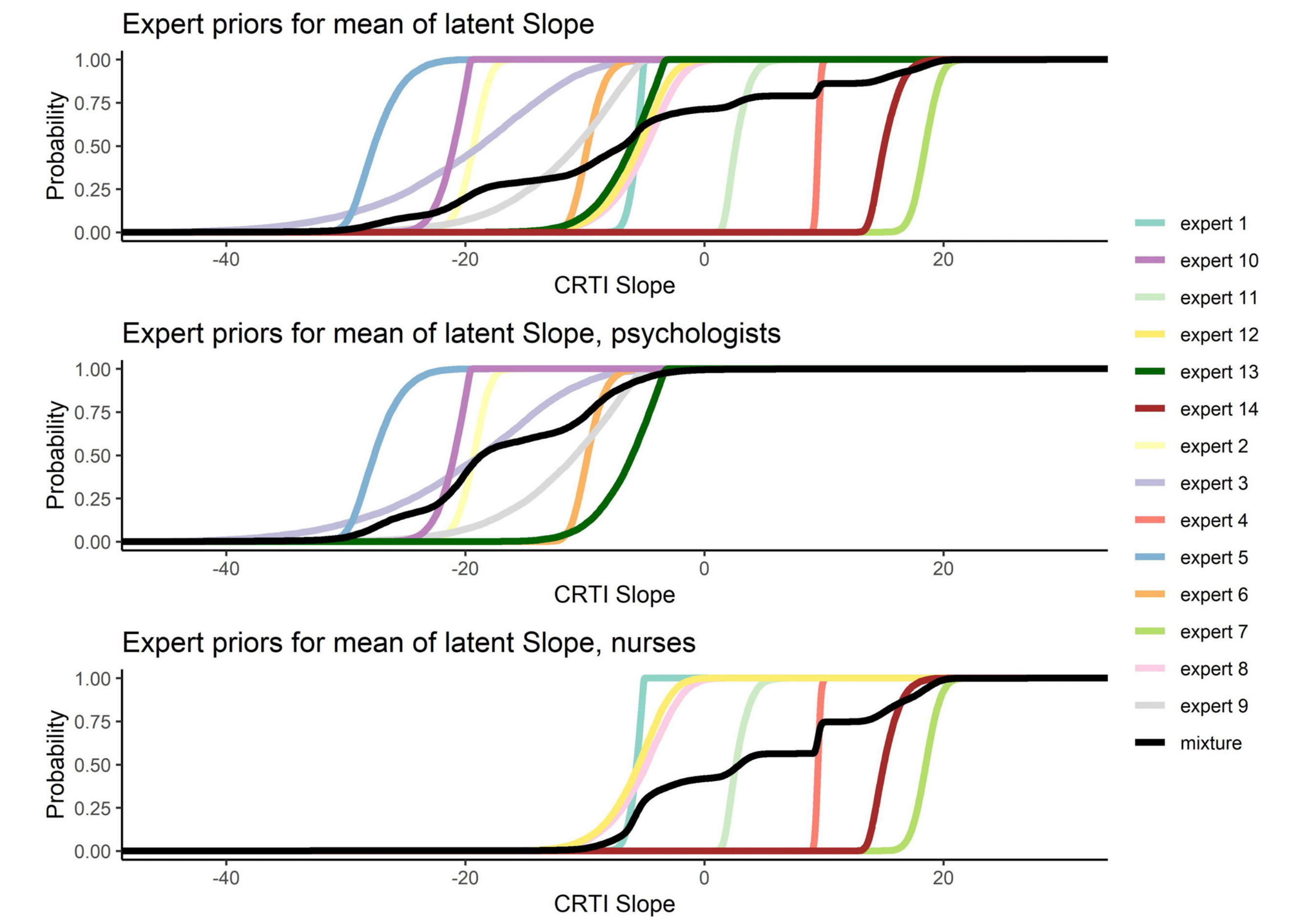


FIGURE 7 | Elicited prior distributions from all experts and the associated mixture priors for all experts, the psychologists' group, and the nurses' group regarding the mean slope of posttraumatic stress symptom (PTSS) development. The cumulative distributions are presented. There was a notable difference in expert judgments between the psychologists' and the nurses' groups.

on the children's development of PTSS. The two groups of experts differed notably in the aggregated elicited judgments: aggregated judgments of the psychologists seemed to align with the data collected by Egberts et al. (2018) while the nurses' judgments seemed to differ more.

With respect to the differences between the two groups of experts, the most remarkable difference was found with respect to the slope parameter. The aggregated views of the groups of experts result in distributions with more uncertainty compared to the individual experts' beliefs. The dispersed views of the experts put together ensure coverage of a larger part of the parameters space than the individual expert judgments do. Interestingly, the more uncertain distributions still clearly present a difference in views regarding the development of PTSS in young burn victims between the nurses' expert group on the one hand and the psychologists' expert group and the data collected by Egberts et al. (2018) on the other. The aggregated judgments from the psychologists assigned almost no probability to the group average PTSS increasing at 1 year post-burn. The aggregated judgments from the nurses, in contrast, assigned a lot of probability to an increase of the group average PTSS at 1 year post-burn. As there is no grounded truth, we cannot conclude which views are a better, or worse representation. However, the results do indicate that the nurses and the psychologists are not in agreement on what happens with respect to the development of PTSS in young burn victims, despite having received similar information about (assessment of) PTSS prior to the elicitation.

The audio recordings of the elicitation settings provide a possible explanation for this important distinction. All psychologists at some point during the elicitation referred to, or specifically mentioned, the construct of PTSS. The group of nurses mentioned several sources of distress, but only two nurses actually referred to PTSS, while one of them judged the 1-year post-burn PTSS to decrease. As burn victims can indeed experience other sources of distress, e.g., related to the development of scar tissue or operations they have to undergo, nurses may have convoluted PTSS with other patient symptoms. This could also explain why the aggregated nurses' view judged the initial PTSS level to be higher for the group average than the aggregated psychologists' view. Overall, the differences possibly

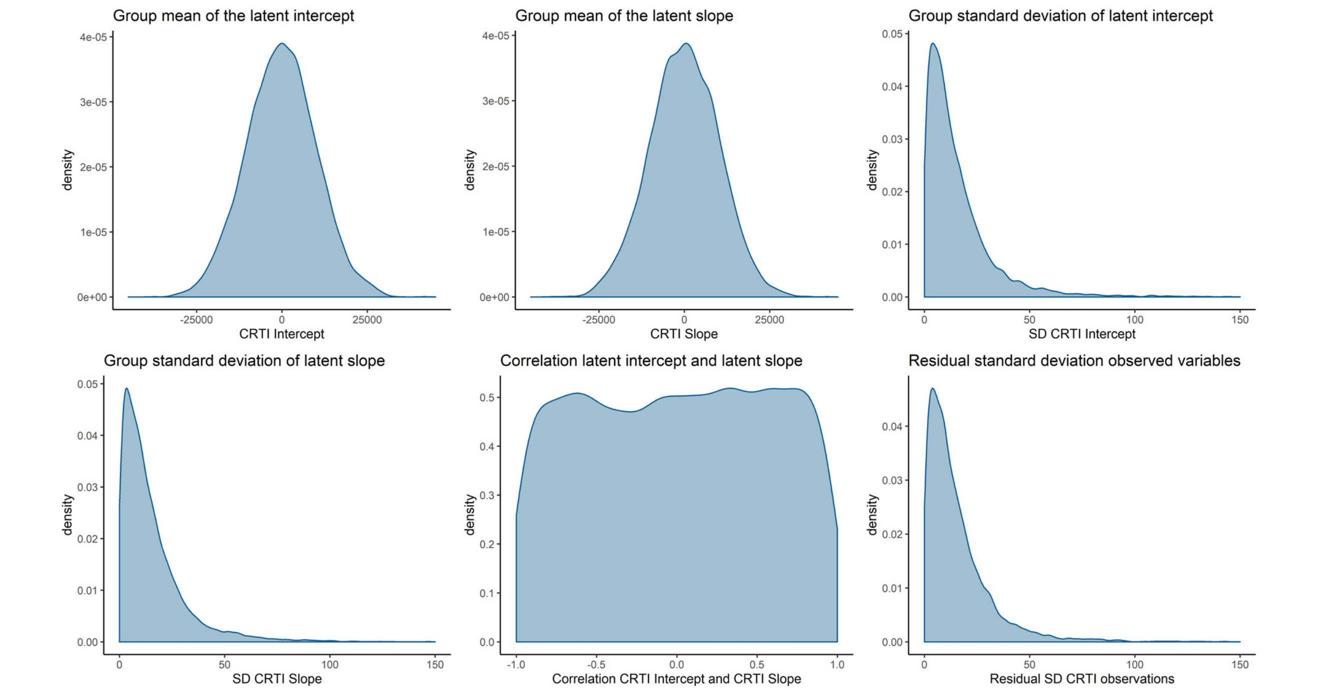


FIGURE 8 | Visual representation of the prior densities that are used to obtain the reference posterior. The prior densities are $\alpha_1 \sim N(0, 10^8)$, $\alpha_2 \sim N(0, 10^8)$, $\psi_{11} \sim half - t(3, 0, 196)$, $\psi_{22} \sim half - t(3, 0, 196)$, $\psi_{21} \sim U(-1, 1)$, and $\theta \sim half - t(3, 0, 196)$.

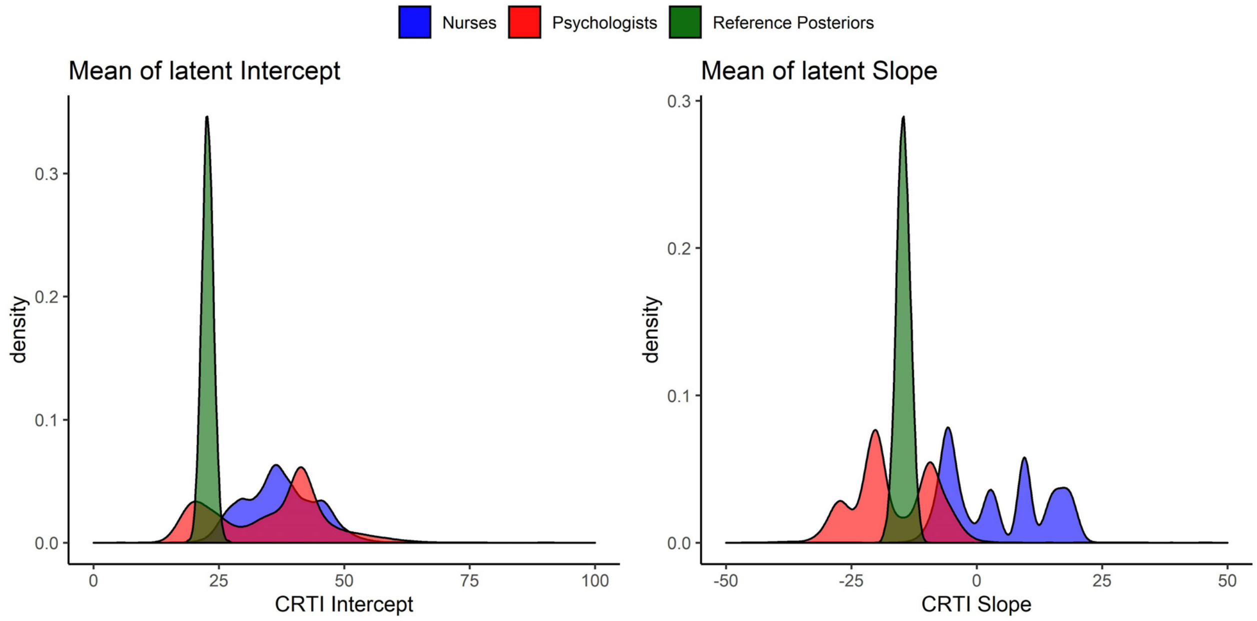


FIGURE 9 | Visual representation of the reference posterior densities for the group mean of latent intercept and slope with the group expert priors for the parameters. The reference posteriors are approximately distributed, $CRTI_{Intercept} \sim N(22.7, 1.3)$, and $CRTI_{Slope} \sim N(-14.6, 1.9)$.

reflect the fact that psychologists are trained to diagnose and treat PTSS, whereas nurses are primarily concerned with procedural and physical care for the patient and are not involved in diagnosing and treating PTSS. In a future study, it could be of interest to investigate the experts' knowledge of the constructs of PTSS and see if this is predictive of KL divergence.

Besides differences between the nurses and the psychologists, we also found a substantial difference between the reference

TABLE 1 | Kullback–Leibler divergences for all individual and mixture priors to the reference posterior.

	Intercept	Slope
Benchmark 1	3.04	3.56
Benchmark 2	8.56	8.39
Nurses	8.19	5.88
Psychologists	1.99	2.18
All	2.72	2.63
Expert 1	42.87	59.18
Expert 2	45.16	25.87
Expert 3	6.71	1.23
Expert 4	72.86	55.38
Expert 5	5.66	98.32
Expert 6	2.10	22.17
Expert 7	79.20	59.61
Expert 8	46.97	4.37
Expert 9	2.48	1.28
Expert 10	43.74	67.55
Expert 11	12.78	64.56
Expert 12	99.94	4.88
Expert 13	0.35	3.62
Expert 14	75.00	74.11

posteriors that provided a representation of the data from Egberts et al. (2018) and the aggregated nurses prior. In **Figure 9**, it can be seen that the psychologists' views overlapped with the reference posteriors. The nurses' views, however, showed almost no overlap with reference posteriors. This could also be assessed numerically, as was done with the KL divergences in **Table 1**. Because the aggregated nurses prior had little overlap with the reference posteriors, the Benchmark 1 priors, i.e., uniform priors that take the information of the measurement instrument into account, outperformed this group in terms of loss of information. This implies that the data collected by Egberts et al. (2018) were better approximated by an uninformed expression of the questionnaire's measurement properties than by the nurses' group prior. The children in the study by Egberts et al. (2018) expressed a lower quantity of PTSS in their self-reported questionnaires compared to the nurses' expert judgments on PTSS for this population.

There can be several explanations for this discrepancy. First, the questionnaire may have resulted in underreporting of symptoms, a view also expressed by one of the experts. In line with this, Egberts et al. (2018) found that mothers gave higher ratings of their child's PTSS compared to the children themselves. On the other hand, mothers' ratings appeared to be influenced by their own symptoms of posttraumatic stress and fathers did not report higher ratings of PTSS compared to their children. Alternatively, the discrepancy could be explained by the elicitation of the expert judgments. Especially the nurses' group reported higher PTSS levels compared to the self-reports, and the previously mentioned convolution of symptoms and lack of specific knowledge about PTSS might be a cause for this observation. In the recordings of the elicitation settings, we found another possible cause. Five of the nurses expressed sentiments that the more severe cases came to mind more easily and therefore might be overrepresented in their beliefs. This is a clear expression of the well-known availability heuristic (Tversky and Kahneman, 1973) that can cause biases in elicitation studies (O'Hagan et al., 2006). In the psychologists' group, only a single expert expressed a similar remark. The availability heuristic, if not remedied, might cause the discrepancy between the reference posteriors and the expert judgments.

The study showed that providing visual feedback on the representation of the experts' beliefs can lead to experts adjusting their input such that obvious incorrect representations of their beliefs are remedied. Unfortunately, it is not possible to validate whether the representation of the experts' beliefs actually corresponds to the "true" beliefs of the expert (O'Hagan et al., 2006; Colson and Cooke, 2018). However, one of the main reasons to use elicitation software is to ameliorate the effects of heuristics and biases by getting experts to actively reflect on the probability distribution that will be used to represent their beliefs. In the recordings, three experts actively reflect on their distributions, adjusting them based on the visual feedback. For this purpose, the elicitation software seems to have worked well. Nevertheless, it seems from our current study that even with the graphical feedback, some experts might still suffer from overconfidence. Expert 11, for instance, stated *"... of course, I have a lot of uncertainty anyway."* However, this does not seem to be reflected in the elicited distribution which has a 99% CI for the latent intercept (27.2, 41.7) and the latent slope (1.2, 5.9). As the experts were only available to us for a limited time, we did not provide a specialized training aimed at elicitation and overcoming heuristics associated with elicitation tasks, which might be a limitation for the current study and the associated (individual-level) results.

This study indicates that aggregating expert judgments could potentially mitigate the severity of individual biases, as one has to rely less on single, possibly overconfident, experts. The aggregation of all experts' judgments or of only the psychologists' judgments leads to less discrepancy between the traditionally collected data and the elicited beliefs in comparison to almost any individual expert and the benchmarks. Aggregating or pooling of expert judgments into a single distribution is common in elicitation studies and can be done in several manners. In our current study, we used opinion pooling with equal weights (O'Hagan et al., 2006, Chapter 9). Alternatively, there is much literature on how expert judgments could be weighted in the aggregation of views. The classical model (Cooke, 1991, Chapter 12) is one of the foremost examples of this. In the classical approach, calibration questions are used to assess the experts. Based on the calibration questions, experts' judgments on the target question or question of interest are weighted to together form the groups' weighted prior beliefs. The calibration questions should be related to the question of interest, and their answers should be known but not to the experts (Colson and Cooke, 2018). It is recommended to have at least eight to 10 calibration questions if dealing with continuous variables (Cooke, 1991, Chapter 12). The experts are elicited concerning the question of interest and the calibration questions. Their answers on the calibration questions are evaluated against the known true values, and the experts are rated on their informativeness and accuracy (Cooke, 1991; Colson and Cooke, 2018). The ratings of the weighting components are based upon the idea of KL divergences (O'Hagan et al., 2006, Chapter 9) such as we used to compare the experts' judgments against the collected data on the question of interest directly. As far as we know, there have not been any studies using the classical approach in the social sciences. Finding calibration questions turns out to be a hard problem, as knowing the true answer to these questions is required. We described the KL divergence between the target question and the experts' judgments, but calibrating experts based on these weight components would be putting emphasis on the traditionally collected data twice. As the traditionally collected data might suffer from biases too, consider for instance the total survey error framework (Groves et al., 2011, Chapter 2) including non-response error and measurement error, this double emphasis might not be desirable. Instead, our equal weights aggregation approach relied on the inclusion of experts with balance in views and diversity in backgrounds (Cooke and Goossens, 1999).

In conclusion, it is possible to express the experts' domain knowledge as prior distributions using the described methodology and compare these elicited distributions to traditionally collected data. The individual expert judgments in general show quite some discrepancy in comparison to traditionally collected data, although there are notable exceptions to this. When considering the mixtures of the groups of experts, the discrepancy becomes less pronounced, especially for the psychologists' group. The psychologists' mixture prior has less KL divergence than mostly any individual expert and notably less KL divergence than Benchmark 1, the uniform prior that takes the information of the measurement instrument into account. The expert judgments add information to the research area, and exploring (dis)similarities between expert judgments and traditional data opens up two exciting avenues for future research. One being the collection of data on the experts that might be predictive for the amount of KL divergence they exhibit with respect to traditionally collected data. The second avenue is the organization of a Delphi-like setting with all experts after the individual judgments are collected and compared with traditional data. The group setting can provide insights into the reasons behind the discrepancies between traditional collected data, individual experts, and groups of experts. If done in a longitudinal manner, this could start a learning cycle in which data and experts converge. Predicting and explaining (dis)similarities between experts' judgments and traditional data such as results of questionnaires can be a potential new line of research for the social sciences.

DATA AVAILABILITY STATEMENT

All related materials for this study, including code and data, can be found on the Open Science Framework (OSF) web page for this project at https://osf.io/y5evf/. The transcripts of the audio recordings include many identifying characteristics with respect to both the experts and patients they described during the elicitation and to preserve privacy, so these are not available.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Ethics Committee of the Faculty of Social and Behavioral Sciences of Utrecht University. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

All authors have been involved in the design of the study and the elicitation procedure. DV programmed the elicitation software. ME arranged the elicitation meetings with the experts. DV and ME conducted all elicitation procedures together. DV wrote and revised the manuscript with contributions and feedback provided by ME, NL, and RS.

FUNDING

This work was supported by Grant NWO-VIDI-452-14-006 from Netherlands Organization for Scientific Research. None of the funders or sponsors of this research had any role in the design and conduct of the study; collection, management, analysis, and interpretation of data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

ACKNOWLEDGMENTS

The authors are grateful to all the experts for their invested time and energy. Additionally, we would like to thank the reviewers for their helpful comments and suggestions that improved the manuscript.

REFERENCES

Alisic, E., Eland, J., Huijbregts, R., and Kleber, R. (2011). *Manual of the Children's Responses to Trauma Inventory - Revised Edition.[Handleiding bij de Schokverwerkingslijst voor Kinderen-Herziene Versie*. Utrecht: Institute for Psychotrauma in Collaboration with Utrecht University.

Alisic, E., Eland, J., and Kleber, R. (2006). *Children's Responses to Trauma Inventory-Revised Version [Schokverwerkingslijst Voor Kinderen-Herziene Versie*. Utrecht: Institute for Psychotrauma in Collaboration with Utrecht University.

Asparouhov, T., and Muthén, B. (2010). *Bayesian Analysis of Latent Variable Models using Mplus*. Available online at: https://www.statmodel.com/download/BayesAdvantages18.pdf (accessed February 6, 2020).

Barons, M. J., Wright, S. K., and Smith, J. Q. (2018). "Eliciting probabilistic judgements for integrating decision support systems," in *Elicitation*, eds L. C.

Dias, A. Morton, and J. Quigley (Berlin: Springer), 445–478. doi: 10.1007/978-3-319-65052-4_17

Bojke, L., Claxton, K., Bravo-Vergel, Y., Sculpher, M., Palmer, S., and Abrams, K. (2010). Eliciting distributions to populate decision analytic models. *Value Health* 13, 557–564. doi: 10.1111/j.1524-4733.2010.00709.x

Bousquet, N. (2008). Diagnostics of prior-data agreement in applied Bayesian analysis. *J. Appl. Stat.* 35, 1011–1029. doi: 10.1080/02664760802192981

Buist, K. L., Dekovic, M., Meeus, W., and van Aken, M. A. (2002). Developmental patterns in adolescent attachment to mother, father and sibling. *J Youth Adolesc.* 31, 167–176. doi: 10.1023/a:1015074701280

Burkner, P.-C. (2019). *Parameterization of Response Distributions in brms*. Available online at: https://cran.r-project.org/web/packages/brms/vignettes/brms_families.html (accessed February 6, 2020).

Catts, H. W., Bridges, M. S., Little, T. D., and Tomblin, J. B. (2008). Reading achievement growth in children with language impairments. *J. Speech Lang. Hear. Res.* 51, 1569–1579. doi: 10.1044/1092-4388(2008/07-0259)

Chang, W., Cheng, J., Allaire, J., Xie, Y., and McPherson, J. (2019). *Shiny: Web Application Framework for r*. Available online at: https://CRAN.R-project.org/package=shiny (accessed February 6, 2020).

Colson, A. R., and Cooke, R. M. (2018). Expert elicitation: using the classical model to validate experts‘ judgments. *Rev. Environ. Econ. Pol.* 12, 113–132. doi: 10.1093/reep/rex022

Cooke, R. M. (1991). *Experts in Uncertainty: Opinion and Subjective Probability in Science*. Oxford: Oxford University Press.

Cooke, R. M., and Goossens, L. H. J. (2008). TU Delft expert judgment data base. *Reliab. Eng. Sys. Saf.* 93, 657–674. doi: 10.1016/j.ress.2007.03.005

Cooke, R. M., and Goossens, L. J. H. (1999). *Procedures Guide for Structured Expert Judgment*. Brussels: Commission of the European Communities.

Dewispelare, A. R., Herren, L. T., and Clemen, R. T. (1995). The use of probability elicitation in the high-level nuclear waste regulation program. *Int. J. Forec.* 11, 5–24. doi: 10.1016/0169-2070(94)02006-b

Dodd, P. J., Yuen, C. M., Sismanidis, C., Seddon, J. A., and Jenkins, H. E. (2017). The global burden of tuberculosis mortality in children: a mathematical modelling study. *Lancet Glob. Health* 5, e898–e906. doi: 10.1016/s2214-109x(17)30289-9

Duncan, T. E., and Duncan, S. C. (2004). An introduction to latent growth curve modeling. *Behav. Ther.* 35, 333–363. doi: 10.1016/s0005-7894(04)80042-x

Egberts, M. R., van de Schoot, R., Geenen, R., and van Loey, N. E. (2018). Mother, father and child traumatic stress reactions after paediatric burn: within-family co-occurrence and parent-child discrepancies in appraisals of child stress. *Burns* 44, 861–869. doi: 10.1016/j.burns.2018.01.003

Elfadaly, F. G., and Garthwaite, P. H. (2017). Eliciting dirichlet and Gaussian copula prior distributions for multinomial models. *Stat. Comput.* 27, 449–467. doi: 10.1007/s11222-016-9632-7

Fischer, K., Lewandowski, D., and Janssen, M. (2013). Estimating unknown parameters in haemophilia using expert judgement elicitation. *Haemophilia* 19, e282–e288. doi: 10.1111/hae.12166

Fisher, R., O'Leary, R. A., Low-Choy, S., Mengersen, K., and Caley, M. J. (2012). A software tool for elicitation of expert knowledge about species richness or similar counts. *Environ. Model. Softw.* 30, 1–14.

Garthwaite, P. H., Al-Awadhi, S. A., Elfadaly, F. G., and Jenkinson, D. J. (2013). Prior distribution elicitation for generalized linear and piecewise-linear models. *J. Appl. Stat.* 40, 59–75. doi: 10.1080/02664763.2012.734794

Goldstein, D. G., and Rothschild, D. (2014). Lay understanding of probability distributions. *Judgm. Decis. Mak.* 9, 1–14.

Gosling, J. P. (2018). "SHELF: the sheffield elicitation framework," in *Elicitation*, eds L. C. Dias, A. Morton, and J. Quigley (Berlin: Springer), 61–93. doi: 10.1007/978-3-319-65052-4_4

Gronau, Q. F., Ly, A., and Wagenmakers, E.-J. (2019). Informed Bayesian t-tests. *Am. Stat.* 74, 1–14.

Groves, R. M., Fowler, F. J. Jr., Couper, M. P., Lepkowski, J. M., Singer, E., and Tourangeau, R. (2011). *Survey methodology*, Vol. 561. Hoboken, NJ: John Wiley & Sons.

Haakma, W., Steuten, L. M., Bojke, L., and IJzerman, M. J. (2014). Belief elicitation to populate health economic models of medical diagnostic devices in development. *Appl. Health Econom. Health Pol.* 12, 327–334. doi: 10.1007/s40258-014-0092-y

Hald, T., Aspinall, W., Devleesschauwer, B., Cooke, R. M., Corrigan, T., Havelaar, A. H., et al. (2016). World Health Organization estimates of the relative contributions of food to the burden of disease due to selected foodborne hazards: a structured expert elicitation. *PloS One* 11:e0145839. doi: 10.1371/journal.pone.0145839

Hampson, L. V., Whitehead, J., Eleftheriou, D., and Brogan, P. (2014). Bayesian methods for the design and interpretation of clinical trials in very rare diseases. *Stat. Med.* 33, 4186–4201. doi: 10.1002/sim.6225

Hampson, L. V., Whitehead, J., Eleftheriou, D., Tudur-Smith, C., Jones, R., Jayne, D., et al. (2015). Elicitation of expert prior opinion: application to the MYPAN trial in childhood polyarteritis nodosa. *PLoS One* 10:e0120981. doi: 10.1371/journal.pone.0120981

Ho, C.-H., and Smith, E. I. (1997). Volcanic hazard assessment incorporating expert knowledge: application to the Yucca Mountain region, Nevada, USA. *Mathem. Geol.* 29, 615–627. doi: 10.1007/bf02769647

James, A., Choy, S. L., and Mengersen, K. (2010). Elicitator: an expert elicitation tool for regression in ecology. *Environ. Model. Softw.* 25, 129–145. doi: 10.1016/j.envsoft.2009.07.003

Kullback, S., and Leibler, R. A. (1951). On information and sufficiency. *Ann. Math. Stat.* 22, 79–86.

Lek, K., and van de Schoot, R. (2018). Development and evaluation of a digital expert elicitation method aimed at fostering elementary school teachers' diagnostic competence. *Front. Educ.* 3:82. doi: 10.3389/feduc.2018.00082

Lek, K., and van de Schoot, R. (2019). How the choice of distance measure influences the detection of prior-data conflict. *Entropy* 21:446. doi: 10.3390/e21050446

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

Little, T. D., Bovaird, J. A., and Slegers, D. W. (2006). Methods for the analysis of change. *Handbook of Personality Development* eds D. K. Mroczek, and T. D. Little (Mahwah, NJ: Erlbaum) 181–211.

Low-Choy, S., James, A., Murray, J., and Mengersen, K. (2012). "Elicitator: a user-friendly, interactive tool to support scenario-based elicitation of expert knowledge," in *Expert knowledge and its application in landscape ecology*, eds A. H. Perera, C. J. Johnston, and C. Ashton Drew (Berlin: Springer), 39–67. doi: 10.1007/978-1-4614-1034-8_3

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

McNeish, D. (2016). Using data-dependent priors to mitigate small sample bias in latent growth models a discussion and illustration using Mplus. *J. Educ. Behav. Stat.* 41, 27–56. doi: 10.3102/1076998615621299

Morris, D. E., Oakley, J. E., and Crowe, J. A. (2014). A web-based tool for eliciting probability distributions from experts. *Environ. Model. Softw.* 52, 1–4. doi: 10.1016/j.envsoft.2013.10.010

Murphy, A. H., and Winkler, R. L. (1974). Subjective probability forecasting experiments in meteorology: some preliminary results. *Bul. Am. Meteorol. Soc.* 55, 1206–1216. doi: 10.1175/1520-0477(1974)055<1206:spfeim>2.0.co;2

Murphy, A. H., and Winkler, R. L. (1984). Probability forecasting in meteorology. *J. Am. Stat. Assoc.* 79, 489–500.

Oakley, J. (2019). *SHELF: Tools to support the sheffield elicitation framework*. Available online at: https://CRAN.R-project.org/package=SHELF (accessed February 6, 2020).

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., et al. (2006). *Uncertain Judgements: Eliciting Experts' Probabilities*. Hoboken, NJ: John Wiley & Sons.

Orth, U., Robins, R. W., and Widaman, K. F. (2012). Life-span development of self-esteem and its effects on important life outcomes. *J. Personal. Soc. Psychol.* 102, 1271–1288. doi: 10.1037/a0025558

Runge, A. K., Scherbaum, F., Curtis, A., and Riggelsen, C. (2013). An interactive tool for the elicitation of subjective probabilities in probabilistic seismic-hazard analysis. *Bul. Seismol. Soc. Am.* 103, 2862–2874. doi: 10.1785/0120130026

Truong, P. N., Heuvelink, G. B., and Gosling, J. P. (2013). Web-based tool for expert elicitation of the variogram. *Comput. Geosci.* 51, 390–399. doi: 10.1016/j.cageo.2012.08.010

Tversky, A., and Kahneman, D. (1973). Availability: a heuristic for judging frequency and probability. *Cogn. Psychol.* 5, 207–232. doi: 10.1016/0010-0285(73)90033-9

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

van de Schoot, R., Sijbrandij, M., Depaoli, S., Winter, S. D., Olff, M., and van Loey, N. E. (2018). Bayesian PTSD-trajectory analysis with informed priors based on a systematic literature search and expert elicitation. *Multivariate Behav. Res.* 53, 267–291. doi: 10.1080/00273171.2017.1412293

van de Schoot, R., Winter, S. D., Ryan, O., Zondervan-Zwijnenburg, M., and Depaoli, S. (2017). A systematic review of Bayesian articles in psychology: the last 25 years. *Psychol. Methods* 22, 217–239.

Veen, D., Stoel, D., Schalken, N., Mulder, K., and van de Schoot, R. (2018). Using the data agreement criterion to rank experts' beliefs. *Entropy* 20:592.

Veen, D., Stoel, D., Zondervan-Zwijnenburg, M., and van de Schoot, R. (2017). Proposal for a five-step method to elicit expert judgement. *Front. Psychol.* 8:2110. doi: 10.3389/fpsyg.2017.02110

Zondervan-Zwijnenburg, M., van de Schoot-Hubeek, W., Lek, K., Hoijtink, H., and van de Schoot, R. (2017). Application and evaluation of an expert judgment elicitation procedure for correlations. *Front. Psychol.* 8:90. doi: 10.3389/fpsyg.2017.00090

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

CONCEPTUAL ANALYSIS
published: 24 November 2020
doi: 10.3389/fpsyg.2020.608045

The Importance of Prior Sensitivity Analysis in Bayesian Statistics: Demonstrations Using an Interactive Shiny App

Sarah Depaoli*, Sonja D. Winter and Marieke Visser

Department of Psychological Sciences, University of California, Merced, Merced, CA, United States

The current paper highlights a new, interactive Shiny App that can be used to aid in understanding and teaching the important task of conducting a prior sensitivity analysis when implementing Bayesian estimation methods. In this paper, we discuss the importance of examining prior distributions through a sensitivity analysis. We argue that conducting a prior sensitivity analysis is equally important when so-called diffuse priors are implemented as it is with subjective priors. As a proof of concept, we conducted a small simulation study, which illustrates the impact of priors on final model estimates. The findings from the simulation study highlight the importance of conducting a sensitivity analysis of priors. This concept is further extended through an interactive Shiny App that we developed. The Shiny App allows users to explore the impact of various forms of priors using empirical data. We introduce this Shiny App and thoroughly detail an example using a simple multiple regression model that users at all levels can understand. In this paper, we highlight how to determine the different settings for a prior sensitivity analysis, how to visually and statistically compare results obtained in the sensitivity analysis, and how to display findings and write up disparate results obtained across the sensitivity analysis. The goal is that novice users can follow the process outlined here and work within the interactive Shiny App to gain a deeper understanding of the role of prior distributions and the importance of a sensitivity analysis when implementing Bayesian methods. The intended audience is broad (e.g., undergraduate or graduate students, faculty, and other researchers) and can include those with limited exposure to Bayesian methods or the specific model presented here.

Keywords: Bayesian statistics, prior distributions, sensitivity analysis, Shiny App, simulation

OPEN ACCESS

Edited by:
Alexander Robitzsch,
University of Kiel, Germany

Reviewed by:
Sara Van Erp,
Utrecht University, Netherlands
Steffen Zitzmann,
Eberhard Karls University of Tübingen, Germany

****Correspondence:***
Sarah Depaoli
sdepaoli@ucmerced.edu

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *18 September 2020*
Accepted: *30 October 2020*
Published: *24 November 2020*

Citation:
Depaoli S, Winter SD and Visser M (2020) The Importance of Prior Sensitivity Analysis in Bayesian Statistics: Demonstrations Using an Interactive Shiny App. Front. Psychol. 11:608045. doi: 10.3389/fpsyg.2020.608045

INTRODUCTION

Through a recent systematic review of the literature in the Psychological Sciences, we know that the use of Bayesian methods is on the rise (van de Schoot et al., 2017). However, this review also highlighted an unnerving fact: Many applied users of Bayesian methods are not properly implementing or reporting the techniques. The goal of this paper is to tackle one of the main issues highlighted in this systematic review—namely, examining the impact of prior distributions through a sensitivity analysis. Understanding the impact of priors, and then making subsequent decisions about these priors, is perhaps the trickiest element of implementing Bayesian methods. Many users of Bayesian estimation methods attempt to avoid this issue by using "diffuse" priors, but this is not always a viable approach because some models need informative priors. The impact of priors (whether diffuse or otherwise) is highly dependent on issues related to model complexity and the structure of the data. Our paper focuses on how to examine the impact of prior distributions in a transparent manner.

As a motivating example, we conducted a small simulation study illustrating the impact of different prior specifications on final model results. This simulation study shows the importance of thoroughly examining the impact of priors through a sensitivity analysis. We also developed an interactive web application (i.e., Shiny App) for users to learn more about the impact of priors and the need for a sensitivity analysis in empirical situations. This App allows users to examine the impact of various prior distribution settings on final model results, ensuring that the user is fully aware of the substantive impact of prior selection. Examining the impact of priors is central to whether Bayesian results are viable, completely understood, and properly conveyed. Our Shiny App aids with further illustrating this issue.

GOALS OF THE CURRENT PAPER

The current paper provides readers with a step-by-step way of thinking about Bayesian statistics and the use of priors. Prior distributions turn out to be one of the most important elements of any Bayesian analysis, largely because of how much weight and influence they can carry regarding final model results and substantive conclusions. Our aims are as follows:

1. Present readers with a friendly introduction to Bayesian methods and the use of priors. We aim to keep the paper accessible to people coming from a wide range of statistical backgrounds, as well as from a variety of different fields.
2. Illustrate the fact that examining the impact of priors is an incredibly important task when interpreting final model results in an applied research setting. We use a small simulation study to illustrate this point.
3. Introduce a new, interactive Shiny App that we developed in order to assist in visualizing important elements of a prior sensitivity analysis.
4. Demonstrate the potential impact of priors through an empirical example using the interactive Shiny App and data that we supply, which provides a tool for readers to explore prior impact in a hands-on setting.
5. Present a set of frequently asked questions regarding priors and a prior sensitivity analysis, as well as candid answers to each question.

INTENDED AUDIENCE AND ORGANIZATION OF THE PAPER

This paper is aimed at novice users of Bayesian methodology. We have designed the paper to benefit students and researchers coming from a wide range of statistical backgrounds. For example, undergraduate students may find the Shiny App useful to experiment with some basics of Bayesian statistics and visualize what different prior settings look like. More advanced graduate students or researchers may find the simulation study as a helpful illustration for capturing the importance of prior sensitivity analyses. In turn, they may also find the application presented in the Shiny App particularly useful to understand the specific impact of priors for the model presented here. The paper and Shiny App have been constructed to benefit students and researchers coming from a wide array of fields within the social and behavioral sciences, and all material to reconstruct the analyses presented here is available online at: https://osf.io/eyd4r/.

The remainder of this paper is organized as follows. The next section highlights the main reasons that one would potentially want to use Bayesian methods in an applied research context. One of the main reasons that we cover in this section is that some researchers may want to incorporate previous knowledge into the estimation process. This is typically done through something called a *prior distribution* (or *prior*), and the section following describes the potential impact of priors. This section is particularly relevant to the Shiny App that we developed, and the issues surrounding priors largely remain at the crux of recognizing when Bayesian methods are misused or inaccurately portrayed.

Next, we present information surrounding the multiple regression model, which is referenced in the subsequent sections. We then present a small simulation study, which is aimed to highlight the impact that different prior settings can have on the accuracy of final model estimates obtained. These results lead into the importance of conducting a prior sensitivity analysis. The following section presents information surrounding our Shiny App, how it works, and how readers can benefit from using it. We highlight how the App can be used to learn more about the important issue of prior sensitivity analysis within Bayesian statistics, and we also provide an interactive platform for readers to gain a deeper understanding of the issues described here. Finally, the paper concludes with a discussion of frequently asked questions regarding prior sensitivity analysis, as well as final thoughts on the importance of transparency within research conducted via the Bayesian estimation framework.

WHY ARE BAYESIAN METHODS USEFUL IN APPLIED RESEARCH?

There are many reasons why a researcher may prefer to use Bayesian estimation to traditional, frequentist (e.g., maximum likelihood) estimation. The main reasons for using Bayesian methods are as follows: (1) the models are too "complex" for traditional methods to handle (see e.g., Depaoli, 2013; Kim et al., 2013; Cieciuch et al., 2014; Depaoli and Clifton, 2015; Zondervan-Zwijnenburg et al., 2019), (2) only relatively small sample sizes are available (see e.g., Zhang et al., 2007; Depaoli et al., 2017a; Zondervan-Zwijnenburg et al., 2019), (3) the researcher *wants* to include background information into the estimation process (see e.g., Zondervan-Zwijnenburg et al., 2017), and (4) there is preference for the types of results that Bayesian methods produce (see e.g., Kruschke, 2013). It is important to note that, regardless of the reasons that Bayesian methods were implemented, a sensitivity analysis of priors is always important to include. In the subsequent sections, we discuss this issue of priors to a greater extent.

WHAT DO WE KNOW ABOUT THE IMPACT OF PRIORS?

The Bayesian literature (using simulation and applied data) has uncovered several important findings surrounding the potential impact of prior distributions on final model results. Some of the literature has shown that prior impact is highly dependent on model complexity, and it is incredibly important to fully examine the influence of priors on final model estimates. In this section, we unpack this issue a bit more, highlighting the reasons one might want to examine their priors.

Priors Can Impact Results (Sometimes in a Big Way!)

One of the reasons why the use of Bayesian methods is considered controversial is the notion that priors can (and do!) impact final model results. What this means in a practical sense is that a researcher can have a very strong opinion about the model parameter values, and this opinion (via the prior) can drive the final model estimates. There are many research scenarios within the Bayesian context where informative (or user-specified) priors have an impact on final model estimates. Some examples include research with models such as the latent growth mixture model (Depaoli et al., 2017b; van de Schoot et al., 2018), the confirmatory factor analytic model (Golay et al., 2013), and logistic regression (Heitjan et al., 2008).

The reverse is true in that the literature has shown that completely diffuse priors can also impact final model results. Although Bayesian theory indicates that large sample sizes can overcome (or swarm) the information in the prior (see e.g., Ghosh and Mukerjee, 1992), some research indicates that diffuse priors can impact final model estimates even with larger sample sizes—sometimes in an adverse manner. Examples of modeling situations where diffuse priors have been shown in simulation to adversely impact final model estimates include probit regression models (Natarajan and McCulloch, 1998), meta-analysis (Lambert et al., 2005), item response theory (Sheng, 2010), structural equation modeling (van Erp et al., 2018)—of which sensitivity analysis guidelines are also provided for structural equation models, latent growth mixture models (Depaoli, 2013), and multilevel structural equation models (Depaoli and Clifton, 2015). In all of these cases, researchers found that diffuse priors had a substantial (negative) impact on the obtained estimates.

Accurate estimates are harder to obtain for some parameters than others. Specifically, more complex models (especially when coupled with smaller sample sizes) can require additional information for certain model parameters in order to supplement flatter likelihoods. For example, in some of our own investigations, variances can be more difficult to estimate than means when the likelihood is relatively flatter (and more peaked for a mean). Models that have many parameters that are difficult-to-estimate may require more informative priors, at least on some model parameters. If a parameter is associated with a flatter likelihood, and diffuse priors are implemented, then there may not be enough information (from the data likelihood or the prior) to produce an accurate estimate. The most common instances where this problem occurs are with more complex models (e.g., mixture models, multilevel models, or latent variable models), but the issue is common enough that the impact of priors should be examined regardless of the informativeness of the prior settings. An important take-away from this should be not to blindly rely on prior settings without understanding their impact, even if they are intended to be diffuse or they are software-defined default priors.

If a prior is used to help incorporate the degree of (un)certainty surrounding a model parameter, then we would expect it to have *some* impact. However, it is really important to understand that impact and account for it when drawing substantive conclusions. Therefore, Bayesian experts often agree that an important, and needed, element of Bayesian estimation is the inclusion of a *sensitivity analysis* of the priors.

WHAT IS A SENSITIVITY ANALYSIS OF PRIORS?

A sensitivity analysis allows the researcher to examine the final model results, based on the original (or reference) prior, in relation to results that would be obtained using different priors. Many Bayesian experts (e.g., Muthén and Asparouhov, 2012; Kruschke, 2015) recommend that a sensitivity analysis should always be conducted, and there has even been a checklist developed (Depaoli and van de Schoot, 2017) that aids in how to conduct and interpret such results in a transparent manner. For applied papers implementing a sensitivity analysis of priors, see: Müller (2012),Depaoli et al. (2017a), or van de Schoot et al. (2018).

The process takes place as follows:

1. The researcher predetermines a set of priors to use for model estimation. These priors can be default priors from the statistical software, or they can be user-specified based on previous knowledge of the model parameters (e.g., based on a simple guess, a meta-analysis of prior literature, interviews with content experts, etc.).

2. The model is estimated, and convergence is obtained for all model parameters.
3. The researcher comes up with a set of "competing" priors to examine; we will describe what this set of priors can look like in the examples below. The point here is *not* to alter the original priors. Rather, it is to examine how robust the original results are when the priors are altered, and the model is re-estimated.[1] It can also be a method used to identify priors that would serve as a poor choice for the model or likelihood—an issue we expand on more in the discussion.
4. Results are obtained for the "competing" priors and then compared with the original results through a series of visual and statistical comparisons.
5. The final model results are written up to reflect the original model results (obtained in Item 1, from the original priors), and the sensitivity analysis results are also presented in order to comment on how robust (or not) the final model results are to different prior settings.

This last point is particularly important. A systematic review of Bayesian statistics in the Psychological Sciences (van de Schoot et al., 2017) unveiled that sensitivity analyses were only reported in 16.2% of the applied studies over the course of 25 years. What this means is that the majority of applied Bayesian papers published in the field did not thoroughly examine the role or impact of priors.

One of the biggest aids for examining the role or impact of priors can be to visually examine the resulting posterior distributions across many different prior settings. We will highlight some important ways to visualize priors and sensitivity analysis results in a subsequent section when introducing our interactive Shiny App.

Visual aids are particularly important here because they can help the researcher to more easily determine: (1) how different or similar the posterior distributions are when different priors are formed, and (2) whether the difference across sets of results (from different prior settings) is *substantively* important. In the end, this latter point is really what matters most. If several sets of priors produce slightly different posterior estimates but the results are substantively comparable, then the results are showing stability (or robustness) across different prior settings. In this case, the researcher can be more confident that the prior setting is not influencing the substantive conclusions in a large way.

One may take these last statements to mean that we are implying the opposite results would be somehow negative or bad. In other words, is it a problem if my sensitivity analysis results show that the resulting posterior changes in substantively meaningful ways when the prior is altered? The answer is NO. There is not necessarily a "problem" here. It is incredibly informative to theory-based research to uncover that results are dependent on the particular theory (i.e., prior) being implemented. This is not a *bad* result at all. It is just one that requires a bit more care when describing. Whatever the results are of the sensitivity analysis (e.g., whether results are stable or not), they should be thoroughly reported in the results and discussion sections of the paper. These findings can be presented in terms of visual depictions of the posteriors from multiple sets of priors, as defined through the sensitivity analysis. Likewise, results can also be presented in statistical form, where percent "bias," or deviation, is computed for parameter estimates obtained under different prior settings.[2] Another alternative when working with diffuse priors could be to report the results across a range of diffuse priors as the main analysis. This tactic might facilitate illustrating the uncertainty surrounding the exact prior specification, especially if various diffuse priors provide varying results.

If the priors are shifted only a small amount in the sensitivity analysis and they result in *very* different results, then it would be beneficial to take a closer look at the model code to ensure everything is properly specified. However, small-to-moderate shifts in the substantive conclusions are not a concern and should just be reported along with the findings and subsequently addressed in the discussion section with respect to learning something about the robustness of results under different prior settings.

Note that the original prior settings are not modified during the sensitivity analysis process. Instead, sensitivity analysis results are presented, and they may be used as evidence that priors should be shifted in some way in a future analysis on another dataset. For transparency reasons, it is important to keep the original prior and not change it because of something that was unveiled in the sensitivity analysis. Doing so would be an instance of Bayesian HARKing (hypothesizing after results are known; Kerr, 1998), which is just as questionable as frequentist HARKing.

PROOF OF CONCEPT SIMULATION: ILLUSTRATING THE IMPACT OF PRIORS

Next, we present a small simulation study illustrating the impact of different prior settings on final model estimates. Since there is no way to know the true value of a population parameter in application, it is not possible to know how much bias estimates contain unless a simulation study is conducted. This simulation study sets the stage for the importance of examining prior impact in application, a concept that we focus on in the interactive Shiny App presented in the following section.

The Model

For illustration purposes, we used the multiple regression model, which is a very common model that is found in the

[1] Several developments have made this step easier by approximating the posterior instead of estimating it directly (e.g., Gustafson and Wasserman, 1995; Roos et al., 2015). These methods have also been implemented in R packages, such as the 'adjustr' package (McCartan, 2020). We do not use this package in our Shiny App as it does not provide the full posterior distribution that we use for our visuals.

[2] We do not refer to the traditional sense of the word "bias," where an estimate is compared to a population value (e.g., in the sense of the comparisons made in the simulation study presented next). Instead, we are referring here to the deviation between two estimates, each obtained as a result of different prior settings. A calculation similar to bias can be implemented, providing the researcher with an indication of the difference between the estimates resulting from the sensitivity analysis. We further illustrate this concept in the section detailing the Shiny App, and we will refer to this concept as "deviation."

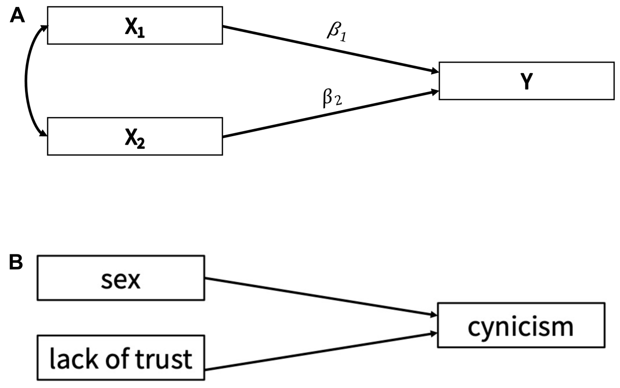


FIGURE 1 | (A) Multiple regression model used in simulation study, with a single outcome variable, *Y*, and two predictors, $X_1 - X_2$. **(B)** Multiple regression model used in the applied example, with an outcome of *Cynicism* and two predictors.

applied psychological literature.[3] In turn, it also acts as a foundation for many other advanced models [e.g., (multilevel) mixed regression models, or latent growth curve models]. These reasons make the multiple regression model a good candidate for demonstration. In addition, we felt this model, even if unfamiliar to the reader, can be conceptually described and understood without having strong background knowledge of the model. Although we limit our discussion to multiple regression, the prior sensitivity analysis principles that we demonstrate can be broadly generalized to other model forms (e.g., growth curve models, confirmatory factor analysis, mixture models).

This model has been used in a variety of research settings within the social and behavioral sciences. For example, it has been used to predict academic achievement (Adeyemo, 2007), self-reassurance (Kopala-Sibley et al., 2013), and sleep quality (Luyster et al., 2011). The base of the model includes a single (continuous) outcome variable that is predicted by several different predictor variables; the model can be found in **Figure 1A**. In this figure, there is a single outcome variable (called "Y"), and two correlated predictors (called "$X_1 - X_2$") with regression weights $\beta_1 - \beta_2$.

Bayesian methods can be implemented in this modeling context in a relatively simple manner. For a basic form of the model, as seen in **Figures 1A,B**, a researcher may be particularly interested in placing informative priors on the regression weights (i.e., the directional paths in the figure) that link the predictors to the outcome. In this case, it may mean that the researcher has a particular idea (or theory) about how the variables relate, as well as how strong of a predictor each variable may be in the model.

Typically, informativeness of a prior is defined by one of three categories: informative, weakly informative, and diffuse. Informative priors are usually conceptualized as priors with a large amount of information surrounding a particular parameter. What this translates to is a large probability mass hovering over a relatively narrowed span of possible values for a parameter to take on. For example, **Figure 2A** illustrates an informative prior, with narrowed variation surrounding a mean value of 75. A weakly informative prior is one that carries more spread, or variation, than an informative prior. **Figure 2B** illustrates a weakly informative prior by highlighting a wider distributional spread. Finally, a diffuse prior is one that offers little-to-no information about the parameter value. One way of conceptualizing this prior form is to use a normal prior with a very wide variance, making it effectively flat across a wide range of values. **Figure 2C** illustrates a diffuse prior setting for the normal distribution. In all three of these plots, the normal prior was centered at 75, but the variance of the priors differed from small (**Figure 2A**) to large (**Figure 2C**).[4]

[3] The multiple regression model is a simple model and, with the use of conjugate priors (described below), the posterior can be analytically derived without the use of MCMC sampling. However, we felt that using a relatively simple modeling context (opposed to a more complicated, latent variable model, for example) would be useful for describing the estimation elements and other concepts that are illustrated here since these more complicated topics can be generalized to using with complex models that require MCMC.

[4] For the purpose of this paper, we will highlight and discuss priors that are normally distributed because they are the most straightforward to illustrate.

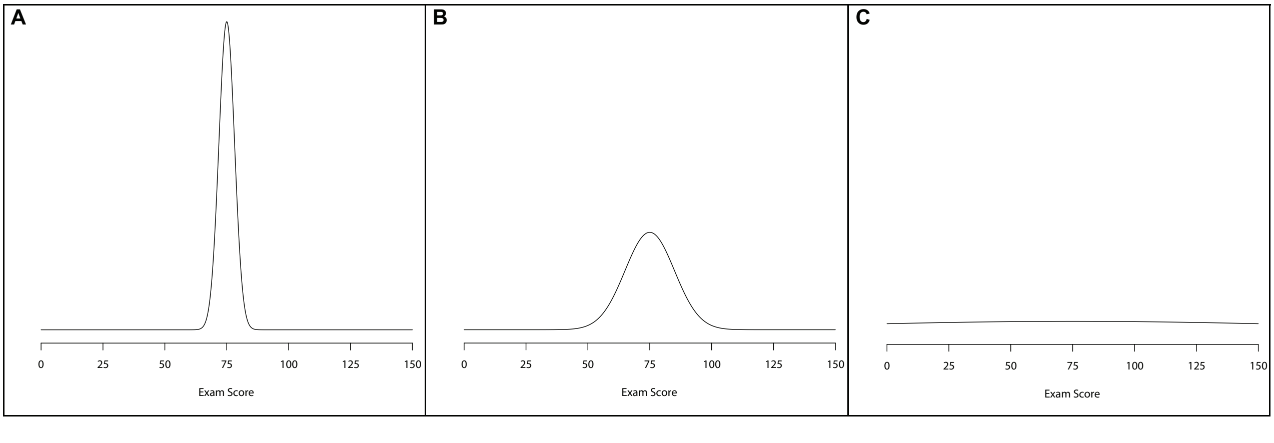


FIGURE 2 | Examples of prior. distributions that are: **(A)** informative, **(B)** weakly informative, and **(C)** diffuse.

Next, we illustrate how priors can impact final model estimates, even for a model as simple as a multiple regression model. Specifically, we conducted a small simulation study illustrating the effect of different prior settings.

Simulation Design

The simulation study used a multiple regression model as displayed in **Figure 1A**. It contained two continuous predictors, a correlation parameter linking these predictors, and a continuous outcome. The population values for these parameters are listed in **Table 1**. In this simulation, we implemented various sets of priors for the regression coefficients linking the two predictors to the outcome. These prior conditions are listed in **Table 1**. Overall, there were 11 prior conditions examined per sample size.

Conditions 1–5 specified informative priors on the regression parameters linking each of the predictors to the outcome. These informative priors were not all *correct* in that some of them had inaccurate mean hyperparameter settings for the prior (i.e., the normal prior was not centered on the population value, rather it was shifted away).[5] Condition 3 is a correct informative prior in that it is centered at the population value and has a relatively narrowed variance. Conditions 1–2 had priors that were shifted downward from the population value, and Conditions 4–5 had priors that were shifted upward.

Conditions 6–10 represented weakly informative priors in that the variance hyperparameter was increased compared to the informative conditions (1–5). The same pattern was exhibited where Condition 8 represented a prior setting with a mean hyperparameter that was accurate to the population value. Conditions 6–7 had mean hyperparameter values that were shifted downward from the truth of the population, and Conditions 9–10 had mean hyperparameters shifted upward.

Finally, Condition 11 represented a diffuse prior, which implemented default settings from M*plus* (Muthén and Muthén, 1998-2017) on the regression parameters. Each of these conditions represented either informative (1–5), weakly informative (6–10), or diffuse priors. Within the informative and weakly informative conditions, we specified (according to the mean hyperparameter) either accurate priors (3 and 8), downward shifted priors (1–2, 6–7), or priors shifted upward from the truth (4–5, 9–10). The goal of these conditions was to highlight the deviation patterns across the sensitivity analysis, with a focus on sensitivity of results to the mean hyperparameter (i.e., accuracy of the mean of the prior) and the variance hyperparameter (i.e., the spread of the prior distribution).

In addition, we also examined the results across three different sample sizes: n = 25, 100, and 1000. These sample sizes ranged from relatively small to relatively large, and they were selected to provide information about how priors impact results differently as sample sizes shift.

TABLE 1 | Population values and simulation conditions for the multiple regression model.

Population values for simulation	
Parameter	**Population value**
Means	
X_1	Fixed to 0[1]
X_2	Fixed to 0
Variances	
X_1	Fixed to 1
X_2	Fixed to 1
Y Intercept	1
Y Resid. Var.	0.5
β_1	1.0
β_2	0.5
Simulation conditions (sample sizes crossed with prior conditions)	
Sample sizes	Prior conditions[2]
n = 25	Informative:
n = 100	(1) $\beta_1 \sim N(0.25, 0.05)$; $\beta_2 \sim N(0.125, 0.05)$
n = 1,000	(2) $\beta_1 \sim N(0.50, 0.05)$; $\beta_2 \sim N(0.250, 0.05)$
	(3) $\beta_1 \sim N(1.00, 0.05)$; $\beta_2 \sim N(0.500, 0.05)$
	(4) $\beta_1 \sim N(2.00, 0.05)$; $\beta_2 \sim N(1.000, 0.05)$
	(5) $\beta_1 \sim N(3.00, 0.05)$; $\beta_2 \sim N(1.500, 0.05)$
	Weakly Informative:
	(6) $\beta_1 \sim N(0.25, 0.1)$; $\beta_2 \sim N(0.125, 0.1)$
	(7) $\beta_1 \sim N(0.50, 0.1)$; $\beta_2 \sim N(0.250, 0.1)$
	(8) $\beta_1 \sim N(1.00, 0.1)$; $\beta_2 \sim N(0.500, 0.1)$
	(9) $\beta_1 \sim N(2.00, 0.1)$; $\beta_2 \sim N(1.000, 0.1)$
	(10) $\beta_1 \sim N(3.00, 0.1)$; $\beta_2 \sim N(1.500, 0.1)$
	Diffuse
	(11) Regression 1 $\sim N(0, 10^{10})$; Slope $\sim N(0, 10^{10})$

Y, the continuous outcome in the model. Resid. Var., residual variance. Predictors 1 and 2 (X_1 and X_2) were both continuous predictors. β_1 = Y on $X_{1.}$ β_2 = Y on $X_{2.}$ [1]The means and variances for the predictors were fixed in the model in order to standardize the predictors. Therefore, estimates are only available for the four remaining parameters. [2]The remaining priors in the model were default diffuse prior settings as implemented in Mplus.

In all, there were 33 cells in this simulation, and we requested 500 iterations per cell. All analyses were conducted in M*plus* version 8.4 (Muthén and Muthén, 1998-2017) using the Bayesian estimation setting with Gibbs sampling. For simplicity, all cells were set up to have a single chain per parameter, with 5,000 iterations in the chain and the first half discarded as the burn-in (i.e., 2,500 iterations were left to form the estimated posterior). Convergence was monitored with the potential scale reduction factor (PSRF, or R-hat; Gelman and Rubin, 1992a,b), and all chains converged for all cells in the design under a setting 1.01 for the convergence criterion. Another index that can be checked is the effective sample size (ESS), which is directly linked to the degree of dependency (or *autocorrelation*) within the chain. Zitzmann and Hecht (2019) recommend that ESSs over 1,000 are required to ensure that there is enough precision in the chain.

However, it is important to keep in mind that priors can be specified using a wide range of distributional forms, including distributions that are not named or not proper distributions (e.g., those that do not integrate/sum to 1.0). We discuss other prior forms for non-normally distributed parameters in the App.

[5] One could argue that if a prior belief dictated a prior that was not centered at the population value that it would be *correct to the theory*. We use the term "correct" in this simulation study to compare a prior that has been centered over the population value (correct) to one that has been shifted away from the population value though a deviant mean hyperparameter (incorrect).

Simulation results indicated that, although the post burn-in portions of the chain were only 2,500 iterations, all of the parameters exceeded the minimum of ESS = 1,000 in the cells examined.[6]

Simulation Findings

Table 2 presents relative percent bias for all model parameters across sample sizes and the 11 prior conditions. Of note, Conditions 3 and 8 represent accurate priors (informative and weakly informative, respectively), and Condition 11 reflects diffuse prior settings. All other priors are either shifted upward or downward, as would be implemented in a sensitivity analysis. Bolded values in the table represent problematic bias levels exceeding ±10% bias.

The most notable finding is how the impact of the priors diminishes as sample size increases. By the time sample size was increased to n = 1,000 (which would be rather large for such a simple model), the prior settings had virtually no impact on findings. However, under the smaller sample sizes, and especially n = 25, we can see a noticeable impact on results. As the priors were shifted for the regression parameters, bias increased in magnitude. This effect occurred in the more extreme conditions even when n = 100, which is not an unreasonable sample size to expect in applied research implementing such a model.

Mean square errors (MSEs) are also presented in **Table 2** for each parameter. MSE represents a measure of variability and bias. Notice that MSE values are quite high for n = 25, but they decrease to a relatively smaller range as sample sizes are increased to n = 100 and beyond. This pattern indicates that sample size has a large role in the efficiency and accuracy of the estimates, as measured through the MSE. In addition, MSEs are much larger for priors that are centered away from the population value.

The practical implication of this simulation highlighted that priors can impact findings (which is indisputable in the Bayesian literature), even when sample sizes are what we might consider to be reasonable. This fact makes sensitivity analyses indispensable when examining the impact of priors on final model results, and examining prior impact is especially important under smaller sample sizes. In practice, researchers do not *know* if subjective priors are accurate to the truth. We argue that researchers should assume that priors have at least some degree of inaccuracy, and they should assess the impact of priors on final model estimates keeping this notion in mind. The only way to truly examine the impact of the prior when working with empirical data is through a sensitivity analysis.

This proof of concept simulation provides a foundation for the Shiny App, which uses empirical data to further illustrate the importance of conducting a sensitivity analysis. In the next section, we present the Shiny App as an educational tool for highlighting the impact of prior settings. A main focus of the App is to illustrate the process of conducting a sensitivity analysis, as well as the type of results that should be examined and reported when disseminating the analysis findings. Specifically, we describe how one would manipulate the settings to examine the impact of priors on final model results. The Shiny App can be used to gain a deeper understanding of the impact of priors, as well as understand the different elements that are needed to properly display sensitivity analysis results.

TABLE 2 | Model parameter estimate percent bias (MSE) for the simulation study.

Condition	*Y* Intercept	*Y* Resid. Var.	β_1	β_2
***n* = 25**				
1	**34.91** (0.0508)	−0.10 (0.2979)	−**40.29** (0.1745)	−**38.50** (0.0474)
2	**21.38** (0.0452)	−0.18 (0.1811)	−**25.66** (0.0774)	−**23.82** (0.0245)
3	**11.90** (0.0403)	−0.34 (0.1239)	−0.05 (0.0115)	1.82 (0.0105)
4	**54.99** (0.0520)	−0.70 (0.5269)	**56.87** (0.3383)	**58.30** (0.0961)
5	**350.76** (0.1468)	−0.68 (14.0520)	**157.51** (2.4857)	**157.86** (0.6300)
6	**22.42** (0.0461)	−0.12 (0.1914)	−**26.50** (0.0919)	−**23.94** (0.0328)
7	**16.66** (0.0434)	−0.18 (0.1513)	−**17.16** (0.0499)	−**14.60** (0.0235)
8	**12.38** (0.0408)	−0.30 (0.1266)	−0.02 (0.0199)	2.48 (0.0184)
9	**30.36** (0.0453)	−0.62 (0.2582)	**36.66** (0.1601)	**38.82** (0.0580)
10	**159.46** (0.0866)	−1.08 (3.3256)	**104.37** (1.1327)	**105.58** (0.3039)
11	**15.37** (0.0426)	−0.20 (0.1451)	0.03 (0.0469)	4.08 (0.0447)
***n* = 100**				
1	4.44 (0.0106)	0.00 (0.0211)	−**13.59** (0.0265)	−**12.58** (0.0111)
2	3.16 (0.0104)	0.02 (0.0194)	−9.06 (0.0160)	−8.06 (0.0086)
3	2.17 (0.0103)	0.04 (0.0183)	−0.26 (0.0076)	0.70 (0.0068)
4	6.37 (0.0107)	0.08 (0.0245)	**17.76** (0.0395)	**18.64** (0.0155)
5	**23.57** (0.0123)	0.14 (0.0908)	**40.44** (0.1742)	**41.22** (0.0502)
6	2.98 (0.0105)	0.00 (0.0192)	−7.55 (0.0151)	−6.46 (0.0094)
7	2.60 (0.0104)	0.00 (0.0187)	−5.12 (0.0119)	−4.04 (0.0087)
8	2.30 (0.0104)	0.02 (0.0184)	−0.30 (0.0091)	0.76 (0.0082)
9	3.52 (0.0105)	0.04 (0.0199)	9.40 (0.0180)	**10.42** (0.0108)
10	7.52 (0.0108)	0.06 (0.0272)	**19.79** (0.0490)	**20.76** (0.0191)
11	2.49 (0.0104)	−0.02 (0.0186)	−0.36 (0.0112)	0.84 (0.0100)
***n* = 1000**				
1	0.27 (0.0010)	−0.36 (0.0020)	−1.33 (0.0012)	−1.56 (0.0010)
2	0.26 (0.0010)	−0.36 (0.0020)	−0.83 (0.0010)	−1.08 (0.0009)
3	0.25 (0.0010)	−0.36 (0.0020)	0.16 (0.0010)	−0.08 (0.0009)
4	0.30 (0.0010)	−0.36 (0.0020)	2.14 (0.0014)	1.90 (0.0010)
5	0.44 (0.0010)	−0.36 (0.0021)	4.12 (0.0027)	3.88 (0.0013)
6	0.26 (0.0010)	−0.36 (0.0020)	−0.59 (0.0010)	−0.84 (0.0010)
7	0.25 (0.0010)	−0.36 (0.0020)	−0.34 (0.0010)	−0.58 (0.0009)
8	0.25 (0.0010)	−0.36 (0.0020)	0.16 (0.0010)	−0.08 (0.0009)
9	0.26 (0.0010)	−0.36 (0.0020)	1.16 (0.0011)	0.92 (0.0010)
10	0.30 (0.0010)	−0.36 (0.0020)	2.16 (0.0014)	1.92 (0.0010)
11	0.25 (0.0010)	−0.36 (0.0020)	0.16 (0.0010)	−0.08 (0.0010)

*Prior conditions (column 1) are described in **Table 1**. MSE, mean square error, which captures a measure of variability accompanied by bias for the simulation estimates. It can be used as a measure that reflects efficiency and accuracy in the simulation results. Bolded values represent percent bias exceeding 10%. Percent bias = [(estimate − population value)/population value] * 100.* β_1 *= Y on* X_1. β_2 *= Y on* X_2.

[6] We selected several cells to examine thoroughly for ESS, all had the lowest sample size (n = 25) and varying degrees of "incorrect" priors. Most parameters had ESS values of 2,500 or nearby, with some parameters lower. However, all ESS values exceeded 1,800 in our investigation. This amount exceeds the recommendation by Zitzmann and Hecht (2019). Therefore, we believe that the chains in the simulation represent adequate precision. ESS values have also been included in the App, which we describe in the example section below.

SENSITIVITY ANALYSIS IN ACTION: AN INTERACTIVE APP

To illustrate the importance and use of prior sensitivity analysis, we created an interactive application using rstan (Stan Development Team, 2020), Shiny (Chang et al., 2020), and RStudio (R Core Team, 2020; RStudio Team, 2020). The App can be accessed online at https://ucmquantpsych.shinyapps.io/sensitivityanalysis/. Alternatively, it is available for download on the Open Science Framework[7]. To run the App on your personal computer, open the *ui.R* and *server.R* files in RStudio and press the "Run App" link in the top-righth and corner of the R Script section of the RStudio window. For more information about Shiny Apps, we refer to RStudio Team (2020).

Our App consists of seven different tabs, with each containing information that will help a user understand how to assess the substantive impact of prior selection. When the App is first loaded, it defaults to the first tab. This tab introduces the App, goes over the main steps of a sensitivity analysis, and describes the other tabs of the App. Within the second tab, a fictional researcher and their study are introduced. Specifically, a researcher has collected a sample of 100 participants to examine whether an individual's sex or lack of trust in others predicts the individual's cynicism (see **Figure 1B** for an illustration of the model). The tab discusses the prior distributions specified by the researcher. While most prior distributions are relatively diffuse (i.e., flat), the researcher specifies an informative prior for the regression effect of cynicism on lack of trust. The remainder of the tab focuses on an evaluation of the posterior results of the original analysis, using trace plots, posterior density plots and histograms, and relevant summary statistics [e.g., posterior mean, SD, 90% highest posterior density interval (HPD interval)].

In the next four tabs, users can specify alternative prior distributions for each parameter in the model: the intercept of cynicism (third tab), the regression effect of cynicism on sex (fourth tab), the effect of cynicism on lack of trust (fifth tab), and the residual variance of cynicism (sixth tab). Within these tabs, the priors for the other parameters are held constant. The user can specify and assess the impact of two alternative prior distributions at a time. Each time a new set of priors is specified, additional analyses are run using the rstan package.[8] The tabs include visual and numerical comparisons that can help assess the impact of the specified prior distributions.

In the seventh tab, users can combine the alternative prior specifications from the previous four tabs to investigate the combined influence of alternative priors on the posterior estimates. Use of the App will be demonstrated in the next section.

Sensitivity Analysis Process

In this section, we will use the Shiny App to execute and report a sensitivity analysis. The first step is to identify the original (comparison) priors that are to be implemented in the investigation. Then the researcher would carry out a sensitivity analysis to examine the robustness of results under different prior specifications. The researcher would specify alternative priors to explore through the sensitivity analysis process. In this section, we will highlight a sensitivity analysis for two parameters in the model, both of which can be captured through the normal distribution. Although there are many distributional forms that priors can take on, the normal distribution is an effective place to start since it is so visually illustrative of the different forms the normal prior can adopt. As a result, we discuss sensitivity analysis in terms of this prior, but it is important to recognize the issues and processes that we highlight can generalize to other distributional forms. For example, a sensitivity analysis for the residual variance of cynicism can also be examined through the App. The prior for this parameter follows an inverse gamma (IG) distribution. In addition to the conjugate distributions (i.e., the prior and posterior distribution are in the same probability distribution family) used in the App, it is also possible to examine non-conjugate priors (e.g., a reference prior). We did not include alternative, non-conjugate, distributions in our App, as we felt it would distract from its main pedagogical purpose. For more information on non-conjugate priors, see Gelman et al. (2014, p. 36+). An example of a write-up for the prior sensitivity analysis can be seen in the **Appendix**.

Specifying Priors on Certain Model Parameters

Priors are specified on all parameters of a model. In this example, we will focus on just two model parameters to illustrate the process of sensitivity analysis. These two parameters are the regression coefficients linking the two predictors to the outcome of *Cynicism*. A separate sensitivity analysis can be conducted on each parameter, and another analysis examines the combined specification of the priors. This latter combined analysis helps to pinpoint the combined impact of a set of alternative priors on all parameters in the model.

Parameter 1: *Cynicism* on *Sex*

The researcher can examine competing prior specifications for the effect of *Cynicism* on *Sex*. For example, if the experts originally assumed that there was no *Sex* effect, then a prior such as $N(0,10)$ could be specified, where the bulk of the distribution is centered around zero. Notice that this prior is weakly informative surrounding zero (i.e., it still contains ample spread about the mean, as opposed to being strictly informative). For the sake of this example, this prior setting can be viewed as the original prior in the analysis.

Alternative prior specifications can be examined through the sensitivity analysis, in order to examine the impact of different priors (perhaps reflecting different substantive theories) on final model results. For example, another theory could state that men (coded as 1) possess higher levels of cynicism than women, suggesting a positive effect. An informative prior centered around a positive value can be explored to examine this prior belief: e.g., $N(5, 5)$. Alternatively, there may be competing research that indicates that men possess lower levels of cynicism than women,

[7] https://osf.io/eyd4r/

[8] Models were estimated through the No-U-Turn sampler (NUTS; Hoffman and Gelman, 2011; Betancourt, 2017) with two chains. Each chain ran for 10,000 iterations, of which 5,000 were discarded as the burn-in.

suggesting a negative effect. An informative prior centered around a negative value can be explored to examine the impact of this prior belief on the posterior results: $N(-10, 5)$. These prior settings would result in an original prior and two alternative specifications such that:

- Original = $N(0, 10)$
- Alternative 1 = $N(5, 5)$
- Alternative 2 = $N(-10, 5)$.

A plot illustrating these prior differences can be found in **Figure 3**.

Parameter 2: *Cynicism* on *Lack of Trust*

For this substantive example predicting cynicism (**Figure 1B**), we can assume that the researchers based their prior distribution specifications on previous research, indicating that *Lack of Trust* had a strong positive relationship with *Cynicism*. Specifically, assume that the original prior (specified by the researchers) was set at $N(6, 1)$, where the value 6 represents the mean hyperparameter (or center) of the distribution and the value 1 represents the variance. This prior density, with a variance hyperparameter of 1, indicates that about 95% of the density falls between 4 and 8. This relatively narrowed prior suggests that the researcher had a relatively strong expectation that a one-point increase in *Lack of Trust* is related to a 4 to 8 point increase in *Cynicism*.

Several competing prior specifications can be imagined for this regression coefficient of *Cynicism* on *Lack of Trust*, each with their own degree of informativeness. The impact of these other prior forms can be examined through a sensitivity analysis. For example, the researcher can examine a diffuse prior distribution, with the intention of downplaying the impact of the prior and emphasizing the data patterns to a larger degree. In this case, a normal distribution can be used as the prior, but the distribution will have a very large spread to coincide with the lack of knowledge surrounding the parameter value. One way of specifying this regression coefficient prior would be as $N(0, 100)$. With such a wide variance (akin to **Figure 2C**), this prior will be largely flat over the parameter space, representing a diffuse prior for this parameter.

Another version of the prior specification can come from an alternative theory on the relationship between *Lack of Trust* and *Cynicism*. Perhaps several experts on the topic of cynicism believe that the degree (or lack) of trust in others has no impact on how cynical a person is. An informative prior centered around zero, with a more narrowed variance compared to the prior described above, reflects this prior belief: $N(0, 5)$.

These prior settings would result in an original prior and two alternative specifications such that:

- Original = $N(6, 1)$
- Alternative 1 = $N(0, 100)$
- Alternative 2 = $N(0, 5)$.

A plot illustrating these prior differences can be found in **Figure 4**.

Examining Priors for Parameter 1 and Parameter 2 Simultaneously

Finally, the combination of each of these alternative prior specifications can also be compared to examine how prior specifications aligned with alternative theories and previous research impact the posterior results. In total, we can use the App to compare six different models at a time.

Assessing Convergence

An alternative prior specification can affect the convergence of parameters in the model. As such, model convergence should always be assessed, even if there were no convergence issues with the original prior specification. A converged chain represents an accurate estimate for the true form of the posterior.

For example, see **Figure 5**, which presents two different plots showing a chain for a single parameter. Each sample pulled from the posterior represents a dot, and these many dots are then connected by a line, which represents the chain. Obtaining stability, or convergence, within the chain is an important element before results can be interpreted. The mean according to the *y*-axis of **Figure 5** represents the mean of the posterior, and the height of the chain represents the amount of variance in the posterior distribution. Convergence is determined by stability in the mean (i.e., horizontal center, according to the *y*-axis) and the variance (i.e., height of the chain). **Figure 5A** shows that there is a great deal of instability in the mean and the variance of this chain.[9] The chain does not have a stable, horizontal center, and the height of the chain is inconsistent throughout. In contrast, **Figure 5B** shows stability in both areas, indicating visually that it appears to have converged. There are statistical tools that can help determine convergence, and they should always accompany visual inspection of plots akin to those in **Figure 5**. Some statistical tools for assessing convergence include the Geweke convergence diagnostic (Geweke, 1992), and the potential scale reduction factor, or R-hat (Gelman and Rubin, 1992a,b; Gelman, 1996; Brooks and Gelman, 1998).

The beginning portion of the chain is often highly dependent on chain starting values (which may be randomly generated within the software). Therefore, this early portion of the chain is often discarded and referred to as the *burn-in* phase. This part of the chain is not representative of the posterior since it can be unstable and highly dependent on the initial value that got the chain started. Only the *post-burn-in* phase (i.e., the phase of the chain beyond the designated burn-in phase) is considered to construct the estimate of the posterior. The user usually defines the length of the burn-in through some statistical diagnostics, while taking into consideration model complexity [e.g., a simple regression model may require a few hundred iterations in the burn-in, but a mixture (latent class) model may require several hundred thousand]. If convergence is not obtained

[9] There are many other elements that should be examined regarding the chains, some of which are levels of autocorrelation and the effective sample size. In the interest of space and the goals of the current tutorial, we refer the reader elsewhere to learn more about these topics. Some helpful resources are: Kruschke (2015) and Depaoli and van de Schoot (2017).

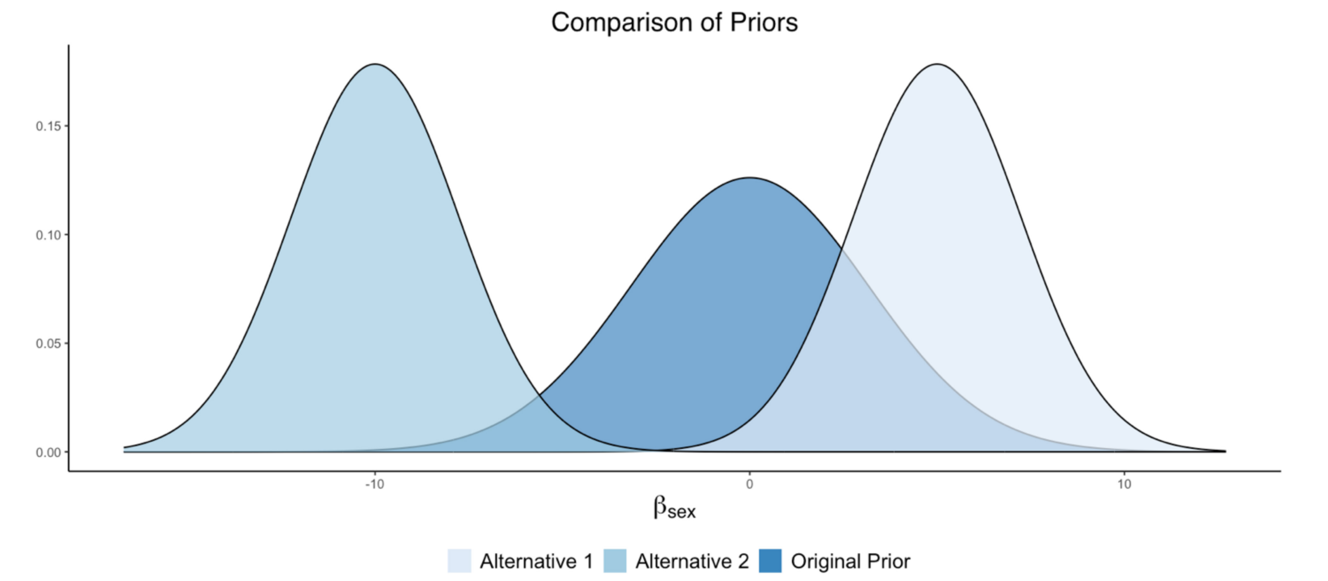


FIGURE 3 | Alternative prior distributions for *Sex* as a predictor of *Cynicism*.

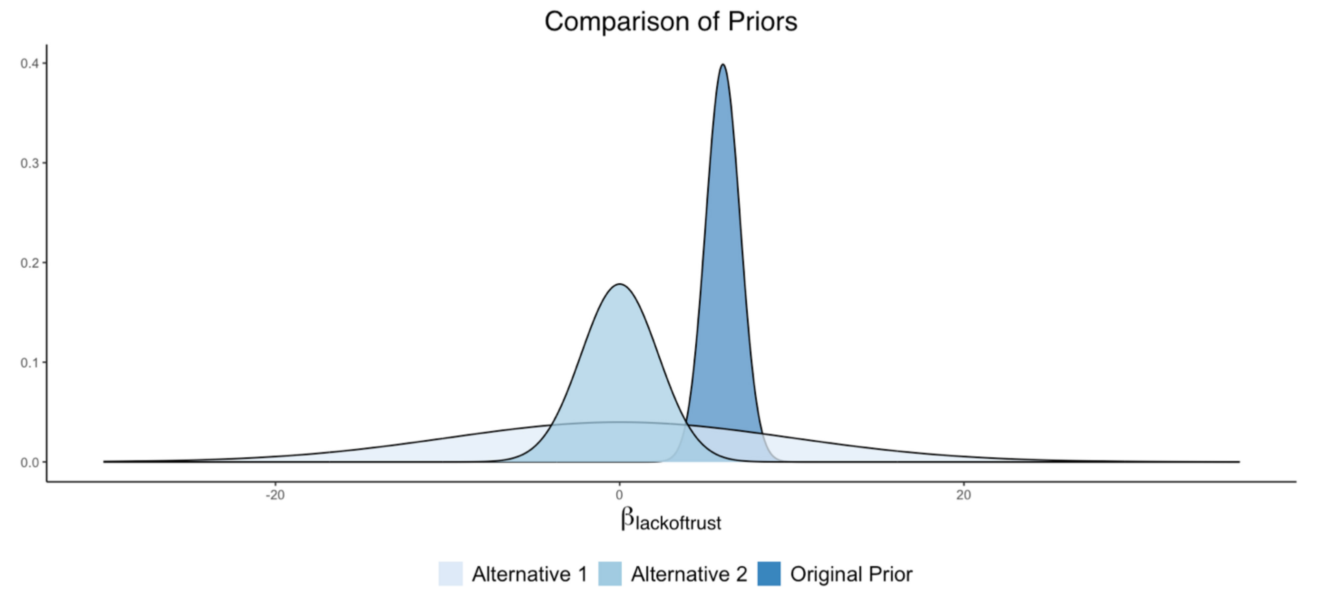


FIGURE 4 | Alternative prior distributions for *Lack of Trust* as a predictor of *Cynicism*.

for a model parameter, then the practitioner can double (or more) the number of iterations to see if the longer chain fixes the issue. If non-convergence still remains, then it may be that the prior is not well suited for the model or likelihood. In the case of a sensitivity analysis, this result could indicate that there is evidence against selecting that particular prior given the current model and likelihood. For more information on convergence and chain length, please see Sinharay (2004) or Depaoli and van de Schoot (2017).

In the App, we evaluated model convergence visually, using trace plots of the posterior chains, and with diagnostics, using R-hat and the ESS.[10] **Figure 6** illustrates that the trace plots, R-hat (<1.01), and ESS (>1,000) for all parameters in the original analysis indicated convergence. For this illustration, **Figure 7** shows the trace plots of an analysis that uses alternative prior specifications for both regression effects: $N(-10, 5)$ for

[10] As mentioned in the section describing the simulation study, the effective sample size (ESS) is directly linked to the degree of dependency (or *autocorrelation*) within each chain. Specifically, the ESS represents the number of independent samples that have the same precision as the total number of autocorrelated samples in the posterior chains. Zitzmann and Hecht (2019) recommend that ESSs over 1,000 are required to ensure that there is enough precision in the chain.

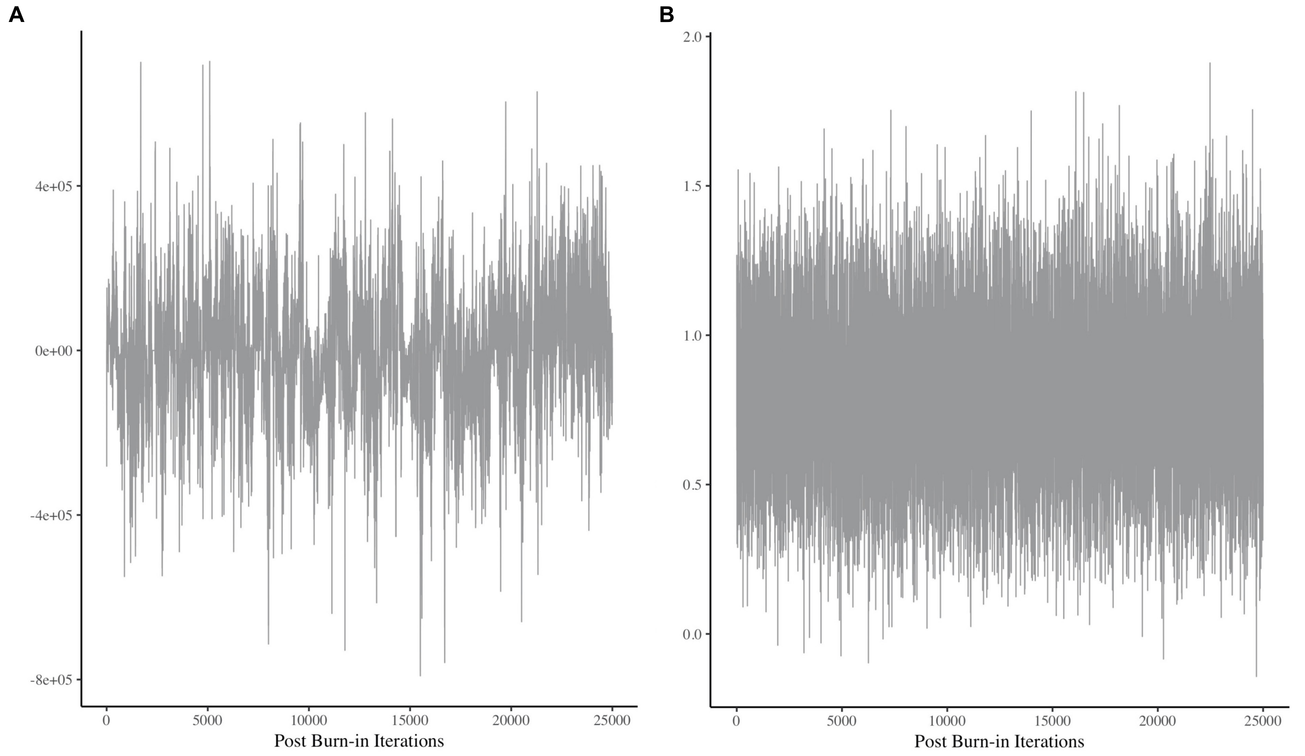


FIGURE 5 | Two chains showing different patterns of (non)convergence. Panel **(A)** shows a great deal of instability throughout the plot, indicating non-convergence. Panel **(B)** shows a relatively stable horizontal mean and variance, indicating convergence. Note that both plots exhibit some degree of autocorrelation, but that is beyond the scope of the current discussion. More information about this issue can be found here: Kruschke (2015) and Depaoli and van de Schoot (2017).

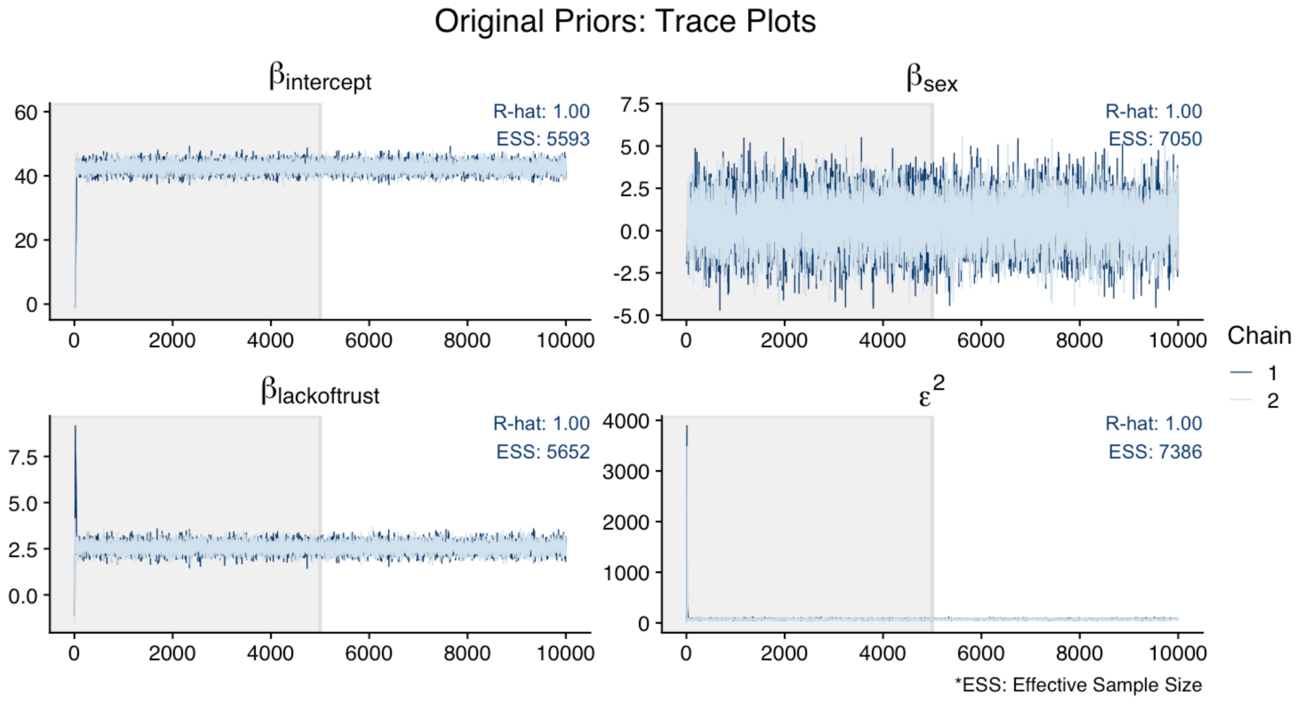


FIGURE 6 | Trace plots of original analysis.

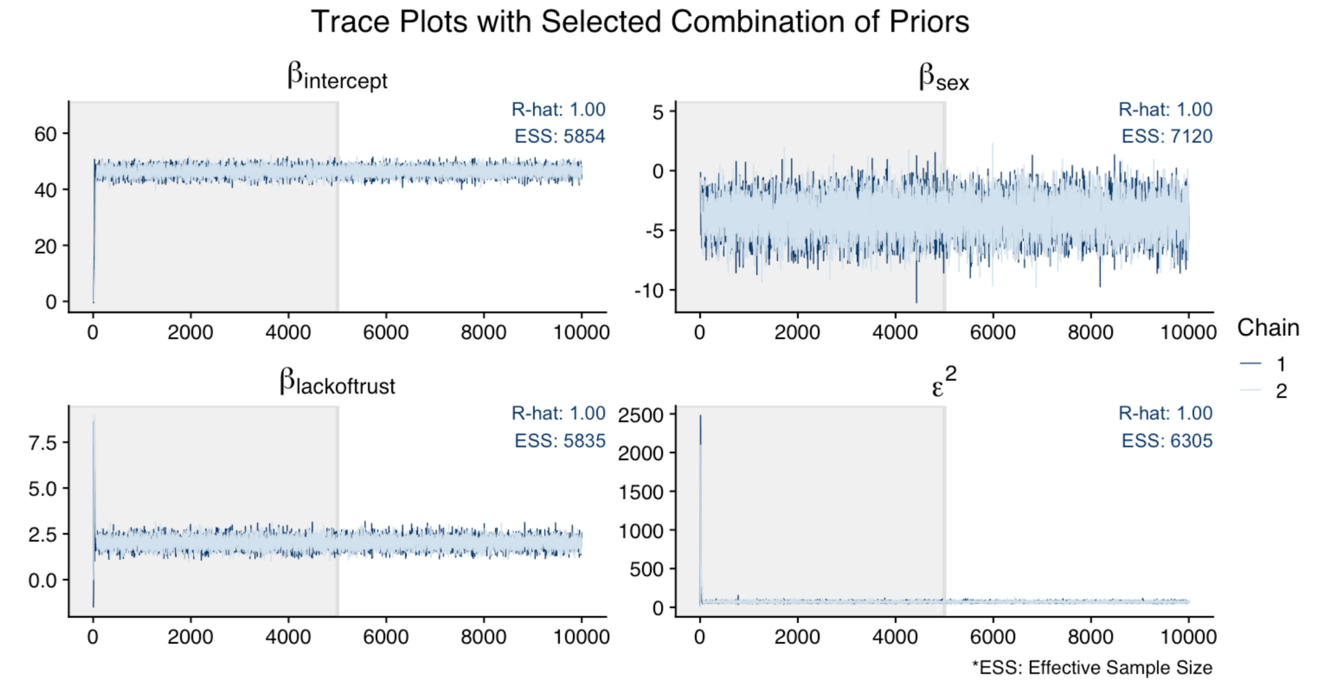


FIGURE 7 | Trace plots of analysis with *N*(–10, 5) prior distribution for *Sex* as a predictor of *Cynicism* and *N*(0, 5) for *Lack of Trust* as a predictor of *Cynicism*.

Sex as a predictor of *Cynicism*, and *N*(0, 5) for *Lack of Trust* as a predictor of *Cynicism*. In this figure, we can see that the trace plot for the effect of *Sex* looks more volatile (though still relatively flat) when using this alternative prior specification; this is most evident by examining the *y*-axis differences across **Figures 6**, **7**. Overall it appears that the alternative priors do not profoundly affect chain convergence, despite some differences with the variance of the chain for the *Cynicism* on *Sex* coefficient (i.e., the variance is wider in **Figure 6** for this parameter).

Inspecting Posterior Density Plots

The next step in the sensitivity analysis is to examine how the alternative prior specifications have affected the posterior distributions of the model parameters. If the posterior distributions are very similar across a range of prior distributions, then it implies that the posterior estimate is robust to different prior distributions. In contrast, if the posterior distribution is drastically altered as a result of an alternative prior, then it shows that the posterior distribution depends more heavily on the specific prior distribution used. For this illustration, we will focus our discussion of the two alternative prior distributions for *Lack of Trust* as a predictor of *Cynicism*. **Figure 8** shows that the posterior distribution for the effect of *Lack of Trust* changes as a result of the alternative prior specifications. Both posterior distributions shift to a lower range of values. This result implies that the posterior distribution of the original analysis is affected by the selected prior distribution and that alternative (more diffuse) prior distributions would have resulted in slightly different posterior distributions. In addition, the posterior distribution of the intercept of *Cynicism* shifts to a higher value for both alternative prior distributions, indicating a substantively different definition of the model intercept (i.e., the average value of *Cynicism* when predictors are zero). Finally, the posterior distributions of *Sex* as a predictor of *Cynicism* does not appear to be affected by the alternative priors for the effect of *Lack of Trust*, while the residual variance of *Cynicism* was impacted.

Comparing the Posterior Estimates

Another way to examine the impact of the prior distribution is to compute the percentage deviation in the average posterior estimate between models with different prior distributions. For this illustration, we will again focus our discussion on the two alternative prior distributions for *Lack of Trust* as a predictor of *Cynicism*. **Figure 9** displays summary statistics of the analyses with the alternative prior specifications, as pulled from the App. The final two columns show the average posterior estimates of the original analysis and the percentage deviation between the original and each alternative analysis. In line with the downward shift of the posterior densities of the effect of *Lack of Trust* across the different prior specifications, the percentage deviation is −23.040% or −24.851%, depending on the alternative prior specification. Another way of capturing the impact of the prior distribution is to compare the 90% HPD intervals and see whether the substantive conclusion about the existence of the effect of *Lack of Trust* changes. In this case, zero is always outside the 90% HPD interval, independent of the prior distribution used in the analysis. Thus, the substantive conclusion regarding the role of *Lack of Trust*

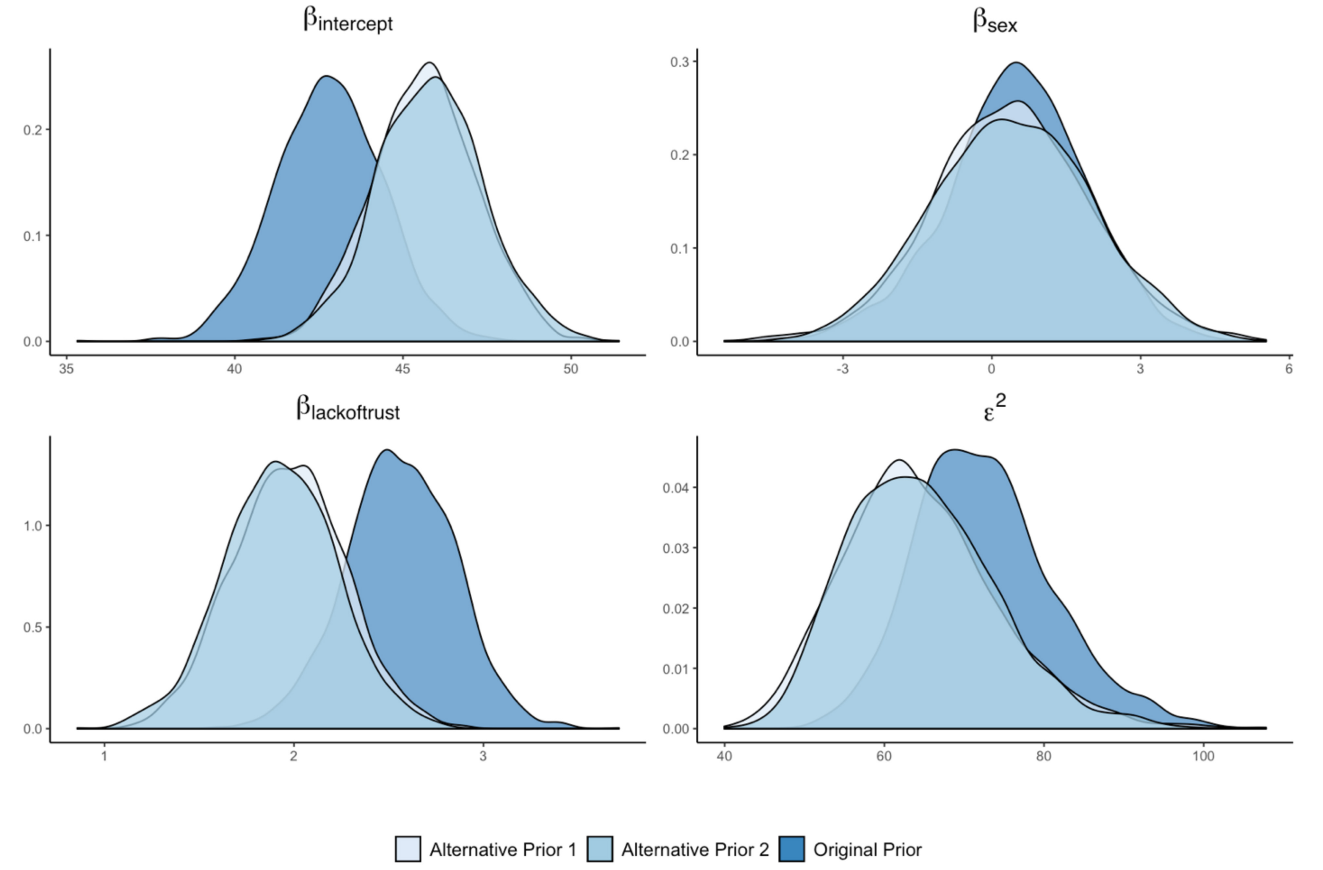


FIGURE 8 | Posterior density plots for original and alternative priors for *Lack of Trust* as a predictor of *Cynicism*.

as a predictor of *Cynicism* does not change across the prior distributions examined here.

Additional Guidelines for Using the App

We constructed the App so that users cannot examine the combination of different priors in the model before specifying and looking at each one separately. This design-based decision was made for pedagogical reasons. We feel that examining each prior separately is helpful when initially learning about prior impact. The practice of modifying a prior setting and tracking how the posterior changes provides a visual learning experience that enhances discussions surrounding sensitivity analyses. However, in practice, the implementation and variation of priors is more complicated. In the final model being estimated, the combination of priors is the main aspect that matters. There is research highlighting that priors in one location in a model can impact results in another location (see e.g., Depaoli, 2012). Because of this, it is important to examine results with the combination of priors implemented all at once. These results reflect the true impact of the prior settings (as opposed to examining a single parameter at a time). Although this App allows the user to examine one prior at a time (as a learning tool), we note that this may not be a feasible practice in some modeling contexts. For example, some item response theory models have thousands of parameters, and it would only be feasible to examine the combination of priors (rather than one at a time).

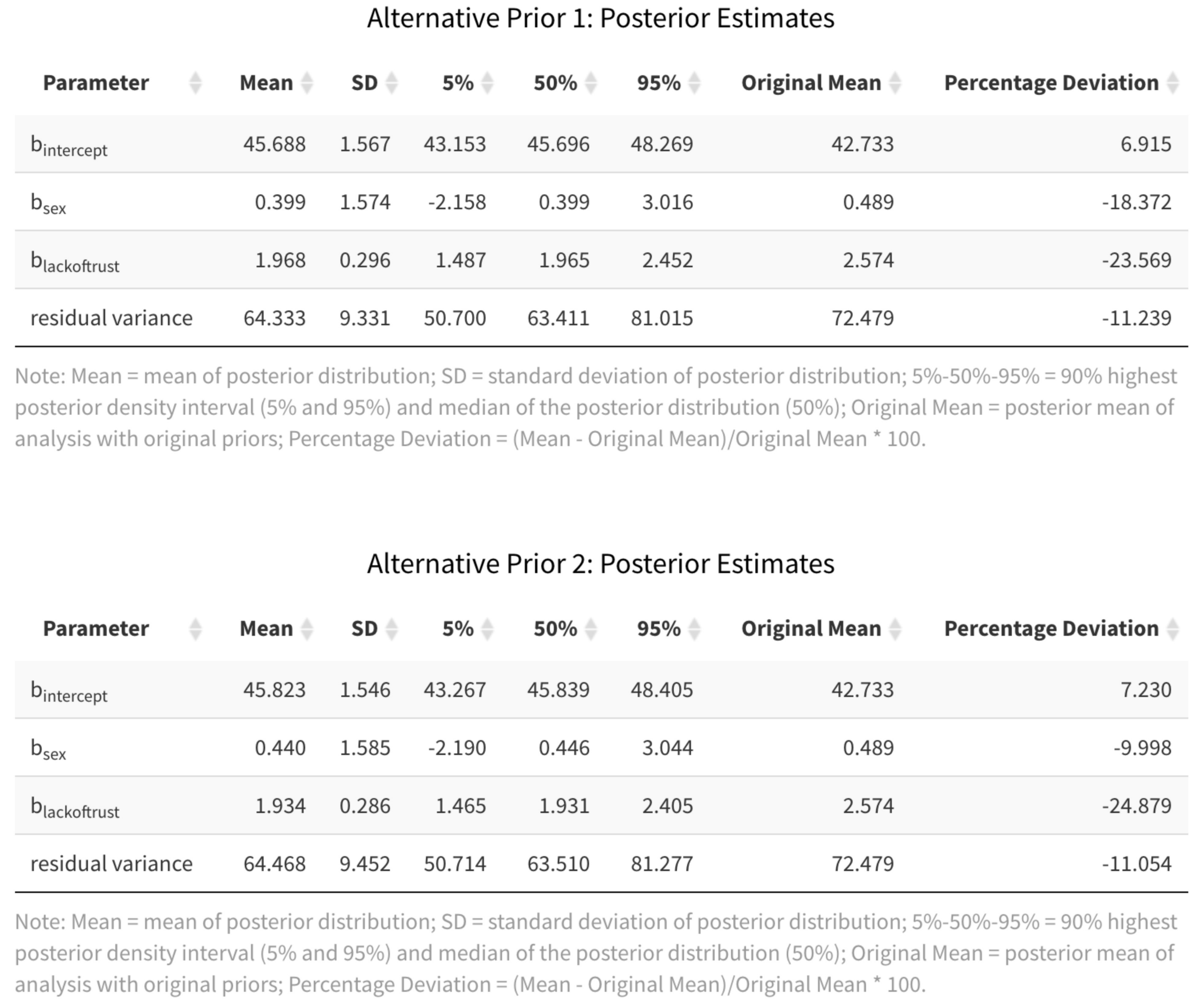
The App was designed to enhance pedagogy surrounding visually demonstrating sensitivity analysis. However, we caution the reader that it is indeed the combination of prior settings that drives the substantive impact of the priors.

CONCLUSION

Our aim was to present examples (via simulation and application) illustrating the importance of a prior sensitivity analysis. We presented a Shiny App that aids in illustrating some of the important aspects of examining sensitivity analysis results. We have formatted the current section to address frequently asked questions (FAQs) in order to provide an at-a-glance view of the most important components for applied researchers to focus on.

Frequently Asked Questions About Prior Sensitivity Analysis

(1) Why is a sensitivity analysis important within the Bayesian framework, and what can we learn from it?



Alternative Prior 1: Posterior Estimates

Parameter	Mean	SD	5%	50%	95%	Original Mean	Percentage Deviation
$b_{intercept}$	45.688	1.567	43.153	45.696	48.269	42.733	6.915
b_{sex}	0.399	1.574	-2.158	0.399	3.016	0.489	-18.372
$b_{lackoftrust}$	1.968	0.296	1.487	1.965	2.452	2.574	-23.569
residual variance	64.333	9.331	50.700	63.411	81.015	72.479	-11.239

Note: Mean = mean of posterior distribution; SD = standard deviation of posterior distribution; 5%-50%-95% = 90% highest posterior density interval (5% and 95%) and median of the posterior distribution (50%); Original Mean = posterior mean of analysis with original priors; Percentage Deviation = (Mean - Original Mean)/Original Mean * 100.

Alternative Prior 2: Posterior Estimates

Parameter	Mean	SD	5%	50%	95%	Original Mean	Percentage Deviation
$b_{intercept}$	45.823	1.546	43.267	45.839	48.405	42.733	7.230
b_{sex}	0.440	1.585	-2.190	0.446	3.044	0.489	-9.998
$b_{lackoftrust}$	1.934	0.286	1.465	1.931	2.405	2.574	-24.879
residual variance	64.468	9.452	50.714	63.510	81.277	72.479	-11.054

Note: Mean = mean of posterior distribution; SD = standard deviation of posterior distribution; 5%-50%-95% = 90% highest posterior density interval (5% and 95%) and median of the posterior distribution (50%); Original Mean = posterior mean of analysis with original priors; Percentage Deviation = (Mean - Original Mean)/Original Mean * 100.

FIGURE 9 | Posterior estimates for the alternative priors for *Lack of Trust* as a predictor of *Cynicism*.

A sensitivity analysis is, in many ways, one of the *most important* elements needed to fully understand Bayesian results in an applied research setting. The simulation study, and the demonstration provided in the Shiny App, showed that priors can have a substantial impact on the posterior distribution. Without a sensitivity analysis, it is not possible to disentangle the impact of the prior from the role that the data play in the model estimation phase. A sensitivity analysis can help the researcher understand the influence of the prior compared to the influence of the data. In other words, this analysis can help to establish how much theory [i.e., through informed theory or lack of theory (e.g., diffuse priors)] influences the final model results, and how much the results are driven by patterns in the sample data.

(2) How many different prior conditions should I test during a sensitivity analysis? In other words, how extensive should the sensitivity analysis be?

There is a running saying (or joke) in statistics that the answer to any statistical question is “it depends.” That saying certainly holds true here. In this case, there is no definitive answer to this question, and it really depends on several factors. The extensiveness of the sensitivity analysis will depend on the complexity of the model, the intended role of the priors (e.g., informative versus diffuse), and the substantive question(s) being asked. There are some general guidelines that we can provide. For example, if diffuse priors are implemented in the original analysis, then it will likely not be relevant to include informative priors in the sensitivity analysis. Instead, the practitioner would be better off testing different forms of diffuse priors. However, if informative priors were used in the original analysis, then it would be advised to examine different forms of the informative priors, as well as diffuse prior settings, in the sensitivity analysis. The practitioner must heavily weigh these different aspects and decide on the scope of the sensitivity analysis accordingly. The main goal here is to understand the impact and role that each prior is playing. There are no set rules for achieving this goal since all research scenarios will differ in substantive ways.

(3) What is the best way to display sensitivity analysis results?

Not to borrow too much from the previous FAQ, but the answer to this current question depends on: (1) what the sensitivity analysis results are showing, (2) model complexity—i.e., the number of model parameters, and (3) the number of conditions examined in the sensitivity analysis. In a case where results are relatively similar across a variety of prior conditions, the researcher may opt to have a couple of sentences indicating the scope of the sensitivity analysis and that results were comparable. However, in a case where results are altered when priors differ (e.g., like some of the examples provide in our Shiny App), the researcher may opt for a larger display of results. This could be provided through visuals, akin to the Shiny App plots we presented (e.g., **Figures 3**, **4**, **8**, **9**), or it may be in a table format indicating the degree of discrepancy in estimates or HPD intervals across parameters. In extreme cases, where there are dozens of parameters crossed with many sensitivity analysis conditions, the researcher may need to put the bulk of the results in an online appendix and just narrate the findings in the manuscript text. Much of this will depend on the degree of the differences observed across the sensitivity analysis, as well as journal space limitations. The important issue is that results must be displayed in some clear fashion (through text, visuals, or tables of results), but what this looks like will depend largely on the nature of the investigation and findings that were obtained.

(4) How should I interpret the sensitivity analysis results?

Sensitivity analysis results are not meant to change or alter the final model results presented. Instead, they are helpful for properly interpreting the impact of the prior settings. This can be valuable for understanding how much influence the priors have, as well as how robust final model estimates are to differences in prior settings—whether they be small or large differences in the priors. Sensitivity analysis results should be reported alongside the final model estimates obtained (i.e., those obtained from the original priors implemented). These results can be used to help bolster the discussion section, as well as make clearer sense of the final estimates. In addition, we discussed an alternative above regarding reporting sensitivity analysis results when diffuse priors are implemented. In this scenario, the practitioner may choose to report results across a range of diffuse priors as the final analysis. This is a strategy that can help illuminate any uncertainty surrounding the exact prior specification if different forms of diffuse priors provide varying results. Finally, if the sensitivity analysis process yields a prior (or set of priors) that produce non-sensical results according to the posterior (e.g., the posterior does not make sense, see Depaoli and van de Schoot, 2017), or results in chains that do not converge, then it may be an indication of a poor prior choice given the model or likelihood. In this case, the prior and results should be described, and it may be useful to describe why this prior setting may not be viable given the poor results that were obtained.

(5) What happens if substantive results differ across prior settings implemented in the sensitivity analysis?

It may initially seem uncomfortable to receive results from the sensitivity analysis that indicate priors have a strong influence on final model estimates. However, this is really not a point of *concern*. Assume sensitivity analysis results indicated that even a slight fluctuation of the prior settings altered the final model results in a meaningful (i.e., substantive) manner. This is an important finding because it may indicate that the exact theory used to drive the specification of the prior (potentially) has a large impact on final model results. Uncovering this finding can help build a deeper understanding about how stable the model (or theory) is. In contrast, if the model results are relatively stable under different prior settings, then this indicates that theory (i.e., the prior) has less of an impact on findings. Either way, the results are interesting and should be fully detailed in the discussion. Understanding the role that priors play will ultimately help lead to more refined and informed theories within the field.

(6) How do I write up results from a sensitivity analysis?

Sensitivity analysis results should be included in the main body of the results section of any applied Bayesian paper. Final model estimates can be reported and interpreted based on the original prior settings implemented. Then the sensitivity analysis can be reported in the context of building a deeper understanding of the impact of the priors. Bayesian results can only be fully understood in the context of the impact of the particular prior settings implemented. After reporting the final model estimates from the original prior settings, a section can be added to the results entitled something like: "Understanding the Impact of the Priors." In this section, visual or table displays of the sensitivity analysis results should be included. Results of the analysis should be described, and some sense of the robustness (or not!) of results to different prior settings should be addressed. These results can then be further expanded upon in the discussion section, and recommendations can be made about what priors the researcher believes should be further explored in subsequent research. The goal is to provide a thorough treatment of the analysis and give readers ample information in order to assess the role of priors in that particular modeling context.

Final Thoughts

As we demonstrated through the simulation study and the Shiny App, priors can have a noticeable impact on the final model results obtained. It is imperative that applied researchers examine the extent of this impact thoroughly and display findings in the final analysis report. Visual aids can be a tremendous asset when presenting sensitivity analysis finding, as they quickly point toward the level of (dis)agreement of results across different prior settings.

A key issue when reporting any analysis, but especially one as complicated as a Bayesian analysis, is transparency. It is important to always be clear about what analyses were conducted, how they were conducted, and how results can be interpreted. This issue of transparency is key within any statistical framework, but it is especially an issue for the Bayesian framework because of how *easy* it is to manipulate results by changing prior settings. Bayesian methods are very useful tools, and it is up to us (i.e., the users, publishers, and consumers of research) to set a precedence of transparency and thoroughness when reporting findings. It is our hope that the Shiny App will play a role in promoting the importance of this issue.

AUTHOR CONTRIBUTIONS

SD conceptualized and wrote the manuscript. SW and MV made the Shiny App. All authors contributed to the article and approved the submitted version.

REFERENCES

Adeyemo, D. A. (2007). Moderating influence of emotional intelligence on the link between academic self-efficacy and achievement of university students. *Psychol. Dev. Soc.* 19, 199–213. doi: 10.1177/097133360701900204

Betancourt, M. (2017). A conceptual introduction to Hamiltonian Monte Carlo. *arXiv [preprint]* Available online at: https://arxiv.org/abs/1701.02434 (accessed September 10, 2020),

Brooks, S. P., and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *J. Comput. Graph. Stat.* 7, 434–455. doi: 10.1080/10618600.1998.10474787

Chang, W., Cheng, J., Allaire, J. J., Xie, Y., and McPherson, J. (2020). *shiny: Web Application Framework for R. R package version 1.5.0*. Available online at: https://CRAN.R-project.org/package=shiny (accessed September 10, 2020).

Cieciuch, J., Davidov, E., Schmidt, P., Algesheimer, R., and Schwartz, S. H. (2014). Comparing results of an exact vs. an approximate (Bayesian) measurement invariance test: a cross-country illustration with a scale to measure 19 human values. *Front. Psychol.* 5:982. doi: 10.3389/fpsyg.2014.00982

Depaoli, S. (2012). Measurement and structural model class separation in mixture-CFA: ML/EM versus MCMC. *Struct. Equat. Model.* 19, 178–203. doi: 10.1080/10705511.2012.659614

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

Depaoli, S., and Clifton, J. (2015). A Bayesian approach to multilevel structural equation modeling with continuous and dichotomous outcomes. *Struct. Equat. Model.* 22, 327–351. doi: 10.1080/10705511.2014.937849

Depaoli, S., Rus, H., Clifton, J., van de Schoot, R., and Tiemensma, J. (2017a). An introduction to Bayesian statistics in health psychology. *Health Psychol. Rev.* 11, 248–264. doi: 10.1080/17437199.2017.1343676

Depaoli, S., and van de Schoot, R. (2017). Improving transparency and replication in Bayesian statistics: the WAMBS-checklist. *Psychol. Methods* 22, 240–261. doi: 10.1037/met0000065

Depaoli, S., Yang, Y., and Felt, J. (2017b). Using Bayesian statistics to model uncertainty in mixture models: a sensitivity analysis of priors. *Struct. Equat. Model.* 24, 198–215. doi: 10.1080/10705511.2016.1250640

Gelman, A. (1996). "Inference and monitoring convergence," in *Markov chain Monte Carlo in practice*, eds W. R. Gilks, S. Richardson, and D. J. Spiegelhalter (New York: Chapman and Hall), 131–143.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D., Vehtari, A., and Rubin, D. B. (2014). *Bayesian Data Analysis (3rd ed.)*. Boca Raton, FL: Chapman and Hall.

Gelman, A., and Rubin, D. B. (1992a). Inference from iterative simulation using multiple sequences. *Stat. Sci.* 7, 457–511. doi: 10.1214/ss/1177011136

Gelman, A., and Rubin, D. B. (1992b). "A single series from the Gibbs sampler provides a false sense of security," in *Bayesian Statistics 4*, eds J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (Oxford: Oxford University Press), 625–631.

Geweke, J. (1992). "Evaluating the accuracy of sampling-based approaches to calculating posterior moments," in *Bayesian Statistics 4*, eds J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (Oxford: Oxford University Press), 169–193.

Ghosh, J. K., and Mukerjee, R. (1992). "Non-informative priors (with discussion)," in *Bayesian Statistics*, 4 Edn, eds J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (Oxford: Oxford University Press), 195–210.

Golay, P., Reverte, I., Rossier, J., Favez, N., and Lecerf, T. (2013). Further insights on the French WISC-IV factor structure through Bayesian structural equation modeling. *Psychol. Assess.* 25, 496–508. doi: 10.1037/a0030676

Gustafson, P., and Wasserman, L. (1995). Local sensitivity diagnostics for Bayesian inference. *Ann. Stat.* 23, 2153–2167. doi: 10.1214/aos/1034713652

Heitjan, D. F., Guo, M., Ray, R., Wileyto, E. P., Epstein, L. H., and Lerman, C. (2008). Identification of pharmacogenetic markers in smoking cessation therapy. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* 147, 712–719. doi: 10.1002/ajmg.b.30669

Hoffman, M. D., and Gelman, A. (2011). The no-U-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *arXiv*[Preprint] Available online at: https://arxiv.org/abs/1111.4246 (accessed September 10, 2020),

Kerr, N. L. (1998). HARKing: hypothesizing after the results are known. *Pers. Soc. Psychol. Rev.* 2, 196–217. doi: 10.1207/s15327957pspr0203_4

Kim, S. Y., Suh, Y., Kim, J. S., Albanese, M., and Langer, M. M. (2013). Single and multiple ability estimation in the SEM framework: a non-informative Bayesian estimation approach. *Multiv. Behav. Res.* 48, 563–591. doi: 10.1080/00273171.2013.802647

Kopala-Sibley, D. C., Zuroff, D. C., Leybman, M. J., and Hope, N. (2013). Recalled peer relationship experiences and current levels of self-criticism and self-reassurance. *Psychol. Psychother.* 86, 33–51. doi: 10.1111/j.2044-8341.2011.02044.x

Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *J. Exp. Psychol. Gen.* 142, 573–603. doi: 10.1037/a0029146

Kruschke, J. K. (2015). *Doing Bayesian Analysis: A Tutorial with R, Jags, and STAN*. San Diego, CA: Academic Press.

Lambert, P. C., Sutton, A. J., Burton, P. R., Abrams, K. R., and Jones, D. R. (2005). How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. *Stat. Med.* 24, 2401–2428. doi: 10.1002/(ISSN)1097-0258

Luyster, F. S., Chasens, E. R., Wasko, M. C. M., and Dunbar-Jacob, J. (2011). Sleep quality and functional disability in patients with Rheumatoid Arthritis. *J. Clin. Sleep Med.* 7, 49–55. doi: 10.5664/jcsm.28041

McCartan, C. (2020). *adjustr: Stan Model Adjustments and Sensitivity Analyses using Importance Sampling. R package version 0.1.1.* Available online at: https://corymccartan.github.io/adjustr/ (accessed October 15, 2020).

Müller, U. K. (2012). Measuring prior sensitivity and prior informativeness in large Bayesian models. *J. Monet. Econ.* 59, 581–597. doi: 10.1016/j.jmoneco.2012.09.003

Muthén, B., and Asparouhov, T. (2012). Bayesian structural equation modeling: a more flexible representation of substantive theory. *Psychol. Methods* 17, 313–335. doi: 10.1037/a0026802

Muthén, L. K., and Muthén, B. (1998-2017). *Mplus user's Guide. Eighth edition*. Los Angeles, CA: Muthén and Muthén.

Natarajan, R., and McCulloch, C. E. (1998). Gibbs sampling with diffuse proper priors: a valid approach to data-driven inference? *J. Comput. Graph. Stat.* 7, 267–277. doi: 10.1080/10618600.1998.10474776

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Roos, M., Martins, T. G., Held, L., and Rue, H. (2015). Sensitivity analysis for Bayesian hierarchical models. *Bayesian Anal.* 10, 321–349. doi: 10.1214/14-ba909

RStudio Team. (2020). *RStudio: Integrated Development for R*. Boston, MA: RStudio, Inc.

Sheng, Y. (2010). A sensitivity analysis of Gibbs sampling for 3PNO IRT models: effects of prior specifications on parameter estimates. *Behaviormetrika* 37, 87–110. doi: 10.2333/bhmk.37.87

Sinharay, S. (2004). Experiences with Markov chain Monte Carlo convergence assessment in two psychometric examples. *J. Educ. Behav. Stat.* 29, 461–488. doi: 10.3102/10769986029004461

Stan Development Team (2020). *RStan: the R Interface to Stan. R package version 2.21.2*. Available online at: http://mc-stan.org/ (accessed September 10, 2020)

van de Schoot, R., Sijbrandij, M., Depaoli, S., Winter, S., Olff, M., and van Loey, N. (2018). Bayesian PTSD-trajectory analysis with informed priors based on a systematic literature search and expert elicitation. *Multiv. Behav. Res.* 53, 267–291. doi: 10.1080/00273171.2017.1412293

van de Schoot, R., Winter, S., Zondervan-Zwijnenburg, M., Ryan, O., and Depaoli, S. (2017). A systematic review of Bayesian applications in psychology: the last 25 years. *Psychol. Methods* 22, 217–239. doi: 10.1037/met0000100

van Erp, S., Mulder, J., and Oberski, D. L. (2018). Prior sensitivity analysis in default Bayesian structural equation modeling. *Psychol. Methods* 23, 363–388. doi: 10.1037/met0000162

Zhang, Z., Hamagami, F., Wang, L., Grimm, K. J., and Nesselroade, J. R. (2007). Bayesian analysis of longitudinal data using growth curve models. *Int. J. Behav. Dev.* 31, 374–383. doi: 10.1177/0165025407077764

Zitzmann, S., and Hecht, M. (2019). Going beyond convergence in Bayesian estimation: why precision matters too and how to assess it. *Struct. Equat. Model. Multidiscipl. J.* 26, 646–661. doi: 10.1080/10705511.2018.1545232

Zondervan-Zwijnenburg, M. A. J., Depaoli, S., Peeters, M., and van de Schoot, R. (2019). Pushing the limits: the performance of ML and Bayesian estimation with small an unbalanced samples in a latent growth model. *Methodology* 15, 31–43. doi: 10.1027/1614-2241/a000161

Zondervan-Zwijnenburg, M. A. J., Peeters, M., Depaoli, S., and van de Schoot, R. (2017). Where do priors come from? Applying guidelines to construct informative priors in small sample research. *Res. Hum. Dev.* 14, 305–320. doi: 10.1080/15427609.2017.1370966

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

APPENDIX

Prior Sensitivity Analysis Example Write Up

The following section represents a hypothetical write-up of sensitivity analysis results, which mimics the example provided in the Shiny App.

For the first step of our sensitivity analysis, we considered the parameters of most substantive interest in our study. In the case of our regression example, we were particularly interested in the regression coefficients associated with *Sex* as a predictor of *Cynicism* and *Lack of Trust* as a predictor of *Cynicism*. After clearly identifying the parameters of interest, we then identified the most appropriate priors for the original (comparison) priors in the analysis. For example, we selected $N(0{,}10)$ as the prior for *Sex* as a predictor of *Cynicism* and $N(6{,}1)$ as the prior for *Lack of Trust* as a predictor of *Cynicism*. The $N(0{,}10)$ priors suggest *Cynicism* and *Sex* are unrelated, and the $N(6{,}1)$ indicates a positive relationship between *Cynicism* and *Lack of Trust*. In addition to selecting priors for the parameters of substantive interest, we also set $N(41{,}10)$ as prior to the intercept and $IG(0.5{,}0.5)$ as the prior for the residual variance.

To understand the impact of different priors on the posterior distribution, we identified a set of alternative priors to compare to each of our original priors. For our regression example, we selected the alternative priors of $N(5{,}5)$ and $N(-10{,}5)$ for *Sex* predicting *Cynicism*. The $N(5{,}5)$ alternative prior suggests that men have a higher degree of cynicism than women, and the $N(-10{,}5)$ alternative prior means men have a lower degree of cynicism than women. Also, we selected alternative priors $N(0{,}100)$ and $N(0{,}5)$ for *Lack of Trust* predicting *Cynicism*. The $N(0{,}100)$ alternative was much more diffuse than the original prior, suggesting a lack of knowledge about the parameter. The $N(0{,}5)$ has a mean of zero, which indicates no relationship between *Cynicism* and *Lack of Trust*. For the intercept, we selected $N(0{,}100)$ and $N(20{,}10)$ as alternative priors. The $N(0{,}100)$ prior is a diffuse, flat prior, and the $N(20{,}10)$ shits the mean of the original prior downward. Both priors suggest lower cynicism values. For the residual variance, we selected $IG(1{,}0.5)$ and $IG(0.1{,}0.1)$ as alternative priors. The $IG(1{,}0.5)$ is more informative than the original prior, and $IG(0.1{,}0.1)$ is more diffuse than the original prior. Finally, we also specified combinations of these alternative priors to understand the combined impact of different priors on model results.

After selecting our alternative priors, we estimated a series of models with different priors. Each model was checked for convergence via visual inspection of the trace plots, as well as through the R-hat diagnostic. In addition, effective sample sizes (ESSs) were also monitored to ensure that autocorrelation was not problematic. The alternative priors selected yielded adequate model convergence and ESS values. Therefore, we moved to the next step of the sensitivity analysis and inspected the posterior density plots. A visual inspection of the posterior density plots revealed a change in the posterior distributions for *Lack of Trust* predicting *Cynicism* when specifying alternative priors. Specifically, the posterior distribution for *Lack of Trust* predicting *Cynicism* shifts to lower values under both alternative priors, suggesting the prior specification impacts the results. The posterior distribution of the intercept and residual variance of *Cynicism* changed depending on the priors specified, which indicates a substantively different interpretation of the intercept depending on the priors. In contrast, the posterior density plots for *Sex* as a predictor of *Cynicism* were relatively similar, regardless of the alternative prior specification.

We also examined how robust the results were by comparing the posterior estimates across models with different prior specifications. If priors have little impact on the results, then there will be a low percentage of deviation in the posterior estimates between models. However, if the priors have a significant effect, then we will see a higher percentage deviation between models. As expected, given the posterior density plots, we see a downward shift in the estimate for *Lack of Trust* as a predictor of *Cynicism* across different prior specifications. Specifically, the percentage deviation is −23.040% or −24.851%, depending on the alternative prior specification.

Further evidence of the impact of the prior on the posterior distribution can be obtained by comparing the 90% highest posterior density (HPD) intervals. If the substantive conclusions regarding a parameter change depending on the prior, then there is evidence of less robust results. In the case of *Lack of Trust* as a predictor of *Cynicism*, zero is always outside the 90% HPD interval, independent of the prior distribution used in the analysis. Thus, the substantive conclusion regarding the role of *Lack of Trust* as a predictor of *Cynicism* does not change across the prior distributions. This is perhaps the most critical finding of the sensitivity analysis. Although some parameters were more readily impacted in the model by the prior distributions specified, the substantive interpretation of model results did not change depending on the prior specified."

CONCEPTUAL ANALYSIS
published: 11 December 2020
doi: 10.3389/fpsyg.2020.611963

Dangers of the Defaults: A Tutorial on the Impact of Default Priors When Using Bayesian SEM With Small Samples

Sanne C. Smid[1]* *and Sonja D. Winter*[2]

[1] Department of Methodology and Statistics, Utrecht University, Utrecht, Netherlands, [2] Department of Psychological Sciences, University of California, Merced, Merced, CA, United States

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

Keywords: Bayesian SEM, default priors, informative priors, small sample size, Shiny app

Edited by:
Christoph Koenig,
Goethe University Frankfurt, Germany

Reviewed by:
Christoph Helm,
University of Teacher Education Zug, Switzerland
Martin Hecht,
Humboldt University of Berlin, Germany

Correspondence:
Sanne C. Smid
s.c.smid@uu.nl

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *29 September 2020*
Accepted: *20 November 2020*
Published: *11 December 2020*

Citation:
Smid SC and Winter SD (2020) Dangers of the Defaults: A Tutorial on the Impact of Default Priors When Using Bayesian SEM With Small Samples. Front. Psychol. 11:611963. doi: 10.3389/fpsyg.2020.611963

Bayesian estimation of Structural Equation Models (SEMs) has gained popularity in the last decades (e.g., Kruschke et al., 2012; van de Schoot et al., 2017), and is more and more often used as a *solution* to problems caused by small sample sizes (e.g., McNeish, 2016a; König and van de Schoot, 2017)[1]. With small samples, frequentist estimation [such as (restricted) Maximum Likelihood or (weighted) least squares estimation] of SEMs can result in non-convergence of the model, which means that the estimator was unable to find the maximum (or minimum) for the derivative of the model parameters. Even when a model converges, simulation studies have shown that the parameter estimates may be inadmissible (e.g., Heywood cases) or inaccurate (i.e., the estimate deviates from the population value; Boomsma, 1985; Nevitt and Hancock, 2004). In contrast

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

to frequentist methods, Bayesian methods do not rely on large sample techniques, which make Bayesian methods an appealing option when only a small sample is available. Within the Bayesian framework, prior distributions need to be specified for all parameters in the model[2]. This additional step may pose a barrier for novice users of Bayesian methods. To make Bayesian SEM accessible to a broad audience, popular software programs for analyzing Bayesian SEMs, such as M*plus* (Muthén and Muthén, (1998–2017)) and the blavaan package (Merkle and Rosseel, 2018) in R (R Core Team, 2018), offer default prior distributions. However, those default prior distributions are not suitable in all cases. When samples are small, the use of solely default priors can result in inaccurate estimates—particularly severely inaccurate variance parameters—unstable results, and a high degree of uncertainty in the posterior distributions (e.g., Gelman, 2006; McNeish, 2016a; Smid et al., 2019b). These three consequences of using default priors with small samples severely limit the inferences that can be drawn about the parameters in the model.

With small samples, the performance of Bayesian estimation highly depends on the prior distributions, whether they are software defaults or specified by the researcher (e.g., Gelman et al., 2014; Kaplan, 2014; McElreath, 2016). McNeish (2016a) discussed that small sample problems (such as non-convergence, inadmissible and inaccurate parameter estimates) cannot be fixed by only switching from a frequentist to a Bayesian estimator. Instead, he argues that if Bayesian methods are used with small samples, “prior distributions must be carefully considered” (McNeish, 2016a, p. 764). This advice is not new: Kass and Wasserman (1996) already warned against relying on default prior settings with small samples. In the quarter-century since that initial warning, Bayesian estimation is increasingly used to deal with small samples (van de Schoot et al., 2017; Smid et al., 2019b). Yet researchers remain stubbornly reliant on default priors, despite clear caution against their use (as shown by McNeish, 2016a; König and van de Schoot, 2017; van de Schoot et al., 2017).

Goals of This Tutorial Paper

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

WHAT IS A SMALL SAMPLE?

Before we continue our discussion of the potential dangers of default priors with small samples, we need to address the question: What exactly *is* a small sample? Whether a sample is small depends on the complexity of the model that is estimated. One way to express the size of a sample is to look at the ratio between the number of observations and the number of unknown parameters in the model (e.g., Lee and Song, 2004; Smid et al., 2019a). A sample could be considered very small when this ratio is 2, which means there are just two observations for each unknown parameter. As SEMs often include many unknown parameters (i.e., factor loadings, intercepts, covariances), samples that may appear relatively large are in fact very small. For example, a confirmatory factor analysis (CFA) model with three latent factors and fifteen observed items consists of 48 unknown parameters: 12 factor loadings (first factor loading fixed at 1 for identification), 15 intercepts, 15 residual variances, three factor variances, and three factor covariances. In this scenario, a sample of 100 participants would still be considered very small (ratio = 2.08). This example demonstrates that general rules of thumb about sample sizes for SEM (e.g., n > 100; Kline, 2015) can be misleading as they do not take into account model complexity. Furthermore, model complexity depends on more than just the number of parameters that are estimated. Other factors that play are role in model complexity are whether the model includes components such as categorical variables, latent factors, multiple groups, or latent classes. A recent review of simulation studies on SEM (Smid et al., 2019b) showed that authors of these simulation papers have widely varying definitions of a “small sample size,” ranging from extremely small (e.g., n = 8 assessed at three time points with one continuous variable; van de Schoot et al., 2015) to what some might consider moderately sized (e.g., n = 200 with 12 ordinal variables; Chen et al., 2015). Thus, assessing whether a sample is (too) small is unfortunately not as easy as checking whether a certain number of participants has been reached, and should be done on an analysis-by-analysis basis.

DANGERS OF THE DEFAULTS

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

First, when samples are small, priors have a relatively larger impact on the posterior than when samples are large. The posterior can be seen as a compromise between the prior and the likelihood. With a larger sample size, the likelihood dominates the posterior (see **Figure 1C**). However, with a small sample size, the likelihood has relatively less weight on the posterior. Accordingly, the prior has relatively more weight on the posterior (see **Figure 1A**).

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

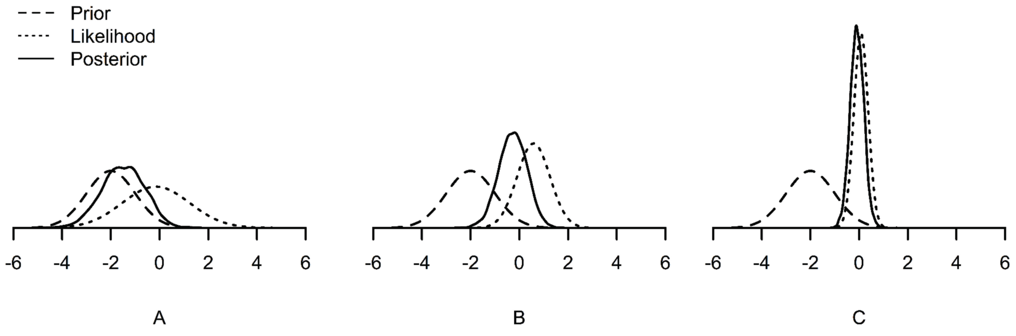


FIGURE 1 | Examples of prior, likelihood and posterior distributions under small **(A)**, medium **(B)**, and large **(C)** sample sizes. The posterior distribution is dominated by the prior under the small sample size **(A)**, and dominated by the likelihood under the large sample size **(C)**.

Therefore, it is of great importance to specify suitable prior distributions when samples are small (e.g., Gelman et al., 2014).

Second, most of the default priors have very wide distributions. For instance, the M*plus* default prior for means and regression coefficients is a Normal distribution with a mean hyperparameter of zero and a variance of 10^{10} (Muthén and Muthén, (1998–2017)). The variance hyperparameter corresponds to a standard deviation of 100.000, meaning, that 68% of the prior distribution contains values between −100.000 and 100.000, and 95% of the prior distribution contains values between −200.000 and 200.000[3]. When such default priors are specified, a wide range of parameter values can be sampled from the posterior during the Bayesian analysis. All those parameter values are therefore considered plausible, which might not always be appropriate. For instance, when measuring mathematical ability on a scale from 0 to 100, values below 0 and above 100 cannot be present in the data. Specifying a default prior with such a wide distribution on the mean of mathematical ability will put a lot of weight on values that are not reasonable (see e.g., Stan Development Team, 2017 p. 131). For small sample sizes, the combination of the relatively larger impact of the prior on the posterior *and* the wide distribution of default priors can lead to extremely incorrect parameter estimates (see e.g., Gelman, 2006; McNeish, 2016a; and the systematic literature review of Smid et al., 2019b).

The third factor that plays a role, is the *false belief* that default priors are non-informative priors which "let the data speak." Default priors can *act* as highly informative priors, as they can heavily influence the posterior distribution and impact the conclusions of a study (see e.g., Betancourt, 2017). As explained by McNeish (2016a, p. 752): *"with small samples, the idea of non-informative priors is more myth than reality (. . .)."* The terminology of informative and non-informative priors can therefore be confusing (see also Bainter, 2017, p. 596). In addition, different software programs use different default priors (see **Table 1**). van Erp et al. (2018, p. 26) investigated the performance of multiple default priors and concluded that, especially with small samples, all investigated default priors performed very differently, and *"that there is not one default prior that performed consistently better than the other priors (. . .)."* The choice of software could thus unintentionally influence the results of a study (see e.g., Holtmann et al., 2016), which is problematic if one is not aware of this. Note that we are not advocating against default priors in general. Default priors can be suitable—even when samples are small—in cases where all values in the prior distribution are reasonable and can occur in the data (for example values around 100,000 or 200,000 are realistic in housing price data, see e.g., LeGower and Walsh, 2017). However, the use of default priors is problematic when researchers assume they let "the data speak" while in reality they "let the default priors speak," meaning that the priors can heavily impact the results without one being aware of this.

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

SHINY APP: THE IMPACT OF DEFAULT PRIORS

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

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

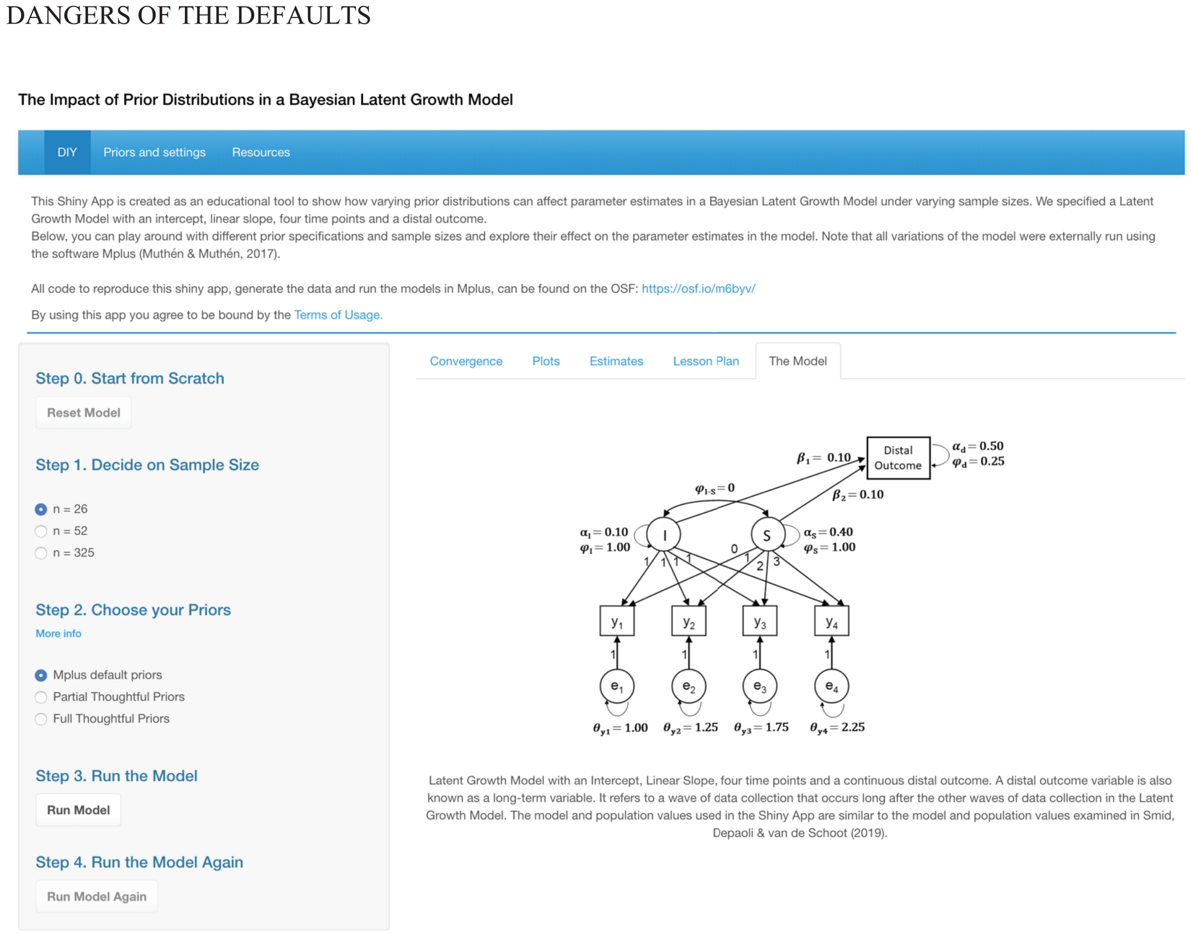
TABLE 1 | Overview of default prior distributions of main parameters for the software program Mplus and the use of Mplus, JAGS and Stan via the R package blavaan.

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

Default priors corresponding to Mplus version 8.4 (see Asparouhov and Muthén, 2010), and blavaan version 0.3-8 (see Merkle, 2019). Prior distributions in Mplus are placed on the variance, while the prior distributions in blavaan are by default placed on the precisions (the inverse of the variance) unless stated otherwise. Abbreviations in order of appearance: N, Normal distribution with hyperparameters mean μ and variance σ^2; I, Identity Matrix; IG, Inverse Gamma; G, Gamma; IW, Inverse Wishart; W, Wishart; B, Beta distribution.

[1] *The prior for the observed and latent variable parameters is placed on the standard deviation (the square root of the variance).*

[2] *In blavaan, three MCMC packages can be used (target = "stan," "stanclassic" and "jags") for the analysis. For all the MCMC packages, the same default priors are specified, with one exception: for target = "jags," a different prior for the covariance is specified.*



DANGERS OF THE DEFAULTS

The Impact of Prior Distributions in a Bayesian Latent Growth Model

This Shiny App is created as an educational tool to show how varying prior distributions can affect parameter estimates in a Bayesian Latent Growth Model under varying sample sizes. We specified a Latent Growth Model with an intercept, linear slope, four time points and a distal outcome.
Below, you can play around with different prior specifications and sample sizes and explore their effect on the parameter estimates in the model. Note that all variations of the model were externally run using the software Mplus (Muthén & Muthén, 2017).

All code to reproduce this shiny app, generate the data and run the models in Mplus, can be found on the OSF: https://osf.io/m6byv/

By using this app you agree to be bound by the Terms of Usage.

Step 0. Start from Scratch

Step 1. Decide on Sample Size

● n = 26
○ n = 52
○ n = 325

Step 2. Choose your Priors

● Mplus default priors
○ Partial Thoughtful Priors
○ Full Thoughtful Priors

Step 3. Run the Model

Step 4. Run the Model Again

Latent Growth Model with an Intercept, Linear Slope, four time points and a continuous distal outcome. A distal outcome variable is also known as a long-term variable. It refers to a wave of data collection that occurs long after the other waves of data collection in the Latent Growth Model. The model and population values used in the Shiny App are similar to the model and population values examined in Smid, Depaoli & van de Schoot (2019).

FIGURE 2 | Main page of the Shiny app, where users can interactively explore the impact of prior settings and sample size in a Bayesian Latent Growth Model.

app were externally run using the software M*plus* (Muthén and Muthén, (1998–2017)) to enhance the user experience[4].

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

The Model, Sample Sizes, and Priors Used in the Shiny App

The model, sample sizes, and prior settings used in the Shiny app are based on Smid et al. (2019a). Specifically, the model is a latent growth model (LGM) with a latent intercept and linear slope, four time points, and a continuous long-term variable (i.e., distal outcome) that is predicted by the latent intercept and slope (see **Figure 3**). A long-term variable is a variable that is collected at a wave of assessment that occurs long after the other waves of assessment in the LGM. An example of a distal outcome is young adult levels of depression that are predicted by conduct and emotional problems at ages 4–16 (Koukounari et al., 2017). Users can select one of three sample sizes: 26, 52, 325, which represent a very small, small, and relatively large sample for the model of interest, which has 13 unknown parameters.

Three different prior specifications are included in the app: one specification using software default priors and two specifications with increasing numbers of thoughtful priors. The default priors that we selected are those specified in M*plus* (Muthén and Muthén, (1998–2017)) and are called "M*plus* default priors" in the Shiny app. The two thoughtful prior specifications, called "Partial Thoughtful Priors" and "Full Thoughtful Priors," were taken from Smid et al. (2019b), details of which are included on the second page of the Shiny app. In short, "Partial Thoughtful Priors" includes informative priors for the mean of the intercept and slope of the LGM, the regression coefficients, and the intercept of the distal outcome. "Full Thoughtful Priors" includes informative priors on all parameters in the model, with the exception of the residual variances. These two specifications reflect scenarios where a researcher has access to prior knowledge regarding some or most of the parameters in the model.

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

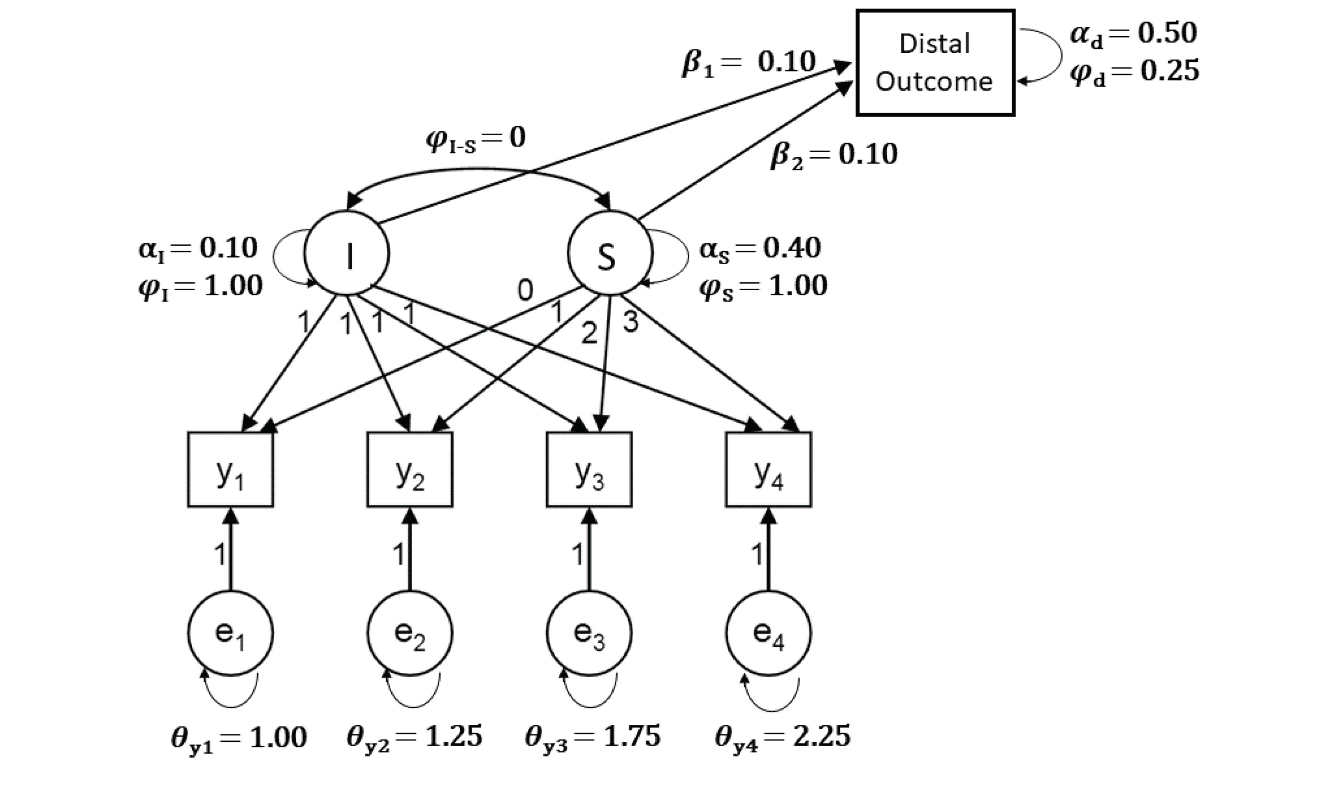


FIGURE 3 | The Latent Growth Model with a distal (long-term) outcome variable that is used in the Shiny app, including population values (model and population values based on Smid et al., 2019a).

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

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

The two thoughtful prior specifications included in the app are just two examples of how thoughtful priors can be included in Bayesian SEM. Other sources that can be used for specifying thoughtful priors include previous research, meta-analyses, or knowledge from experts in the field (for in-depth discussions of these topics, we refer to Zondervan-Zwijnenburg et al., 2017; Lek and van de Schoot, 2018; van de Schoot et al., 2018). Even if prior knowledge is not readily available, researchers can think about impossible and implausible values for the parameters and specify prior distributions that only contain information about the typical range of the parameters. To illustrate this idea, imagine that the distal outcome of the LGM shown in **Figure 3** was measured with a questionnaire that had a range from 0 to 20. A researcher could use this information to specify a prior for the intercept of the distal outcome that makes values outside of that range highly improbable [e.g., $N(10, 15)$]. For some parameters, it may be challenging to identify prior hyperparameters that will exclude implausible values. For example, the inverse Gamma distribution is often used as a prior for the (residual) variance parameters. The parameters of this distribution, called shape and scale, are not as easily interpreted and thoughtfully specified as the mean and variance of a normal distribution. Fortunately, methods for specifying thoughtful prior hyperparameters for the inverse Gamma distribution have been suggested (e.g., Zitzmann et al., 2020). Alternatively, researchers may decide to switch to a different distribution altogether (van Erp et al., 2018). Examples include the half-Cauchy prior (Gelman, 2006; Polson and Scott, 2012) or reference priors such as Jeffrey's prior (Tsai and Hsiao, 2008).

Using the Shiny App as a Teacher

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

GUIDELINES: HOW TO RECOGNIZE A (MIS)BEHAVING PRIOR?

To formulate suitable prior distributions *and* to check afterward whether the priors are "behaving," information is needed about the reasonable range of values for the parameters in the model. This information can be based on previous studies, the scale or questionnaire that is used, or expert knowledge from the field. In our reading list (available via https://osf.io/pnmde), we provide an overview of relevant literature on how to specify suitable priors based on multiple sources of information. Below, we discuss four ways to identify a (mis)behaving prior after conducting a Bayesian analysis (see also **Table 2**), by inspecting for all parameters the (a) effective sample size, (b) trace plots, (c) prior-likelihood-posterior distributions, and (d) the posterior standard deviation and 95% highest posterior density.

Effective Sample Size

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

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

TABLE 2 | Possible signs of "misbehaving" priors.

Effective sample size

- Low effective sample size (i.e., $< 1,000$) can be a first indication that the priors are problematic

Trace plots

- Spikes: shape of alien communication captured in a sci-fi movie instead of a fat caterpillar
- Highly improbable values for the parameter on the y-axis based on information about the reasonable range of values about parameters
- Chains that are not overlapping

Prior-likelihood-posterior comparison

- Substantial deviation between prior, likelihood and/or posterior: e.g., a posterior that is much narrower or wider than the prior and likelihood, while taking into account the amount of information in the prior (i.e., level of informativeness of the prior) and in the likelihood (i.e., sample size)

Posterior SD and 95% HPD

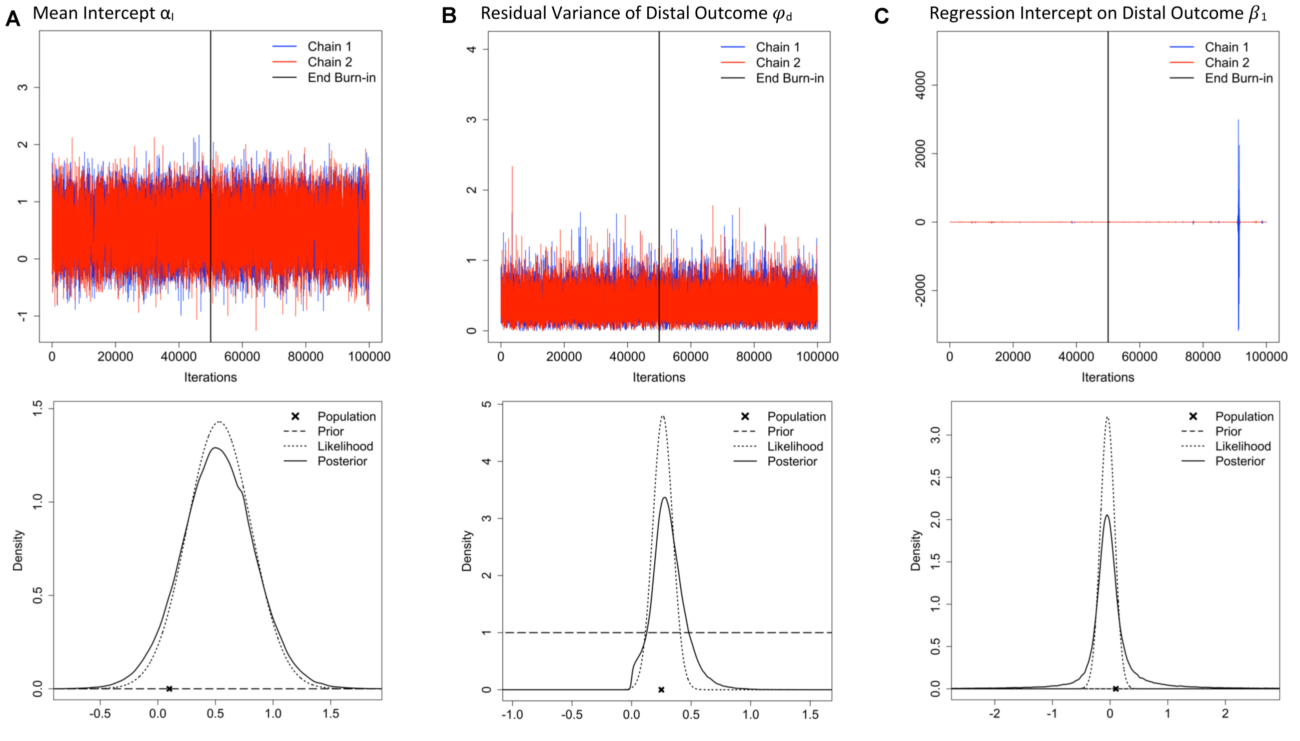
- Much smaller or larger posterior SD or 95% HPD than expected based on the amount of information in the prior (i.e., level of informativeness of the prior) and in the likelihood (i.e., sample size) distribution. Autocorrelation is undesirable as it increases the uncertainty in posterior estimates. If autocorrelation within the chains is low, then the ESS approaches the total number of samples in the posterior chains, and the posterior distribution will be more precise and more likely to approximate the parameter estimate well (Zitzmann and Hecht, 2019). If autocorrelation within the chains is high, a larger number of samples will be necessary to reach an adequate ESS. A low ESS can be the first indicator that there might be a misbehaving prior. Multiple recommendations have been made about how to assess whether the ESS is *too* low: Zitzmann and Hecht (2019) recommend that ESSs should ideally be over 1,000 to ensure that there is enough precision in the chain. It is also possible to compute a lower bound for the number of effective samples required using a desired level of precision and the credible interval level of interest (Vats et al., 2019; Flegal et al., 2020). Finally, it can also be helpful to look at the ratio of the ESS to the total number of samples, where a ratio < 0.1 indicates that there are high levels of autocorrelation in the chains (although this does not necessarily indicate that the posterior distribution is not precise; Gabry et al., 2019). A low ESS can serve as the first clue that something might be wrong, but even if all ESSs appear acceptable, plots and posterior estimates should be inspected to further confirm if priors are behaving.

Trace Plots

Three characteristics of a trace plot can indicate a misbehaving prior. First, the shape of the trace plot: If the multiple chains are well-behaved, the chains should resemble the hungry caterpillar after 6 days of eating (see **Figure 4A**). A misbehaving prior can result in trace plots that exhibit spikes, closely resembling alien communication captured in a sci-fi movie (**Figure 4C**). Second, do the values that are covered by the posterior make sense for this parameter, or is the y-axis stretched to cover unrealistic values? Even when subtle spikes are present (**Figure 4B**), the y-axis range could show that the chains are drawing improbable values from the posterior distribution and should be given extra attention. Third, a lack of overlap of the chains can indicate a misbehaving prior. When the chains do not overlap, it indicates that they are sampling from different parts of the posterior distribution and are not converging toward the same location.

Prior-Likelihood-Posterior Comparison

One important aspect of our Shiny app is that the prior, likelihood, and posterior distributions are visualized to make



Prior = N(0, 10^{10})
Posterior SD = 0.32
95% HPD = -0.10; 1.15

Prior = IG(-1, 0)*
Posterior SD = 0.14
95% HPD = 0.02; 0.58

Prior = N(0, 10^{10})
Posterior SD = 94.64
95% HPD = -1.54; 1.73

FIGURE 4 | Traceplots; prior, likelihood, posterior plots; posterior standard deviation (SD) and 95% highest posterior density interval (HPD) for three parameters: mean intercept **(A)**, residual variance of the distal outcome **(B)** and the regression effect of the slope on the distal outcome **(C)** under sample size $n = 26$ and M*plus* default priors (examples retrieved from the Shiny app). *The M*plus* default prior for residual variance parameters is IG(−1, 0), which is improper (i.e., does not integrate to 1) and has a constant density of 1 on the interval (−∞, ∞) (Asparouhov and Muthén, 2010).

comparisons across different priors and sample size settings easy[6]. When there is a substantial deviation between the prior, likelihood and posterior distributions, results should be interpreted with caution, especially when the sample size is small. Researchers should decide how much impact of the prior and likelihood on the posterior is desirable. Is it preferable that the posterior is a compromise between the prior and likelihood, or that the posterior is dominated by one of two? For instance, when the likelihood and the prior deviate a lot, one might not want to trust the posterior results[7]. In case of small samples, the results might especially be driven by the prior distributions. This is only desirable when researchers trust the specified prior distributions, not when they are defaults of the software program. **Figure 4** shows the prior-likelihood-posterior comparison for three parameters. Although the prior distributions (dashed lines) look completely flat, default prior distributions were used for all parameters. In **Figure 4A**, the posterior (solid line) closely follows the likelihood distribution (dotted line), which is desirable here because the default prior (dashed line) is specified and we do not want it to impact the posterior much. In **Figures 4B,C**, the posteriors seem to have tails that are too fat (kurtotic) compared to the likelihood distribution and the flat default priors, and results should therefore be inspected further.

Posterior SD and 95% HPD

The posterior standard deviation (SD) and 95% credible (or highest posterior density; HPD) interval can be inspected to assess whether the estimates are unusually certain or uncertain. Uncertainty is demonstrated by a large posterior SD and a wide 95% HPD.

Available information about reasonable values for the parameters as well as the amount of information in the prior and likelihood should be used to assess whether the level of (un)certainty of the posterior is reasonable. For instance, in **Figure 4C**, a posterior SD of 94.64 is reported, which is a much higher value than would be expected for a regression estimate and implies that some very extreme values were likely sampled from the posterior. This level of uncertainty is also reflected by the extreme spikes in the trace plot and the kurtotic posterior distribution. The parameters depicted in **Figure 4** illustrate that the combination of a non-informative prior and a small sample size does not always lead to problems across all parameters in a model. It is important to note that even if it appears that the priors of the main parameter(s) of interest are behaving well, a misbehaving prior that is located elsewhere in the model may lead to inaccuracies in the posterior estimates of the main parameters. For example, in a multilevel SEM with a between-level covariate effect, the between-level variance estimate may not be of substantive interest. However, a supposedly non-informative prior [IG(0.001, 0.001)] for the between-level variance parameter can turn into a misbehaving prior when the amount of variance located at the between-level is large (Depaoli and Clifton, 2015). In a simulation study, Depaoli and Clifton (2015) showed that this misbehaving prior resulted in a biased posterior estimate of the between-level covariate effect. A researcher who only inspected the trace plot for the between-level covariate effect may not have realized that their results were negatively affected by a prior placed on between-level variance parameter. For that reason, it is critical to always examine all parameters in the SEM.

What to Do If You Suspect a Misbehaving Prior?

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

Reporting of Bayesian SEM

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

AN ILLUSTRATION: THE IMPACT OF DEFAULT PRIORS

To illustrate the impact of prior settings and sample size—and the informative behavior of default priors with a small sample size—we retrieved the trace plots, prior-likelihood-posterior plots, and posterior SDs from the Shiny app for a single parameter: the regression effect of the distal outcome regressed on the linear slope (β_2 in **Figure 3**). The plots (**Figure 5**) show signs of a misbehaving prior when samples are small (n = 26, or 52 for this model) when default priors are used. Specifically, the trace plots exhibit spikes that reach highly improbable values for the regression coefficient, the plots have a stretched y-axis, and show chains that are not overlapping. Moreover, the prior-likelihood-posterior plots for the two small sample sizes show that the posterior distribution (solid line) is wider than the likelihood estimate (dotted line). Overall, the plots displayed in **Figure 5** show that default priors, which are assumed to be non-informative, can impact the results when samples are small.

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

[7] For readers interested in the impact of so-called prior-data conflict, we refer to simulation studies by Depaoli (2014); Holtmann et al. (2016), and Smid et al. (2019a).

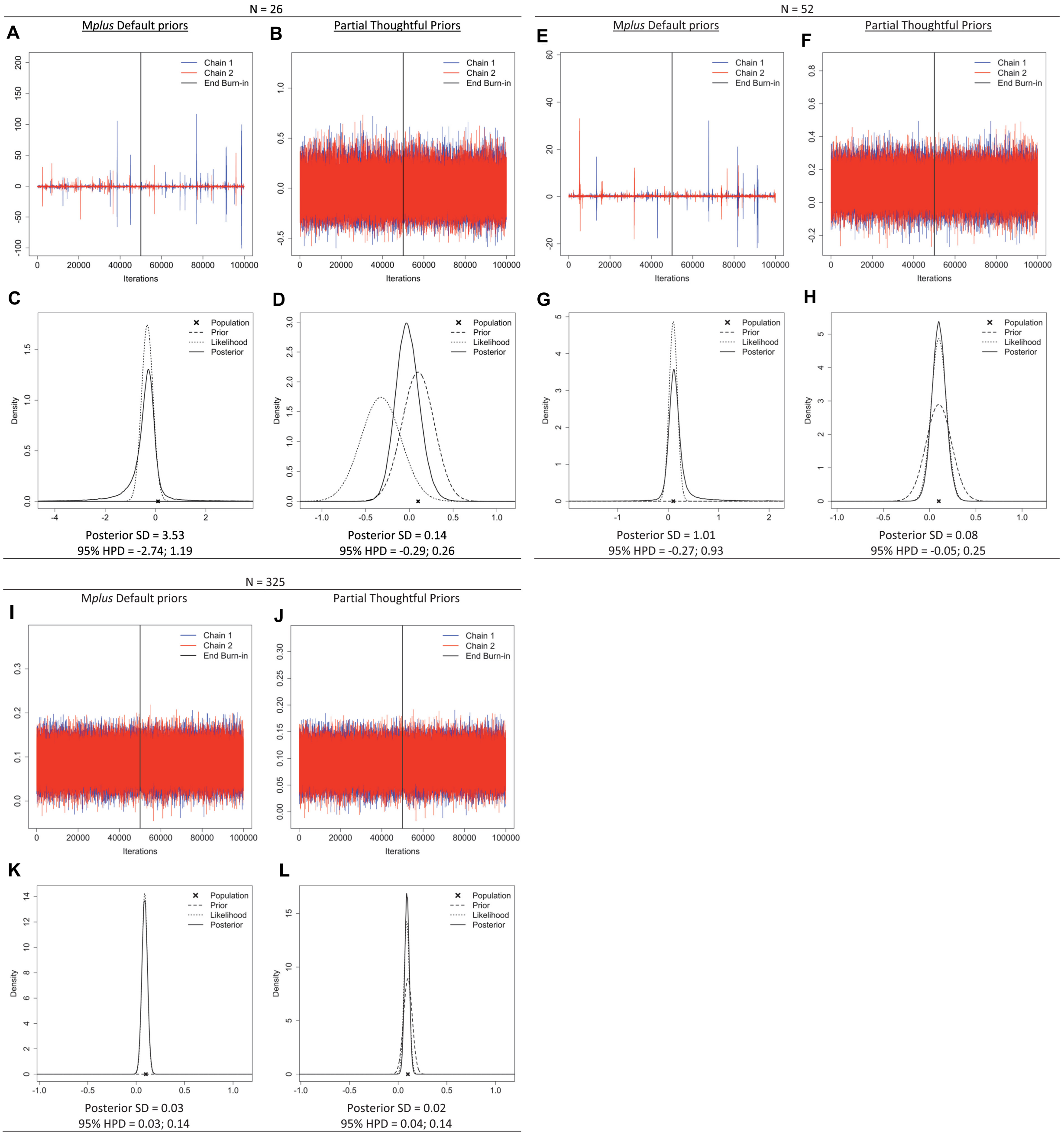


FIGURE 5 | Trace plots; prior, likelihood, posterior plots; posterior standard deviation (SD) and 95% highest posterior density intervals (HPD) for regression coefficient β_2 under sample sizes *n = 26, 52, 325* when M*plus* default priors and partial thoughtful priors are specified. **(A,B,E,F,I,J)** Trace plot. **(C,D,G,H,K,L)** Prior, Likelihood, Posterior Plot.

Options for improving model estimation include increasing the sample size or specifying suitable priors for the parameters.

SUMMARY

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

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

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

AUTHOR CONTRIBUTIONS

SS designed the tutorial manuscript and shiny app, and further developed the idea of the shiny app with SW. SW worked out the code for the shiny app with input and feedback from SS. SS took the lead in writing the manuscript. SW wrote the "Shiny App" section and provided feedback on the manuscript. Both authors contributed to the article and approved the submitted version.

FUNDING

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

REFERENCES

Asparouhov, T., and Muthén, B. (2010). *Bayesian Analysis of Latent Variable Models Using Mplus*. Available online at: http://www.statmodel.com/download/BayesAdvantages18.pdf (accessed October 6, 2020).

Bainter, S. A. (2017). Bayesian estimation for item factor analysis models with sparse categorical indicators. *Multivar. Behav. Res.* 52, 593–615. doi: 10.1080/00273171.2017.1342203

Betancourt, M. (2017). *How the Shape of a Weakly Informative Prior Affects Inferences*. Available online at: https://mc-stan.org/users/documentation/case-studies/weakly_informative_shapes.html (accessed October 6, 2020).

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

Chen, J., Zhang, D., and Choi, J. (2015). Estimation of the latent mediated effect with ordinal data using the limited-information and Bayesian full-information approaches. *Behav. Res. Methods* 47, 1260–1273. doi: 10.3758/s13428-014-0526-3

Darnieder, W. F. (2011). *Bayesian Methods for Data-Dependent Priors*. Doctoral dissertation, Ohio State University, Columbus, OH.

Depaoli, S. (2014). The impact of "inaccurate" informative priors for growth parameters in Bayesian growth mixture modeling. *Struct. Equ. Modeling* 21, 239–252. doi: 10.1080/10705511.2014.882686

Depaoli, S., and Clifton, J. P. (2015). A bayesian approach to multilevel structural equation modeling with continuous and dichotomous outcomes. *Struct. Equ. Modeling* 22, 327–351. doi: 10.1080/10705511.2014.937849

Depaoli, S., and van de Schoot, R. (2017). Improving transparency and replication in Bayesian statistics: the WAMBS-checklist. *Psychol. Methods* 22, 240–261. doi: 10.1037/met0000065

Flegal, J. M., Hughes, J., Vats, D., and Dai, N. (2020). *mcmcse: Monte Carlo Standard Errors for MCMC*. R Package Version 1.4-1. Available online at: https://CRAN.R-project.org/package=mcmcse

Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., and Gelman, A. (2019). Visualization in Bayesian workflow. *J. R. Stat. Soc. Ser. A Stat. Soc.* 182, 389–402. doi: 10.1111/rssa.12378

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (Comment on Article by Browne and Draper). *Bayesian Anal.* 1, 515–534. doi: 10.1214/06-ba117a

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2014). *Bayesian Data Analysis*, 3rd Edn. Boca Raton, FL: CRC Press.

Gelman, A., and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Stat. Sci.* 7, 457–472. doi: 10.1214/ss/1177011136

Geyer, C. J. (1992). Practical Markov chain monte carlo Author(s). *Stat. Sci.* 7, 473–483.

Holtmann, J., Koch, T., Lochner, K., and Eid, M. (2016). A comparison of ML, WLSMV, and Bayesian methods for multilevel structural equation models in small samples: a simulation study. *Multivar. Behav. Res.* 51, 661–680. doi: 10.1080/00273171.2016.1208074

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

Kass, R. E., and Wasserman, L. (1996). The selection of prior distributions by formal rules. *J. Am. Stat. Assoc.* 91, 1343–1370. doi: 10.1080/01621459.1996.10477003

Kline, R. B. (2015). *Principles and Practice of Structural Equation Modeling*, 4th Edn. New York, NY: Guilford Press.

König, C., and van de Schoot, R. (2017). Bayesian statistics in educational research: a look at the current state of affairs. *Educ. Rev.* 70, 1–24. doi: 10.1080/00131911.2017.1350636

Koukounari, A., Stringaris, A., and Maughan, B. (2017). Pathways from maternal depression to young adult offspring depression: an exploratory longitudinal mediation analysis. *Int. J. Methods Psychiatr. Res.* 26:e1520. doi: 10.1002/mpr.1520

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

Kruschke, J. K., Aguinis, H., and Joo, H. (2012). The time has come Bayesian methods for data analysis in the organizational sciences. *Organ. Res. Methods* 15, 722–752. doi: 10.1177/1094428112457829

Lee, S.-Y., and Song, X.-Y. (2004). Evaluation of the bayesian and maximum likelihood approaches in analyzing structural equation models with small sample sizes. *Multivariate Behav. Res.* 39 653–686. doi: 10.1207/s15327906mbr3904_4

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

Lek, K., and van de Schoot, R. (2018). Development and evaluation of a digital expert elicitation method aimed at fostering elementary school teachers' diagnostic competence. *Front. Educ.* 3:82. doi: 10.3389/feduc.2018.00082

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

McNeish, D. (2016a). On using bayesian methods to address small sample problems. *Struct. Equ. Modeling* 23, 750–773. doi: 10.1080/10705511.2016.1186549

McNeish, D. (2016b). Using data-dependent priors to mitigate small sample bias in latent growth models: a discussion and illustration using Mplus. *J. Educ. Behav. Stat.* 41, 27–56. doi: 10.3102/1076998615621299

Merkle, E. (2019). *Prior Distributions*. Available online at: https://faculty.missouri.edu/~{}merklee/blavaan/prior.html (accessed October 6, 2020).

Merkle, E. C., and Rosseel, Y. (2018). blavaan: Bayesian structural equation models via Parameter expansion. *J. Stat. Softw.* 85, 1–30.

. Muthén, L. K., and Muthén, B. O. (1998–2017). *Mplus User's Guide*, 8th Edn. Los Angeles, CA: Muthén & Muthén.

Nevitt, J., and Hancock, G. R. (2004). Evaluating small sample approaches for model test statistics in structural equation modeling. *Multivar. Behav. Res.* 39, 439–478. doi: 10.1207/S15327906MBR3903_3

Polson, N. G., and Scott, J. G. (2012). On the half-Cauchy prior for a global scale parameter. *Bayesian Anal.* 7, 887–902. doi: 10.1214/12-ba730

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. Available online at: https://www.R-project.org/ (accessed October 6, 2020).

Smid, S. C., Depaoli, S., and van de Schoot, R. (2019a). Predicting a distal outcome variable from a latent growth model: ML versus Bayesian estimation. *Struct. Equ. Modeling* 27, 1–23. doi: 10.1080/10705511.2019.1604140

Smid, S. C., McNeish, D., Miočević, M., and van de Schoot, R. (2019b). Bayesian versus frequentist Estimation for structural equation models in small sample contexts: a systematic review. *Struct. Equ. Modeling* 27, 1–31. doi: 10.1080/10705511.2019.1577140

Stan Development Team (2017). *Stan Modeling Language: User's Guide and Reference Manual. Version 2.17.0*. Stan Development Team. doi: 10.1080/10705511.2015.1044653

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

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

van de Schoot, R., Sijbrandij, M., Depaoli, S., Winter, S. D., Olff, M., and Van Loey, N. E. (2018). Bayesian PTSD-trajectory analysis with informed priors based on a systematic literature search and expert elicitation. *Multivar. Behav. Res.* 53, 267–291. doi: 10.1080/00273171.2017.1412293

van de Schoot, R., Winter, S., Ryan, O., Zondervan-Zwijnenburg, M., and Depaoli, S. (2017). A systematic review of Bayesian papers in psychology: the last 25 years. *Psychol. Methods* 22, 217–239. doi: 10.1037/met0000100

van Erp, S. J., Mulder, J., and Oberski, D. L. (2018). Prior sensitivity analysis in default Bayesian structural equation modeling. *Psychol. Methods* 23, 363–388. doi: 10.1037/met0000162

Vats, D., Flegal, J. M., and Jones, G. L. (2019). Multivariate output analysis for Markov chain monte carlo. *Biometrika* 106, 321–337. doi: 10.1093/biomet/asz002

Wagenmakers, E.-J., Lee, M., Lodewyckx, T., and Iverson, G. J. (2008). "Bayesian versus frequentist inference," in *Statistics for Social and Behavioral Sciences. Bayesian evaluation of informative hypotheses*, eds H. Hoijtink, I. Klugkist, and P. Boelen (New York, NY: Springer-Verlag), 181–207. doi: 10.1007/978-0-387-09612-4_9

Zitzmann, S., and Hecht, M. (2019). Going beyond convergence in Bayesian estimation: why precision matters too and how to assess it. *Struct. Equ. Modeling* 26, 646–661. doi: 10.1080/10705511.2018.1545232

Zitzmann, S., Lüdtke, O., Robitzsch, A., and Hecht, M. (2020). On the performance of Bayesian approaches in small samples: a comment on Smid, McNeish, Mioèeviæ, and van de Schoot (2020). *Struct. Equ. Modeling* 1–11. doi: 10.1080/10705511.2020.1752216 [Epub ahead of print].

Zondervan-Zwijnenburg, M., Peeters, M., Depaoli, S., and Van de Schoot, R. (2017). Where do priors come from? Applying guidelines to construct informative priors in small sample research. *Res. Hum. Dev.* 14, 305–320. doi: 10.1080/15427609.2017.1370966

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 21 January 2021
doi: 10.3389/fpsyg.2020.608287

Systematic Parameter Reviews in Cognitive Modeling: Towards a Robust and Cumulative Characterization of Psychological Processes in the Diffusion Decision Model

N.-Han Tran[1]*, Leendert van Maanen[2], Andrew Heathcote[3] and Dora Matzke[4]

[1] Department of Human Behavior, Ecology and Culture, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany, [2] Department of Experimental Psychology, Utrecht University, Utrecht, Netherlands, [3] Department of Psychology, University of Tasmania, Hobart, TAS, Australia, [4] Psychological Methods, Department of Psychology, University of Amsterdam, Amsterdam, Netherlands

OPEN ACCESS

Edited by:
Sarah Depaoli,
University of California, Merced, CA,
United States

Reviewed by:
John K. Kruschke,
Indiana University Bloomington,
United States
Daniel W. Heck,
University of Marburg, Germany

***Correspondence:**
N.-Han Tran
han_tran@eva.mpg.de

Specialty section:
This article was submitted to
Quantitative Psychology and
Measurement,
a section of the journal
Frontiers in Psychology

Received: 19 September 2020
Accepted: 16 December 2020
Published: 21 January 2021

Citation:
Tran N-H, van Maanen L, Heathcote A
and Matzke D (2021) Systematic
Parameter Reviews in Cognitive
Modeling: Towards a Robust and
Cumulative Characterization of
Psychological Processes in the
Diffusion Decision Model.
Front. Psychol. 11:608287.
doi: 10.3389/fpsyg.2020.608287

Parametric cognitive models are increasingly popular tools for analyzing data obtained from psychological experiments. One of the main goals of such models is to formalize psychological theories using parameters that represent distinct psychological processes. We argue that systematic quantitative reviews of parameter estimates can make an important contribution to robust and cumulative cognitive modeling. Parameter reviews can benefit model development and model assessment by providing valuable information about the expected parameter space, and can facilitate the more efficient design of experiments. Importantly, parameter reviews provide crucial—if not indispensable—information for the specification of informative prior distributions in Bayesian cognitive modeling. From the Bayesian perspective, prior distributions are an integral part of a model, reflecting cumulative theoretical knowledge about plausible values of the model's parameters (Lee, 2018). In this paper we illustrate how systematic parameter reviews can be implemented to generate informed prior distributions for the Diffusion Decision Model (DDM; Ratcliff and McKoon, 2008), the most widely used model of speeded decision making. We surveyed the published literature on empirical applications of the DDM, extracted the reported parameter estimates, and synthesized this information in the form of prior distributions. Our parameter review establishes a comprehensive reference resource for plausible DDM parameter values in various experimental paradigms that can guide future applications of the model. Based on the challenges we faced during the parameter review, we formulate a set of general and DDM-specific suggestions aiming to increase reproducibility and the information gained from the review process.

Keywords: Bayesian inference, cognitive modeling, diffusion decision model, prior distributions, cumulative science

1. INTRODUCTION

With an expanding recent appreciation of the value of quantitative theories that make clear and testable predictions (Lee and Wagenmakers, 2014; Oberauer and Lewandowsky, 2019; Navarro, 2020), cognitive models have become increasingly popular. As a consequence, open science and reproducibility reforms have been expanded to include modeling problems. Lee et al. (2019) proposed a suite of methods for robust modeling practices largely centred on the pre- and postregistration of models. In the interest of cumulative science, we believe that the development and assessment of cognitive models should also include systematic quantitative reviews of the model parameters. Several model classes, including multinomial processing trees (Riefer and Batchelder, 1988), reinforcement learning models (Busemeyer and Stout, 2002), and evidence-accumulation models (Donkin and Brown, 2018), have now been applied widely enough that sufficient information is available in the literature to arrive at a reliable representation of the distribution of the parameter estimates. In this paper, we describe a systematic parameter review focusing on the latter class of models.

A systematic quantitative characterization of model parameters provides knowledge of the likely values of the model parameters and has various benefits. First, it can promote more precise and realistic simulations that help to optimally calibrate and design experiments, avoiding unnecessary experimental costs (Gluth and Jarecki, 2019; Heck and Erdfelder, 2019; Kennedy et al., 2019; Pitt and Myung, 2019; Schad et al., 2020). Second, knowledge about the parameter space can be crucial in maximum-likelihood estimation where an informed guess of the starting point of optimization is often key to finding the globally best solution (Myung, 2003). Third,—and most important for the present paper—systematic quantitative parameter reviews provide crucial information for the specification of informative prior distributions in Bayesian cognitive modeling. The *prior distribution* is a key element of Bayesian inference; it provides a quantitative summary of the likely values of the model parameters in the form of a probability distribution. The prior distribution is combined with the incoming data through the likelihood function to form the *posterior distribution*. The prior distribution is an integral part of Bayesian models, and should reflect theoretical assumptions and cumulative knowledge about the relative plausibility of the different parameter values (Vanpaemel, 2011; Vanpaemel and Lee, 2012; Lee, 2018). Prior distributions play a role both in parameter estimation and model selection. By assigning relatively more weight to plausible regions of the parameter space, informative prior distributions can improve parameter estimation, particularly when the data are not sufficiently informative, for instance due to a small number of observations. Even as the number of observations grows, informative priors remain crucial for Bayesian model selection using Bayes factors (Jeffreys, 1961; Kass and Raftery, 1995). Unfortunately, the theoretical and practical advantages of the prior have been undermined by the common use of vague distributions (Trafimow, 2005; Gill, 2014).

The goal of this paper is to illustrate how a systematic quantitative parameter review can facilitate the specification of informative prior distributions. To this end, we first introduce the Diffusion Decision Model (DDM; Ratcliff, 1978; Ratcliff and McKoon, 2008), a popular cognitive model for two-choice response time tasks (see Ratcliff et al., 2016, for a recent review). Using the DDM as a case study, we will then outline how we used a systematic literature review in combination with principled data synthesis and data quantification using distribution functions to construct informative prior distributions. Lastly, based on the challenges we faced during the parameter review, we formulate a set of general and DDM-specific suggestions about how to report cognitive modeling results, and discuss the limitations of our methods and future directions to improve them.

1.1. Case Study: The Diffusion Decision Model

In experimental psychology, inferences about latent cognitive processes from two-choice response time (RT) tasks are traditionally based on separate analyses of mean RT and the proportion of correct responses. However, these measures are inherently related to each other in a speed-accuracy trade-off. That is, individuals can respond faster at the expense of making more errors. Evidence-accumulation models of choice RT and accuracy have provided a solution for this conundrum because they allow for the decomposition of speed-accuracy trade-off effects into latent variables that underlie performance (Ratcliff and Rouder, 1998; Donkin et al., 2009a; van Maanen et al., 2019). These models assume that evidence is first extracted from the stimuli and then accumulated over time until a decision boundary is reached and a response initiated. Among the many evidence-accumulation models, the DDM is the most widely applied, not only in psychology, but also in economics and neuroscience, accounting for experiments ranging from decision making under time-pressure (Voss et al., 2008; Leite et al., 2010; Dutilh et al., 2011), prospective memory (Horn et al., 2011; Ball and Aschenbrenner, 2018) to cognitive control (Gomez et al., 2007; Schmitz and Voss, 2012).

Figure 1 illustrates the DDM. Evidence (i.e., gray line) fluctuates from moment to moment according to a Gaussian distribution with standard deviation s, drifting until it reaches one of two boundaries, initiating an associated response. The DDM decomposes decision making in terms of four main parameters corresponding to distinct cognitive processes: (1) the mean rate of evidence accumulation (drift rate v), representing subject ability and stimulus difficulty; (2) the separation of the two response boundaries (a), representing response caution; (3) the mean starting point of evidence accumulation (z), representing response bias; and (4) mean non-decision time (T_{er}), which is the sum of times for stimulus encoding and response execution. RT is the sum of non-decision time and the time to diffuse from the starting point to one of the boundaries. A higher drift rate leads to faster and more accurate responses. However, responses can also be faster because a participant chooses to be less cautious and thus decreases their boundary

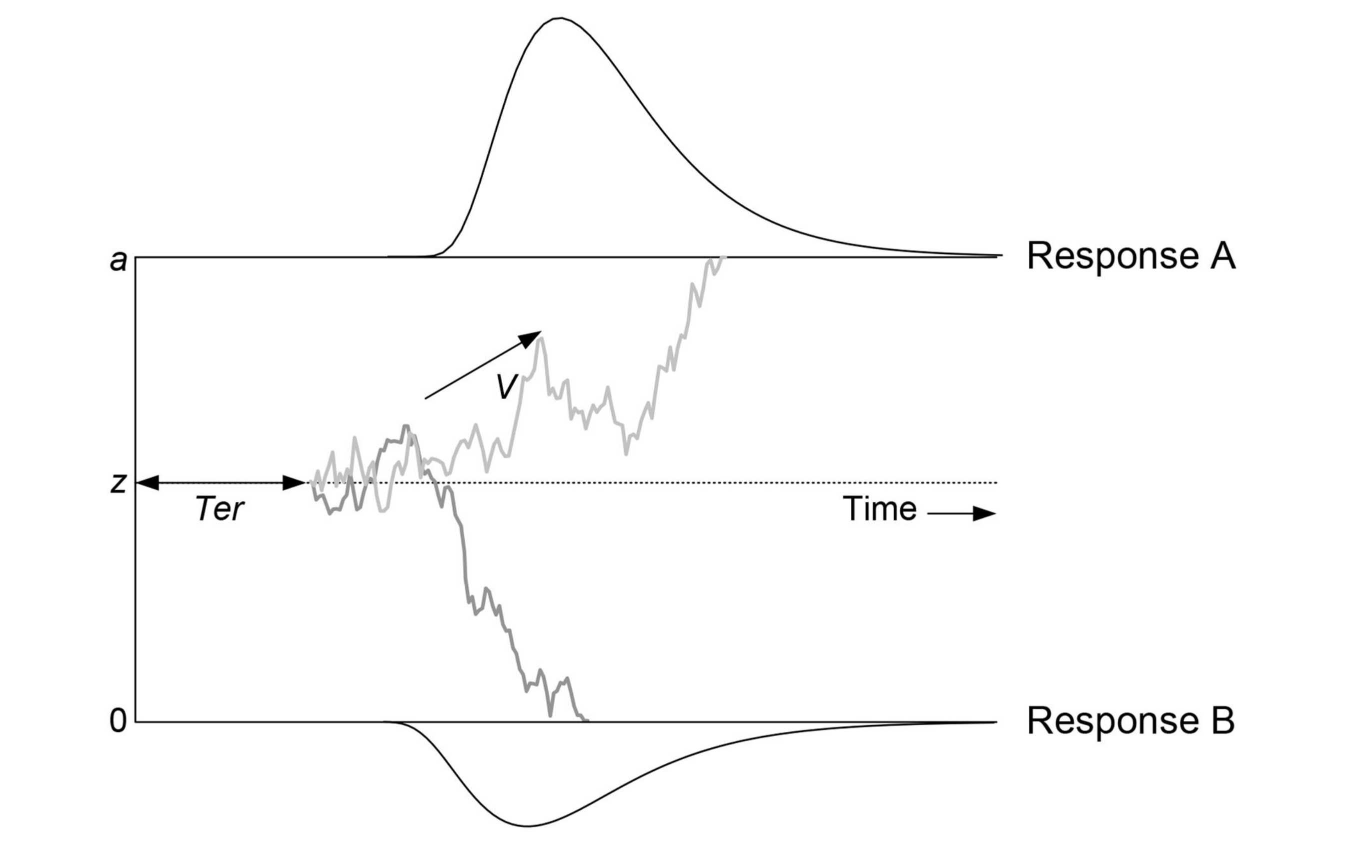


FIGURE 1 | The Diffusion Decision Model (DDM; taken with permission from Matzke and Wagenmakers, 2009). The DDM assumes that noisy information is accumulated over time from a starting point until it crosses one of the two response boundaries and triggers the corresponding response. The gray line depicts the noisy decision process. "Response A" or "Response B" is triggered when the corresponding boundary is crossed. The DDM assumes the following main parameters: drift rate (v), boundary separation (a), mean starting point (z), and mean non-decision time (T_{er}). These main parameters can vary from trial to trial: across-trial variability in drift rate (s_v), across-trial variability in starting point (s_z), and across-trial variability in non-decision time ($s_{T_{er}}$). Starting point can be expressed relative to the boundary in order to quantify bias, where $z_r = \frac{z}{a} = 0.5$ indicates unbiased responding. Similarly, across-trial variability in starting point can be expressed relative to the boundary: $s_{z_r} = \frac{s_z}{a}$.

separation, which will reduce RT but increase errors, causing the speed-accuracy trade-off. Starting accumulation closer to one boundary than the other creates a bias toward the corresponding response. Starting points z is therefore most easily interpreted in relation to boundary separation a, where the relative starting point, also known as bias, is given by $z_r = \frac{z}{a}$. Drift rate can vary from trial to trial according to a Gaussian distribution with standard deviation s_v. Both non-decision time and starting point are assumed to be uniformly distributed across trials, with range $s_{T_{er}}$ and s_z, respectively, where s_z can be expressed relative to a: $s_{z_r} = \frac{s_z}{a}$. One parameter of the accumulation process needs to be fixed to establish a scale that makes the other accumulation-related parameters identifiable (Donkin et al., 2009b). Most commonly this scaling parameters is the moment-to-moment variability of drift rate (s), usually with a value fixed to 0.1 or 1.

Fitting the DDM and many other evidence-accumulation models to experimental data is difficult because of the complexity of the models and the form of their likelihood resulting in high correlations among the parameters (i.e., "sloppiness"; Gutenkunst et al., 2007; Gershman, 2016). Informative prior distribution can ameliorate some of these problems. The growing popularity of cognitive modeling has led to extensive application of the DDM to empirical data (Theisen et al., 2020), providing us with a large number of parameter estimates to use for constructing informative prior distributions. In 2009, Matzke and Wagenmakers presented the first quantitative summary of the DDM parameters based on a survey of parameter estimates found in 23 applications. However, their survey is now outdated and was not as extensive or systematic as the approach taken here.

2. MATERIALS AND METHODS

All analyses were written in R or R Markdown (Allaire et al., 2018; R Core Team, 2020). The extraced parameter estimates and the analysis code are available on GitHub (http://github.com/nhtran93/DDM_priors) and the project's Open Science Framework (OSF) site: https://osf.io/9ycu5/.

2.1. Literature Search

The literature search was conducted according to the PRISMA guidelines (Moher et al., 2009). Every step was recorded and the inclusion as well as rejection of studies adhered strictly to the pre-specified inclusion criteria. Results from different search engines were exported as BibTex files, maintained with reference management software and exported into separate Microsoft Excel spreadsheets.

2.1.1. Search Queries

The literature search was commenced and completed in December 2017. It consisted of cited reference searches and independent searches according to pre-specified queries. Searches in all databases were preformed three times in order to ensure reproducibility. Four electronic databases were searched with pre-specified queries: (National Library of Medicine, 2017), PsycInfo (American Psychological Association, 2017), Web of Science (WoS, 2017), and Scopus (Elsevier, 2017). A preliminary search of the four databases served to identify relevant search strings, which were different for each database (see the **Supplementary Materials** or https://osf.io/9ycu5/ for details). The searches began from the publication year of the seminal paper by Ratcliff (1978). The cited reference searches were based on Ratcliff and McKoon (2008), Wiecki et al. (2013), and Palmer et al. (2005), and were performed in both Scopus and Web of Science. These key DDM papers were selected to circumvent assessing an unfeasible number of over $3,000$ cited references to the seminal Ratcliff article, with a potentially high number of false positives (in terms of yielding papers that reported parameter estimates), while still maintaining a wide search covering various areas of psychology and cognitive neuroscience.

2.1.2. Inclusion and Exclusion Criteria

All duplicated references were excluded. After obviously irrelevant papers—judged based on title and abstract—were excluded, the full-texts were acquired to determine the inclusion or exclusion of the remaining articles. Articles were included in the literature review if they (i) used the standard DDM according to Ratcliff (1978) and Ratcliff and McKoon (2008) with or without across-trial variability parameters; and (ii) reported parameter estimates based on empirical data from humans. Articles were excluded if (i) they reported reviews; and (ii) the parameter estimates were based on animal or simulation studies. We also excluded articles that did not report parameter estimates (neither in tables nor in graphs) and articles that estimated parameters in the context of a regression model with continuous predictors that resulted in estimates of intercepts and regression slopes instead of single values of the model parameters.

2.1.3. Data Extraction

The data extraction spreadsheet was pilot-tested using six articles and adjusted accordingly. The following parameter estimates were extracted: drift rate (v), boundary separation (a), starting point (z) or bias ($z_r = \frac{z}{a}$), non-decision time (T_{er}), across-trial variability in drift rate (s_v), across-trial range in starting point (s_z) or relative across-trial starting point ($s_{z_r} = \frac{s_z}{a}$), and across-trial range in non-decision time ($s_{T_{er}}$). Parameter estimates were obtained from tables as well as from graphs using the GraphClick software (Arizona, 2010). Whenever possible, we extracted parameter estimates for each individual participant; otherwise we extracted the mean across participants or in Bayesian hierarchical applications the group-level estimates. When the DDM was fit multiple times with varying parameterizations to the same data within one article, we used the estimates corresponding to the model identified as best by the authors, with a preference for selections made based on the AIC (Akaike, 1973, 1974), in order to identify the best trade-off between goodness-of-fit and parametric complexity (Myung and Pitt, 1997). When the DDM was applied to the same data across different articles, we extracted the parameter estimates from the first application; if the first application did not report parameter estimates, we used the most recent application that reported parameter estimates. Finally, articles that obtained estimates using the `EZ` (Wagenmakers et al., 2007) or `EZ2` (Grasman et al., 2009) methods, or the `RWiener` R package (Wabersich and Vandekerckhove, 2014), which all fit the simple diffusion model estimating only the four main DDM parameters (Stone, 1960), were excluded due to concerns about potential distortions caused by ignoring across-trial parameter variability (Ratcliff, 2008). Note that we did not automatically exclude all articles without across-trial variability parameters. For articles that did not use `EZ`, `EZ2`, or `RWiener`, but reported models without across-trial variability parameters, we assumed that the author's choice of fixing these parameters to zero was motivated by substantive or statistical reasons and not by the limitations of the estimation software, and hence we included them in the parameter review.

2.2. Parameter Transformations

Once extracted, parameter estimates had to be transformed in a way that makes aggregation across articles meaningful. In this section we report issues that arose with respect to these transformations and the solutions that we implemented. A detailed explanation of the transformations can be found in the **Supplementary Materials**.

2.2.1. Within-Trial Variability of Drift Rate

In all of the studies we examined, the accumulation-related parameters were scaled relative to a fixed value of the moment-to-moment variability in drift rate (typically $s = 0.1$ or $s = 1$). This decision influences the magnitude of all parameter estimates except those related to non-decision time. Once we determined s for each article, we re-scaled the affected parameter estimates to $s = 1$. Articles that used the DMAT software (Vandekerckhove and Tuerlinckx, 2008) for parameter estimation were assumed to use the DMAT default of $s = 0.1$, and articles that used HDDM (Wiecki et al., 2013) or fast-DM (Voss and Voss, 2007) were assumed to use the default setting of 1. Articles (co-) authored by Roger Ratcliff were assigned $s = 0.1$.[1] We excluded 25 articles because the scaling parameter was not reported and even if we assumed the scaling parameter to be s, its value could not be determined.

[1] Based on personal communication with Roger Ratcliff.

2.2.2. Measurement (RT) Scale

Although the measurement (i.e., RT) scale influences the magnitude of the parameter estimates, none of the articles mentioned explicitly whether the data were fit on the seconds or milliseconds scale. Moreover, researchers did not necessarily report all estimates on the same RT scale. For instance, T_{er} or $s_{T_{er}}$ were sometimes reported in milliseconds, whereas the other parameters were reported in seconds. Whenever possible, we used axis labels, captions and descriptions in figures and tables, or the default setting of the estimation software to determine the RT scale. Articles that used the DMAT, HDDM, or fast-DM were assumed to use the default setting of seconds and we assigned an RT scale of seconds to papers authored by Roger Ratcliff [2] even if T_{er} was reported in milliseconds. We also evaluated the plausibility of the reported estimates with respect to the second or millisecond scale by computing a rough estimate of the expected RT for each experimental condition as $E(RT) = (a - z)/v$. We then used the following two-step decision rule to determine the RT scale of each parameter:

1. Determine the RT scale of T_{er}: If estimated T_{er} was smaller than 5, we assumed that T_{er} was reported in seconds; otherwise we assumed that T_{er} was reported in milliseconds.
2. Determine RT scale of remaining parameters: If $E(RT)$ was smaller than 10, we assumed that the remaining parameters were reported in seconds; otherwise we assumed that the remaining parameters were reported in milliseconds.

Once we determined the RT scale for each parameter, we re-scaled the parameter estimates to the seconds scale. Individual parameters estimates that were considered implausible after the transformation (i.e., outside of the parameter bounds, such as a negative a) were checked manually. In particular, we checked for (1) inconsistencies in the magnitude across the parameter estimates within articles (e.g., a value of a indicative of seconds vs. a value of T_{er} indicative of milliseconds); (2) reporting or typographic errors; (3) extraction errors; and (4) errors in determining the measurement scale, which typically reflected the use of non-standard experiments or special populations. In a number of cases we also revisited and whenever necessary reconsidered the assigned value of s. We removed all parameter estimates from 13 articles that reported implausible estimates reflecting ambiguous or inconsistent RT scale descriptions or clear reporting errors.

2.2.3. Starting Point and Bias

We expressed all starting point z and starting point variability s_z estimates relative to a. As the attributions of the response options to the two response boundaries is arbitrary, the direction of the bias (i.e., whether z_r is greater or less than 0.5) is arbitrary. As these attributions cannot be made commensurate over articles with different response options, values of z_r cannot be meaningfully aggregated over articles. As a consequence, bias, z_r, and its complement, $1 - z_r$, are exchangeable for the purpose of our summary. We therefore used both values in order to create a single "mirrored" distribution. This distribution is necessarily symmetric with a mean of 0.5, but retains information about variability in bias.[3]

2.2.4. Drift Rate

There are two ways in which drift rates v can be reported. In the first, positive drift rates indicate a correct response (e.g., "word" response to a "word" stimulus and "non-word" response to a "non-word" stimulus in a lexical decision task) and negative rates indicate an incorrect response. In the second, positive drift rates correspond to one response option (e.g., "word" response) and negative rates to the other option (e.g., "non-word"). Here, we adopt the former—accuracy coding—method in order to avoid ambiguity regarding the arbitrary attribution of boundaries to response options. We do so by taking the absolute values of the reported drift rates to construct the prior distribution. Readers who wish to adopt the latter—response coding—method, should appropriately mirror our accuracy-coded priors around 0.

2.3. Generating Informative Prior Distributions for the DDM Parameters

After post-processing and transforming the parameter estimates, we combined each parameter type across articles and experimental conditions within each study into separate univariate distributions. We then attempted to characterize these empirical parameter distributions with theoretical distributions that provided the best fit to the overall shape of the distributions of parameter estimates.

2.3.1. Parameter Constraints

In many applications of the DDM, researchers impose constraints on the parameter estimates across experimental manipulations, conditions, or groups, either based on theoretical grounds or the results of model-selection procedures. After extracting all parameters from the best fitting models, we identified parameters that were constrained across within- and between-subject manipulations, conditions, or groups within each study. For the purpose of constructing the prior distributions we only considered these fixed parameters once and did not repeatedly include them in the empirical distributions. For instance, a random dot motion task with three difficulty conditions may provide only one estimate for a constrained parameter (i.e., non-decision time), but three parameters for an unconstrained parameter (i.e., drift rate).

2.3.2. Synthesis Across Articles

Most studies reported parameter estimates aggregated across participants, with only eight reporting individual estimates. Before collapsing them with the aggregated estimates, individual estimates were averaged across participants in each study. Parameter estimates were equally weighted when combined across studies as details necessary for weighting them according to their precision were typically not available. We will revisit this decision in the Discussion.

[2] Based on personal communication with Roger Ratcliff.

[3] The bias z_r parameters estimated using the HDDM software (Wiecki et al., 2013) are coded as $1 - z_r$ in our parameter review. Note that this has no influence on the resulting prior distribution as we used both z_r and $1 - z_r$ to create the prior.

In the results reported in the main body of this article, we aggregated the parameter estimates across all research domains (e.g., neuroscience, psychology, economics), populations (e.g., low/high socioeconomic status, clinical populations), and tasks (e.g., lexical decision, random dot motion tasks). In the **Supplementary Materials**, we provide examples of prior distributions derived specifically for two of the most common tasks in our database (i.e., lexical decision and random dot motion tasks) and priors restricted to non-clinical populations. Data and code to generate such task and population-specific priors are available in the open repository, so that interested readers can construct priors relevant to their specific research questions.

2.3.3. Distributions

A full characterization of the distribution of model parameters takes into account not only the parameters' average values and variability but also their correlations across participants (e.g., people with lower drift rates may have higher thresholds) and potentially even their correlations across studies or paradigms using multilevel structures. Although multivariate prior distributions would be optimal to represent correlations across participants, they require individual parameter estimates for the estimation of the covariance matrices. As only eight studies reported individual parameter estimates, we were restricted to use univariate distributions.

We attempted to characterize the aggregated results using a range of univariate distribution functions that respected the parameter types' bounds (e.g., non-decision time T_{er} must be positive) and provided the best fit to the overall shape of the empirical distributions. We first considered truncated normal, lognormal, gamma, Weibull, and truncated Student's t distribution functions. However, in some cases the empirical distributions clearly could not be captured by the univariate distributions and were contaminated by outliers due to non-standard tasks, special populations, and possible reporting errors that we not identified during the post-processing steps. We therefore also considered characterizing the empirical distribution using mixture distributions. Mixtures were chosen from the exponential family of distributions that respected the theoretical bounds of the parameter estimates. In particular we used mixtures of two gamma distributions, and truncated normals mixed with either a gamma, lognormal, or another truncated normal distribution. Specifically, we focused on normal mixtures because we assume a finite variance for the parameters and thus the Gaussian distributions represents the most conservative probability distribution to assign to the parameter distributions (for further information see the principles of maximum entropy; Jaynes, 1988).

The univariate and mixture distributions were fit to the empirical distributions using maximum-likelihood estimation (Myung, 2003), with additional constraints on upper and/or lower bounds. For (mirrored) bias z_r and s_{z_r}, which are bounded between 0 and 1, we used univariate truncated normal and truncated t distributions on $[0, 1]$. A lower bound of zero was imposed on all other parameters. We then used AIC weights (wAIC; Wagenmakers and Farrell, 2004) to select the theoretical distributions that struck the best balance between goodness-of-fit and simplicity. A table of the AIC and wAIC values for all fitted univariate and mixture distributions and the code to reproduce this table, can be found in the open repository on GitHub or the OSF.

We propose that the wAIC-selected distributions can be used as informative prior distributions for the Bayesian estimation of the DDM parameters. For simplicity, for parameters where a mixture was the best-fitting distribution, we propose as prior the distribution component that best captures the bulk of the parameter estimates as indicated by the highest mixture weight. We will revisit this choice in the Discussion.

3. RESULTS

Figure 2 shows the PRISMA flow diagram corresponding to our literature search. The total of 196 relevant articles (i.e., "Reported estimates" in **Figure 2**) covered a wide range of research areas from psychology and neuroscience to medicine and economics. We excluded 38 references because they did not report the scaling parameters and we were unable to reverse engineer them or because of inconsistent RT scale descriptions or clear reporting errors. Thus, we extracted parameter estimates from a total of 158 references. The most common paradigms were various perceptual decision-making tasks (e.g., random dot motion task; 37 references), lexical decision tasks (33), and recognition memory tasks (17). A total of 29 references included clinical groups and 26 references used Bayesian estimation methods.

The histograms in **Figure 3** show the empirical distributions of the parameter estimates. The red lines show the best fitting theoretical distributions or the dominant theoretical distribution components with the highest mixture weight (i.e., the proposed informative prior distributions). The black lines show the non-dominant mixture distribution components. Note that in most cases the the mixture served to inflate the distributions' tails while preserving a single mode.

Table 1 gives an overview of the informative prior distributions, the corresponding upper and lower bounds (see column "T-LB" and "T-UB"), and whenever appropriate also the mixture weight of the dominant distribution component. The table also shows the upper and lower bounds of the parameter estimates collected from the literature (see column "E-LB" and "E-UB"); these bounds can be used to further constrain parameter estimation by providing limits for prior distributions and bounded optimization methods.

The results of the model comparisons are available at https://osf.io/9ycu5/. For drift rate v, the selected model was a mixture of a zero-bounded truncated normal and a lognormal distribution (wAIC = 0.4), with the mixture weight, and the location and scale of the dominant truncated normal component shown in the first row of **Table 1**.[4] For boundary separation a, the selected model was a mixture of gamma distributions (wAIC = 0.76), with the shape and scale parameters of the dominant gamma component shown in the second row of **Table 1**. For non-decision time T_{er}

[4] The location and scale parameters of the truncated normal distribution refer to μ and σ and not to its expected value and variance.

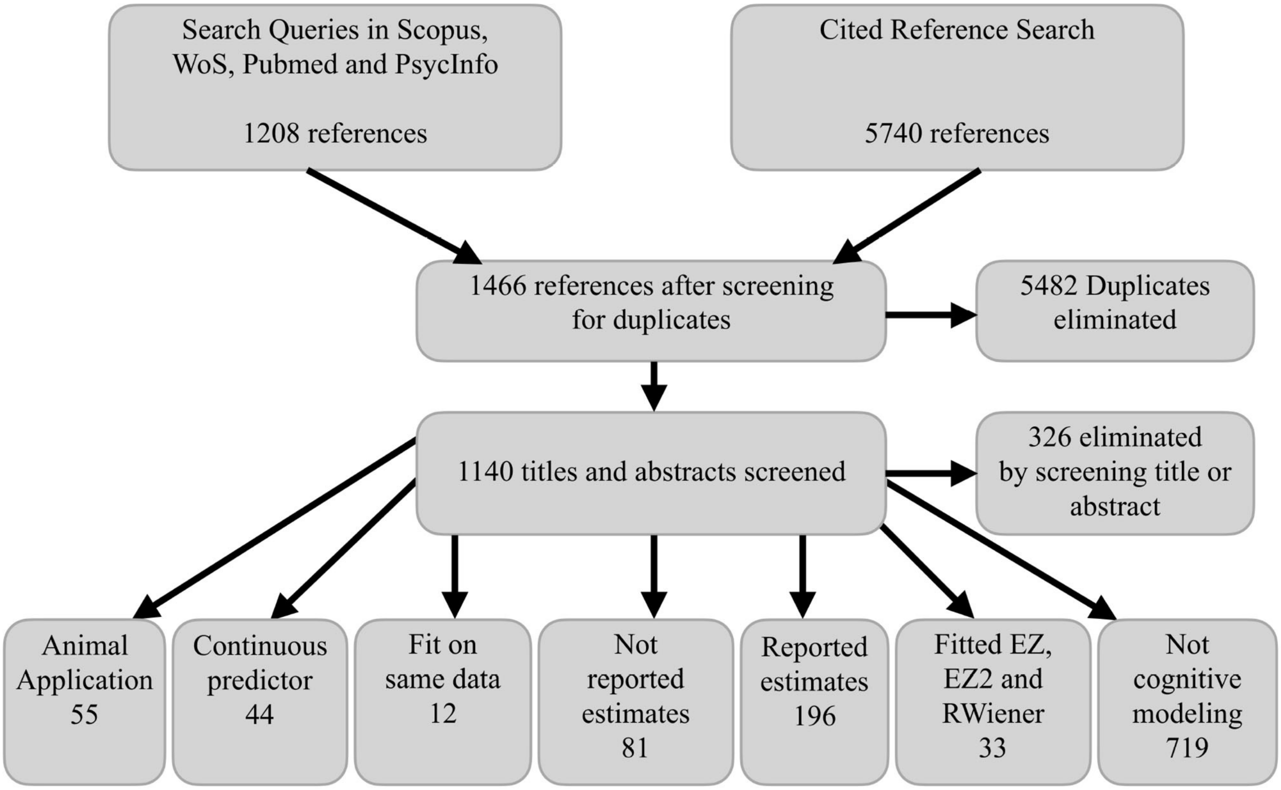

Search Queries in Scopus, WoS, Pubmed and PsycInfo
1208 references
Cited Reference Search
5740 references
1466 references after screening for duplicates
5482 Duplicates eliminated
1140 titles and abstracts screened
326 eliminated by screening title or abstract
Animal Application 55
Continuous predictor 44
Fit on same data 12
Not reported estimates 81
Reported estimates 196
Fitted EZ, EZ2 and RWiener 33
Not cognitive modeling 719

FIGURE 2 | PRISMA flow diagram. WoS, Web of Science. `RWiener` refers to the R package from Wabersich and Vandekerckhove (2014). `EZ` and `EZ2` refer to estimation methods for the simple DDM developed by Wagenmakers et al. (2007) and Grasman et al. (2009), respectively.

and across-trial variability in non-decision time $s_{T_{er}}$, the selected model was a zero-bounded truncated t distribution (wAIC = 1 for both T_{er} and $s_{T_{er}}$). For mirrored bias z_r, the selected model was a truncated t distribution on $[0, 1]$ (wAIC = 1.0). For across-trial variability in drift rate s_v, the selected model was a mixture of a zero-bounded truncated normal and a gamma distribution (wAIC = 0.35), where the truncated normal had the highest mixture weight. Lastly, for s_{z_r}, the selected model was a truncated normal distribution on $[0, 1]$ (wAIC = 0.74).

4. DISCUSSION

The increasing popularity of cognitive modeling has led to extensive applications of models like the Diffusion Decision Model (DDM) across a range of disciplines. These applications have the potential to provide substantial information about the plausible values of the parameters of cognitive models. We believe that for cognitive models where sufficient information are available in the literature, a systematic quantitative characterization of model parameters can be a very useful addition to existing modeling practices. Parameter reviews can benefit modeling practices in various ways, from facilitating parameter estimation to enabling more precise and realistic simulations to improve study design and calibrate future experiments (Gluth and Jarecki, 2019; Heck and Erdfelder, 2019; Pitt and Myung, 2019).

Here, we used the DDM as example case of how a systematic quantitative parameter review can be incorporated into modeling practices to provide informative prior distributions for the model parameters. Our empirical distributions of the parameter estimates were largely consistent with those of Matzke and Wagenmakers (2009), but because our sample was much larger we were better able to capture the tails of the parameter distributions. Although, for simplicity, here we suggested single-component distributions as priors, the full mixture distributions that we selected could also be used. Bayesian DDM software, such as the Dynamic Models of Choice software (DMC; Heathcote et al., 2019), can be easily adapted to use any form of univariate prior, including mixtures. In most cases the mixture served to inflate the distributions' tails while preserving a single mode. However, aggregation over heterogeneous studies naturally carries with it the possibility of creating multi-modal prior distributions, as illustrated by the results for s_v in **Figure 3**. If the data proved sufficiently uninformative that such multi-modality carried through to the posterior, caution should be exercised in reporting and interpreting measures of central tendency.

Inferring the parameters of complex cognitive models like the DDM from experimental data is challenging because their parameters are often highly correlated. The cumulative

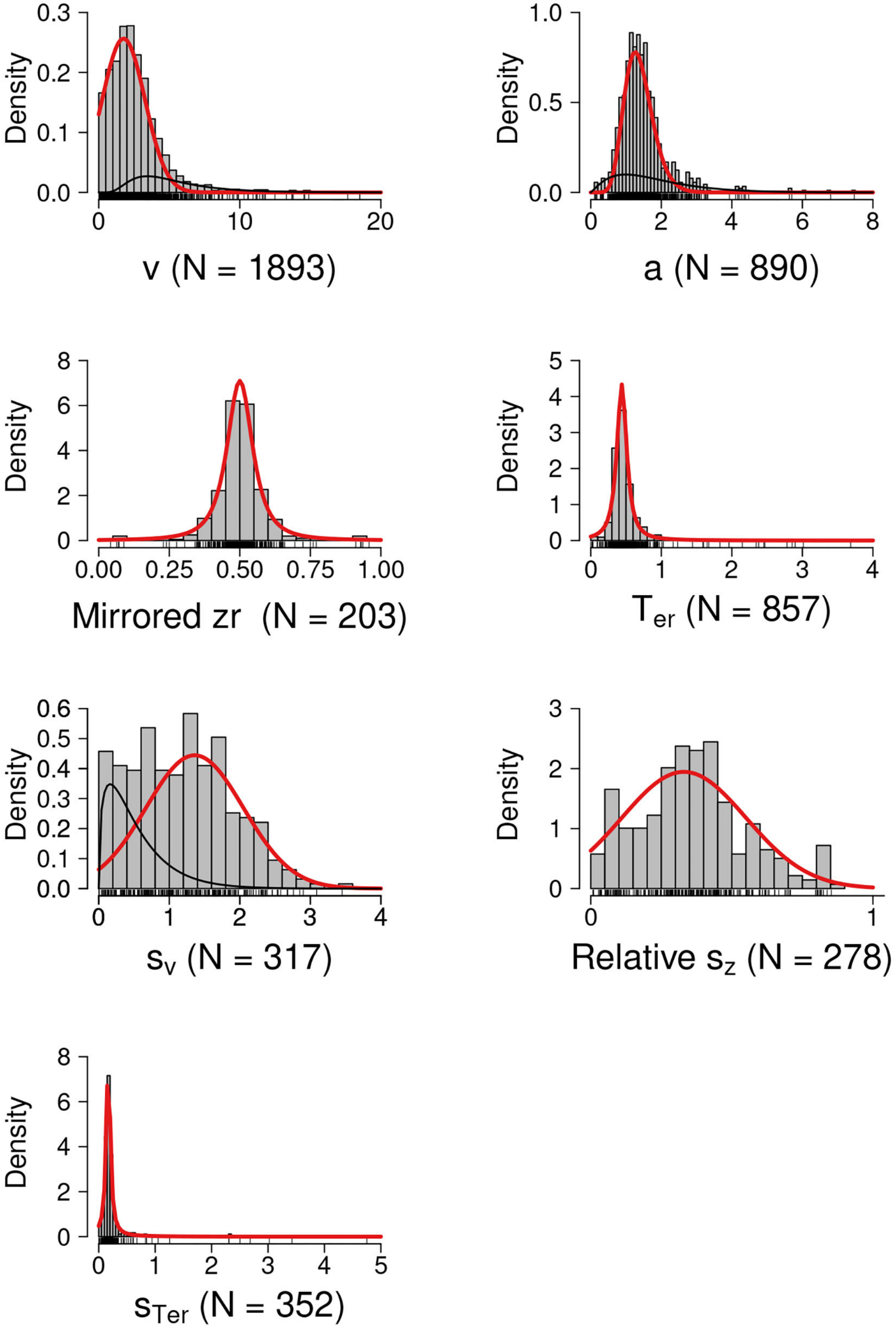


FIGURE 3 | Prior Distributions for the DDM Parameters. The red lines show the best fitting theoretical distributions or the dominant theoretical distribution components with the highest mixture weight (i.e., the proposed informative prior distributions). The black lines show the non-dominant distribution components. *N*, number of unique estimates.

TABLE 1 | Informative prior distributions.

DDM parameter	*N*	Distribution	Weight	Location/Shape	Scale	df	T-LB	T-UB	E-LB	E-UB
v	1,893	Truncated normal* & lognormal	0.85	1.76	1.51		0	+ Inf	0.01	18.51
a	890	Gamma* & gamma	0.76	11.69	0.12		0	+ Inf	0.11	7.47
Mirrored z_r	203	Truncated t	–	0.5	0.05	1.85	0	1	0.04	0.96
T_{er}	857	Truncated t	–	0.44	0.08	1.32	0	+ Inf	0	3.69
s_v	317	Truncated normal* & gamma	0.75	1.36	0.69		0	+ Inf	0	3.45
s_{z_r}	278	Truncated normal	–	0.33	0.22		0	1	0.01	0.85
$s_{T_{er}}$	352	Truncated t	–	0.17	0.04	0.88	0	+ Inf	0	4.75

N, *The number of unique estimates; Weight, The mixture weight of the dominant distribution component; df, Degrees of freedom; T-LB, Theoretical lower bound of the prior distribution; T-UB, Theoretical upper bound of the prior distribution; E-LB, Lower bound of the empirical parameter estimates; E-UB, Upper bound of the empirical parameter estimates; *, Dominant distribution component.*

knowledge distilled into parameter estimates from past research can practically benefit both traditional optimization-based methods (e.g., maximum likelihood) and Bayesian estimation. In the former case, parameter reviews can provide informed guesses for optimization starting points as well as guidance for configuring bounded optimization methods. Even when powerful and robust optimization algorithms (e.g., particle swarm methods) are used, reasonable initial values and bounds can increase time efficiency and are often helpful for avoiding false convergence on sub-optimal solutions. In the latter—Bayesian case—parameter reviews can facilitate the use of informative prior distributions, which benefits both Bayesian model selection and parameter estimation.

Informative priors are essential for Bayesian model selection using Bayes factors (Jeffreys, 1961; Kass and Raftery, 1995). Unlike other model-selection methods like the Deviance Information Criteria (DIC; Spiegelhalter et al., 2014) and Widely Applicable Information Criterion (WAIC; Vehtari et al., 2017) that depend only on posterior samples, Bayes factors depend crucially on the prior distribution even when large amounts of data are available. This is because the marginal likelihood of the competing models is obtained by taking a weighted average of the probability of the data across all possible parameter settings with the weights given by the parameters' prior density. The workflow outlined here may therefore facilitate the more principled use of prior information in Bayesian model selection in the context of evidence-accumulation models (for recent developments, see Evans and Annis, 2019; Gronau et al., 2020).

In terms of Bayesian estimation, the extra constraint provided by informative priors can benefit some parameters more than others. In the DDM, for example, the across-trial variability parameters are notoriously difficult to estimate (Boehm et al., 2018; Dutilh et al., 2019). This has led to calls for these parameters to be fixed to zero (i.e., use the simple diffusion model; Stone, 1960) to improve the detection of effects on the remaining parameters (van Ravenzwaaij et al., 2017). Informative priors may provide an alternative solution that avoids the potential systematic distortion caused by ignoring the variability parameters (Ratcliff and McKoon, 2008) and enables the study of effects that cannot be accommodated by the simple diffusion model, such as differences between correct and error RTs (Damaso et al., submitted). Of course, in extreme cases, the central tendency of informative prior distributions may provide guidelines for fixing difficult-to-estimate model parameters to a constant (e.g., Matzke et al., 2020).

Information about the empirical distribution of parameter estimates, both in terms of the main body and the tails of the distributions, can especially benefit design optimization and parameter estimation in non-standard and difficult to access populations (e.g., Shankle et al., 2013; Matzke et al., 2017). For example, in clinical populations long experimental sessions are often impossible due to exhaustion or attention lapses. Expenses can also be constraining, such as with studies using costly fMRI methods. Therefore, data are often scarce, with a total number of trials as low as 100 reported in some DDM applications (e.g., O'Callaghan et al., 2017). In these cases, experimental designs can be optimized, and parameter estimation improved, with the aid of informative parameter distributions that put weight on plausible parts of the parameter space. Moreover, informative priors can also increase sampling efficiency and speed up the convergence of MCMC routines.

Ideally, informative prior distributions for cognitive models should be based on prior information extracted from experimental paradigms (or classes of paradigms) and participant populations relevant to the research question at hand, although care should be exercised in the latter case where group members fall along a continuum (e.g., age or the severity of a clinical diagnosis). In reality, constructing such highly specific priors might not always be feasible, either because of a paucity of relevant parameter values reported in the literature, or when new paradigms or populations are studied. However, we believe that using informative priors based on a range of broadly similar paradigms and heterogeneous populations is better than using vague priors, as long as appropriate caution is exercised in cases when the data are not sufficiently informative and hence the prior dominates inferences about the model parameters. Here, we presented informative prior distributions for parameter estimates aggregated across paradigms and populations and also provided paradigm-specific priors for the two most popular tasks in our database (e.g., lexical decision and random dot motion task) and priors for non-clinical populations.

The variance of the paradigm/population-specific priors showed a general decreasing tendency. The decrease in variance was relatively small for the priors based on non-clinical populations, and was restricted to the main DDM parameters. For the lexical decision task, the variance of all of the priors decreased relative to the overall priors. For the random dot motion task, with the exception of v, all variances decreased, albeit the decrease was negligible for a. To summarize, for the tasks and groups we examined here, paradigm and population heterogeneity appears to introduce additional variability in the parameter estimates, but the degree of additional variability strongly depends on the type of parameter. Our open repository provides the data and code to generate informed prior distributions for any selection of studies included in the database so that researchers can construct informative prior distributions relevant to their own research questions. Naturally, paradigm/population-specific priors are only sensible when a sufficiently large number of parameter estimates are available in the database to fit the theoretical (mixture) distributions, or when researchers can augment the repository with estimates from additional studies.

Despite their usefulness, systematic quantitative parameter reviews are not without their pitfalls. Using available cumulative knowledge from past literature always has to be viewed in light of the file drawer problem (Rosenthal, 1979). Many researchers have not published their non-significant results, therefore the literature is biased, and thus the parameter estimates retrieved from the literature might be biased toward specific model settings that converged or led to significant results. Furthermore, some cognitive models are too new and have not been widely applied to empirical data, so past literature might not provide researchers with a sufficiently reliable representation of the distribution of the parameter estimates. Therefore, cognitive modelers may not always be able to incorporate our proposed quantitative parameter review into their workflow, and should carefully weigh out the feasibility and benefits of such an endeavor.

4.1. Recommendations for Reporting Cognitive Modeling Results

Our literature review revealed a wide variety of reporting practices, both in terms of *what* researcher report and *how* they report their modeling results. The diversity of reporting practices is likely to reflect differences between disciplines and is in itself not problematic. However, we believe that the full potential of cumulative science can only be realized if authors provide sufficient information for others to interpret and reproduce their results. We endorse code and data sharing, and—following Lee et al. (2019)—we strongly urge researchers to provide sufficiently precise mathematical and statistical descriptions of their models, and to post-register exploratory model developments. In what follows, we reflect on the challenges we faced in performing the systematic parameter review, and formulate a set of general and DDM-specific suggestions that aim to increase computational reproducibility and the expected information gain from parameter reviews. Although our recommendations are certainly not exhaustive and do not apply to all model classes, we hope that they provide food for thought for cognitive modelers in general and RT modelers in particular.

4.1.1. Model Parameterization and Scaling

The following recommendations are aimed at supporting well-informed choices about which model and which model parameters to include in a parameter review. Most parametric cognitive models can be parameterized in various ways. First, some cognitive models require fixing one (or more) parameters to make the model identifiable (Donkin et al., 2009b; van Maanen and Miletić, 2020). In the DDM, modelers typically fix the moment-to-moment variability of drift rate s to 0.1 or 1 for scaling purposes. Note, however, that the exact value of the scaling parameter is arbitrary, and—depending on the application—one may chose to estimate s from the data and use other parameters for scaling. We stress the importance of explicitly reporting which parameters are used for scaling purposes and the value of the scaling parameter(s) because the chosen setting influences the magnitude of the other parameter estimates. Another scaling issue relates to the measurement units of the data. For example, RTs are commonly measured in both seconds and milliseconds. Although the measurement scale influences the magnitude of the parameter estimates, none of the articles included in the present parameter review explicitly reported the measurement unit of their data. Further, articles did not consistently report all parameter estimates on the same RT scale (i.e., all parameter estimates reported in seconds, but T_{er} reported in milliseconds). Hence, we urge researchers to make an explicit statement on this matter and whenever possible stick to the same measurement unit throughout an article to avoid any ambiguity.

Second, in cognitive models one parameter is sometimes expressed as a function of one or more other parameters. The DDM, for instance, can be parameterized in terms of *absolute* starting point z or *relative* starting point $z_r = \frac{z}{a}$ (i.e., bias). The choice between z and z_r depends on the application but can also reflect default software settings. Although the two parametrizations are mathematically identical and have no consequences for the magnitude of the other parameters, it is clearly important to communicate which parameterization is used in a given application. Third, in many applications, researchers impose constraints on the model parameters across experimental manipulations, conditions, or groups. Such constraints sometimes reflect practical or computational considerations, but preferably they are based on a priori theoretical rationale (e.g., threshold parameters cannot vary based on stimulus properties that are unknown before a trial commences; Donkin et al., 2009a) or the results of model-selection procedures (e.g., Heathcote et al., 2015; Strickland et al., 2018). Regardless of the specific reasons for parameter constraints, we urge modelers to clearly communicate which parameters are hypothesized to reflect the effect(s) of interest, and so which are fixed and which are free to vary across the design. Moreover, we recommend researchers to report the competing models (including the parametrization) that were entertained to explain the data, and indicate the grounds on which a given model was chosen as best, such as AIC (Akaike, 1981), BIC

(Schwarz, 1978), DIC (Spiegelhalter et al., 2002, 2014), WAIC (Watanabe, 2010), or Bayes factors (Kass and Raftery, 1995). We note that parameter reviews are also compatible with cases where there is uncertainty about which is the best model, through the use of Bayesian model averaging (Hoeting et al., 1999). In this approach, the parameter estimates used in the review are averaged across the models in which they occur, weighted by the posterior probability of the models.

4.1.2. Model Estimation

In the face of the large number of computational tools available to implement cognitive models and the associated complex analysis pipelines, researchers have numerous choices on how to estimate model parameters. For instance, a variety of DDM software is available, such as fast-DM (Voss and Voss, 2007), HDDM (Wiecki et al., 2013), DMC (Heathcote et al., 2019), DMAT (Vandekerckhove and Tuerlinckx, 2008), using a variety of estimation methods, such as maximum likelihood, Kolmogorov-Smirnov, chi-squared minimization (Voss and Voss, 2007), quantile maximum probability (Heathcote and Brown, 2004), or Bayesian Markov chain Monte Carlo (MCMC; e.g., Turner et al., 2013) techniques. We encourage researchers to report the software they used, and whenever possible, share their commented code to enable computational reproducibility (McDougal et al., 2016; Cohen-Boulakia et al., 2017). Knowledge about the estimation software can also provide valuable information about the parametrization and scaling issues described above.

4.1.3. Parameter Estimates and Uncertainty

We recommend researchers to report all parameter estimates from their chosen model and not only the ones that are related to the experimental manipulation or the psychological effect of interest. In the DDM in particular, this would mean reporting the across-trial variability parameters, and not only the main parameters (i.e., drift rate, boundary separation, starting point, and non-decision time), even if only a subset of parameters is the focus of the study. Ideally, in the process of aggregation used to create prior distributions, estimates should be weighted by their relative uncertainty. The weighing should reflect the uncertainty of the individual estimates resulting from fitting the model to finite data and—if average parameters are used, as was the case here—also sampling error reflecting the sample size used in each study. Although we had access to the sample sizes, most studies reported parameter estimates averaged across participants without accounting for the uncertainty of the individual estimates. Moreover, the few studies that reported individual estimates provided only point estimates and failed to include measures of uncertainty. As a proxy to participant-level measures of uncertainty one may use the number of trials that provide information for the estimation of the various model parameters. However, this approach requires a level of detail about the experimental design and the corresponding model specification (including the number of excluded trials per participant) that was essentially never available in the surveyed studies.

Given these problems with reporting, we have decided to give equal weights to all (averaged) parameter estimates regardless of the sample size. The reason for this decision was that studies with large sample sizes typically used a small number of trials and likely resulted in relatively imprecise individual estimates, whereas studies with small sample sizes typically used a large number of trials and likely resulted in relatively precise individual estimates. We reasoned that as a result of this trade-off, the equal weighting may not be necessarily unreasonable. To remedy this problem in future parameter reviews, we urge researchers to either report properly weighted group average estimates or report individual estimates along with measures of uncertainty, let these be (analytic or bootstrapped) frequentist standard errors and confidence intervals (e.g., Visser and Poessé, 2017), or Bayesian credible intervals and full posterior distributions (Jeffreys, 1961; Lindley, 1965; Eberly and Casella, 2003).

4.1.4. Individual Parameters and Correlations

Ideally, researchers should report parameter estimates for each individual participant. In the vast majority of the studies examined here, only parameters averaged over participants were available. This means that we were unable to evaluate correlations among parameter estimates reflecting individual differences. Such correlations are likely quite marked. For example, in the DDM a participant with a higher drift rate, which promotes accuracy, is more likely to be able to afford to set a lower boundary and still maintain good performance, so a negative correlation between rates and boundaries might be expected. Access to individual parameters would allow estimation of these correlations, and thus enable priors to reflect this potentially important information. As we discuss below, the failure to report individual estimates brings with it important limitations on what can be achieved with the results of systematic parameter reviews.

4.2. Limitations and Future Directions

The approach to parameter reviews taken here—obtaining values from texts, tables, and graphs from published papers and performing an aggregation across studies—has the advantage of sampling estimates that are representative of a wide variety of laboratories, paradigms, and estimation methods. Indeed, for the priors presented in **Figure 3** we included a few studies with much longer RTs than are typically fit with the DDM (e.g., Lerche and Voss, 2019). The larger parameter values from these studies had the effect of broadening the tails of the fitted distributions so they represent the full variety of estimates reported in the literature.

However, this approach has a number of limitations beyond those related to the vagaries of incomplete reporting practices just discussed. The first limitation is related to the aggregation of parameter estimates over different designs. The most straightforward example concerns including parameters from studies with long RTs. The solution is equally straightforward: only including studies with RTs that fall in the range of interest specific to a particular application. A related but more subtle issue occurs in our DDM application where the meaning of the magnitude of the response bias (z_r) parameter is design specific, and so it is difficult to form useful aggregates over different paradigms. To take a concrete example, a bias toward "word" responses over "non-word" responses in a lexical-decision paradigm cannot be made commensurate with a bias favoring "left" over "right" responses in a random dot motion paradigm. Our approach—forming an aggregate with maximum uncertainty by assuming either direction is equally likely (i.e., mirroring the values)—removes any information about the average direction while at least providing some information about variability in bias. Although this approach likely overestimates the variability of the bias estimates, we believe that overestimation is preferable to underestimation which might result in an overly influential prior distribution. Again, this problem can again be avoided by constructing priors based on a more specific (in this case task-specific) aggregation. Our online data repository reports raw starting point and bias estimates, which combined with the design descriptions from the original papers could be used to perform such an aggregation. The priors for the lexical decision task reported in the **Supplementary Materials** provide an example that did not require us to mirror the bias estimates. We note, however, that we had to exclude a paper where it was unclear which response was mapped to which DDM boundary, so we would add a reporting guideline that this choice be spelled out. We also note that similar problems with aggregation over different designs are likely to occur for other parameter types and also beyond the DDM, for instance in evidence-accumulation models such as the Linear Ballistic Accumulator (Brown and Heathcote, 2008). For instance, if one decomposes drift rates in the DDM into the average over stimuli and "stimulus bias" (i.e., the difference in rates between the two stimulus classes; White and Poldrack, 2014), then the same issue applies, but now with respect stimuli rather than responses.

The second limitation—which is related to incomplete reporting, but is harder to address within a traditional journal format—concerns obtaining a full multivariate characterization of the prior distribution of parameters that takes into account correlations among parameters as well as their average values and variability. Because most estimates reported in the literature are averages over participants, we were restricted to providing separate univariate characterizations of prior distributions for each parameter. To the degree that the implicit independence assumption of this approach is violated[5] problems can arise. Continuing the example of negatively correlated rates and boundaries, although a higher value of both separately may be quite probable, both occurring together may be much less likely that the product of their individual probabilities that would be implied by independence.

Problems related to this limitation arise, for example, if in planning a new experiment one were to produce synthetic data by drawing parameter combinations independently from the univariate priors in **Figure 3**, potentially producing simulated participants with parameter values that are unlikely in a real experiment. With Bayesian methods, ignoring the correlations among parameters can compromise the efficiency of MCMC samplers and complicate the interpretation of Bayes factors because the resulting uni-variate priors will assign mass to implausible regions of the parameter space. Although standard Bayesian MCMC samplers used for evidence-accumulation models have not taken account of these population correlations, a new generation of samplers is appearing that does (Gunawan et al., 2020). This development underscores the need for future systematic parameter reviews to move in the direction of multivariate characterizations. This may be achieved by revisiting the original data sets, which due to open science practices are becoming increasingly available, refitting the DDM, and then using the resulting individual parameter estimates to form multivariate priors. This future direction will be time consuming and computationally challenging, and will no doubt bring with it new methodological problems that we have not addressed here. Nevertheless, we believe that the long-term gains for cognitive modeling will make this enterprise worthwhile.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://osf.io/9ycu5/and GitHub: https://github.com/nhtran93/DDM_priors.

AUTHOR CONTRIBUTIONS

N-HT designed the study, collected the data, performed the analyses, and wrote the manuscript. N-HT, LvM, and DM designed the study and the analyses. DM and AH provided critical feedback and helped shape the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the Utrecht University and the Max Planck Society. AH was supported by the ARC Discovery Project grant no. DP200100655. DM was supported by a Veni grant (451-15-010) from the Netherlands Organization of Scientific Research (NWO).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2020.608287/full#supplementary-material

[6] To be clear, we are not talking about correlations among parameters within a participant, which are a consequence of the mathematical form of the model's likelihood and the particular parameterization adopted for the design. Rather, we are addressing correlations at the population level, i.e., across participants. Although the two types of correlations can be related, they are not the same and in our experience can sometimes differ very markedly.

REFERENCES

Akaike, H. (1973). "Information theory and an extension of maximum likelihood principle," in *Proceedings of the Second International Symposium on Information Theory* (New York, NY), 267–281.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automat. Control* 19, 716–723. doi: 10.1109/TAC.1974.1100705

Akaike, H. (1981). Likelihood of a model and information criteria. *J. Econometr*. 16, 3–14. doi: 10.1016/0304-4076(81)90071-3

Allaire, J., Cheng, J., Xie, Y., McPherson, J., Chang, W., Allen, J., et al. (2018). *rmarkdown: Dynamic Documents for R*. R package version 1.

American Psychological Association (2017). *Psycinfo*. Available online at: https://www.apa.org/pubs/databases/psycinfo/ (accessed December 26, 2017).

Arizona (2010). *GraphClick*.

Ball, B. H., and Aschenbrenner, A. J. (2018). The importance of age-related differences in prospective memory: Evidence from diffusion model analyses. *Psychon. Bull. Rev*. 25, 1114–1122. doi: 10.3758/s13423-017-1318-4

Boehm, U., Annis, J., Frank, M. J., Hawkins, G. E., Heathcote, A., Kellen, D., et al. (2018). Estimating across-trial variability parameters of the diffusion decision model: expert advice and recommendations. *J. Math. Psychol*. 87, 46–75. doi: 10.1016/j.jmp.2018.09.004

Brown, S. D., and Heathcote, A. (2008). The simplest complete model of choice response time: linear ballistic accumulation. *Cogn. Psychol*. 57, 153–178. doi: 10.1016/j.cogpsych.2007.12.002

Busemeyer, J. R., and Stout, J. C. (2002). A contribution of cognitive decision models to clinical assessment: decomposing performance on the Bechara gambling task. *Psychol. Assess*. 14, 253–262. doi: 10.1037/1040-3590.14.3.253

Cohen-Boulakia, S., Belhajjame, K., Collin, O., Chopard, J., Froidevaux, C., Gaignard, A., et al. (2017). Scientific workflows for computational reproducibility in the life sciences: status, challenges and opportunities. *Future Generat. Comput. Syst*. 75, 284–298. doi: 10.1016/j.future.2017.01.012

Damaso, K., Williams, P., and Heathcote, A. (submitted). What does a (hu)man do after (s)he makes a fast versus slow error, and why?

Donkin, C., Averell, L., Brown, S., and Heathcote, A. (2009a). Getting more from accuracy and response time data: Methods for fitting the linear ballistic accumulator. *Behav. Res. Methods* 41, 1095–1110. doi: 10.3758/BRM.41.4.1095

Donkin, C., and Brown, S. D. (2018). "Response times and decision-making," in *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience, Volume 5: Methodology, 4th Edn*., eds E. J. Wagenmakers and J. T. Wixted (John Wiley and Sons, Inc.), 349–382.

Donkin, C., Brown, S. D., and Heathcote, A. (2009b). The overconstraint of response time models: rethinking the scaling problem. *Psychon. Bull. Rev*. 16, 1129–1135. doi: 10.3758/PBR.16.6.1129

Dutilh, G., Annis, J., Brown, S. D., Cassey, P., Evans, N. J., Grasman, R. P., et al. (2019). The quality of response time data inference: a blinded, collaborative sssessment of the validity of cognitive models. *Psychon. Bull. Rev*. 26, 1051–1069. doi: 10.3758/s13423-017-1417-2

Dutilh, G., Krypotos, A.-M., and Wagenmakers, E.-J. (2011). Task-related versus stimulus-specific practice. *Exp. Psychol*. 58, 434–442. doi: 10.1027/1618-3169/a000111

Eberly, L. E., and Casella, G. (2003). Estimating Bayesian credible intervals. *J. Stat. Plann. Inference* 112, 115–132. doi: 10.1016/S0378-3758(02)00327-0

Elsevier (2017). *Scopus*. Available online at: https://www.scopus.com/home.uri (accessed December 12, 2017).

Evans, N. J., and Annis, J. (2019). Thermodynamic integration via differential evolution: a method for estimating marginal likelihoods. *Behav. Res. Methods* 51, 930–947. doi: 10.3758/s13428-018-1172-y

Gershman, S. J. (2016). Empirical priors for reinforcement learning models. *J. Math. Psychol*. 71, 1–6. doi: 10.1016/j.jmp.2016.01.006

Gill, J. (2014). *Bayesian Methods:A Social and Behavioral Sciences Approach*. New York, NY: Chapman and Hall. doi: 10.1201/b17888

Gluth, S., and Jarecki, J. B. (2019). On the importance of power analyses for cognitive modeling. *Comput. Brain Behav*. 2, 266–270. doi: 10.1007/s42113-019-00039-w

Gomez, P., Ratcliff, R., and Perea, M. (2007). A model of the Go/No-Go task. *J. Exp. Psychol. Gen*. 136, 389–413. doi: 10.1037/0096-3445.136.3.389

Grasman, R. P., Wagenmakers, E.-J., and van der Maas, H. L. (2009). On the mean and variance of response times under the diffusion model with an application to parameter estimation. *J. Math. Psychol*. 53, 55–68. doi: 10.1016/j.jmp.2009.01.006

Gronau, Q. F., Heathcote, A., and Matzke, D. (2020). Computing bayes factors for evidence-accumulation models using Warp-III bridge sampling. *Behav. Res. Methods* 52, 918–937. doi: 10.3758/s13428-019-01290-6

Gunawan, D., Hawkins, G. E., Tran, M. N., Kohn, R., and Brown, S. D. (2020). New estimation approaches for the hierarchical Linear Ballistic Accumulator model. *J. Math. Psychol*. 96:102368. doi: 10.1016/j.jmp.2020.102368

Gutenkunst, R. N., Waterfall, J. J., Casey, F. P., Brown, K. S., Myers, C. R., and Sethna, J. P. (2007). Universally sloppy parameter sensitivities in systems biology models. *PLoS Comput. Biol*. 3:e189. doi: 10.1371/journal.pcbi.0030189

Heathcote, A., and Brown, S. (2004). Reply to Speckman and Rouder: a theoretical basis for QML. *Psychon. Bull. Rev*. 11, 577–578. doi: 10.3758/BF03196614

Heathcote, A., Lin, Y.-S., Reynolds, A., Strickland, L., Gretton, M., and Matzke, D. (2019). Dynamic models of choice. *Behav. Res. Methods* 51, 961–985. doi: 10.3758/s13428-018-1067-y

Heathcote, A., Loft, S., and Remington, R. W. (2015). Slow down and remember to remember! A delay theory of prospective memory costs. *Psychol. Rev*. 122, 376–410. doi: 10.1037/a0038952

Heck, D. W., and Erdfelder, E. (2019). Maximizing the expected information gain of cognitive modeling via design optimization. *Comput. Brain Behav*. 2, 202–209. doi: 10.1007/s42113-019-00035-0

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. *Stat. Sci*. 14, 382–401. doi: 10.1214/ss/1009212519

Horn, S. S., Bayen, U. J., and Smith, R. E. (2011). What can the diffusion model tell Us about prospective memory? *Can. J. Exp. Psychol*. 65, 69–75. doi: 10.1037/a0022808

Jaynes, E. T. (1988). "The relation of Bayesian and maximum entropy methods," in *Maximum Entropy and Bayesian Methods in Science and Engineering*, eds G. J. Erickson and C. Smith (Dordrecht: Kluwer Academic Publishers), 25–29. doi: 10.1007/978-94-009-3049-0_2

Jeffreys, H. (1961). *Theory of Probability, 3rd Edn*. Oxford, UK: Oxford University Press.

Kass, R. E., and Raftery, A. E. (1995). Bayes factors. *J. Am. Stat. Assoc*. 90, 773–795. doi: 10.1080/01621459.1995.10476572

Kennedy, L., Simpson, D., and Gelman, A. (2019). The experiment is just as important as the likelihood in understanding the prior: a cautionary note on robust cognitive modeling. *Comput. Brain Behav*. 2, 210–217. doi: 10.1007/s42113-019-00051-0

Lee, M. D. (2018). "Bayesian methods in cognitive modeling," in *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience, Volume 5: Methodology, 4th Edn*., eds E. J. Wagenmakers and J. T. Wixted (New York, NY: John Wiley Sons, Inc.), 37–84.

Lee, M. D., Criss, A. H., Devezer, B., Donkin, C., Etz, A., Leite, F. P., et al. (2019). Robust modeling in cognitive science. *Comput. Brain Behav*. 2, 141–153. doi: 10.1007/s42113-019-00029-y

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

Leite, F. P., Ratcliff, R., Lette, F. P., and Ratcliff, R. (2010). Modeling reaction time and accuracy of multiple-alternative decisions. *Attent. Percept. Psychophys*. 72, 246–273. doi: 10.3758/APP.72.1.246

Lerche, V., and Voss, A. (2019). Experimental validation of the diffusion model based on a slow response time paradigm. *Psychol. Res*. 83, 1194–1209. doi: 10.1007/s00426-017-0945-8

Lindley, D. V. (1965). *Introduction to Probability Theory and Statistics From a Bayesian Point of View*. Cambridge: Cambridge University Press.

Matzke, D., Hughes, M., Badcock, J. C., Michie, P., and Heathcote, A. (2017). Failures of cognitive control or attention? The case of stop-signal deficits in schizophrenia. *Attent. Percept. Psychophys*. 79, 1078–1086. doi: 10.3758/s13414-017-1287-8

Matzke, D., Logan, G. D., and Heathcote, A. (2020). A cautionary note on evidence-accumulation models of response inhibition in the stop-signal paradigm. *Comput. Brain Behav*. 3, 269–288. doi: 10.1007/s42113-020-00075-x

Matzke, D., and Wagenmakers, E.-J. (2009). Psychological interpretation of the ex-Gaussian and shifted Wald parameters: a diffusion model analysis. *Psychon. Bull. Rev*. 16, 798–817. doi: 10.3758/PBR.16.5.798

McDougal, R. A., Bulanova, A. S., and Lytton, W. W. (2016). Reproducibility in computational neuroscience models and simulations. *IEEE*

Trans. Biomed. Eng. 63, 2021–2035. doi: 10.1109/TBME.2016.2539602

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., and Group, T. P. (2009). Preferred reporting items for systematic reviews and meta-analyses: the prisma statement. *PLoS Med*. 6:e1000097. doi: 10.1371/journal.pmed.1000097

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

Myung, I. J., and Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: a Bayesian approach. *Psychon. Bull. Rev*. 4, 79–95. doi: 10.3758/BF03210778

National Library of Medicine (2017). *Empirical Priors for DDM & LBA*.

Navarro, D. J. (2020). If mathematical psychology did not exist we might need to invent it: a comment on theory building in psychology. *PsyArXiv*. doi: 10.31234/osf.io/ygbjp

Oberauer, K., and Lewandowsky, S. (2019). Addressing the theory crisis in psychology. *Psychon. Bull. Rev*. 26, 1596–1618. doi: 10.3758/s13423-019-01645-2

O'Callaghan, C., Hall, J. M., Tomassini, A., Muller, A. J., Walpola, I. C., Moustafa, A. A., et al. (2017). Visual hallucinations are characterized by impaired sensory evidence accumulation: Insights from hierarchical drift diffusion modeling in Parkinson's disease. *Biol. Psychiatry Cogn. Neurosci. Neuroimaging* 2, 680–688. doi: 10.1016/j.bpsc.2017.04.007

Palmer, J., Huk, A. C., and Shadlen, M. N. (2005). The effect of stimulus strength on the speed and accuracy of a perceptual decision. *J. Vis*. 5, 376–404. doi: 10.1167/5.5.1

Pitt, M. A., and Myung, J. I. (2019). Robust modeling through design optimization. *Comput. Brain Behav*. 2, 200–201. doi: 10.1007/s42113-019-00050-1

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna: R Core Team.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychol. Rev*. 85, 59–108. doi: 10.1037/0033-295X.85.2.59

Ratcliff, R. (2008). The EZ diffusion method: Too EZ? *Psychon. Bull. Rev*. 15, 1218–1228. doi: 10.3758/PBR.15.6.1218

Ratcliff, R., and McKoon, G. (2008). The diffusion decision model: theory and data for two-choice decision tasks. *Neural Comput*. 20, 873–922. doi: 10.1162/neco.2008.12-06-420

Ratcliff, R., and Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychol. Sci*. 9, 347–356. doi: 10.1111/1467-9280.00067

Ratcliff, R., Smith, P. L., Brown, S. D., and McKoon, G. (2016). Diffusion decision model: current issues and history. *Trends Cogn. Sci*. 20, 260–281. doi: 10.1016/j.tics.2016.01.007

Riefer, D. M., and Batchelder, W. H. (1988). Multinomial modeling and the measurement of cognitive processes. *Psychol. Rev*. 95, 318–339. doi: 10.1037/0033-295X.95.3.318

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychol. Bull*. 86, 638–641. doi: 10.1037/0033-2909.86.3.638

Schad, D. J., Betancourt, M., and Vasishth, S. (2020). Toward a principled Bayesian workflow in cognitive science. *Psychol. Methods*. doi: 10.1037/met0000275

Schmitz, F., and Voss, A. (2012). Decomposing task-switching costs with the diffusion model. *J. Exp. Psychol. Hum. Percept. Perform*. 38, 222–250. doi: 10.1037/a0026003

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Stat*. 6, 461–464. doi: 10.1214/aos/1176344136

Shankle, W. R., Hara, J., Mangrola, T., Hendrix, S., Alva, G., and Lee, M. D. (2013). Hierarchical Bayesian cognitive processing models to analyze clinical trial data. *Alzheimers Dement*. 9, 422–428. doi: 10.1016/j.jalz.2012.01.016

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *J. R. Stat. Soc. Ser. B* 64, 583–616. doi: 10.1111/1467-9868.00353

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2014). The deviance information criterion: 12 years on. *J. R. Stat. Soc. Ser. B* 76, 485–493. doi: 10.1111/rssb.12062

Stone, M. (1960). Models for choice-reaction time. *Psychometrika* 25, 251–260. doi: 10.1007/BF02289729

Strickland, L., Loft, S., Remington, R. W., and Heathcote, A. (2018). Racing to remember: a theory of decision control in event-based prospective memory. *Psychol. Rev*. 125, 851–887. doi: 10.1037/rev0000113

Theisen, M., Lerche, V., von Krause, M., and Voss, A. (2020). Age differences in diffusion model parameters: a meta-analysis. *Psychol. Res*. doi: 10.1007/s00426-020-01371-8

Trafimow, D. (2005). The ubiquitous Laplacian assumption: reply to Lee and Wagenmakers (2005). *Psychol. Rev*. 112, 669–674. doi: 10.1037/0033-295X.112.3.669

Turner, B. M., Sederberg, P. B., Brown, S. D., and Steyvers, M. (2013). A method for efficiently sampling from distributions with correlated dimensions. *Psychol. Methods* 18, 368–384. doi: 10.1037/a0032222

van Maanen, L., and Miletić, S. (2020). The interpretation of behavior-model correlations in unidentified cognitive models. *Psychon. Bull. Rev*. doi: 10.3758/s13423-020-01783-y

van Maanen, L., van der Mijn, R., van Beurden, M. H. P. H., Roijendijk, L. M. M., Kingma, B. R. M., Miletić, S., et al. (2019). Core body temperature speeds up temporal processing and choice behavior under deadlines. *Sci. Rep*. 9:10053. doi: 10.1038/s41598-019-46073-3

van Ravenzwaaij, D., Donkin, C., and Vandekerckhove, J. (2017). The EZ diffusion model provides a powerful test of simple empirical effects. *Psychon. Bull. Rev*. 24, 547–556. doi: 10.3758/s13423-016-1081-y

Vandekerckhove, J., and Tuerlinckx, F. (2008). Diffusion model analysis with MATLAB: a DMAT primer. *Behav. Res. Methods* 40, 61–72. doi: 10.3758/BRM.40.1.61

Vanpaemel, W. (2011). Constructing informative model priors using hierarchical methods. *J. Math. Psychol*. 55, 106–117. doi: 10.1016/j.jmp.2010.08.005

Vanpaemel, W., and Lee, M. D. (2012). Using priors to formalize theory: optimal attention and the generalized context model. *Psychon. Bull. Rev*. 19, 1047–1056. doi: 10.3758/s13423-012-0300-4

Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Stat. Comput*. 27, 1413–1432. doi: 10.1007/s11222-016-9696-4

Visser, I., and Poessé, R. (2017). Parameter recovery, bias and standard errors in the linear ballistic accumulator model. *Brit. J. Math. Stat. Psychol*. 70, 280–296. doi: 10.1111/bmsp.12100

Voss, A., Rothermund, K., and Brandtstädter, J. (2008). Interpreting ambiguous stimuli: separating perceptual and judgmental biases. *J. Exp. Soc. Psychol*. 44, 1048–1056. doi: 10.1016/j.jesp.2007.10.009

Voss, A., and Voss, J. (2007). Fast-DM: a free program for efficient diffusion model analysis. *Behav. Res. Methods* 39, 767–775. doi: 10.3758/BF03192967

Wabersich, D., and Vandekerckhove, J. (2014). The RWiener package: an R package providing distribution functions for the Wiener diffusion model. *R J*. 6, 49–56. doi: 10.32614/RJ-2014-005

Wagenmakers, E.-J., and Farrell, S. (2004). AIC model selection using Akaike weights. *Psychon. Bull. Rev*. 11, 192–196. doi: 10.3758/BF03206482

Wagenmakers, E.-J., Van Der Maas, H. L. J., and Grasman, R. P. P. P. (2007). An EZ-diffusion model for response time and accuracy. *Psychon. Bull. Rev*. 14, 3–22. doi: 10.3758/BF03194023

Watanabe, S. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *J. Mach. Learn. Res*. 11, 3571–3594. Available online at: https://dl.acm.org/doi/10.5555/1756006.1953045. doi: 10.5555/1756006.1953045

White, C. N., and Poldrack, R. A. (2014). Decomposing bias in different types of simple decisions. *J. Exp. Psychol. Learn. Mem. Cogn*. 40, 385–398. doi: 10.1037/a0034851

Wiecki, T. V., Sofer, I., and Frank, M. J. (2013). HDDM: hierarchical Bayesian estimation of the drift-diffusion model in Python. *Front. Neuroinform*. 7:14. doi: 10.3389/fninf.2013.00014

Wo, S. (2017). *Web of Science*. Available online at: http://www.webofknowledge.com/ (accessed December 26, 2017).

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 25 January 2021
doi: 10.3389/fpsyg.2020.611267

Prior Specification for More Stable Bayesian Estimation of Multilevel Latent Variable Models in Small Samples: A Comparative Investigation of Two Different Approaches

Steffen Zitzmann[1]*, *Christoph Helm*[2] *and Martin Hecht*[3]

[1] Hector Research Institute of Education Sciences and Psychology, University of Tübingen, Tübingen, Germany, [2] Institute for the Management and Economics of Education, University of Teacher Education Zug, Zug, Switzerland, [3] Department of Psychology, Humboldt-Universität zu Berlin, Berlin, Germany

OPEN ACCESS

Edited by:
Christoph Koenig,
Goethe University Frankfurt, Germany

Reviewed by:
Esther Ulitzsch,
University of Kiel, Germany
Alexander Naumann,
Leibniz Institute for Research and Information in Education (DIPF), Germany

***Correspondence:**
Steffen Zitzmann
steffen.zitzmann@uni-tuebingen.de

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *28 September 2020*
Accepted: *22 October 2020*
Published: *25 January 2021*

Citation:
Zitzmann S, Helm C and Hecht M (2021) Prior Specification for More Stable Bayesian Estimation of Multilevel Latent Variable Models in Small Samples: A Comparative Investigation of Two Different Approaches. Front. Psychol. 11:611267. doi: 10.3389/fpsyg.2020.611267

Bayesian approaches for estimating multilevel latent variable models can be beneficial in small samples. Prior distributions can be used to overcome small sample problems, for example, when priors that increase the accuracy of estimation are chosen. This article discusses two different but not mutually exclusive approaches for specifying priors. Both approaches aim at stabilizing estimators in such a way that the Mean Squared Error (MSE) of the estimator of the between-group slope will be small. In the first approach, the MSE is decreased by specifying a slightly informative prior for the group-level variance of the predictor variable, whereas in the second approach, the decrease is achieved directly by using a slightly informative prior for the slope. Mathematical and graphical inspections suggest that both approaches can be effective for reducing the MSE in small samples, thus rendering them attractive in these situations. The article also discusses how these approaches can be implemented in M*plus*.

Keywords: Bayesian estimation, Markov chain Monte Carlo, multilevel modeling, structural equation modeling, small sample

As van de Schoot et al. (2017) pointed out, the number of applications of Bayesian approaches is growing quickly, mainly because software that is easy to use such as M*plus* (Muthén and Muthén, 2012) is providing Bayesian estimation as an option. Bayesian approaches can be beneficial in several respects, for example, by offering greater flexibility (e.g., Hamaker and Klugkist, 2011; Muthén and Asparouhov, 2012; Lüdtke et al., 2013) or fewer estimation problems (e.g., Hox et al., 2012; Depaoli and Clifton, 2015; Zitzmann et al., 2016), particularly when latent variable models are estimated. One major difference between Bayesian and traditional Maximum Likelihood (ML) estimation is that the former not only uses the information from the data at hand (i.e., the likelihood function) but combines it with additional information from what is called the prior distribution. Inferences are based on the result of this combination, that is, the posterior distribution. Scholars have advised researchers against the use of default priors in an automatic fashion and have encouraged them to specify priors on their own (e.g., McNeish, 2016; Smid et al., 2020). This may be an obstacle to some researchers. However, the prior can also be considered a feature of Bayesian estimation that can be used to improve estimation by choosing a favorable prior—a task that is particularly challenging but also particularly worth pursuing when the sample size is small.

The choice of prior has received a lot of attention in the methodological literature (e.g., Natarajan and Kass, 2000; Gelman, 2006; Chung et al., 2013), and scholars have made different suggestions about how priors can be specified in advantageous ways. Only recently, Smid et al. (2020) discussed how priors can be "thoughtfully" constructed on the basis of previous knowledge about the parameter of interest (e.g., on the basis of a previous study or a meta-analysis) in order to reduce small-sample bias. However, it has been argued that the variability of an estimator should not be ignored when evaluating the quality of a method (e.g., Greenland, 2000; Zitzmann et al., 2020), particularly when the sample size is small. Therefore, other suggestions for specifying the prior have been aimed at reducing the Mean Squared Error (MSE), which combines bias and variability: MSE $=$ bias^2 $+$ variability. One such approach was proposed by Zitzmann et al. (2015), who focused on the between-group slope in multilevel latent variable modeling. The authors suggested that researchers should suitably modify the estimator of the group-level variance of the predictor variable because this will result in a more stable (i.e., more accurate) estimator of the slope. To this end, a slightly informative prior is specified for the group-level variance of the predictor to pull the variance estimates away from zero (i.e., the indirect approach). By doing so, the estimates of the slope will not be too large, and the MSE of the estimator of the slope will be reduced. Notably, in contrast to Smid et al.'s (2020) suggestion, the prior does not need to match previous knowledge or the true value of the parameter in the population. Rather, an incorrect prior whose location deviates from the parameter in the population might reduce the MSE even more than a correct prior will. Zitzmann et al. (2015) found that for a standardized predictor (standardized at Level 1), a slightly informative inverse gamma prior for the group-level variance provided a somewhat biased but much more accurate (because it had a smaller MSE) estimator in small samples. Alternatively, to reduce the MSE of the estimator of the slope, one can specify a slightly informative prior directly for the slope in order to shrink the estimates and thereby ensure that they will not be too large (i.e., the direct approach).

In the present article, we mathematically work out the idea behind the direct approach for a simple multilevel latent variable model, and we contrast this approach with the indirect approach and with ML. Then, we graphically show the benefits that both approaches have over ML when the sample size is small. Finally, we discuss how these approaches can be implemented in M*plus*.

1. EXAMPLE MODEL

Before we go into detail, we present an example model that we will use later to illustrate the different strategies. The model was suggested by Lüdtke et al. (2008) as one way to yield (asymptotically) unbiased estimates of between-group slopes in contextual studies (see also Asparouhov and Muthén, 2019). To this end, on the group level in the model, the dependent variable Y is predicted by a latent variable (i.e., the latent group mean) instead of the unreliable manifest group mean of the predictor variable, which is why the model was named the *multilevel latent covariate model* (Lüdtke et al., 2008). Such latent group means have become part of many more complex multilevel structural equation models that are commonly applied in research practice (see Preacher et al., 2010, 2016, for overviews of such models).

More specifically, the individual-level predictor X splits into two uncorrelated and normally distributed parts: a between-group part X_b, which is the latent group mean, and a within-group part X_w, which is the individual deviation from X_b. For a person $i = 1, \ldots, n$ in group $j = 1, \ldots, J$, the decomposition thus reads:

$$X_{ij} = X_{b,j} + X_{w,ij} \tag{1}$$

$X_{b,j}$ is distributed around μ_X with variance τ_X^2, whereas the deviation $X_{w,ij}$ has variance σ_X^2. Hereafter, we will also call σ_X^2 and τ_X^2 the within-group and between-group variances of X, respectively.

Applying Raudenbush and Bryk's (2002) notation, the regression at the individual level reads:

$$\text{Level 1: } Y_{ij} = \beta_{0j} + \beta_w X_{w,ij} + \varepsilon_{ij} \tag{2}$$

where β_w is the (fixed) within-group slope that describes the relationship between the predictor and the dependent variable at the individual level, and the ε_{ij} are normally distributed residuals. The residual variance is σ_Y^2. At the group level, the intercept β_{0j} is regressed on X_b:

$$\text{Level 2: } \beta_{0j} = \alpha + \beta_b X_{b,j} + \delta_j \tag{3}$$

where α is the overall intercept, and β_b is the between-group slope (i.e., the relationship between X and Y at the group level). The δ_j are normally distributed residuals with variance τ_Y^2.

Here, we focus on the between-group slope (β_b), which is of great interest in many applications of multilevel models (e.g., in the analysis of contextual effects). When the data are balanced (i.e., equal numbers of persons per group), the ML estimator of β_b is given by:

$$\hat{\beta}_b = \frac{\hat{\tau}_{YX}}{\hat{\tau}_X^2} \tag{4}$$

where $\hat{\tau}_X^2$ and $\hat{\tau}_{YX}$ are sample estimators of the group-level variance of X and the group-level covariance of X and Y, respectively.

Some statistical properties of the ML estimator in Equation 4 need to be discussed first to be able to compare this estimator with the Bayesian estimators later on. First, by using the first-order Taylor expansion (e.g., Casella and Berger, 2001; see also Grilli and Rampichini, 2011) and ignoring terms involving higher order factors such as $\frac{1}{n^2(n-1)}$ or $\frac{1}{n^2}$ for better readability, then the bias of $\hat{\beta}_b$ can roughly be approximated by:

$$\mathrm{E}\left(\hat{\beta}_b\right) - \beta_b \approx -\frac{2}{J-1}\left\{-\frac{2\left(1-\rho_X\right)}{n\rho_X} + \frac{1-\rho_X}{n\rho_X}\left(1+\frac{\beta_w}{\beta_b}\right)\right\}\beta_b \tag{5}$$

where $\rho_X = \frac{\tau_X^2}{\tau_X^2+\sigma_X^2}$ is the intraclass correlation (ICC) of X.[1] This equation could be simplified further, but we continue to use this expression here to emphasize formal similarities with the biases of the Bayesian estimators (see below). However, even in its current form, it is evident from Equation (5) that the bias critically depends on the sample size (because J occurs in the denominator) and that the bias is generally non-zero in small samples. However, if we let J become large, the bias diminishes because $\frac{1}{J-1}$ becomes very small—a property of the estimator that we refer to as "asymptotic unbiasedness." In a similar way, we can yield an approximation of the variability of $\tilde{\beta}_b$:

$$\mathrm{Var}\left(\hat{\beta}_b\right) \approx \frac{1}{J-1}\left\{\left[\frac{\rho_Y}{\rho_X} + \frac{1-\rho_X}{n\rho_X}\left(\frac{\rho_Y}{\rho_X} + \frac{1-\rho_Y}{1-\rho_X}\right)\right]\frac{\tau_Y^2+\sigma_Y^2}{\tau_X^2+\sigma_X^2} + \left[-1-\frac{2\left(1-\rho_X\right)}{n\rho_X}\frac{\beta_w}{\beta_b}\right]\beta_b^2\right\} \tag{6}$$

where $\rho_Y = \frac{\tau_Y^2}{\tau_Y^2+\sigma_Y^2}$ is the ICC of Y. Similar to the bias, the variability depends on the sample size in such a way that the variability will be small when J is large. Because the MSE of $\hat{\beta}_b$ is the sum of the squared bias and the variability,

$$\mathrm{MSE}\left(\hat{\beta}_b\right) \approx \left[\mathrm{E}\left(\hat{\beta}_b\right) - \beta_b\right]^2 + \mathrm{Var}\left(\hat{\beta}_b\right) \tag{7}$$

this measure will be small as well. Taken together, the more information the data provide, the more the overall accuracy of the estimator improves.

Whereas the asymptotic properties are favorable, the ML estimator tends to be biased in small samples, and it has high variability and thus a large MSE in these situations (e.g., McNeish, 2017). This challenges the usefulness of the ML estimator when the sample size is small because the result from a single study might be highly inaccurate. Therefore, scholars have called for alternative estimators that are less variable and thus more accurate (i.e., they have a smaller MSE), although they might be more biased than ML. In the multilevel literature, such estimators have been suggested by Chung et al. (2013), Greenland (2000), Grilli and Rampichini (2011), and Zitzmann et al. (2015), for example. Next, we develop the direct strategy, and recap the indirect strategy of specifying the prior.

2. THE DIRECT STRATEGY

We refer to the first strategy as the *direct strategy* because the prior is specified directly for the between-group slope (β_b). To illustrate, we assume a normal prior, which can be formalized as:

$$\beta_b \sim \mathrm{N}\left(a, b\right) \tag{8}$$

which should be read as "β_b is normally distributed with mean a and variance b." However, for better interpretability, we employ another, more convenient parameterization. Instead of a and b, we use the terms β_0 and $\frac{\tau_Y^2}{\nu_0\tau_X^2}$:

$$\beta_b \sim \mathrm{N}\left(\beta_0, \frac{\tau_Y^2}{\nu_0\tau_X^2}\right) \tag{9}$$

As we will show, β_0 and ν_0 can be meaningfully interpreted.

One way of expressing the likelihood for the slope is:

$$\beta_b \sim \mathrm{N}\left(\hat{\beta}_b, \frac{\hat{\tau}_Y^2}{J\hat{\tau}_X^2}\right) \tag{10}$$

where $\hat{\tau}_Y^2$ and $\hat{\tau}_X^2$ are the sampling variances of τ_Y^2 and τ_X^2, respectively. If we combine the prior in Equation (9) with the likelihood, we obtain the following posterior:

$$\beta_b \sim \mathrm{N}\left(\frac{\nu_0}{\nu_0+J}\beta_0 + \frac{J}{\nu_0+J}\hat{\beta}_b, \frac{J}{\nu_0+J}\frac{\hat{\tau}_Y^2}{J\hat{\tau}_X^2}\right) \tag{11}$$

which is also a normal distribution. The mean of this distribution defines the Bayesian Expected A Posteriori (EAP) estimator, which is the standard choice for a point estimator in Bayesian estimation (Note that the Bayes module in M*plus* uses the median of the posterior). With $w = \frac{J}{\nu_0+J}$, this Bayesian estimator can also be expressed as:

$$\bar{\beta}_b = \left(1-w\right)\beta_0 + w\hat{\beta}_b \tag{12}$$

As can be seen from the equation, the estimator is simply the weighted average of the mean of the prior (β_0) and $\hat{\beta}_b$, which suggests straightforward interpretations for the parameters of the prior. One may think of β_0 as the *prior guess* for the between-group slope and ν_0 as the *prior sample size* (see also Hoff, 2009). These interpretations are substantiated by the observation that the larger ν_0, the smaller w, and the more the estimates shrink toward β_0. Less technically speaking, when we are more confident in β_0, the prior will gain more weight, and the posterior will shift to the mean of the prior. However, when we choose ν_0 to be very small, w will be close to 1, and $\bar{\beta}_b$ will be similar to $\hat{\beta}_b$, which justifies the view that the modified estimator includes the original ML estimator as a limiting case. Notice that the prior guess does not need to represent previous knowledge about β_b. Rather, it could be set to a value that is much smaller than what previous studies have suggested and also much smaller than the parameter in the population. However, such an "incorrect" prior guess might still be beneficial, particularly when the sample size is small.

To be able to compare the properties of the Bayesian estimator with the ML estimator and with the Bayesian estimator from the second strategy of specifying the prior, we again use the Taylor expansion, and we ignore terms involving higher order factors. A rough approximation of the bias of $\bar{\beta}_b$ is then given by:

$$\mathrm{E}\left(\bar{\beta}_b\right) - \beta_b \approx \left(1-w\right)\beta_0 + \left\{-\left(1-w\right) - \frac{2w}{J-1}\left[-\frac{2\left(1-\rho_X\right)}{n\rho_X} + \frac{1-\rho_X}{n\rho_X}\left(1+\frac{\beta_w}{\beta_b}\right)\right]\right\}\beta_b \tag{13}$$

[1] The ICC quantifies the amount of the total variance that can be attributed to differences between the groups (e.g., Snijders and Bosker, 2012).

Similar to the ML estimator, $\bar{\beta}_b$ is generally biased when the sample size is small. However, the bias vanishes when J approaches infinity because w approaches 1, and $\frac{1}{J-1}$ approaches 0 (asymptotic unbiasedness). Moreover, if ν_0 is set to a value close to 0, the bias will become similar to the bias of $\hat{\beta}_b$.

The variability of $\bar{\beta}_b$ can be approximated as:

$$\mathrm{Var}\left(\bar{\beta}_b\right) \approx \frac{w^2}{J-1}\left\{\left[\frac{\rho_Y}{\rho_X}+\frac{1-\rho_X}{n\rho_X}\left(\frac{\rho_Y}{\rho_X}+\frac{1-\rho_Y}{1-\rho_X}\right)\right]\frac{\tau_Y^2+\sigma_Y^2}{\tau_X^2+\sigma_X^2} + \left[-1-\frac{2\left(1-\rho_X\right)}{n\rho_X}\frac{\beta_w}{\beta_b}\right]\beta_b^2\right\} \quad (14)$$

With a very large J, the variability becomes negligibly small, and the same holds for the MSE. However, the more interesting questions are: How does the MSE of $\bar{\beta}_b$ depend on the prior parameters (β_0, ν_0) when the sample size is small, and how must they be chosen such that the MSE will be smaller than the MSE of ML? Before we compare the different choices for (β_0, ν_0), we present another strategy for specifying the prior. Alternatively to specifying the prior directly for the between-group slope, one can also specify a prior for the group-level variance of the predictor, thereby also modifying the estimator of the slope. We call this strategy the *indirect strategy*.

3. THE INDIRECT STRATEGY

The principle that underlies the indirect strategy was discovered in the early years of Structural Equation Modeling (SEM), where models were fit on the basis of the variances and covariances of variables. One observation was that when the sample size was small, covariance matrices tended to be on the border of positive definiteness (e.g., a variance estimate close to 0, correlations close to −1 or 1; e.g., van Driel, 1978; Dijkstra, 1992; Kolenikov and Bollen, 2012). Hence, estimators of slope parameters tended to have high variability and thus also a large MSE. This led Yuan and Chan (2008) to develop the ridge technique to mitigate such problems as it modifies the estimator of the covariance matrix by adding a small value to the main diagonal (see also Yuan and Chan, 2016; Yang and Yuan, 2019). The main idea behind this technique can also be adapted for Bayesian estimation. Papers by Chung et al. (2013), Chung et al. (2015), or Zitzmann et al. (2015) are good examples of this. By means of simulation, Zitzmann et al. (2015) verified that specifying a slightly informative prior for the group-level variance of the predictor that pulls estimates of this variance slightly away from zero can increase the accuracy of the estimator of the between-group slope by reducing its MSE. Note that pulling the variance estimates away from zero corresponds to adding a value to these estimates. A formal argument for why such a prior reduces the MSE was only recently presented by Zitzmann et al. (2020). For reasons of completeness and comparability with the two previously presented estimators, we illustrate the strategy here once more, using the example model from above.

Rather than beginning with the assumption of a normal prior for the between-group slope, we begin with a gamma prior for the inverse of the group-level variance of the predictor variable (τ_X^2):

$$\frac{1}{\tau_X^2} \sim \mathrm{Gamma}\left(a, b\right) \quad (15)$$

where a and b are the parameters of the gamma distribution.[2] Equation (15) reads "τ_X^2 is inverse-gamma distributed." As for the normal prior in the previous section, we employ a reparameterization for better interpretability later on. If we set a to $\frac{\nu_0}{2}$ and b to $\frac{\nu_0\tau_0^2}{2}$, the prior reads:

$$\frac{1}{\tau_X^2} \sim \mathrm{Gamma}\left(\frac{\nu_0}{2}, \frac{\nu_0\tau_0^2}{2}\right) \quad (16)$$

where, as we will show, τ_0^2 and ν_0 have interpretations similar to those of the parameters of the (reparameterized) normal prior.

The likelihood for the inverse of the group-level variance can be written as:

$$\frac{1}{\tau_X^2} \sim \mathrm{Gamma}\left(\frac{J}{2}, \frac{J\hat{\tau}_X^2}{2}\right) \quad (17)$$

where $\hat{\tau}_X^2$ is the sample variance. Combined with the prior in Equation (16), we yield the inverse gamma posterior:

$$\frac{1}{\tau_X^2} \sim \mathrm{Gamma}\left(\frac{\nu_0+J}{2}, \frac{\nu_0\tau_0^2+J\hat{\tau}_X^2}{2}\right) \quad (18)$$

As Zitzmann et al. (2020) showed in their Appendix C, the mean of this distribution can be approximated as:

$$\bar{\tau}_X^2 \approx \left(1-w\right)\tau_0^2 + w\hat{\tau}_X^2 \quad (19)$$

where $w = \frac{J}{\nu_0+J}$. This equation defines the Bayesian EAP estimator of τ_X^2. It is interesting to note that the equation resembles Equation (12). The right-hand side of the equation is also a weighted average, and τ_0^2 and ν_0 can be thought of as the prior guess and the prior sample size, respectively (see Hoff, 2009; Lüdtke et al., 2018; Zitzmann et al., 2020).

Adding a prior for τ_X^2 also has consequences for the estimator of the between-group slope (β_b). Replacing the denominator in Equation (4) ($\hat{\tau}_X^2$) with $\bar{\tau}_X^2$ results in:

$$\tilde{\beta}_b = \frac{\hat{\tau}_{YX}}{\left(1-w\right)\tau_0^2 + w\hat{\tau}_X^2} \quad (20)$$

This new estimator is indicated by a tilde (∼) in order to better differentiate it from the ML estimator and from the Bayesian estimator that results from the direct strategy of specifying the prior (Equation 12).

[2] The inverse of a variance is sometimes also referred to as the precision in the statistical literature (e.g., Hoff, 2009).

To derive some properties of $\tilde{\beta}_b$, we apply exactly the same reasoning that led to the respective properties of $\hat{\beta}_b$ and $\bar{\beta}_b$. Accordingly, the bias of $\tilde{\beta}_b$ is roughly:

$$\mathrm{E}\left(\tilde{\beta}_b\right) - \beta_b \approx (f-1)\,\beta_b - \frac{2wf^2}{J-1}\left\{-wf\left[1+\frac{2\,(1-\rho_X)}{n\rho_X}\right]+1 + \frac{1-\rho_X}{n\rho_X}\left(1+\frac{\beta_w}{\beta_b}\right)\right\}\beta_b \tag{21}$$

where f is used as an abbreviation for the ratio $\frac{\tau_X^2}{(1-w)\tau_0^2+w\tau_X^2}$.[3] Notice that the equation implies that $\tilde{\beta}_b$ is generally biased when the sample size is finite, whereas the bias diminishes when J approaches infinity (asymptotic unbiasedness). Moreover, the bias becomes similar to the biases of $\bar{\beta}_b$ and $\hat{\beta}_b$ when we let ν_0 become very small.

The variability of $\tilde{\beta}_b$ is:

$$\begin{aligned}\mathrm{Var}\left(\tilde{\beta}_b\right) \approx \frac{f^2}{J-1}&\left\{\left[\frac{\rho_Y}{\rho_X}+\frac{1-\rho_X}{n\rho_X}\left(\frac{\rho_Y}{\rho_X}+\frac{1-\rho_Y}{1-\rho_X}\right)\right]\frac{\tau_Y^2+\sigma_Y^2}{\tau_X^2+\sigma_X^2}\right.\\ &+\left[2wf\left(wf\left(1+\frac{2\,(1-\rho_X)}{n\rho_X}\right)\right.\right.\\ &\left.\left.-\,2\left(1+\frac{1-\rho_X}{n\rho_X}\left(1+\frac{\beta_w}{\beta_b}\right)\right)\right)+1\\ &\left.\left.+\frac{2\,(1-\rho_X)}{n\rho_X}\frac{\beta_w}{\beta_b}\right]\beta_b^2\right\}\end{aligned} \tag{22}$$

Similar to the previous equation, we can easily infer that when J is large, the variability will be small and, thus, the MSE, which combines bias and variability, will be small as well—an observation that once again demonstrates the role of the sample size in determining the accuracy of estimations. However, as mentioned above, it is much more interesting to ask how the prior parameters τ_0^2 and ν_0 must be chosen such that the MSE will be reduced in comparison with ML in small samples.

4. COMPARING THE MSEs IN SMALL SAMPLES

In this section, we investigate the MSE of the different strategies for specifying priors in small samples for different choices of the prior parameters, using the example model from above to simulate data that are typical in psychology. Because it is difficult to infer from the equations how the MSEs compare with each other, they were plotted against the sample size to allow for graphical comparisons.

In accordance with Lüdtke et al. (2008), we considered the case of standardized variables (standardized at Level 1), and we assumed that the between-group slope (β_b) was 0.7 in the population. Moreover, we set the number of persons per group (n) to 5, which is not uncommon in many subdisciplines of psychology, including organizational, personality, and social psychology. The ICC of the predictor was 0.1, which could be considered small- to medium-sized compared with typical ICCs (Snijders and Bosker, 2012; Zitzmann et al., 2015). The sample size at the group level (J) was varied from 20 to 60 groups because these numbers represent small sample sizes (e.g., Hox et al., 2012; see also Hox et al., 2010) and the aim was to compare the estimators in these situations.

Figure 1 depicts a normalized version of the MSE, the Root Mean Squared Error (RMSE), for five different estimators of the slope. The first estimator in the figure is the ML estimator (solid black line). The second estimator (blue dashed line) is the Bayesian estimator that results when the direct strategy is combined with a correct prior for β_b (i.e., the prior guess, β_0, equals the parameter in the population). Because β_b was 0.7 in the population, a correct prior for β_b was specified by setting β_0 equal to this value. The third estimator (blue dotted line) also resulted from the direct strategy. However, β_0 was set to 0 (and thus well below 0.7) in order to shrink estimates that were too large toward zero. The fourth estimator (red dashed line) resulted from the indirect strategy with a correct prior for the group-level variance of the predictor (τ_X^2). The prior guess (τ_0^2) was set to 0.1, which was the value of τ_X^2 in the population.[4] The fifth estimator (red dotted line) resulted from the indirect strategy as well. However, β_0 was set to 1, which was above the parameter in the population. Thus, estimates of the variance were pulled away from zero, and, therefore, the estimates of the slope were shrunken. The three different panels of **Figure 1** show the RMSEs for different values of ν_0: 0.1 (upper left), 1.0 (upper right), and 5.0 (lower left). The first two values can be considered choices that are only slightly informative, whereas the latter is more informative and was used here to illustrate what happens to the RMSE when the priors become more informative.

As can be seen in the **Figure 1**, the different estimators tended to provide different RMSEs. The RMSE was largest for the ML estimator, whereas the RMSE was reduced when a Bayesian estimator was used. The reduction was particularly pronounced when J was very small. In addition and more important, the extent of the reduction also depended on the strategy for specifying the prior and the choices for the prior parameters. Although the direct strategy reduced the RMSE overall, the RMSE was slightly smaller when this strategy was combined with an incorrect prior (i.e., $\beta_0 = 0$) than with a correct prior (i.e., $\beta_0 = 0.7$). Moreover, the choice of a larger ν_0 was associated with a smaller RMSE. However, the smallest RMSEs emerged when the indirect strategy was used with an incorrect prior (i.e., $\tau_0^2 = 1$, which was also the upper bound of τ_X^2 due to standardization). With a larger value of $\nu_0 = 1$, the RMSE was reduced relative to a ν_0 of 0.1. However, setting ν_0 to 5 did not yield an RMSE that was even smaller. Rather, the RMSE was slightly larger than with a ν_0 of 1 because the bias induced by the prior outweighed the variability in the computation of the RMSE. Additional results

[3] We would like to state that we recognized a typo in the bias formula of Zitzmann et al.'s (2020) original publication. There should be a minus sign in front of the first term of the curly-bracketed expression. Equation (21) presents the corrected formula. However, the numerical results on which Figure 3 in Zitzmann et al. (2020) was based were not affected by the typo because these results were generated from formulas that were correct and also provided even more precise approximations than the ones presented in the article (because terms with higher order factors were not omitted).

[4] Because of the standardization, τ_X^2 is equal to the value of the ICC.

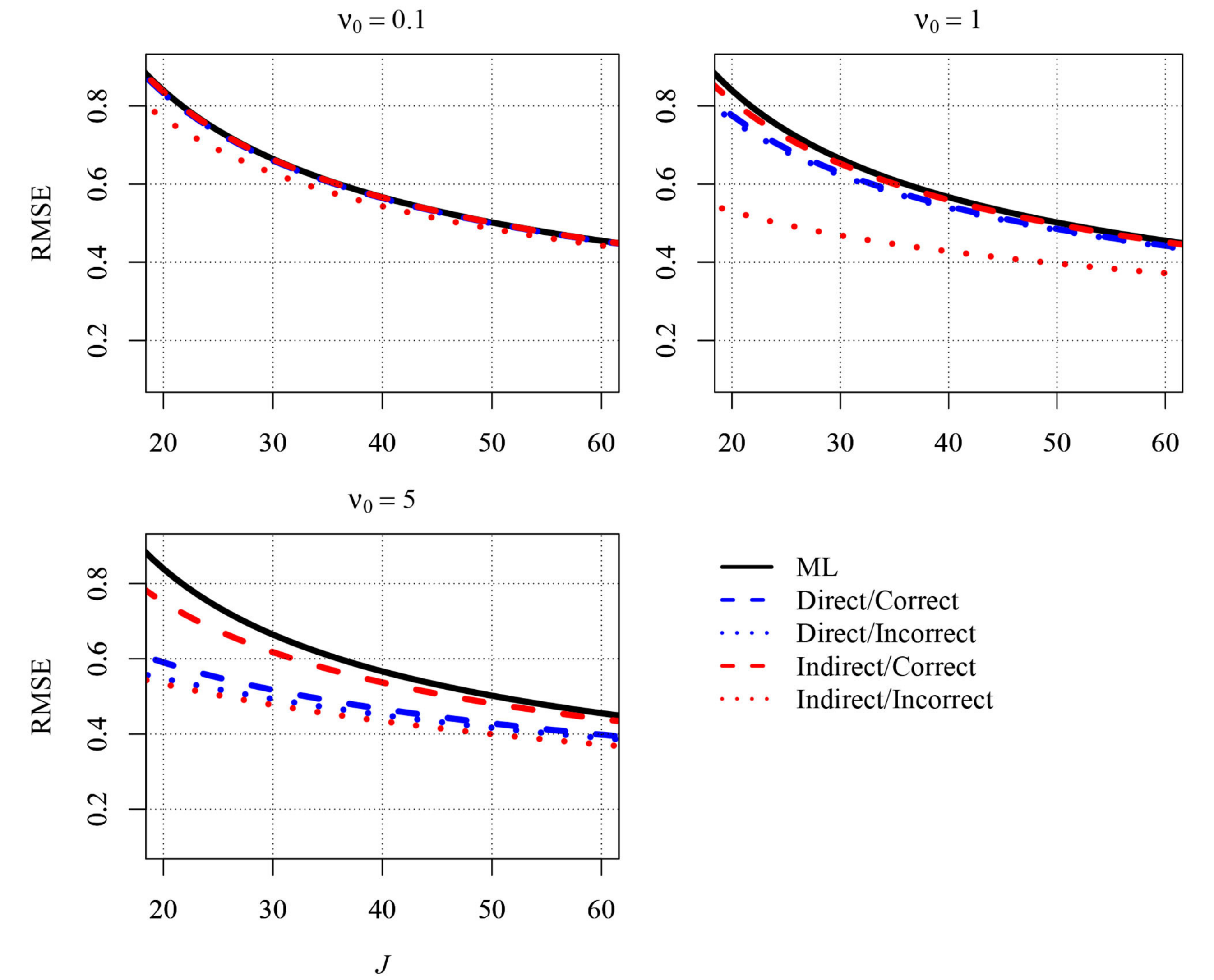


FIGURE 1 | The analytically derived Root Mean Squared Error (RMSE) in estimating the between-group slope for the direct and the indirect approach as a function of the sample size at the group level (J) and the prior distribution. Results are shown for $n = 5$ persons per group and an intraclass correlation of ICC = 0.1. ML, maximum likelihood; direct, the prior was specified directly for the between-group slope; indirect, the prior was specified for the group-level variance of the predictor variable; correct, correct prior (i.e., the prior guess equaled the value of the parameter in the population); incorrect, incorrect prior (i.e., the prior guess deviated from the parameter in the population); ν_0, prior sample size.

are presented in the **Appendix**. **Figure A1** shows the RMSEs of the different estimators for a larger number of 10 persons per group, whereas **Figure A2** shows the RMSEs for a higher ICC of .2. Although the RMSEs were smaller in **Figures A1**, **A2** compared with **Figure 1**, the big picture was similar overall: The different estimators provided different RMSEs. All Bayesian estimators provided smaller RMSEs than the ML estimator in very small samples except the indirect strategy with an incorrect informative prior.

To sum up, both strategies for specifying the prior offer attractive ways to obtain more accurate estimators of the between-group slope in small samples when used with slightly informative priors. Especially when no previous knowledge exists about the parameters, the choice of a relatively small prior guess for the between-group slope or a relatively large prior guess for the group-level variance of the predictor could be useful when these choices are combined with a small ν_0 in the low one-digit range. Although somewhat biased, the resulting Bayesian estimators of the slope were found to be more accurate than ML when the sample size was small.

5. DISCUSSION

It has been argued that Bayesian approaches can be beneficial when the sample size is small because prior distributions can be used to increase estimation accuracy. In the present article, we focused on the between-group slope because this parameter is often of interest in multilevel latent variable modeling. Two approaches for specifying priors can be distinguished, both of which are aimed at reducing the MSE of the estimator of the between-group slope: In the first approach, a slightly informative prior is specified directly for the slope, whereas in the indirect approach, the MSE is reduced by using a slightly informative prior for the group-level variance of the predictor variable. In the present article, we worked out the former approach mathematically and compared it with the indirect approach and with ML. Graphical inspections suggested that both approaches can be very effective in reducing the MSE compared with ML in small samples, rendering them attractive for researchers. We would like to add that these approaches are not mutually exclusive and that researchers can also apply them simultaneously by specifying slightly informative priors for the slope as well as for the group-level variance of the predictor variable. To provide initial information about how such a simultaneous application of the two approaches performs, we conducted an additional simulation study with 20 to 60 groups, 5 persons per group, and an ICC of the predictor variable of 0.1. **Figure 2** depicts the RMSE for five different estimators of the slope. The first estimator is the ML estimator (solid black line). The second estimator (blue dashed line) is the Bayesian estimator that resulted when the direct strategy and the indirect strategy were simultaneously applied and combined with correct priors for the between-group slope and the group-level variance of the predictor, respectively. The third estimator (blue dotted

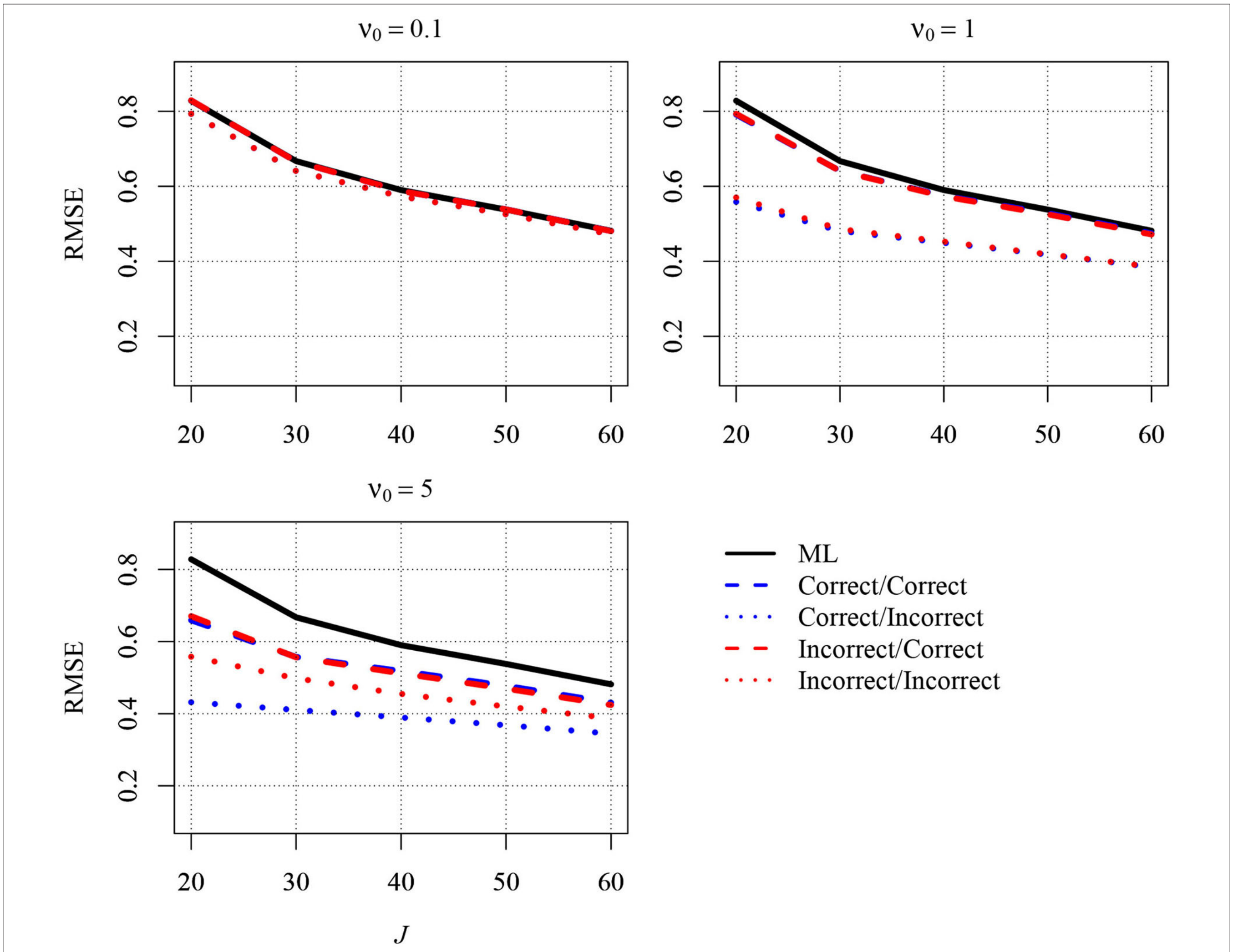


FIGURE 2 | The simulated Root Mean Squared Error (RMSE) in estimating the between-group slope for the combined approach as a function of the sample size at the group level (J) and the prior distribution. Results are shown for $n = 5$ persons per group and an intraclass correlation of ICC = 0.1. ML, maximum likelihood; correct/correct, correct priors (i.e., the prior guesses equaled the values of the parameters in the population) were specified for the between-group slope and the group-level variance of the predictor variable; correct/incorrect, a correct prior was specified for the between-group slope, and an incorrect prior (i.e., the prior guess deviated from the parameter in the population) was specified for the group-level variance of the predictor variable; incorrect/correct, an incorrect prior was specified for the between-group slope, and a correct prior was specified for the group-level variance of the predictor variable; incorrect/incorrect, incorrect priors were specified for the between-group slope and the group-level variance of the predictor variable; ν_0, prior sample size.

line) also resulted from combining the two strategies. However, whereas the direct strategy was combined with a correct prior, the indirect strategy was combined with an incorrect prior. The fourth estimator (red dashed line) resulted from simultaneously applying the direct strategy with an incorrect prior and the indirect strategy with a correct prior. The fifth estimator (red dotted line) resulted from the simultaneous application of the two strategies as well. However, both strategies were combined with incorrect priors. The three different panels of the figure show the RMSEs for different values of the prior sample size. Again, the RMSE was largest for the ML estimator, whereas the RMSE was reduced when a Bayesian estimator was used, particularly when this estimator was combined with slightly informative priors and the sample size was small. Thus, the overall finding from this simulation confirmed that a simultaneous application of the two approaches (i.e., specifying slightly informative priors for the slope as well as for the group-level variance of the predictor variable) can also be beneficial. However, because the consequences of such a use could not be studied exhaustively here, it would be interesting to conduct a more thorough simulation on this topic in future research.

Although our findings were generally favorable and could be considered a successful "proof of concept," a word of caution is nevertheless needed. Our demonstrations were very limited. For example, the specific conditions we studied do not completely reflect real data. Future research should consider a wider range of conditions for more conclusive findings. Moreover, the example model we used was overly simple. Realistic models typically involve more than one predictor and also multiple indicators per construct. However, one can derive the Bayesian estimators analogously in this more general multivariate case. Zitzmann (2018) even showed that in a multilevel SEM with two latent predictors with three indicators each, a slightly informative inverse Wishart prior for the covariance matrix of the predictors led to more accurate estimators of the between-group slopes, particularly when the samples size was small. Finally, the MSEs of the estimators we derived were only rough approximations. These approximations can nevertheless be useful for deriving hypotheses about which prior works well under which condition.

Before we come to M*plus*, we wish to acknowledge that parameter stabilization does not require Bayesian estimation. In fact, the idea of using slightly informative priors is similar to using techniques from the frequentist framework (Hastie et al., 2009). For example, the weighting parameter (w) of the Bayesian estimator in Equation (12) has an effect similar to that achieved by the penalty in regularized SEM (e.g., Jacobucci et al., 2016), and the weighting parameter in Equation (19) corresponds with the tuning parameter in ridge generalized least squares (e.g., Yuan and Chan, 2016) and regularized consistent partial least squares estimation (e.g., Jung and Park, 2018). Despite the existence of these methods, we employed Bayesian estimation here for reasons of convenience and because this type of estimation is an option in M*plus*, which is the software that many researchers use to fit multilevel latent variable models.

M*plus* does not use Bayesian estimation as the default, and users must request it by setting ESTIMATOR to BAYES. Next, to yield a more accurate estimator of the between-group slope by using a slightly informative prior for this parameter, users must specify such a prior manually. In M*plus*, normal priors are parameterized as in Equation (8), where a is the mean and b is the variance. Thus, to specify a normal prior with the prior guess (β_0) and the prior sample size (ν_0) equaling 0 and 1, respectively, users must compute a and b first. Given $a = \beta_0$ and $b = \frac{\tau_Y^2}{\nu_0 \tau_X^2}$, we yield an a of 0 and a b of $\frac{\tau_Y^2}{\tau_X^2}$. Because τ_Y^2 and τ_X^2 are unknown, they need to be replaced with, for example, their sample estimates. Assuming that these estimates are $\hat{\tau}_Y^2 = 0.15$ and $\hat{\tau}_X^2 = 0.1$, then the prior is specified by the following line of code:

```
MODEL PRIORS:
    Name of slope ~ N(0, 1.5);
```

Our findings suggest that this prior increases the accuracy of estimation in small samples. Choosing an even smaller value for b can also be useful in these situations. Alternatively, one could also specify a slightly informative prior for the group-level variance of the predictor. To be able to do this, users must compute the parameters a and b in Equation (15) because M*plus* uses this parameterization of the inverse gamma prior. Setting both τ_0^2 and ν_0 to 1 results in $a = b = \frac{1}{2}$, using $a = \frac{\nu_0}{2}$ and $b = \frac{\nu_0 \tau_0^2}{2}$. The following code line implements the prior with M*plus*:

```
MODEL PRIORS:
    Name of variance ~ IG(0.5, 0.5);
```

For a standardized predictor, this prior is quite effective when the sample size is small. Specifying somewhat larger values (e.g., by setting $\nu_0 = 2$) might increase estimation accuracy even further (Depaoli and Clifton, 2015).

To conclude, we worked out and discussed Bayesian approaches that perform better than ML in small samples, and we offered some practical guidance on how to implement these approaches with M*plus*. We hope that this article will help researchers in the field of psychology move beyond using Bayesian estimation as "just another estimator" and will help them make choices that are beneficial when their aim is to fit multilevel latent variable models and the sample size is small.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

SZ: writing, mathematical derivations, and graphic design. CH: writing. MH: writing and lead. All authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

We acknowledge support by Open Access Publishing Fund of University of Tübingen.

REFERENCES

Asparouhov, T., and Muthén, B. O. (2019). Latent variable centering of predictors and mediators in multilevel and time-series models. *Struct. Equat. Model*. 26, 119–142. doi: 10.1080/10705511.2018.1511375

Casella, G., and Berger, R. L. (2001). *Statistical Inference, 2nd Edn*. Pacific Grove, CA: Duxbury Press.

Chung, Y., Gelman, A., Rabe-Hesketh, S., Liu, J., and Dorie, V. (2015). Weakly informative prior for point estimation of covariance matrices in hierarchical models. *J. Educ. Behav. Stat*. 40, 136–157. doi: 10.3102/1076998615570945

Chung, Y., Rabe-Hesketh, S., Dorie, V., Gelman, A., and Liu, J. (2013). A nondegenerate penalized likelihood estimator for variance parameters in multilevel models. *Psychometrika* 78, 685–709. doi: 10.1007/s11336-013-9328-2

Depaoli, S., and Clifton, J. P. (2015). A Bayesian approach to multilevel structural equation modeling with continuous and dichotomous outcomes. *Struct. Equat. Model*. 22, 327–351. doi: 10.1080/10705511.2014.937849

Dijkstra, T. K. (1992). On statistical inference with parameter estimates on the boundary of the parameter space. *Brit. J. Math. Stat. Psychol*. 45, 289–309. doi: 10.1111/j.2044-8317.1992.tb00994.x

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

Greenland, S. (2000). Principles of multilevel modelling. *Int. J. Epidemiol*. 29, 158–167. doi: 10.1093/ije/29.1.158

Grilli, L., and Rampichini, C. (2011). The role of sample cluster means in multilevel models. *Methodology* 7, 121–133. doi: 10.1027/1614-2241/a000030

Hamaker, E. L., and Klugkist, I. (2011). "Bayesian estimation of multilevel models," in *Handbook of Advanced Multilevel Analysis*, eds J. J. Hox and J. K. Roberts (New York, NY: Routledge), 137–161.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Edn*. New York, NY: Springer.

Hoff, P. D. (2009). *A First Course in Bayesian Statistical Methods*. New York, NY: Springer. doi: 10.1007/978-0-387-92407-6

Hox, J. J., Maas, C. J. M., and Brinkhuis, M. J. S. (2010). The effect of estimation method and sample size in multilevel structural equation modeling. *Stat. Neerland*. 64, 157–170. doi: 10.1111/j.1467-9574.2009.00445.x

Hox, J. J., van de Schoot, R., and Matthijsse, S. (2012). How few countries will do? Comparative survey analysis from a Bayesian perspective. *Surv. Res. Methods* 6, 87–93. doi: 10.18148/srm/2012.v6i2.5033

Jacobucci, R., Grimm, K. J., and McArdle, J. J. (2016). Regularized structural equation modeling. *Struct. Equat. Model*. 23, 555–566. doi: 10.1080/10705511.2016.1154793

Jung, S., and Park, J. (2018). Consistent partial least squares path modeling via regularization. *Front. Psychol*. 9:174. doi: 10.3389/fpsyg.2018.00174

Kolenikov, S., and Bollen, K. A. (2012). Testing negative error variances: is a Heywood case a symptom of misspecification? *Sociol. Methods Res*. 41, 124–167. doi: 10.1177/0049124112442138

Lüdtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., and Muthén, B. O. (2008). The multilevel latent covariate model: a new, more reliable approach to group-level effects in contextual studies. *Psychol. Methods* 13, 203–229. doi: 10.1037/a0012869

Lüdtke, O., Robitzsch, A., Kenny, A., and Trautwein, U. (2013). A general and flexible approach to estimating the social relations model using Bayesian methods. *Psychol. Methods* 18, 101–119. doi: 10.1037/a0029252

Lüdtke, O., Robitzsch, A., and Wagner, J. (2018). More stable estimation of the STARTS model: a Bayesian approach using Markov chain Monte Carlo techniques. *Psychol. Methods* 23, 570–593. doi: 10.1037/met0000155

McNeish, D. (2016). On using Bayesian methods to address small sample problems. *Struct. Equat. Model*. 23, 750–773. doi: 10.1080/10705511.2016.1186549

McNeish, D. (2017). Small sample methods for multilevel modeling: a colloquial elucidation of REML and the Kenward-Roger correction. *Multivar. Behav. Res*. 5, 661–670. doi: 10.1080/00273171.2017.1344538

Muthén, B. O., and Asparouhov, T. (2012). Bayesian structural equation modeling: a more flexible representation of substantive theory. *Psychol. Methods*, 17, 313–335. doi: 10.1037/a0026802

Muthén, L. K., and Muthén, B. O. (2012). *Mplus User's Guide, 7th Edn*. Los Angeles, CA: Muthén Muthén.

Natarajan, R., and Kass, R. E. (2000). Reference Bayesian methods for generalized linear mixed models. *J. Am. Stat. Assoc*. 95, 227–237. doi: 10.1080/01621459.2000.10473916

Preacher, K. J., Zhang, Z., and Zyphur, M. J. (2016). Multilevel structural equation models for assessing moderation within and across levels of analysis. *Psychol. Methods* 21, 189–205. doi: 10.1037/met0000052

Preacher, K. J., Zyphur, M. J., and Zhang, Z. (2010). A general multilevel SEM framework for assessing multilevel mediation. *Psychol. Methods* 15, 209–233. doi: 10.1037/a0020141

Raudenbush, S. W., and Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods. Advanced Quantitative Techniques in the Social Sciences*, 2nd Edn. Thousand Oaks, CA: Sage.

Smid, S. C., McNeish, D., Miočević, M., and van de Schoot, R. (2020). Bayesian versus frequentist estimation for structural equation models in small sample contexts: a systematic review. *Struct. Equat. Model*. 27, 131–161. doi: 10.1080/10705511.2019.1577140

Snijders, T. A. B., and Bosker, R. J. (2012). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*, 2nd Edn. Los Angeles, CA: Sage.

van de Schoot, R., Winter, S. D., Ryan, O., Zondervan-Zwijnenburg, M., and Depaoli, S. (2017). A systematic review of Bayesian articles in psychology: the last 25 years. *Psychol. Methods* 22, 217–239. doi: 10.1037/met0000100

van Driel, O. P. (1978). On various causes of improper solutions in maximum likelihood factor analysis. *Psychometrika* 43, 225–243. doi: 10.1007/BF02293865

Yang, M., and Yuan, K.-H. (2019). Optimizing ridge generalized least squares for structural equation modeling. *Struct. Equat. Model*. 26, 24–38. doi: 10.1080/10705511.2018.1479853

Yuan, K.-H., and Chan, W. (2008). Structural equation modeling with near singular covariance matrices. *Comput. Stat. Data Anal*. 52, 4842–4858. doi: 10.1016/j.csda.2008.03.030

Yuan, K.-H., and Chan, W. (2016). Structural equation modeling with unknown population distributions: ridge generalized least squares. *Struct. Equat. Model*. 23, 163–179,. doi: 10.1080/10705511.2015.1077335

Zitzmann, S. (2018). A computationally more efficient and more accurate stepwise approach for correcting for sampling error and measurement error. *Multivar. Behav. Res*. 53, 612–632. doi: 10.1080/00273171.2018.1469086

Zitzmann, S., Lüdtke, O., and Robitzsch, A. (2015). A Bayesian approach to more stable estimates of group-level effects in contextual studies. *Multivar. Behav. Res*. 50, 688–705. doi: 10.1080/00273171.2015.1090899

Zitzmann, S., Lüdtke, O., Robitzsch, A., and Hecht, M. (2020). On the performance of Bayesian approaches in small samples: a comment on Smid, McNeish, Miočević, and van de Schoot (2020). *Struct. Equat. Model*. doi: 10.1080/10705511.2020.1752216. [Epub ahead of print].

Zitzmann, S., Lüdtke, O., Robitzsch, A., and Marsh, H. W. (2016). A Bayesian approach for estimating multilevel latent contextual models. *Struct. Equat. Model*. 23, 661–679. doi: 10.1080/10705511.2016.1207179

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

APPENDIX

FURTHER RESULTS

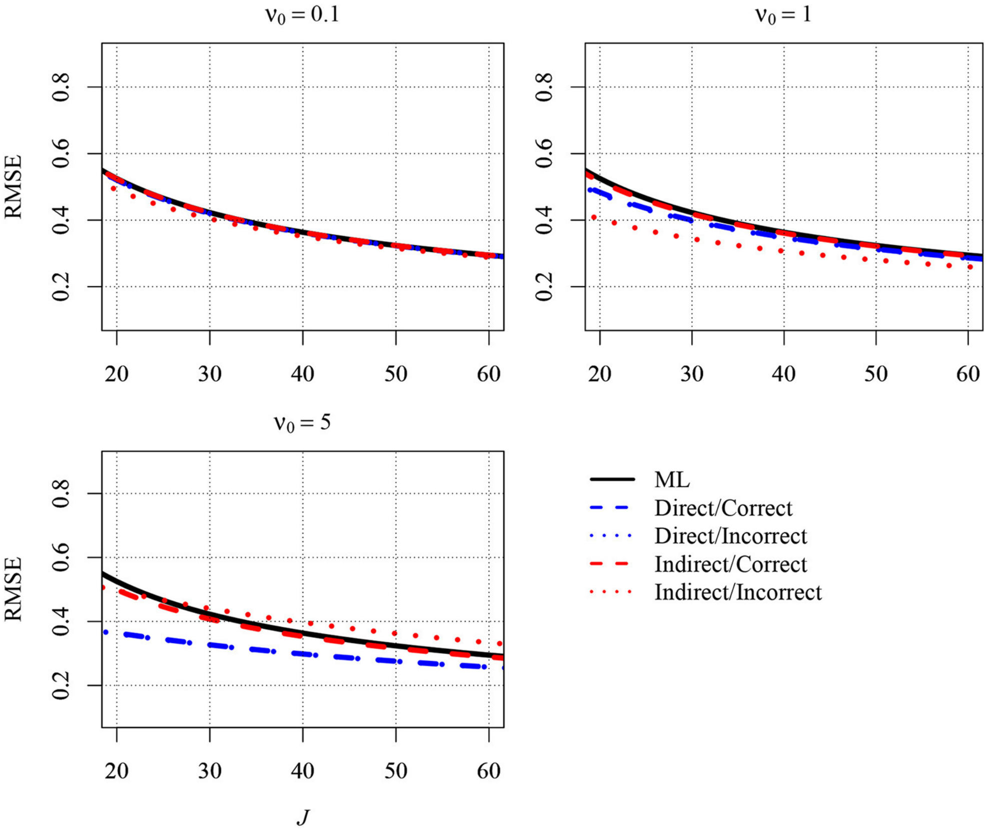


FIGURE A1 | The analytically derived Root Mean Squared Error (RMSE) in estimating the between-group slope for the direct and the indirect approach as a function of the sample size at the group level (*J*) and the prior distribution. Results are shown for $n = 10$ persons per group and an intraclass correlation of ICC = 0.1. ML, maximum likelihood; direct, the prior was specified directly for the between-group slope; indirect, the prior was specified for the group-level variance of the predictor variable; correct, correct prior (i.e., the prior guess equaled the value of the parameter in the population); incorrect, incorrect prior (i.e., the prior guess deviated from the parameter in the population); ν_0, prior sample size.

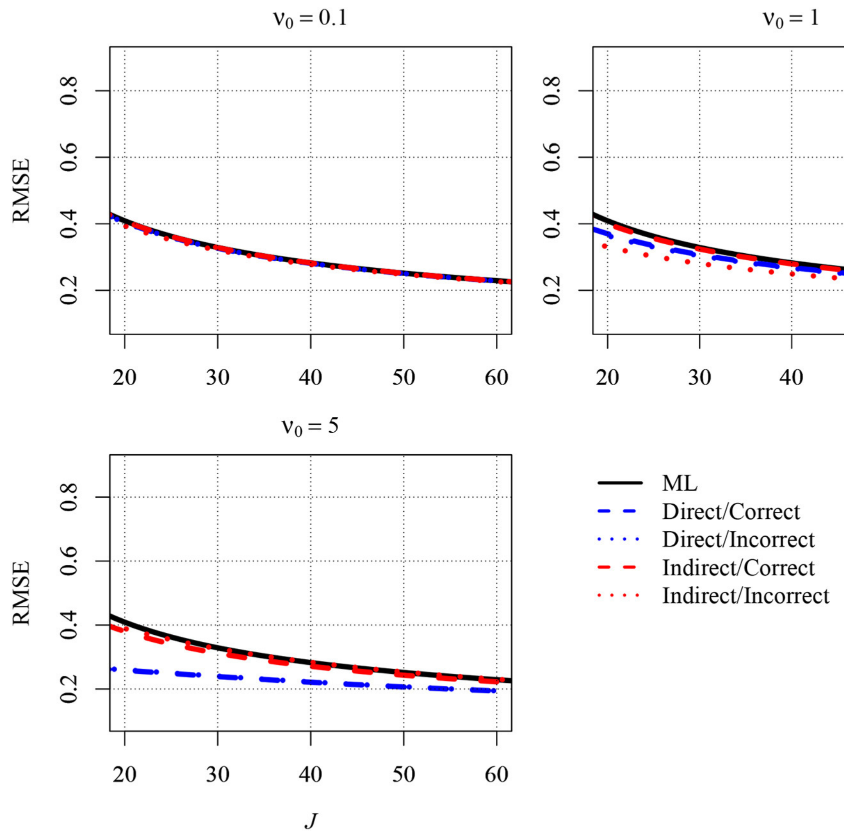


FIGURE A2 | The analytically derived Root Mean Squared Error (RMSE) in estimating the between-group slope for the direct and the indirect approach as a function of the sample size at the group level (J) and the prior distribution. Results are shown for a small number of $n = 5$ persons per group and an intraclass correlation of ICC = 0.2. ML, maximum likelihood; direct, the prior was specified directly for the between-group slope; indirect, the prior was specified for the group-level variance of the predictor variable; correct, correct prior (i.e., the prior guess equaled the value of the parameter in the population); incorrect, incorrect prior (i.e., the prior guess deviated from the parameter in the population); ν_0, prior sample size.

From Data to Causes III: Bayesian Priors for General Cross-Lagged Panel Models (GCLM)

Michael J. Zyphur[1], Ellen L. Hamaker[2], Louis Tay[3], Manuel Voelkle[4], Kristopher J. Preacher[5], Zhen Zhang[6,7], Paul D. Allison[8], Dean C. Pierides[9], Peter Koval[10] and Edward F. Diener[11,12]*

[1] Department of Management and Marketing, The University of Melbourne, Parkville, VIC, Australia, [2] Department of Methodology and Statistics, Utrecht University, Utrecht, Netherlands, [3] Department of Psychological Sciences, Purdue University, West Lafayette, IN, United States, [4] Department of Psychology, Humboldt University of Berlin, Berlin, Germany, [5] Department of Psychology and Human Development, Humboldt University of Berlin, Berlin, Germany, [6] Cox School of Business, Southern Methodist University, Dallas, TX, United States, [7] W.P. Carey School of Business, Arizona State University, Tempe, AZ, United States, [8] Department of Sociology, University of Pennsylvania, Philadelphia, PA, United States, [9] Stirling Management School, University of Stirling, Stirling, United Kingdom, [10] Melbourne School of Psychological Sciences, The University of Melbourne, Parkville, VIC, Australia, [11] Department of Psychology, The University of Utah, Salt Lake City, UT, United States, [12] Department of Psychology, University of Virginia, Charlottesville, VA, United States

OPEN ACCESS

Edited by:
Christoph Koenig,
Goethe University Frankfurt, Germany

Reviewed by:
Axel Mayer,
Bielefeld University, Germany
Fredrik Falkenström,
Linköping University, Sweden

***Correspondence:**
Michael J. Zyphur
mzyphur@unimelb.edu.au

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: 30 September 2020
Accepted: 11 January 2021
Published: 15 February 2021

Citation:
Zyphur MJ, Hamaker EL, Tay L, Voelkle M, Preacher KJ, Zhang Z, Allison PD, Pierides DC, Koval P and Diener EF (2021) From Data to Causes III: Bayesian Priors for General Cross-Lagged Panel Models (GCLM). Front. Psychol. 12:612251. doi: 10.3389/fpsyg.2021.612251

This article describes some potential uses of Bayesian estimation for time-series and panel data models by incorporating information from prior probabilities (i.e., priors) in addition to observed data. Drawing on econometrics and other literatures we illustrate the use of informative "shrinkage" or "small variance" priors (including so-called "Minnesota priors") while extending prior work on the general cross-lagged panel model (GCLM). Using a panel dataset of national income and subjective well-being (SWB) we describe three key benefits of these priors. First, they shrink parameter estimates toward zero or toward each other for time-varying parameters, which lends additional support for an income → SWB effect that is not supported with maximum likelihood (ML). This is useful because, second, these priors increase model parsimony and the stability of estimates (keeping them within more reasonable bounds) and thus improve out-of-sample predictions and interpretability, which means estimated effect should also be more trustworthy than under ML. Third, these priors allow estimating otherwise under-identified models under ML, allowing higher-order lagged effects and time-varying parameters that are otherwise impossible to estimate using observed data alone. In conclusion we note some of the responsibilities that come with the use of priors which, departing from typical commentaries on their scientific applications, we describe as involving reflection on how best to apply modeling tools to address matters of worldly concern.

Keywords: panel data model, Granger causality (VAR), Bayesian, shrinkage estimation, small-variance priors

FROM DATA TO CAUSES III: BAYESIAN PRIORS FOR GENERAL CROSS-LAGGED PANEL DATA MODELS (GCLM)

Panel data models track multiple independent units N over multiple occasions of measurement T with parameters typically estimated by frequentist methods (e.g., Arellano, 2003; Baltagi, 2013; Little, 2013; Allison, 2014; Hsiao, 2014; Hamaker et al., 2015). This approach to causal inference was recently illustrated by Zyphur et al. (2020a,b), showing the benefits of a general cross-lagged panel model (GCLM) specified as a structural equation model (SEM) and estimated by maximum likelihood. However, moving away from such frequentist estimators, time-series, and panel data models can be extended to allow additional flexibility in data and model structures, thereby enhancing the range of applications and practical usefulness of models such as the GCLM.

In the current article we do this by showing how Bayesian estimation and inference can expand the range of available model specifications because Bayesian approaches allow including information from prior probabilities (i.e., priors) as well as observed data when estimating parameters (for general discussions see Gill, 2008; Gelman et al., 2014). Prior probabilities can be specified in various ways when estimating panel data models (e.g., Schuurman et al., 2016) including weakly informative priors to improve the stability of estimates (keeping them within more reasonable bounds; Lüdtke et al., 2018), but here we illustrate the use of informative "small variance" or "shrinkage" priors for parameters and/or parameter differences using an approach that follows from existing work (see Muthén and Asparouhov, 2012; Asparouhov et al., 2015; Zyphur and Oswald, 2015). This approach to informative priors "shrinks" parameter estimates toward zero or toward each other while allowing estimates to deviate from these priors as a function of observed data.

In this article we endeavor to show how, in the context of panel data models, such priors have many benefits, helping to solve the problem of "how to build models that are flexible enough to be empirically relevant . . . but not so flexible as to be seriously over-parameterized" (Koop and Korobilis, 2010, p. 269). In brief, these priors allow many parameters to be estimated while at the same time minimizing model complexity, shrinking parameter estimates toward zero, and/or toward each other by inducing a strong positive correlation among parameters (i.e., reducing parameter differences; Korobilis, 2013). Two key benefits of this prior specification and of Bayesian estimation and inference more generally are as follows.

First, the priors increase generalizability by reducing variance in a classic bias-variance trade-off, which is important for practically applying results from panel data models by reducing overfitting (Korobilis, 2013). Second, they allow estimating models that are under-identified in frequentist approaches due to limited T and/or N, such as when estimating time-varying unit effects and multiple lagged effects (see Canova, 2007; Koop and Korobilis, 2010; Canova and Ciccarelli, 2013; Giannone et al., 2015). By using informative priors, under-identified parameters need not be strictly constrained to zero or equality over time as would be required with frequentist estimators, thus allowing model results to be more sensitive to observed data patterns when compared to models that constrain parameters to zero or equality over time.

In what follows, we illustrate these benefits by first reviewing the GCLM and its identification in SEM under frequentist estimators. We then describe Bayesian estimation and inference, focusing on the benefits of small-variance priors. Using Gallup World Poll data from Diener et al. (2013) used in Zyphur et al.'s articles, we then fit various models to illustrate the benefits of our Bayesian approach. In so doing, we support different conclusions than the original two articles on the GCLM, which revealed no causal effects among income and subjective well-being (SWB). With a Bayesian approach, we show a positive short-run and long-run effect of income on SWB, but not the reverse. We conclude with brief thoughts on panel data models, including the importance of using them to study processes that are of serious worldly concern. Before continuing we emphasize that our effort here is to illustrate some of the logic and potential uses of prior probabilities for time-series and panel data models, rather than provide a comprehensive overview of priors in longitudinal data models. Other work on priors, sensitivity analyses, and reporting standards exists and we advise interested authors to further explore these topics (e.g., Depaoli and van de Schoot, 2017; Smid et al., 2020), including specifically in the domain of panel data models similar to the GCLM (Lüdtke et al., 2018).

THE GENERAL CROSS-LAGGED PANEL MODEL (GCLM)

The GCLM is specified for a unit i at an occasion t with two variables $x_{i,t}$ and $y_{i,t}$ (for additional insight see Zyphur et al., 2020a,b). Parenthetical superscripts (x) and (y) indicate the equation in which a coefficient belongs; subscripts x and y indicate the predictor with which a coefficient is associated; and h indicates a lag or lead, such as $y_{i,t-h}$. With this notation, the general model is shown as follows (for $t > 1$):

$$\begin{aligned} x_{i,t} &= \alpha_t^{(x)} + \lambda_t^{(x)}\eta_i^{(x)} + \beta_{x1}^{(x)}x_{i,t-1} + \delta_{x1}^{(x)}u_{i,t-1}^{(x)} + \beta_{y1}^{(x)}y_{i,t-1} \\ &\quad +\delta_{y1}^{(x)}u_{i,t-1}^{(y)} + u_{i,t}^{(x)} \end{aligned} \tag{1}$$

$$\begin{aligned} y_{i,t} &= \alpha_t^{(y)} + \lambda_t^{(y)}\eta_i^{(y)} + \beta_{y1}^{(y)}y_{i,t-1} + \delta_{y1}^{(y)}u_{i,t-1}^{(y)} + \beta_{x1}^{(y)}x_{i,t-1} \\ &\quad +\delta_{x1}^{(y)}u_{i,t-1}^{(x)} + u_{i,t}^{(y)} \end{aligned} \tag{2}$$

wherein $u_{i,t}$ is an impulse capturing random events that are meant to mimic random assignment to levels of a variable, with variance ψ_{u_t} and contemporaneous covariance or "co-movement"$\psi_{u_t}^{(xy)}$; $\alpha_t$is an occasion effect at a time t; $\eta_i$is a unit effect capturing stable factors over time, with$\eta_i^{(x)} \sim N(0, \psi_\eta^{(x)})$, $\eta_i^{(y)} \sim N(0, \psi_\eta^{(y)})$, and covariance$\psi_\eta^{(xy)}$; λ_t is a time-varying unit effect; $\beta_{x1}^{(x)}$and $\beta_{y1}^{(y)}$ are autoregressive (AR) effects of past impulses on the same variable (with coefficients on lagged predictors taking a form $\beta_{yh}^{(y)}$, wherein h is the lag); $\delta_{x1}^{(x)}$ and $\delta_{y1}^{(y)}$ are moving average or MA effects of past impulses on the same variable; $\beta_{y1}^{(x)}$ and $\beta_{x1}^{(y)}$ are cross-lagged or CL effects of past impulses on another variable; and $\delta_{y1}^{(x)}$ and $\delta_{x1}^{(y)}$ are cross-lagged moving average or CLMA effects of past impulses among different variables[1]. With

[1] In order to identify a scale for η we fix one of each $\lambda_t^{(x)}$ and $\lambda_t^{(y)}$ terms to unity. In our previous papers and in our online Excel file that automates Mplus input, we did this for the final occasion $\lambda_6 = 1$. Choosing this or any other occasion is an arbitrary decision, but in the current paper we set $\lambda_1 = 1$ in order to facilitate some of the Bayesian prior specifications we describe.

this logic, we interpret at least three kinds of effects: (1) total effects of a variable on itself combine AR and MA terms to show the short-run persistence of impulses [e.g., $\beta_{y1}^{(y)} + \delta_{y1}^{(y)}$] such that a process is more mean-reverting as these terms tend towards zero; (2) Granger-causal effects of impulses that combine all CL and CLMA terms to show short-run or direct effects among different variables over time [e.g., $\beta_{x1}^{(y)} + \delta_{x1}^{(y)}$]; and (3) impulse responses map the change in a system across all parameters due to an impulse [e.g., a change along $u_{i,t}^{(y)}$], showing long-run or total effects of an impulse across all variables in a system over time (see Zyphur et al., 2020a).

We map this general model structure onto the following SEM:

$$\mathbf{y}_i = \mathbf{\Lambda}\eta_i \tag{3}$$

$$\boldsymbol{\eta}_i = \boldsymbol{\alpha} + \mathbf{B}\boldsymbol{\eta}_i + \boldsymbol{\zeta}_i \tag{4}$$

with all terms as follows for an AR(1)MA(1)CL(1)CLMA(1) model and a single unit effect for each of k observed variables at T occasions: $\mathbf{y}_i$is a kT length vector of observed variables; $\mathbf{\Lambda}$ is a $kT \times (2kT + k)$ matrix, mapping kT observed variables onto kT latent analogs; $\boldsymbol{\eta}_i$ is a $2kT + k$ length vector, with kT terms mapped to $\mathbf{y}_i$, kT impulses, and k unit effects; $\boldsymbol{\alpha}$ is a $2kT + k$ length vector with kT occasion effects only; $\mathbf{B}$ is a $(2kT + k) \times (2kT + k)$ matrix with kT unities to map kT observed variables to kT impulses, kT time-varying unit effects, $2k$ AR and MA terms, and $2k(k - 1)$ CL and CLMA terms; and $\boldsymbol{\zeta}_i$ is a $2kT + k$ length vector with covariance matrix $\mathbf{\Psi}$ containing k unit effect variances, $k(k - 1)/2$ unit effect covariances, kT impulse variances, and $kT(k - 1)/2$ co-movements.

As with time-series and panel data models in general, the GCLM requires choosing different numbers of unit effects and the following lag orders: p lags in an AR(p) model; q lags in an MA(q) model; c lags in an CL(c) model; l lags in an CLMA(l) model. Substantive and statistical checking should inform these choices, with an emphasis on conservative models that balance theory and contextual knowledge with model fit (Armstrong et al., 2015; Green and Armstrong, 2015). In Zyphur et al. (2020a) this was done by modeling income $x_{i,t}$ and SWB $y_{i,t}$ for $N = 135$ countries and $T = 6$ years from 2006 to 2011 (see Diener et al., 2013). After substantive and statistical checking, an AR(1)MA(2)CL(1)CLMA(1) model was chosen for income $x_{i,t}$ (adding a higher-order MA term $\delta_{x2}^{(x)} u_{i,t-2}^{(x)}$ to Eq. 1), and an AR(1)MA(1)CL(1)CLMA(1) model was chosen for SWB $y_{i,t}$ (fitting with Eq. 2).

Descriptive statistics are in Zyphur et al. (2020a) and results are in **Table 1** as the maximum-likelihood or "ML" model these authors estimated. **Table 2** shows Granger causality tests from the four steps discussed by these authors, with AIC and BIC values showing that eliminating CL and CLMA effects improves model fit (by decreasing AIC and BIC values, indicating better model quality as a trade-off between fit and parsimony). This fails to support any form of Granger causality using the logic from Zyphur et al. (2020a). Finally, impulse responses in **Figures 1A–D** show very weak support for long-run effects with *CI*s that include zero, and an unexpected negative SWB → income effect. In sum, these results are counter to those originally presented by

TABLE 1 | Model results.

Parameters	Estimates (*SE*s or posterior *SD*s) (Ranges for time-varying parameters)		
	ML	Bayes 1	Bayes 2
SWB → SWB AR/MA Terms $\beta_{y1}^{(y)}$ and $\delta_{y1}^{(y)}$			
$\beta_{y1}^{(y)}$	0.39 (0.36)	0.34 (0.21) [0.30, 0.39]	0.34 (0.21) [0.29, 0.39]
$\delta_{y1}^{(y)}$	0.19 (0.32)	0.15 (0.19) [0.08, 0.22]	0.16 (0.19) [0.08, 0.23]
$\beta_{y1}^{(y)} + \delta_{y1}^{(y)}$	0.58* (0.09)	0.49* (0.08) [0.38, 0.62]	0.49* (0.08) [0.38, 0.62]
Income → Income AR/MA terms $\beta_{x1}^{(x)}$ and $\delta_{x1}^{(x)}$			
$\beta_{x1}^{(x)}$	0.96* (0.13)	0.97* (0.03) [0.94, 1.0]	0.97* (0.03) [0.94, 1.0]
$\delta_{x1}^{(x)}$	−0.33 (0.25)	−0.27* (0.07) [−0.30, −0.21]	−0.26* (0.07) [0.30, −0.21]
$\delta_{x2}^{(x)}$	0.06 (0.09)	0.02 (0.06) [−0.03, 0.08]	0.01 (0.04) [−0.07, 0.11]
$\delta_{x.}^{(x)}$	−0.27 (0.19)	−0.25* (0.10) [−0.32, −0.13]	−0.26* (0.08) [−0.35, −0.10]
$\beta_{x1}^{(x)} + \delta_{x.}^{(x)}$	0.69* (0.20)	0.72* (0.08) [0.62, 0.87]	0.72* (0.07) [0.59, 0.90]
Income → Subjective well-being CL/CLMA terms $\beta_{x1}^{(y)}$ and $\delta_{x1}^{(y)}$			
$\beta_{x1}^{(y)}$	0.13 (0.32)	0.23 (0.20) [0.19, 0.26]	0.24 (0.20) [0.20, 0.26]
$\delta_{x1}^{(y)}$	0.01 (0.25)	−0.03 (0.19) [−0.07, 0.01]	−0.03 (0.19) [−0.08, 0.002]
$\beta_{x1}^{(y)} + \delta_{x1}^{(y)}$	0.14 (0.16)	0.22 (0.12) [0.12, 0.25]	0.22 (0.12) [0.12, 0.24]
Subjective well-being → Income CL/CLMA terms $\beta_{y1}^{(x)}$ and $\delta_{y1}^{(x)}$			
$\beta_{y1}^{(x)}$	−0.10 (0.07)	0.01 (0.03) [−0.001, 0.03]	0.01 (0.02) [−0.001, 0.03]
$\delta_{y1}^{(x)}$	0.08 (0.07)	−0.02 (0.05) [−0.04, 0.01]	−0.02 (0.05) [−0.04, 0.01]
$\beta_{y1}^{(x)} + \delta_{y1}^{(x)}$	−0.02 (0.04)	−0.01 (0.05) [−0.04, 0.01]	−0.01 (0.04) [−0.04, 0.01]
Co-movement in impulses $\psi_{u_t}^{(xy)}$ as correlations			
$\psi_{u_1}^{(xy)}$	0.64 (0.59)	0.87* (0.31)	0.87* (0.30)
$\psi_{u_2}^{(xy)}$	0.45* (0.21)	0.44* (0.17)	0.44* (0.17)
$\psi_{u_3}^{(xy)}$	0.003 (0.13)	0.05 (0.13)	0.05 (0.13)
$\psi_{u_4}^{(xy)}$	−0.02 (0.12)	0.05 (0.13)	0.06 (0.13)
$\psi_{u_5}^{(xy)}$	0.32* (0.14)	0.41* (0.11)	0.41* (0.11)
$\psi_{u_6}^{(xy)}$	0.11 (0.13)	0.14 (0.11)	0.15 (0.11)
Unit effect variances $\psi_{\eta}^{(y)}$and$\psi_{\eta}^{(x)}$, and covariance $\psi_{\eta}^{(xy)}$ as a correlation			
$\psi_{\eta}^{(y)}$	1.01	1.01	1.01
$\psi_{\eta}^{(x)}$	0.40	0.37	0.37
$\psi_{\eta}^{(xy)}$	0.96* (0.06)	0.86* (0.18)	0.86* (0.18)
Time-varying unit effects ("factor loadings") $\lambda_t^{(y)}$ and $\lambda_t^{(x)}$ as correlations			
$\lambda_1^{(y)}$	0.96* (0.06)	0.92* (0.10)	0.91* (0.10)
$\lambda_2^{(y)}$	0.48 (0.32)	0.47* (0.18)	0.47* (0.18)
$\lambda_3^{(y)}$	0.48 (0.32)	0.44* (0.17)	0.44* (0.17)
$\lambda_4^{(y)}$	0.46 (0.30)	0.51* (0.18)	0.51* (0.17)
$\lambda_5^{(y)}$	0.52 (0.30)	0.45* (0.17)	0.45* (0.17)
$\lambda_6^{(y)}$	0.46 (0.33)	0.44* (0.18)	0.45* (0.17)
$\lambda_1^{(x)}$	0.73* (0.25)	0.69* (0.19)	0.68* (0.19)
$\lambda_2^{(x)}$	−0.03 (0.22)	−0.01 (0.06)	−0.01 (0.06)

(Continued)

TABLE 1 | Continued

Parameters	Estimates (*SE*s or posterior *SD*s) (Ranges for time-varying parameters)		
	ML	Bayes 1	Bayes 2
$\lambda_3^{(x)}$	0.15 (0.09)	−0.01 (0.05)	−0.01 (0.05)
$\lambda_4^{(x)}$	0.16* (0.07)	0.02 (0.05)	0.02 (0.05)
$\lambda_5^{(x)}$	0.15* (0.08)	0.01 (0.06)	0.01 (0.06)
$\lambda_6^{(x)}$	0.16* (0.08)	−0.01 (0.05)	−0.01 (0.05)
Fit indices			
k / pD	54	50.69	50.78
PPP	–	0.32	0.32
DIC	–	829.23	827.65

ML, maximum likelihood; SWB, subjective well-being; AR, autoregressive; MA, moving average; CL, cross-lagged; CLMA, cross-lagged moving average; BIC, Bayes information criterion; DIC, deviance information criterion; k / pD, the number of model parameters, which is exactly k under ML and is estimated as pD for Bayes models.

** p < 0.05.*

Diener et al. (2013), who found a positive income → SWB effect as well as a positive SWBincome effect, which Zyphur et al. (2020a) proposed was likely due to failing to control for unit effects $\eta_i^{(x)}$ and $\eta_i^{(y)}$ (and their covariance $\psi_\eta^{(xy)}$).

However, the model chosen by Zyphur et al., was limited by their reliance on a frequentist estimator. Although these estimators are common and inferences based on them may be sound in many cases, estimators such as ML rely on only observed data rather than also incorporating prior information about parameters (van de Schoot et al., 2017). Specifically, time-varying unit effects $\lambda_t\eta_i$ and AR/MA terms rely on $kT(T-1)/2$ observed auto-covariances for estimation. On the other hand, unit effect covariances $\psi_\eta^{(xy)}$, CL/CLMA terms, and impulse co-movements $\psi_{u_t}^{(xy)}$ rely on $k(k-1)T^2/2$ observed cross-covariances for estimation. In turn, for an SEM to be identified the number of observed auto- and cross-covariances (associated with T) must grow with the number of time-varying unit effects $\lambda_t\eta_i$ and the p, q, c, and l lag orders for AR, MA, CL, and CLMA terms (for a general discussion of identification see Bollen, 1989). Also, with many estimated parameters, the N required to assure asymptotic assumptions are met for ML also increases. Furthermore, even with large T and N, some models may not be identified and may violate asymptotic assumptions, such as if AR, MA, CL, and CLMA effects are time-varying, which we can show by modifying Eqs. 1 and 2 with a t subscript as follows (for $t > 1$):

$$x_{i,t} = \alpha_t^{(x)} + \lambda_t^{(x)}\eta_i^{(x)} + \beta_{x1,t}^{(x)}x_{i,t-1} + \delta_{x1,t}^{(x)}u_{i,t-1}^{(x)} + \beta_{y1,t}^{(x)}y_{i,t-1} + \delta_{y1,t}^{(x)}u_{i,t-1}^{(y)} + u_{i,t}^{(x)} \quad (5)$$

$$y_{i,t} = \alpha_t^{(y)} + \lambda_t^{(y)}\eta_i^{(y)} + \beta_{y1,t}^{(y)}y_{i,t-1} + \delta_{y1,t}^{(y)}u_{i,t-1}^{(y)} + \beta_{x1,t}^{(y)}x_{i,t-1} + \delta_{x1,t}^{(y)}u_{i,t-1}^{(x)} + u_{i,t}^{(y)} \quad (6)$$

This model allows for "regime changes" as changes in effects over time (Stock and Watson, 1996, 2009), which is reasonable given the fact that people, organizations, and entire economies are complex dynamic systems that are always in flux (Williams and Cook, 2016). However, Eqs. 5 and 6 imply that there are now $T-1$ unique parameters for *each* AR, MA, CL, and CLMA term, and these proliferate rapidly as k increases, such that the total number of time-varying AR, MA, CL, and CLMA effects is $(T-1)[2k + 2k(k-1)]$. For example, with $k = 4$ observed variables and $T = 10$ occasions of measurement, Eqs. 5 and 6 imply a model with 288 β and δ terms, requiring large N. Furthermore, this large number of terms is based on lag orders that are limited to the simplest $p = q = c = l = 1$ case, which will not always hold in practice and, when it does not, will put substantial requirements on observed data and the estimates derived from them.

Clearly, for GCLMs like that in Eqs. 5 and 6 and for panel data models more generally, parameter identification and overfitting as well as meeting ML assumptions may be difficult (Lüdtke et al., 2018), especially as lag orders and the number of unit effects grow. Due to this problem, parameter estimates—and therefore Granger causality tests and impulse responses—may have reduced generalizability and the number of parameters that can be estimated are limited by N and T. This is unfortunate for many reasons, such as difficulty in supporting hypotheses due to moderate N. Also, ironically, the parameter restrictions required to achieve model identification run counter to the impetus for panel data models like ours, which is partly to overcome the "incredible" identifying assumptions typically found in regression models (see Sims, 1980, 1986). In order to provide a solution to these problems, we now describe a Bayesian approach to estimation and inference.

TABLE 2 | Granger Causality Tests and Δ R^2.

ML	Bayes 1	Bayes 2
AIC / BIC	DIC	DIC
Step 1: Derive fit of full model		
845.94 / 1002.82	829.23	827.65
Step 2: Constraint all income → SWB effects		
841.86 / 990.03	835.48	833.85
Step 3: Constrain all SWB → Income effects		
844.08 / 992.25	820.83	818.71
Step 4: Constraining all CL/CLMA terms		
842.62 / 984.98	833.35	831.70

ML, maximum-likelihood; SWB, subjective well-being; AIC, Akaike's information criterion; BIC, Bayes information criterion; DIC, Deviance information criterion.

BAYESIAN ESTIMATION AND INFERENCE

There are two key differences between Bayesian and frequentist estimation. The first and perhaps primary difference is that whereas frequentist probabilities apply to data (or events), Bayesian probabilities apply to parameters (or hypotheses; Zyphur and Oswald, 2015). The implication is that instead of

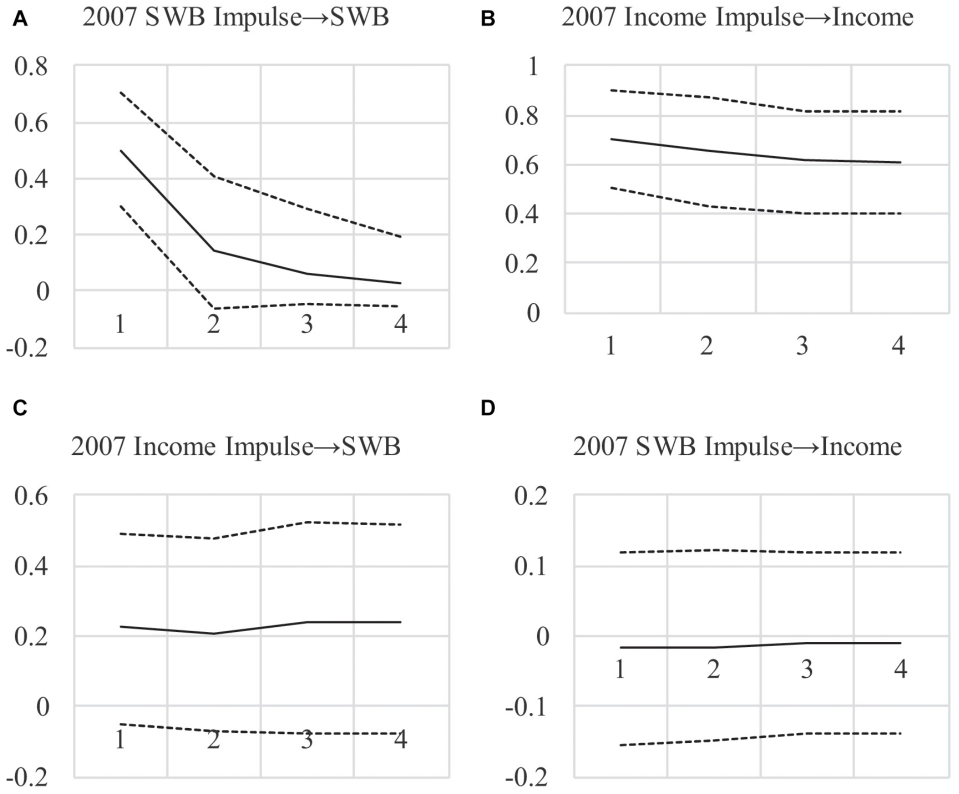


FIGURE 1 | (A–D) Impulse Response Functions for AR(1)MA(2) Model Under Maximum-Likelihood. Impulses begin in 2007, showing the effect of a 1-unit impulse in 2007 over the next 4 years with 95% confident intervals.

representing relative frequencies, probabilities represent degrees of belief or knowledge (Howson and Urbach, 2006). The classic idea is that Bayesian probabilities are meant to be inductive, allowing direct probabilistic inferences about parameters in a model θ given observed data *Y* (Hacking, 2001; Jaynes, 2003). With this orientation, Bayesian estimation and inference are done in order to represent degrees of uncertainty around parameters, measured by a "posterior" probability distribution $f(\theta|Y)$. The mean, median, or mode of this distribution is used to describe specific parameter point estimates and variance in the distribution is used to describe uncertainty in parameters for hypothesis testing. For example, the *SD* of a parameter distribution can be used to approximate a frequentist *SE* for the computation of Bayesian *p*-values (for discussion, see Muthén and Asparouhov, 2012; Zyphur and Oswald, 2015). In all cases, posterior distributions are meant to represent knowledge or beliefs about parameters, with hypothesis tests serving to inform knowledge or beliefs about parameters based on model results.

The second difference between frequentist and Bayesian methods is how such results are derived, which is to say how a posterior distribution $f(\theta|Y)$ is estimated. Unlike frequentist estimation, Bayesian estimators must directly incorporate two sources of information to estimate parameters in a model $_\theta$: prior probabilities of parameters $f(\theta)$ that serve to indicate the knowledge or beliefs about parameters before estimation; and the probability of observed data *Y* given parameter estimates $f(Y|\theta)$, which can be understood as a likelihood. The result is posterior probabilities $f(\theta|Y)$, which are then used for inference. The proportional relation (∝) among these terms can be shown as follows (see Muthén and Asparouhov, 2012):

$$f(\theta|Y) \propto f(\theta)f(Y|\theta) \tag{7}$$

wherein model results $f(\theta|Y)$ are derived based on both information in the priors $f(\theta)$ and the data *Y* in the form of the model likelihood $f(Y|\theta)$.

The result of this logic is that Bayesian estimators are justified based on the degree to which they satisfy the rule in Eq. 7, which is designed to be a logically consistent system for updating prior knowledge or beliefs with additional data (Zyphur and Oswald, 2015). This is very much unlike frequentist estimators, which are justified based on asymptotic theories that describe how estimators perform when, for example, a sample size grows to infinity and/or a study is conducted an infinite number of times. One result of this difference between Bayesian and frequentist logics is that frequentist estimators like ML satisfy assumptions only as $N \to \infty$, which creates problems for SEM with many parameters and small *N* (Anderson and Gerbing, 1984; MacCallum et al., 1996). Conversely, because Bayesian estimation requires only that the rule in Eq. 7 be followed, models with many parameters and small *N* are not problematic apart from the way that small *N* exacts an appropriate toll by increasing levels of uncertainty in $f(\theta|Y)$ (rather than also violating assumptions about the estimator in relation to N; for insight into the importance of priors in such cases see Smid et al., 2020). The point is that as long as estimation follows the rule in Eq. 7, then a sample size N is always appropriate even if it makes reducing uncertainty in a posterior distribution $f(\theta\ |Y)$ difficult.

Given the focus on Eq. 7, a key question to answer for a Bayesian approach is how to choose prior probabilities for model parameters in $f(\theta)$. Typically, "uninformative" or "diffuse" priors are used for$f(\theta)$ in an attempt to eliminate their influence on posteriors $f(\theta|Y)$ (Gelman et al., 2014). The point of these priors can be conceptualized as "flattening" probability (i.e., "leveling" belief or knowledge) across the range of possible parameter values in θ. This is akin to being agnostic about specific parameter values (i.e., having no strong prior knowledge or beliefs), which is meant to result in reducing the influence of priors $f(\theta)$ during estimation. In turn, such priors produce strong agreement among Bayesian and frequentist estimates as N increases, which is sensible because as priors' influence decreases, posteriors are increasingly dominated by the likelihood $f(Y|\theta)$ that many frequentist methods maximize (the reader can see this by conceptually by removing the prior from Eq. 7). In turn, statistical modeling programs such as Mplus often use various kinds of diffuse priors by default, such as a prior for a regression slope with a variance that is, practically speaking, infinity, such as $\beta \sim N(0, 10^{10})$ (Asparouhov and Muthén, 2010; Muthén, 2010). The reader can intuit how this prior is uninformative by recognizing that the mean of the distribution 0 has virtually no greater probability than a value of 100 for β, because $\beta \sim N(0, 10^{10})$ implies an extremely flat probability distribution (i.e., approximately equal belief or knowledge for any specific value of β).

Conversely, priors become informative and increasingly influential as they become increasingly dense around specific parameter values, such as a small-variance prior for a regression slope $\beta \sim N(0, 0.01)$ (Muthén and Asparouhov, 2012; Zyphur and Oswald, 2015). In this case, the density of the prior distribution is high around the value 0, and during estimation this pulls estimates of β toward 0 (Gill, 2008; Gelman et al., 2014). Thus, informative priors that favor null parameter values effectively "shrink" parameter estimates toward 0, which is useful because this increases generalizability by reducing the tendency to overfit model estimates to an observed dataset (McNeish, 2015). As Giannone et al. (2015) note, priors such as these "are successful because they effectively reduce the estimation error while generating only relatively small biases in the estimates of the parameters" (p. 436). Of course it is notable that alternative small-variance priors can be chosen—as we note further below with references to relevant work that the reader may consult—our choice of small-variance priors here follows from existing work using these in the psychology and organizational literature (see Muthén and Asparouhov, 2012; Zyphur and Oswald, 2015).

Furthermore, because Bayesian estimation relies on prior probabilities $f(\theta)$ and the likelihood $f(Y|\theta)$, priors behave more like observed data when they favor specific parameter values—whatever these might be. By this we mean that in a model with small-variance priors, parameters will be identified as a function of the information in observed data *and* the priors, so that even if there is insufficient information in a dataset to identify a parameter, the small-variance prior may serve to help identification. This can be understood by considering that as priors $f(\theta)$ become more informative, this is akin to a reduction in the number of parameters that are freely estimated in a Bayesian model (symbolized as *pD*). In turn, a diffuse prior such as $\beta \sim N(0, 10^{10})$ offers little help in identifying estimates of β without sufficient information in the likelihood $f(Y|\theta)$ to do so. On the other hand, a small-variance prior such as $\beta \sim N(0, 0.01)$ may allow estimating β even when there is insufficient information in the model likelihood to do so (e.g., if a likelihood is relatively "flat" across a range of values for β; Asparouhov et al., 2015). This is because a model with a small-variance prior for the β does not "freely" estimate it in a frequentist sense, but instead combines the prior $\beta \sim N(0, 0.01)$ with the data Y in the form of the likelihood $f(Y|\theta)$.

In sum, informative priors, such as small-variance priors, are useful because they can shrink estimates to avoid overfitting, thereby increasing generalizability, while at the same time helping to identify parameters that otherwise may not be estimable due to insufficient information in a dataset Y. Furthermore, these priors can serve to operationalize prior knowledge or beliefs about parameters, while allowing data to update the priors to produce results that combine these two sources of information. As previously noted, this is consistent with the interest of an informal Bayesian who seeks to use panel data models to change knowledge or beliefs about the ways in which variables are causally related over time (Granger, 1980).

Priors for Time-Series and Panel Data Models

Due to their ability to address overfitting and non-identified parameters, informative priors have become popular in time-series and panel data modeling, particularly in a vector autoregressive or VAR framework (for discussions, see Canova, 2007; Koop and Korobilis, 2010; Giannone et al., 2015). To illustrate this, the approach we use here relies on small-variance priors for parameters as well as parameter differences for time-varying terms. As examples, consider that higher-order lags may be shrunk toward zero, such as a second-order MA effect: $\delta^{(y)}_{y2} \sim N(0, 0.01)$; or, differences in time-varying parameters may be shrunk toward each other, such as AR effects at different occasions: $(\beta^{(x)}_{x1,t} - \beta^{(x)}_{x1,t+1}) \sim N(0, 0.01)$. Although the former approach may be somewhat familiar (especially in the econometric VAR community), the latter approach is more novel and is designed for cases wherein similar parameters are expected to have small differences. To understand priors such as $(\beta^{(x)}_{x1,t} - \beta^{(x)}_{x1,t+1}) \sim N(0, 0.01)$, it may be useful to connect this to terms associated with an SEM (e.g., Eqs. 3 and 4). Specifically, a prior distribution for regression terms in a matrix $\mathbf{B}$, or $f(\mathbf{B})$, may be parameterized as $f(\mathbf{B}) \sim MVN(0, \mathbf{\Psi_B})$, with the covariance matrix $\mathbf{\Psi_B}$ having diagonal elements that imply a diffuse prior distribution (e.g., 1000) and off-diagonal elements that imply large covariances among the parameters (e.g., 999.95). Taken together, the large on-diagonal values imply that, on average, parameter values will be largely driven by the data Y, but the large off-diagonal values operationalize a prior expectation of very small parameter differences, thus shrinking parameters toward each other during estimation (*without* also shrinking them toward zero).

This approach with small-variance priors is a simplification of others, such as state-space models with hierarchical priors (see Koop and Korobilis, 2010; Korobilis, 2013). Although these other methods can be approximated using our approach in various ways (for insight see Chow et al., 2010), our goal is not to extend these other methods but instead to provide an introduction to using small-variance priors for panel data models in SEM within a very user-friendly framework. For this our Mplus input and output are available in **Supplementary Material** with the required data from Zyphur et al.'s online materials so that the reader can freely experiment with priors in GCLMs (notably R users can convert our basic GCLM code into Lavaan using the R program Mplus2lavaan, available here: https://rdrr.io/cran/lavaan/man/mplus2lavaan.html).

The interested reader may also want to examine more technical on the choice of small-variance priors after exploring our article (e.g., Canova, 2007; Canova and Ciccarelli, 2013), especially that which covers the level of prior informativeness in the form of prior variances (e.g., Giannone et al., 2015). Related work also exists in psychology showing that weakly informative priors can help stabilize model parameters in models similar to the GCLM (Lüdtke et al., 2018), which as we show offers important insights that helps motivate some small-variance prior specifications. For pedagogical purposes, we set prior variances at 0.01 (i.e., a prior *SD* of 0.1) in order to express somewhat strong prior expectations that parameters are close to the mean and to be consistent with existing work on small-variance priors (Muthén and Asparouhov, 2012; Zyphur and Oswald, 2015), but in practice researchers may use sensitivity analyses to examine informativeness or they may use automated techniques to determine prior variances (e.g., Giannone et al., 2015).

THE GCLM WITH SMALL-VARIANCE PRIORS

In order to show how a Bayesian approach to estimation and inference can benefit time-series and panel data models (or other models), we now modify the GCLM presented previously and we alter the way it has been estimated by using small-variance priors. We begin with time-varying parameters that incorporate small-variance priors for differences in parameter estimates over time (sometimes called time-varying effects models or TVEMs) and then we proceed to a more traditional form of small-variance "Minnesota" prior for higher-order lags in panel data models—named for the location of the central bank and economists who pioneered the approach.

Time-Varying Parameters

Our general panel data model from Eqs. 1 and 2 can be usefully extended by allowing time-varying AR, MA, CL, and CLMA effects, which we show as follows (for $t > 2$):

$$x_{i,t} = \alpha_t^{(x)} + \lambda_t^{(x)}\eta_i^{(x)} + \beta_{x1,t}^{(x)}x_{i,t-1} + \delta_{x1,t}^{(x)}u_{i,t-1}^{(x)} + \delta_{x2,t}^{(x)}u_{i,t-2}^{(x)} + \beta_{y1,t}^{(x)}y_{i,t-1} + \delta_{y1,t}^{(x)}u_{i,t-1}^{(y)} + u_{i,t}^{(x)} \quad (8)$$

$$y_{i,t} = \alpha_t^{(y)} + \lambda_t^{(y)}\eta_i^{(y)} + \beta_{y1,t}^{(y)}y_{i,t-1} + \delta_{y1,t}^{(y)}u_{i,t-1}^{(y)} + \beta_{x1,t}^{(y)}x_{i,t-1} + \delta_{x1,t}^{(y)}u_{i,t-1}^{(x)} + u_{i,t}^{(y)} \quad (9)$$

wherein all terms are as described previously. This kind of specification is important because researchers have found that some of the greatest improvements in fit and prediction come from allowing time-varying parameters (a type of non-stationarity; Sims and Zha, 2006). However, in our case of $k = 2$ and $T = 6$, this model is not identified with a frequentist estimator because of the many time-varying terms. For example, income $x_{i,t}$ has 19 parameters that rely on only 15 auto-covariances for estimation: five time-varying unit effects λ_t; one unit effect variance$\psi_\eta^{(x)}\psi_\eta^{(x)}$; five AR terms; and eight MA terms. Also, even the SWB variable with only an MA(1) specification has 16 unique parameters that rely on 15 auto-covariances, meaning the model is under-identified for both x and y. Yet, even if the model were identified, the abundance of parameters might overfit the data, producing results that are not as generalizable—a problem that frequentist estimators can produce in panel data models like Eqs. 8 and 9. Furthermore, given our modest sample size $N = 135$, estimating so many parameters calls into question the asymptotic justification for ML in relation to the number of parameters estimated.

In order to increase model parsimony and identify the model while at the same time helping to address asymptotic concerns related to the ML estimator used in Zyphur et al. (2020a), we take a Bayesian approach with small-variance priors for *differences* in AR, MA, CL, and CLMA terms, with priors as follows (for $t > 1$) to allow differences in parameters over time by "shrinking" these differences (i.e., by helping parameters remain similar over time):

$$\text{AR effects for income}: \ (\beta_{x1,t}^{(x)} - \beta_{x1,t+1}^{(x)}) \sim N(0, 0.01)$$

$$\text{MA effects (first} - \text{order) for income}: \ (\delta_{x1,t}^{(x)} - \delta_{x1,t+1}^{(x)}) \sim N(0, 0.01)$$

$$\text{MA effects (second} - \text{order) for income}: \ (\delta_{x2,t}^{(x)} - \delta_{x2,t+1}^{(x)}) \sim N(0, 0.01)$$

$$\text{CL effects for income}: \ (\beta_{y1,t}^{(x)} - \beta_{y1,t+1}^{(x)}) \sim N(0, 0.01)$$

$$\text{CLMA effects for income}: \ (\delta_{y1,t}^{(x)} - \delta_{y1,t+1}^{(x)}) \sim N(0, 0.01)$$

$$\text{AR effects for SWB}: \ (\beta_{y1,t}^{(y)} - \beta_{y1,t+1}^{(y)}) \sim N(0, 0.01)$$

$$\text{MA effects for SWB}: \ (\delta_{y1,t}^{(y)} - \delta_{y1,t+1}^{(y)}) \sim N(0, 0.01)$$

$$\text{CL effects for SWB}: \ (\beta_{x1,t}^{(y)} - \beta_{x1,t+1}^{(y)}) \sim N(0, 0.01)$$

$$\text{CLMA effects for SWB}: \ (\delta_{x1,t}^{(y)} - \delta_{x1,t+1}^{(y)}) \sim N(0, 0.01)$$

Where in all terms are as above and the expected differences in each set of parameters are set to a small value. This prior specification implies that the GCLM under ML is not nested within this model with time-varying lagged effects—although it still provides an interesting opportunity to compare results. Specifically, with a mean of 0 and a variance of 0.01, these normally distributed priors set a roughly 68% probability that the parameters are within ++/−. One of each other over time. This is akin to relatively strong prior beliefs that the parameters are similar over time, which can be understood in relation to a prior $f(\mathbf{B}) \sim MVN(0, \mathbf{\Psi_B})$, with the covariance matrix $\mathbf{\Psi_B}$ having large diagonal and off-diagonal elements—implying diffuse priors while at the same time imposing an expectation of similar parameter values over time. Conveniently, observed data will test the veracity of this expectation by pulling posteriors away from these priors if this is warranted by the data (Muthén and Asparouhov, 2012).

To continue, we can increase model parsimony further by setting small-variance priors for time-varying unit effects, which we illustrate in two ways. First, recall that SWB is often highly stable (see Easterlin, 1995, 2001; Diener and Lucas, 1999; Clark et al., 2008). Indeed, in psychology, it is common to assume a form of mean-stationarity for η_i by setting $\lambda_t \equiv 1$ (e.g., Hamaker et al., 2015). This assumption of constant effects is so common that it is the default for multilevel models and most fixed-effects approaches (see Nezlek, 2012a,b; Allison, 2014; Hoffman, 2015). Our results in **Table 1** for the ML model support this, showing similar effects for $\lambda_t^{(y)}$ over T. Therefore, we set the following small-variance priors to operationalize an expectation of mean-stationarity for $\eta_i^{(y)}$ (for $t> 1$): $(\lambda_t^{(y)} - \lambda_{t+1}^{(y)}) \sim N(0, .01)$.

This kind of prior specification—operationalizing theory and past findings—is in the spirit of Minnesota priors (see Koop and Korobilis, 2010). In this tradition, econometricians often assume small if any unit effects for variables like income (Canova and Ciccarelli, 2013). Instead, trends are often treated as stochastic rather than deterministic—as noted in Zyphur et al. (2020b)—which is supported by results in **Table 1** for the ML model, showing weak time-varying effects $\lambda_t^{(x)}$. Conveniently, Bayesian priors allow a model that incorporates time-varying unit effects but simultaneously bets against them, so to speak. To put this into practice, we use a prior that assumes no unit effects (for $t > 1$) $\lambda_t^{(x)} \sim N(0, 0.01)$. This prior has multiple benefits: it shrinks unit effects toward zero; it reduces the number of parameters that are freely estimated; and it allows unit effects to manifest in posteriors as a function of the observed data—in part by leaving income's unit effect variance $\psi_\eta^{(x)}$ unrestricted[2].

Furthermore, these prior specifications on the factor loadings of the latent unit effects help to resolve a dilemma that other researchers may experience when using a relatively small sample (here $N = 135$) in the presence of modest unit effects variances (for an overview and relevant simulations see Lüdtke et al., 2018). Specifically, the default non-noninformative or diffuse priors in Mplus can cause estimation problems with unit effect variances and their factor loadings, which we encountered with variances tending to zero and loadings that were incredibly large when estimating the GCLM with a Bayes estimator and the default priors in Mplus (we omit results but the reader can find them in our online materials in the file "AR(1)MA(2) (Step 1, Full Model) Bayes.out"). One solution to this problem is imposing a mean-stability assumption by restricting the factor loadings to equality over time (after the $t = 1$ occasion, which resolves the problem with the parameter estimates as shown in the file "AR(1)MA(2) (Step 1, Full Model) Bayes_mean stability.out"). However, the small-variance priors we describe here allow avoiding the mean-stability assumption while also stabilizing the variance and factor loadings estimates.

In sum, the above combination of small-variance priors minimizes model complexity due to time-varying parameters while at the same time allowing the estimation of all parameters even when they are not identified with frequentist estimators or because of other estimation problems. Using a Bayes approach, we estimate the model in Eqs. 8 and 9 with the above priors using a Markov Chain Monte Carlo (MCMC) method with a Gibbs sampler in Mplus. For this and other models that follow, estimation is done with at least 10,000 iterations in two chains—these were thinned by retaining every 50th estimate (for a total of 500,000 iterations) to assure convergence within the 10,000 estimates and eliminate autocorrelation across the iterations.

Convergence is checked by examining the quality of chain mixing with the *estimated* or *potential scale reduction* (*PSR*) factor, with values of 1.05 or less typically used as a cut-off (see Gill, 2008, pp. 478–482; Asparouhov and Muthén, 2010). We also use Kolmogorov-Smirnov tests that compute p-values for parameter differences between chains, testing convergence for each parameter separately (while allowing for a Type-I error rate of 0.05 across all p-values). Model fit is evaluated by the posterior-predictive probability or p-value (*PPP*), which indicates the relative fit of model-generated data versus observed data, with values of 0.50 being optimal and values greater than 0.05 typically considered acceptable (Muthén and Asparouhov, 2012). Comparisons of models may be done using the deviance information criterion (DIC) as a relative index of model quality (balancing fit and parsimony), with smaller values indicating a better model. The DIC is useful because it is uniquely sensitive to the number of estimated parameters *pD*, which is a function of the number of unrestricted parameters and the amount of information provided by priors (see Asparouhov et al., 2015), and thus this value will typically not be an integer value as in the ML case where priors do not exist. Consistent with other approaches, we use the *SD* of posterior distributions to compute Bayesian analogs of two-tailed p-values (Zyphur and Oswald, 2015). For impulse responses, we use 95% credibility intervals with the highest posterior density, which are similar to bootstrap *CI*s (Rubin, 1981). For all parameters not explicitly mentioned, we use default uninformative/diffuse priors in Mplus (Asparouhov and Muthén, 2010), which is done for convenience and to keep the reader focused on the bespoke priors specification used here.

[2] As a form of sensitivity analysis, we also estimated a model with a small-variance prior for differences in income's unit effects, setting $(\lambda_t^{(x)} - \lambda_{t+1}^{(x)}) \sim (0, 0.01)$ (for $t> 1$). We observed no notable differences in model results with this set of priors versus the prior $(\lambda_t^{(x)}) \sim (0, 0.01)$.

Model results are in **Table 1** under the "Bayes 1" model, showing acceptable fit ($PPP = 0.32$). For concision, we report the averages and ranges of time-varying AR, MA, CL, and CLMA terms (readers can examine full results in our online materials). For example, AR effects for income have four terms associated with each occasion of measurement that is endogenous to all lagged effects, with $\beta_{x1,3}^{(x)} = 0.97$, $\beta_{x1,4}^{(x)} = 0.97$, $\beta_{x1,5}^{(x)} = 0.94$, and $\beta_{x1,6}^{(x)} = 1.0$, for which **Table 1** shows the mean 0.97, the range [0.94, 1.0], and the average posterior $SD = 0.03$ ($p < 0.001$). As **Table 1** shows, averaged terms are similar to their frequentist counterparts in many cases, such as the AR effect for the ML model $\beta_{x1}^{(x)} = 0.96$ versus the Bayesian average 0.97. Furthermore, ranges are relatively small for Bayesian estimates, indicating little difference in most parameters over time under the combination of small-variance priors and observed data used here.

However, a noticeable change occurs in the level of uncertainty around parameters. For example, the AR effect for the ML model has an $SE = 0.13$, whereas the Bayes average has an $SD = 0.03$. This reduction in uncertainty is expected for two reasons. First, allowing parameters to vary over time can increase their fit to the data at each occasion, reducing uncertainty around the estimates as a function of better fit to the covariance for any two occasions. Second, as **Table 1** shows, the total number of parameters estimated in the Bayes model is slightly smaller than the ML-based model (54 versus an estimated $pD = 50.69$ in the Bayes model), ostensibly because of the small-variance priors. In turn, although time-varying effects are allowed, the Bayes model appears to be slightly more parsimonious, implying less uncertainty for the entire model, which on average should result in smaller Bayesian posterior *SD*s than ML-based *SE*s.

An interesting consequence of this uncertainty reduction is that Granger-causality tests and impulse responses show different results for the income → SWB effect, supporting it much more strongly. In the ML-based model, the income → SWB effect is $\beta_{x1}^{(y)} + \delta_{x1}^{(y)} = 0.14$, $SE = 0.16$, $p = 0.40$; whereas in the Bayes model it becomes $\beta_{x1}^{(y)} + \delta_{x1}^{(y)} = 0.22$, $SD = 0.12$, $p = 0.06$. However, rather than relying on *p*-values, we test Granger-causality using the DIC. As shown in **Table 2**, the DIC for the full model is 829.23, and eliminating the income → SWB CL and CLMA terms increases this to 835.48, indicating reduced model quality and therefore supporting an income → SWB effect. Alternatively, removing the SWB → income CL and CLMA terms, the DIC falls to 820.83, indicating improved model quality and therefore failing to support an SWB → income effect. Finally, testing for income-SWB feedback by constraining all CL and CLMA terms also reduces model quality with a DIC of 833.35, providing support for feedback effects. Yet, this raises the question of whether the income → SWB effect is driving the larger DIC value when eliminating all CL and CLMA terms in order to test for feedback.

To investigate this and to show long-run effects, we examined impulse responses (see **Figures 2A–D**)[3]. The differences between the ML-based and Bayesian impulse responses are notable, with much less uncertainty around income's persistence over time (the top-right figure). Also, impulse responses show a larger effect for income → SWB and much less uncertainty around the estimate, with 95% credibility intervals encompassing zero only at the margins (consistent with $p = 0.06$). Furthermore, the SWB → income effect is approximately zero across all periods. These findings lend more credibility to a positive long-run income → SWB effect when compared to the frequentist estimates in **Figures 1A–D**, and less credibility to a long-run SWB → income effect. The results also imply that the lower DIC value when testing feedback is due to the income → SWB effect rather than the opposite, arguing against income-SWB feedback.

In sum, the small-variance priors we use allow model specifications that are plausible yet under-identified with frequentist methods. By allowing effects to vary over time, we provide a better fit to the observed data and reduce the uncertainty around estimates, pointing to an effect of income on SWB that appears to be long-lasting. Indeed, when eliminating the SWB → income CL and CLMA effects, which is warranted based on the decrease in the DIC, we show an income → SWB effect combining CL and CLMA terms: $\beta_{x1}^{(y)} + \delta_{x1}^{(y)} = 0.24$ with a posterior $SD = 0.12$, $p = 0.04$. Furthermore, this effect with a one-tailed test has a $p = 0.02$ and the 95% credibility interval in **Figures 2A–D** exclude zero. The implication is that a positive impulse to national income may have a positive immediate and long-run effect on SWB, neither of which was found in the ML-based analyses in **Tables 1**, **2**, and **Figures 1A–D**, because of the restrictions on the effects that were required. For our "informal Bayesian," this implies updated knowledge or belief about a causal income → SWB effect, which may be used to inform policy decisions.

Reducing Lag Orders

To further tackle overparameterization and provide an additional tool for estimating models that may be under-identified, small-variance priors can be applied to high-order lags and unit effects. As with time-varying parameters, the issue is that estimating many lagged effects and time-varying unit effects can overfit observed data while also making models under-identified with frequentist estimators. This is important because, for prediction, "[e]vidence favors Bayesian estimation of an equation with high-order lags rather than restricted models arrived at by classical testing methods" (Allen and Fildes, 2001, p. 335; Stock and Watson, 2001).

To illustrate this approach while keeping our results both concise and comparable to those presented thus far, we specify the same models for both income and SWB (Eqs. 8 and 9), but set small-variance priors on the second-order MA lag for income. In Zyphur et al. (2020a) the authors appear compelled to choose a single model for income, comparing the results of AR(1)MA(1), AR(1)MA(2), AR(2)MA(1), and AR(2)MA(2) models for $x_{i,t}$. Conveniently, a Bayes estimator changes the nature of this choice by allowing higher-order lags to have small-variance priors with means of zero, reflecting an expectation of no higher-order lagged

[3] Time-varying AR, MA, CL, and CLMA effects imply a unique impulse response for each impulse over time. We calculate impulse responses based on the first available impulse given the model lag order MA(2). For discussion of impulse responses for time-varying parameters and Bayesian estimators, see Koop (1996).

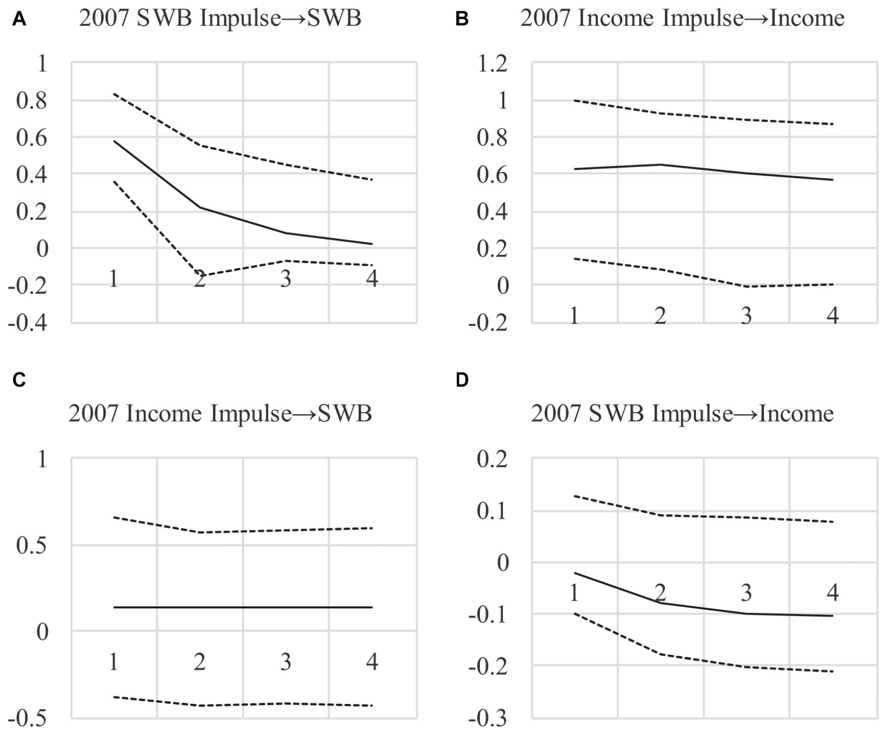


FIGURE 2 | (A–D) Impulse Response Functions for AR(1)MA(2) Model With Bayesian Small-Priors. Impulse begin in 2007, showing the effect of a 1-unit impulse in 2007 over the next 4 years, with 95% credibility intervals (with the highest posterior density).

effects while allowing them to emerge as a function of the data. The result is an ability to retain time-varying effects for AR, MA, CL, and CLMA terms while also allowing them to have many lags that minimally add to model complexity due to the use of small-variance priors. To show this, we set the following small-variance prior for the time-varying $\delta_{x2,t}^{(x)}$ term in Eq. 8: $\delta_{x2,t}^{(x)} \sim N(0, 0.01)$. This small-variance prior allows the second-order MA terms to vary over time while shrinking them toward zero and keeping the number of estimated parameters manageable.

The results of this model are shown in **Table 1** under the "Bayes 2" model, with Granger causality tests in **Table 2**. As **Table 1** shows, the fit of the model improves over the previous "Bayes 1" model that allowed the second-order MA term $\delta_{x2,t}^{(x)}$ to be unrestricted with a small-variance prior on differences with $(\delta_{x2,t}^{(x)} - \delta_{x2,t+1}^{(x)}) \sim N(0, 0.01)$. Specifically, the DIC falls from 829.23 to 827.65, indicating some improvement by shrinking second-order MA terms toward zero while still allowing them to be time-varying.

Interestingly, this second Bayes model also fits the data better than two others that may also seem warranted and of interest to researches exploring different MA structures for income. The first is a model wherein the same small-variance prior is applied but the second-order MA term is constrained to equality over time, with an effect $\delta_{x2}^{(x)}$ as in the original ML model in **Table 1** and$\delta_{x2}^{(x)} \sim (0, 0.01)$. The DIC for this model increases to 831.46, arguing for the time-varying specification with the same null small-variance prior $\delta_{x2,t}^{(x)} \sim N(0, 0.01)$ in the Bayes 2 model in **Table 1**. The second model that seems plausible is one that takes the prior expectation of no effect as the actual model specification, fixing the second-order MA term to zero (i.e., fixing $\delta_{x2,t}^{(x)} \equiv 0$), resulting in a purely MA(1) specification for income. This model has a DIC that increases to 830.63, again favoring the MA(2) specification with the small-variance null prior on the second-order lagged MA effect $\delta_{x2,t}^{(x)} \sim (0, 0.01)$. In sum, the small-variance priors that allow time-varying effects outperform other plausible specifications in this case, and allow researchers to operationalize an expectation of no higher-order lagged effects while still allowing results to be pulled away from this prior expectation as a function of the data.

Given the improved fit of the second Bayes model in terms of the DIC, it is interesting to note that, again, Granger-causality tests under this model show an increase in the DIC when removing the income → SWB effect (see **Table 2**, Bayes 2 model), with the full model DIC = 827.65, but with all income → SWB CL and CLMA terms eliminated the DIC increases to 833.85. This is consistent with the overall income → SWB effect, which again is $\beta_{x1}^{(y)} + \delta_{x1}^{(y)} = 0.22$, $SD = 0.12$, $p = 0.06$. Furthermore, as before, constraining the SWB → income CL and CLMA terms to zero improves model fit with DIC = 818.71, failing to support Granger-causality in this direction. Also, feedback effects appear to exist with DIC = 831.70 under a model with no CL and CLMA terms, but this appears to be entirely due to the income → SWB effect, which is supported by impulse responses, which we omit because they are very similar to **Figures 2A–D**.

In sum, there are at least three benefits of the informative, small-variance priors that we use here. First, they shrink estimates toward zero or toward each other for time-varying parameters, which in our model reduces uncertainty substantially and thereby supports an income → SWB effect that could not be supported with an ML estimator. This is useful because, second, the increase in model parsimony due to the model priors also increases generalizability, which means the income → SWB effect is also more trustworthy than under ML. Third, the priors we use allow estimation that would be impossible with a frequentist estimator, allowing higher-order AR, MA, CL, and CLMA effects while also using small-variance priors that reduce the need to choose amongst models that have different lag orders for these terms. These three benefits are in addition to those that apply to Bayesian estimation more generally, including not only its fit with an "informal Bayesian" using panel data models to make inferences under uncertainty, but also computational efficiencies of Bayesian estimation (see Chib, 2008; Muthén and Asparouhov, 2012; Zyphur and Oswald, 2015). For more details, the reader may consult additional work on Bayesian analysis for panel data models (e.g., Sims, 1980, 1986; Canova, 2007; Koop and Korobilis, 2010; Korobilis, 2013; Giannone et al., 2015; Schuurman et al., 2016).

DISCUSSION

In their recent series "From Data to Causes," Zyphur et al. (2020a) described the GCLM parameters and their relationship to Granger causality and intervention planning via impulse responses, with all terms estimated via ML in an SEM framework. These authors also compared their approach to others, noting the benefits of dynamic models that make the future conditional on the past while controlling for unit effects, thus addressing issues with static approaches including latent curve models (i.e., latent growth or trajectory models; Zyphur et al., 2020b). However these authors did not acknowledge shortcomings of their frequentist estimation method and thus in the current article we extended the GCLM to the case of Bayesian estimation and inference, showing the usefulness of small-variance priors for both parameter estimates and parameter differences in models that would otherwise have high dimensions that produce generalizability and/or estimation problems. The result is that here we were able to estimate time-varying parameters while shrinking higher-order lagged effects and time-varying unit effects for income toward zero, reducing parameter uncertainty and allowing us to support an income → SWB effect that does not receive support under ML estimation.

With such Bayesian approaches to time-series and panel data modeling, researchers have a set of powerful tools for doing the practical work that defines the applied social sciences. This work has various characteristics that often center on theorizing and empirically studying causal effects, such as the income → SWB effect, which we support in the current study. For any applied science, the point of such a finding—and research more generally—is a practical affair, with researchers seeking to develop understandings of the world that can guide action, such as organizational or public policy interventions (Cartwright and Hardie, 2012). In turn, the point of these interventions is to create specific kinds of outcomes, such as improving SWB by helping poor nations to develop their economies in order to increase income. To these ends, a benefit of small-variance priors and methods of "shrinkage" more generally is to improve generalizability so that such inferences can have a greater chance of working in real-world situations.

However, there are various dangers associated with using models such as ours uncritically. One danger is the well-known problem of exactly how a researcher or policy maker should derive priors—what sources of information should be used for this purpose—and how the choice of different prior specifications may affect results. These topics have received substantial attention in Bayesian literature and we encourage the interested reader to engage with this work (again the interested reader may consult excellent work on these and other topics; e.g., Depaoli and van de Schoot, 2017; Smid et al., 2020). As we noted previously our use of the specific small-variance prior of $\sim N(0{,}0.01)$ was used for example purposes and to fit with previous literature (see Muthén and Asparouhov, 2012; Zyphur and Oswald, 2015). Future work may investigate other potential types of small-variance priors to complement the existing and ever-growing body of work on the use of priors for Bayesian analysis of time-series and panel data models (e.g., Lüdtke et al., 2018).

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

MZ organized and led the research including analyses and initial manuscript drafting. EH, LT, MV, KP, ZZ, PA, DP, PK, and ED contributed substantial insights during the phases of model development and interpretation as well as generating conclusions in the discussion section. All authors contributed directly to manuscript writing and revisions during many collective rounds of editing and write-up.

FUNDING

This research was supported by Australian Research Council's Future Fellowship Scheme (project FT140100629).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.612251/full#supplementary-material

REFERENCES

Allen, P. G., and Fildes, R. (2001). "Econometric forecasting," in *Principles of Forecasting: A Handbook for Researchers and Practitioners*, ed. J. S. Armstrong (Norwell: Kluwer Academic Publishers).

Allison, P. D. (2014). Maximum likelihood for dynamic panel models with cross-lagged effects. *Panel Data Econometr.* 17, 547–581. doi: 10.1016/b978-0-12-815859-3.00017-2

Anderson, J. C., and Gerbing, D. W. (1984). The effect of sampling error on convergence, improper solutions, and goodness-of-fit indices for maximum likelihood confirmatory factor analysis. *Psychometrika* 49, 155–173. doi: 10.1007/bf02294170

Arellano, M. (2003). *Panel Data Econometrics*. Oxford: Oxford University Press.

Armstrong, J. S., Green, K. C., and Graefe, A. (2015). Golden rule of forecasting: Be conservative. *J. Bus. Res.* 68, 1717–1731. doi: 10.1016/j.jbusres.2015.03.031

Asparouhov, T., and Muthén, B. (2010). *Bayesian Analysis Using Mplus: Technical Implementation*. URL: Retrieved from: www.statmodel.com/download/Bayes3.pdf

Asparouhov, T., Muthén, B., and Morin, A. J. (2015). Bayesian Structural Equation Modeling With Cross-Loadings and Residual Covariances: Comments on Stromeyer et al. *J. Manag.* 41, 1561–1577. doi: 10.1177/0149206315591075

Baltagi, B. H. (2013). "Dynamic panel data models," in *Handbook of Research Methods and Applications on Empirical Macroeconomics*, eds N. Hashimzade and M. Thornton (Cheltenham: Edward Elgar), 229–248. doi: 10.4337/9780857931023.00016

Bollen, K. A. (1989). *Structural Equations With Latent Variables*. New York: Wiley. doi: 10.1002/9781118619179

Canova, F. (2007). *Methods for Applied Macroeconomic Research*. Princeton: Princeton University Press. doi: 10.1515/9781400841028

Canova, F., and Ciccarelli, M. (2013). Panel vector autoregressive models: A survey. *Adv. Econometr.* 32, 205–246. doi: 10.1108/s0731-905320130000031006

Cartwright, N., and Hardie, J. (2012). *Evidence-Based Policy: A Practical Guide To Doing It Better*. Oxford: Oxford University Press.

Chib, S. (2008). "Panel data modelling and inference: A Bayesian primer," in *The Econometrics of Panel Data: Fundamentals and Recent Developments in Theory And Practice*, eds L. Mátyás and P. Sevestre (Berlin: Springer-Verlag), 479–516. doi: 10.1007/978-3-540-75892-1_15

Chow, S.-M., Ho, M.-H. R., Hamaker, E. L., and Dolan, C. V. (2010). Equivalence and differences between structural equation modeling and state-space modeling techniques. *Struct. Equat. Model.* 17, 303–332. doi: 10.1080/10705511003661553

Clark, A. E., Frijters, P., and Shields, M. A. (2008). Relative income, happiness, and utility: An explanation for the Easterlin paradox and other puzzles. *J. Econom. Liter.* 46, 95–144. doi: 10.1257/jel.46.1.95

Depaoli, S., and van de Schoot, R. (2017). Improving transparency and replication in bayesian statistics: The WAMBS-checklist. *Psychol. Methods* 22, 240–261. doi: 10.1037/met0000065

Diener, E., and Lucas, R. E. (1999). "Personality and subjective well-being," in *Well-Being: Foundations of Hedonic Psychology*, eds D. Kahneman, E. Diener, and N. Schwarz (New York: Sage), 213–229.

Diener, E., Tay, L., and Oishi, S. (2013). Rising income and the subjective well-being of nations. *J. Person. Soc. Psychol.* 104, 267–276. doi: 10.1037/a0030487

Easterlin, R. A. (1995). Will raising the incomes of all increase the happiness of all? *J. Econ. Behav. Organiz.* 27, 35–47. doi: 10.1016/0167-2681(95)00003-b

Easterlin, R. A. (2001). Income and happiness: Towards a unified theory. *Econ. J.* 214, 465–484. doi: 10.1111/1468-0297.00646

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2014). *Bayesian Data Analysis*. New York: CRC Press. doi: 10.1201/b16018

Giannone, D., Lenza, M., and Primiceri, G. E. (2015). Prior selection for vector autoregressions. *Rev. Econ. Statist.* 97, 436–451. doi: 10.1162/rest_a_00483

Gill, J. (2008). *Bayesian methods: A Social and Behavioral Science Approach*, 2nd Edn. Boca Raton: Chapman & Hall.

Granger, C. W. J. (1980). Testing for causality, a personal viewpoint. *J. Econ. Dynam. Contr.* 2, 329–352. doi: 10.1016/0165-1889(80)90069-X

Green, K. C., and Armstrong, J. S. (2015). Simple versus complex forecasting: The evidence. *J. Bus. Res.* 68, 1678–1685. doi: 10.1016/j.jbusres.2015.03.026

Hacking, I. (2001). *An Introduction to Probability and Inductive Logic*. Cambridge: Cambridge University Press.

Hamaker, E. L., Kuiper, R. M., and Grasman, R. P. (2015). A critique of the cross-lagged panel model. *Psychol. Methods* 20, 102–116. doi: 10.1037/a0038889

Hoffman, L. (2015). *Longitudinal Analysis: Modeling Within-Person Fluctuations and Change*. New York: Routledge.

Howson, C., and Urbach, P. (2006). *Scientific Reasoning: the Bayesian Approach*. Chicago: Open Court Publishing.

Hsiao, C. (2014). *Analysis of Panel Data*, 3rd Edn. Cambridge: Cambridge University Press.

Jaynes, E. T. (2003). *Probability Theory: The Logic of Science*. Cambridge: Cambridge university press. doi: 10.1017/CBO9780511790423

Koop, G. (1996). Parameter uncertainty and impulse response analysis. *J. Econom.* 72, 135–149. doi: 10.1016/0304-4076(94)01717-4

Koop, G., and Korobilis, D. (2010). Bayesian multivariate time series methods for empirical macroeconomics. *Foundat. Trends Econometr.* 3, 267–358. doi: 10.1561/0800000013 doi: 10.1561/0800000013

Korobilis, D. (2013). Hierarchical shrinkage priors for dynamic regressions with many predictors. *Int. J. Forecast.* 29, 43–59. doi: 10.1016/j.ijforecast.2012.05.006

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

Lüdtke, O., Robitzsch, A., and Wagner, J. (2018). More stable estimation of the STARTS model: A Bayesian approach using Markov chain Monte Carlo techniques. *Psychol. Methods* 23, 570–593. doi: 10.1037/met0000155

MacCallum, R. C., Browne, M. W., and Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychol. Methods* 1:130. doi: 10.1037/1082-989x.1.2.130

McNeish, D. M. (2015). Using Lasso for predictor selection and to assuage overfitting: A method long overlooked in behavioral sciences. *Multivar. Behav. Res.* 50, 474–481.

Muthén, B. (2010). *Bayesian Analysis In Mplus: A Brief Introduction*. Retrieved from: www.statmodel.com/download/IntroBayesVersion%203.pdf

Muthén, B., and Asparouhov, T. (2012). Bayesian structural equation modeling: a more flexible representation of substantive theory. *Psychol. Methods* 17:313. doi: 10.1037/a0026802

Nezlek, J. B. (2012a). *Diary Methods for Social and Personality Psychology*. London: Sage.

Nezlek, J. B. (2012b). "Multilevel modeling of diary-style data," in *Handbook of Research Methods for Studying Daily Life*, eds M. R. Mehl and T. S. Conner (New York: Guilford Press), 357–383.

Schuurman, N. K., Grasman, R. P., and Hamaker, E. L. (2016). A comparison of inverse-wishart prior specifications for covariance matrices in multilevel autoregressive models. *Multivar. Behav. Res.* 51, 1–22. doi: 10.1007/s12480-016-0007-6

Sims, C. A. (1980). Macroeconomics and reality. *Econometrica* 48, 1–48. doi: 10.4337/9781784719029.00006

Sims, C. A. (1986). Are forecasting models usable for policy analysis? *Feder. Res. Bank .Minneap. Q. Rev.* 10, 2–16.

Sims, C. A., and Zha, T. (2006). Were there regime switches in U.S. monetary policy? *Am. Econ. Rev.* 96, 54–81. doi: 10.1257/000282806776157678

Smid, S. C., McNeish, D., Miočević, M., and van de Schoot, R. (2020). Bayesian versus frequentist estimation for structural equation models in small sample contexts: A systematic review. *Struct. Equat. Model.* 27, 131–161. doi: 10.1080/10705511.2019.1577140

Stock, J. H., and Watson, M. W. (1996). Evidence on structural instability in macroeconomic time series relations. *J. Bus. Econ. Stat.* 14, 11–30.

Stock, J. H., and Watson, M. W. (2001). Vector autoregressions. *J. Econ. Perspect.* 15, 101–115. doi: 10.1257/jep.15.4.101

Stock, J. H., and Watson, M. W. (2009). "Forecasting in dynamic factor models subject to structural instability," in *The Methodology and Practice of Econometrics: Festschrift in Honor of D. F. Hendry*, eds N. Shephard, and J. Castle (New York, NY: Oxford University Press), 173–205.

Rubin, D. B. (1981). The bayesian bootstrap. *Ann. Statist.* 9, 130–134.

van de Schoot, R., Winter, S. D., Ryan, O., Zondervan-Zwijnenburg, M., and Depaoli, S. (2017). A systematic review of Bayesian articles in psychology: The last 25 years. *Psychol. Methods* 22, 217–239. doi: 10.1037/met0000100

Williams, J. W., and Cook, N. M. (2016). Econometrics as evidence? Examining the 'causal' connections between financial speculation and commodities prices. *Soc. Stud. Sci.* 46, 701–724. doi: 10.1177/0306312716658980

Zyphur, M. J., Allison, P. D., Tay, L., Voelkle, M. C., Preacher, K. J., Zhang, Z., et al. (2020a). From data to causes I: Building a general cross-lagged panel model (GCLM). *Organiz. Res. Methods* 23, 651–687. doi: 10.1177/1094428119847278

Zyphur, M. J., and Oswald, F. L. (2015). Bayesian Estimation and Inference A User's Guide. *J. Manag.* 41, 390–420. doi: 10.1177/0149206313501200

Zyphur, M. J., Voelkle, M. C., Tay, L., Allison, P. D., Preacher, K. J., Zhang, Z., et al. (2020b). From data to causes II: Comparing approaches to panel data analysis. *Organizat. Res. Methods* 23, 688–716. doi: 10.1177/1094428119847280

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Psychology

ORIGINAL RESEARCH
published: 24 March 2021
doi: 10.3389/fpsyg.2021.620802

Systematically Defined Informative Priors in Bayesian Estimation: An Empirical Application on the Transmission of Internalizing Symptoms Through Mother-Adolescent Interaction Behavior

Susanne Schulz[1]*, *Mariëlle Zondervan-Zwijnenburg*[2], *Stefanie A. Nelemans*[1], *Duco Veen*[3,4], *Albertine J. Oldehinkel*[5], *Susan Branje*[1] *and Wim Meeus*[1]

[1] Youth and Family, Utrecht University, Utrecht, Netherlands, [2] Methodology and Statistics, Utrecht University, Utrecht, Netherlands, [3] Julius Global Health, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, Netherlands, [4] Optentia Research Program, North-West University, Potchefstroom, South Africa, [5] Interdisciplinary Center Psychopathology and Emotion Regulation, University of Groningen, University Medical Center Groningen, Groningen, Netherlands

OPEN ACCESS

Edited by:
Christoph Koenig,
Goethe University Frankfurt, Germany

Reviewed by:
Luiz Carvalho,
Fundação Getúlio Vargas, Brazil
Martin Hecht,
Humboldt University of Berlin, Germany

****Correspondence:***
Susanne Schulz
s.schulz@uu.nl

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *23 October 2020*
Accepted: *23 February 2021*
Published: *24 March 2021*

Citation:
Schulz S, Zondervan-Zwijnenburg M, Nelemans SA, Veen D, Oldehinkel AJ, Branje S and Meeus W (2021) Systematically Defined Informative Priors in Bayesian Estimation: An Empirical Application on the Transmission of Internalizing Symptoms Through Mother-Adolescent Interaction Behavior. Front. Psychol. 12:620802. doi: 10.3389/fpsyg.2021.620802

Background: Bayesian estimation with informative priors permits updating previous findings with new data, thus generating cumulative knowledge. To reduce subjectivity in the process, the present study emphasizes how to systematically weigh and specify informative priors and highlights the use of different aggregation methods using an empirical example that examined whether observed mother-adolescent positive and negative interaction behavior mediate the associations between maternal and adolescent internalizing symptoms across early to mid-adolescence in a 3-year longitudinal multi-method design.

Methods: The sample consisted of 102 mother-adolescent dyads (39.2% girls, M_{age} T1 = 13.0). Mothers and adolescents reported on their internalizing symptoms and their interaction behaviors were observed during a conflict task. We systematically searched for previous studies and used an expert-informed weighting system to account for their relevance. Subsequently, we aggregated the (power) priors using three methods: linear pooling, logarithmic pooling, and fitting a normal distribution to the linear pool by means of maximum likelihood estimation. We compared the impact of the three differently specified informative priors and default priors on the prior predictive distribution, shrinkage, and the posterior estimates.

Results: The prior predictive distributions for the three informative priors were quite similar and centered around the observed data mean. The shrinkage results showed that the logarithmic pooled priors were least affected by the data. Most posterior estimates were similar across the different priors. Some previous studies contained extremely specific information, resulting in bimodal posterior distributions for the analyses with linear pooled prior distributions. The posteriors following the fitted normal priors and default priors were very similar. Overall, we found that maternal, but not adolescent, internalizing symptoms predicted subsequent mother-adolescent interaction behavior, whereas negative interaction behavior seemed to predict subsequent internalizing symptoms. Evidence regarding mediation effects remained limited.

Conclusion: A systematic search for previous information and an expert-built weighting system contribute to a clear specification of power priors. How information from multiple previous studies should be included in the prior depends on theoretical considerations (e.g., the prior is an updated Bayesian distribution), and may also be affected by pragmatic considerations regarding the impact of the previous results at hand (e.g., extremely specific previous results).

Keywords: intergenerational transmission, internalizing psychopathology, mother-adolescent interaction, informative priors, linear pool, logarithmic pool, Bayesian estimation, longitudinal mediation analysis

INTRODUCTION

New studies and analyses in social sciences are theoretically and empirically grounded in previous knowledge that has often accumulated in decades of research. While there is overall agreement that this process is essential to generate strong hypotheses, findings from previous studies are rarely integrated into new analyses. Accounting for such previous findings in subsequent analyses by means of informative priors in Bayesian estimation allows researchers to draw more precise conclusions and obtain insight into the relation between previous knowledge and the current data.

Bayesian estimation with informative priors increases the precision of the posterior distributions by updating previous information with new data and thus gradually accumulating knowledge. While the frequentist approach regards parameters of interests as unknown, but assumes that there is only one true parameter value in the population, the Bayesian approach regards parameters of interest as uncertain and describes them with a probability distribution (van de Schoot et al., 2014). By combining previous information with new data from the analyses, Bayesian estimation allows researchers to make assumptions about model parameters, such as curtailing or excluding certain parameter values (Zondervan-Zwijnenburg et al., 2017). To date, most empirical studies rely on diffuse or naive prior distributions, such as default software settings, that do not account for the available previous knowledge (e.g., van de Schoot et al., 2017). Simulation studies and mathematical demonstrations indicated that using informative priors that are derived from previous studies, meta-analyses, or experts, outperformed frequentist approaches and approaches using diffuse priors in terms of decreased relative bias, improved estimation accuracy (e.g., decreased RMSE values), and increased power when samples were too small for complex analyses (Smid et al., 2019; Zitzmann et al., 2020). However, if informative priors are not chosen carefully or are weakly defined, Bayesian estimation methods may perform poorly and result in biased estimates (Depaoli, 2013; Holtmann et al., 2016). Therefore, a systematic and transparent approach is essential when specifying informative priors (Zondervan-Zwijnenburg et al., 2017; van de Schoot et al., 2021). The present study highlights the use of different approaches to systematically define informative priors and the integration of new data to answer novel research questions.

Weighting Previous Studies

If previous designs are not consistent with the new study, for example due to different populations or different assessments, this can raise potential bias and inflated type I errors (Hobbs et al., 2011; Viele et al., 2014). Previous findings should therefore strongly inform the posterior distributions when they are based on designs that are comparable to the present study, and weakly when they are not. To ensure that previous findings do not outweigh the current data and dominate the posterior distributions, power priors that downweigh previous data by determining the amount of relevant information have been recommended (Ibrahim and Chen, 2000). Specifically, a power prior takes the likelihood of the information from the previous study to the power δ, where δ is a value between 0 (ignore the previous data completely) and 1 (treat the data as equal to the current data and fully include the evidence). For normal distributions, when delta $\delta \neq 0$, raising the likelihood to the power δ is equal to dividing the variance from the previous study by δ and using it as the prior variance σ_0^2. Traditionally, power priors include the use of unknown weights, which have been criticized to over-attenuate the influence of previous data (Neelon and O'Malley, 2010) as they do not capture the extent to which previous findings are applicable to the present design and data.

Previous studies can be more or less similar to a specific study's design and thus provide stronger or weaker input for priors than other studies. Meta-analyses, for example, quantify existing information, and thus provide accumulated, more robust evidence than single studies. However, they also include a wide range of different methodological designs, such as different participant age ranges or assessment methods, and thus cannot provide strong input for specific parameter estimates if individual participant data is not available. Empirical studies that closely reflect the research questions and design of the new study that is to be conducted provide the strongest input for informative priors, but are more susceptible to potential estimation errors, biases, or chance findings than meta-analyses. How much weight a particular study receives, should therefore depend on a range of aspects that correspond to the study's design at hand. Longitudinal studies, for example, involve different considerations than cross-sectional studies, such as temporal ordering and the lengths of intervals between time points. If the study at hand employs a longitudinal design, findings from studies with repeated measurements would receive more weight than studies that solely include measurements at the same time point. Only a previous study with data that can be considered exchangeable with the new data should receive a weight of 1. To determine how much an individual study deviates from the new data, we therefore propose to determine each study's individual weight for the construction of power priors. Studies with lower relevance obtain lower scores for δ, which means that their variance will be inflated. The larger the variance (i.e., uncertainty), the smaller the impact of a previous study on the specified prior distribution, and therefore also on the posterior distribution.

A carefully constructed and justified weighting scheme that is tailored to the specific research question is essential when specifying informative priors. To avoid arbitrary and subjective decisions, expert knowledge can inform this process (Bolsinova et al., 2017; van de Schoot et al., 2018; Veen et al., 2020). Expert knowledge as input for prior distributions has been previously used to estimate the size of parameters for which no data was available (e.g., Hald et al., 2016) or to complement existing data (e.g., van de Schoot et al., 2018). Our proposed method includes quantifying and weighing previous information, systematically collecting and justifying all decisions, visualizing informative priors, and conducting sensitivity analyses to compare the impact of different priors on the posterior estimates (Zondervan-Zwijnenburg et al., 2017). This can be beneficial beyond a pure meta-analytical approach that solely quantifies previous information. As such, Bayesian estimation with informative priors allows researchers to update previous information by combining it with new data. This cumulative process gradually decreases the uncertainty of parameter estimates (König and van de Schoot, 2018). In the current study, we used expert knowledge to define inclusion criteria and create an appropriate weighting scheme for all included previous studies.

Aggregating Previous Studies

If multiple studies contain information on one parameter, the previous information needs to be aggregated into one distribution. Three aggregation methods are: (1) linear pooling, (2) logarithmic pooling, and (3) a normal distribution fitted to the linear pool.

Linear Pooling

The linear pool of distributions sums the densities provided by the different studies, resulting in a mixture prior (Genest and Zidek, 1986). The linear pool directly represents the previous studies by combining them without any modifications to the initial information. One way to obtain the linear pool is to run multiple Bayesian analyses: one for each prior specification. Subsequently, the posterior samples can be aggregated (see Zondervan-Zwijnenburg et al., 2017). This method can be applied in any software package that allows for Bayesian estimation with customizable prior specifications. However, as parameter estimates within a model are not independent, this method becomes impractical in a model in which multiple parameters have multiple sources of previous information. In more advanced Bayesian software such as Stan (Carpenter et al., 2017), the linear pool of previous studies can be programmed at once. A difficulty that remains is that a linear pool becomes multimodal when the different prior likelihoods diverge. Multimodality is complex for estimation and interpretation. It may cause non-convergence, and it can be odd to consider, for example, 0.2 and 0.5 equally plausible values, but 0.35 a value with low probability. There is the possibility that this scenario occurs when local maxima have previously been found.

Logarithmic Pooling

Whereas the linear pool sums distributions, the logarithmic (a.k.a. geometric) pool multiplies them. In practice this means that extreme modes originating from only one study can be compensated by their multiplication with other studies that allocate less probability to this area. In this manner, the logarithmic pool emphasizes the common range of parameter values. Logarithmic pools are typically unimodal and less dispersed than linear pools (Genest and Zidek, 1986). The logarithmic pool can also be considered a Bayesian updating procedure, in which the first[1] study is the initial prior. A potential disadvantage of the logarithmic pool, however, is that if one previous study places near-zero probability to a range of values, the multiplication by near-zero probability will predominate in the pooled distribution. de Carvalho et al. (2020) define pooled distribution and their parameters for sets of common distributions. When the pooled distribution is a common distribution as well, the prior can be easily specified in software packages that allow for Bayesian estimation with custom prior distributions.

Normal Distribution Fitted to the Linear Pool

Another alternative to including a potentially bimodal linear pool, is to obtain the normal distribution best fitting to this pooled distribution. In this method, the previous studies are considered to be samples from an underlying normal distribution. By fitting a normal distribution to the results of the previous studies, we aim to retrieve the underlying normal distribution of the parameter. When the underlying previous studies have different means, the fitted normal distribution will

[1] Note that just as in multiplication in general, the order of updating is irrelevant for the final outcome.

have a variance larger than that of the underlying studies. Once the hyperparameters of the fitted normal distribution are obtained, the normal prior distribution can be specified in any software package that allows for Bayesian estimation with custom prior distributions.

Conducting sensitivity analyses with different priors, including diffuse default priors, allows us to compare findings and highlight the robustness of our model results if priors are modified (van Erp et al., 2018). The current study will compare the results of these three pooling methods and diffuse default priors on the posterior distributions in an empirical example that examined whether mother-adolescent interaction behavior meditates the associations between maternal and adolescent internalizing behavior.

Empirical Application: Mother-Adolescent Interaction Behavior as Mediator in the Transmission of Internalizing Symptoms

Adolescence is a crucial period for the development of internalizing problems, such as symptoms of anxiety or depression, which increase adolescents' risk for psychopathological disorders, school dropout, and unemployment in later life (Kessler et al., 2012; Clayborne et al., 2019). Maternal internalizing symptoms are among the most salient predictors of adolescent internalizing symptoms (e.g., Goodman and Gotlib, 1999; Connell and Goodman, 2002). Genetic similarities cannot fully explain these associations (Natsuaki et al., 2014; Eley et al., 2015) and specific patterns of how mothers and adolescents interact may be another mechanism through which maternal internalizing symptoms are associated with adolescent internalizing symptoms (Goodman and Gotlib, 1999). Specifically, internalizing symptoms might render mothers less sensitive to their children's needs, more emotionally unavailable, and more irritated, which can suppress mothers' expression of positive interaction behavior and increase their expression of negative, hostile and angry interaction behavior toward the adolescent (Lovejoy et al., 2000). Such diminished positive and heightened negative interaction behavior may in turn undermine the adolescents' self-esteem and emotion-regulation, make them feel helpless, and prompt negative self-evaluations, which render them more sensitive to internalizing symptoms (Gottman et al., 1997; Garber and Flynn, 2001). Hence, it is likely that maternal interaction behavior underlies the transmission of internalizing symptoms from mothers to adolescents.

Transactional theories (e.g., Sameroff, 2009) suggest that adolescents are not only influenced by their parents, but also influence their parents. Hence, associations between maternal and adolescent internalizing symptoms are likely to be bidirectional (Hughes and Gullone, 2010; Wilkinson et al., 2013). Adolescent internalizing symptoms can disrupt interactional processes in the family (Sheeber et al., 2001; Berg-Nielsen et al., 2002) and thus, similarly, predict changes in mother-adolescent interaction behavior (e.g., Nelemans et al., 2014), which in turn prompt maternal internalizing symptoms. It is thus important to include bidirectional associations between maternal and adolescent internalizing symptoms when investigating the mediating role of mother-adolescent interaction behavior. Similarly, as social interactions include two partners who continuously regulate and react to each other's behaviors (Fogel, 1993), it is essential to examine not only maternal interaction behavior toward adolescents, but also adolescents' interaction behavior toward mothers. However, most studies to date are based on the assumption that associations between maternal and adolescent internalizing symptoms are unidirectional from mothers to adolescents and only driven by maternal interaction behavior toward adolescents. If potential effects from adolescents to mothers are ignored, alleged mediation effects may be spurious. Fully understanding the mediating role of mother-adolescent positive and negative interaction behavior in the transmission of internalizing symptoms thus requires a model that reflects reciprocal associations between mothers and adolescents. In this study, we will investigate whether mother-adolescent interaction behavior underlies the intergenerational transmission of internalizing symptoms, including associations from both mothers to adolescents and from adolescents to mothers.

Several studies have been conducted to support each pathway in the theoretically proposed mediation model (see **Supplementary Material** for a systematic and critical review of previous literature). Findings from meta-analyses on mother-child interactions indeed indicated associations of maternal interaction behavior with maternal internalizing symptoms (Lovejoy et al., 2000; McCabe, 2014) and child internalizing symptoms (McLeod et al., 2007a,b; Yap et al., 2014; Pinquart, 2017). Observational, longitudinal assessments in adolescence best reflect our study's design and thus provide strong specific information. The few studies that meet these criteria, however, remain inconsistent regarding whether maternal internalizing predict both subsequent positive (Simons et al., 1993; Feng et al., 2007) and negative interaction behavior (Feng et al., 2007) as well as whether interaction behavior predicts subsequent adolescent internalizing symptoms (Hofer et al., 2013; Milan and Carlone, 2018) or not (Feinberg et al., 2007; Schwartz et al., 2012). Studies on reversed associations from adolescents to mothers remain scarce and the one available study found that adolescent interaction behavior did not predict maternal internalizing symptoms (Milan and Carlone, 2018).

The Present Study

This study applied a systematic approach to defining informative priors in Bayesian estimation to highlight the role of Bayesian estimation in integrating and cumulating empirical knowledge. We compared the effects of three different kinds of informative priors on the posterior distribution using an empirical illustration: Specifically, we examined whether observed mother-adolescent positive and negative interaction behavior mediate associations between maternal and adolescent internalizing symptoms, using a multi-method longitudinal design (see **Figure 1**). To increase the precision of our results, we systematically searched and weighed findings from previous studies, using an expert-designed weighting and scoring system,

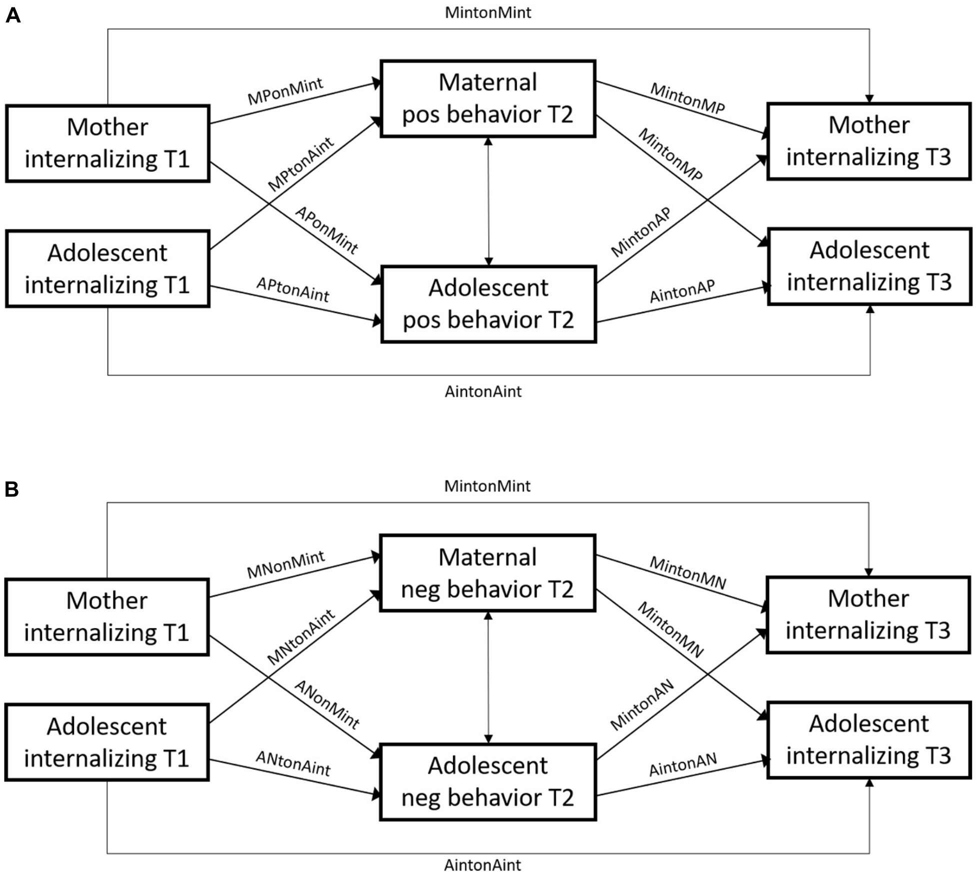


FIGURE 1 | Conceptual SEM models examining the mediating effects of positive (model **A**) and negative interaction behavior (model **B**) in the associations between maternal and adolescent internalizing symptoms. M, maternal; A, adolescent; pos, positive; neg, negative.

and synthesized the information into linear pool, logarithmic pool, and fitted normal prior distributions. Furthermore, we conducted sensitivity analyses to compare the impact of informative and diffuse priors on the mediating effects of mother-adolescent interaction behavior in the transmission of internalizing symptoms. This allowed us to identify the role of different priors and the robustness of our results.

MATERIALS AND METHODS

All relevant materials, documents, and syntax files are available at https://osf.io/c37mv.

Participants

The sample consisted of 102 mother-adolescent dyads (39% girls, M_{age} T_1 = 13.0, SD_{age} = 0.51) who were part of a larger sample of families participating in the ongoing Research on Adolescent Development And Relationships Young (RADAR-Y) study. All participants were assessed in annual home visits. Most adolescents (95%) and their mothers (91%) were of Dutch origin. They predominantly lived with both biological parents (86%) in medium to high socioeconomic status households (91%), based on parents' occupation level.

Sample attrition was low across all time points (1–7%), with 94 mother-adolescent dyads who participated at the first time point remaining in the study at the third time point. Mothers and adolescents who dropped out of the study did not significantly differ from those who remained in the study on most of the study or background measures (ANOVA p-values $\geq$ 0.056). However, mothers who dropped out of the study showed more negative interaction behavior at the second time point, $F(1,87) = 4.67$, $p = 0.033$, than mothers who remained in the study.

Procedure

The present study used three time points from early to mid-adolescence, when adolescents were on average approximately 13, 14, and 15 years of age. Families were recruited through 230 randomly selected elementary schools in the central and western regions of Netherlands. Of those initially selected (N = 1,544), families who did not fulfil the full family requirements (n = 364), could not be contacted or withdrew their participation (n = 569), or did not provide written consent for all family members (n = 114) were excluded. Of those 497 families who participated at the first time point, a subsample of 102 randomly selected mother-adolescent dyads participated in an interaction task.

During annual home assessments, adolescents and their mothers completed a series of questionnaires and subsequently participated in a conflict interaction task. The conflict task consisted of a 10-min videotaped interaction between adolescents and their mothers, during which they discussed a topic of frequent disagreement, explained their individual thoughts, and presented a solution to the conflict. Prior to the task, adolescents and their mothers agreed upon a topic, chosen out of a series of suggested subjects or an own subject. The interviewer ensured that a topic was chosen, but was otherwise absent during the topic selection and the actual conflict task. Adolescents and mothers were compensated for their participation at each time point. The study procedure was approved by the Medical Research Ethics Committee of the University Medical Center Utrecht.

Measures

Adolescent Internalizing Symptoms

We assessed adolescent internalizing symptoms as a combined score of self-reported anxiety and depression symptoms. Anxiety symptoms were measured with the Screen for Child Anxiety Related Emotional Disorders (SCARED; Birmaher et al., 1997), which consists of 38 items (e.g., "I get really frightened for no reason at all") on a 3-point scale (1 = almost never, 3 = often). Depression symptoms were measured with 2nd edition of the Reynolds Adolescent Depression Scale (RADS-2; Reynolds, 2000), which consists of 23 items (e.g., "I feel that no one cares about me") on a 4-point scale (1 = almost never, 4 = often). As anxiety and depression symptoms correspond to the same higher-order latent factor of internalizing symptoms within a hierarchical structure of psychopathology (Achenbach, 1966; Lahey et al., 2017), total anxiety and depression scores were averaged after a multiple imputation procedure to form a total internalizing symptom score for each participant. The anxiety, depression, and total internalizing scales showed high internal consistency across all time points (α = 0.91–0.96). Higher scores indicated higher levels of adolescent internalizing symptoms.

Maternal Internalizing Symptoms

We assessed maternal internalizing symptoms with the anxious/depressed, withdrawn, and somatic complaints syndrome scales of the Adult Self Report (ASR; Achenbach and Rescorla, 2003). The syndrome scales consist of 18 items (e.g., "I feel lonely"), 9 items (e.g., "I keep from getting involved with others"), and 12 items (e.g., "I feel tired without good reason), respectively, that are measured on a 3-point scale (0 = not true, 2 = very true or often true). The total internalizing scale showed high internal consistency across all time points (α = 0.90–0.91). Higher scores indicated higher levels of maternal internalizing symptoms.

Maternal and Adolescent Interaction Behavior

Rating scales were adapted from the *Family Interaction Task* coding system (Weinfield et al., 1999, 2002). We observed maternal and adolescent positive interaction behavior toward the other by coding verbal and nonverbal expressions/displays of maternal emotional involvement during the conflict task. Verbal expressions include showing interest, listening, responding, and understanding. Nonverbal expressions included smiling, interested attitude, nodding, maintained eye contact. We observed maternal and adolescent negative interaction behavior toward the interaction partner by coding how hostile and angry the mother or adolescent behaved during the conflict task. Maternal negative behaviors included blaming, rejecting, mocking, and exerting negative facial expressions or physical reactions. Adolescent negative behaviors included sighing and groaning, pouting, refusing to cooperate, criticizing, and exerting negative facial expressions or physical reactions.

Three independent raters coded maternal and adolescent interaction behavior toward the other on a 5-point scale (1 = low score on the relevant interaction behavior, 5 = high score on the relevant interaction behavior). All raters underwent extensive training before coding a random selection of the sample. Higher scores of positive interaction behavior indicated more common, appropriate, and consistent use of these verbal and nonverbal expressions, while higher scores of negative interaction behavior indicated higher levels of negative, hostile behaviors. Interrater agreements using intraclass correlations (ICC) based on 15% of the sample showed acceptable agreement for maternal interaction behavior (ICCs = 0.80–0.89) and adolescent interaction behavior (ICCs = 0.86–0.87).

Prior Distributions From Previous Knowledge

For the regression paths in our models, we implemented two search strategies (see **Figure 2** for a flowchart on study inclusion): a search for meta-analyses and reviews, and a search for empirical studies.

Meta-Analyses and Systematic Reviews

We conducted a literature search in Web of Science for all meta-analyses and systematic reviews published until December 2019, based on a combination of key words that reflected the target sample (child, adolescent) and their parents (parent*, maternal, and mother), internalizing symptoms (anxi*, depress*, and internalizing), as well as positive and negative behaviors (positive, negative, affect, warmth, hostil*, and rejection) during the interaction (interaction*, relation*, and parenting). Meta-analyses were selected if they (a) included studies on adolescence, and (b) assessed positive and/or negative interaction behavior, as defined for our sample, from mother or parent toward adolescent and/or from adolescent toward mother or parent. This search strategy identified 388 studies, of which 7 meta-analyses and 1 systematic review were included in this study. Some meta-analyses showed substantial overlap in studies. In these cases, we only included the meta-analysis that scored highest on the scoring scheme (i.e., most comparable to our design) to avoid biasing the results. This led to a final inclusion of 4 meta-analyses, of which 2 focused on the associations between maternal internalizing symptoms and mother-adolescent interaction behavior and 2 focused on the associations between mother-adolescent interaction behavior and adolescent internalizing symptoms. The systematic review

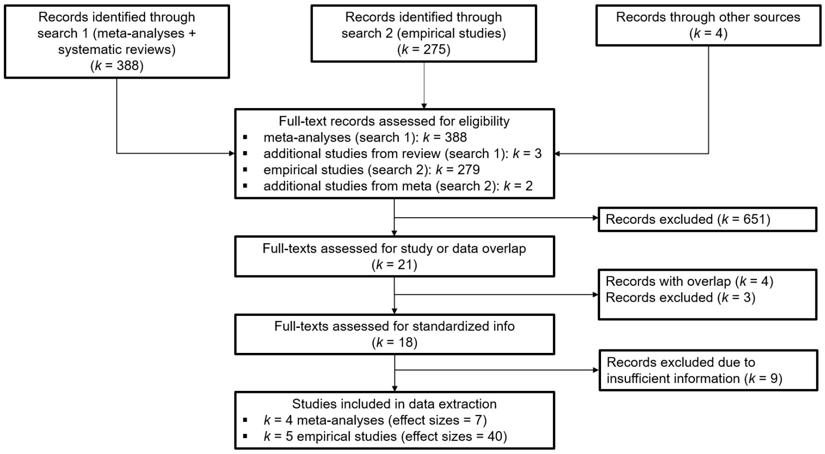

Records identified through search 1 (meta-analyses + systematic reviews) (k = 388)
Records identified through search 2 (empirical studies) (k = 275)
Records through other sources (k = 4)
Full-text records assessed for eligibility
meta-analyses (search 1): k = 388
additional studies from review (search 1): k = 3
empirical studies (search 2): k = 279
additional studies from meta (search 2): k = 2
Records excluded (k = 651)
Full-texts assessed for study or data overlap (k = 21)
Records with overlap (k = 4)
Records excluded (k = 3)
Full-texts assessed for standardized info (k = 18)
Records excluded due to insufficient information (k = 9)
Studies included in data extraction
k = 4 meta-analyses (effect sizes = 7)
k = 5 empirical studies (effect sizes = 40)

FIGURE 2 | Flow chart for study inclusion from search 1 (meta-analyses and systematic reviews) and search 2 (empirical studies) based on the PRISMA guidelines.

that was included identified 3 additional empirical studies that were not included in the meta-analyses and mainly focused on associations that were not investigated in any meta-analysis (e.g., associations between adolescent internalizing symptoms and adolescent interaction behavior). One of these empirical studies did not provide standardized information and was thus excluded.

Empirical Studies

Our second search strategy to identify relevant studies was twofold: First, we conducted a literature search in Web of Science for all empirical studies that were not included in the meta-analyses published from January 2012[2] until March 2020 using the same search string as for the meta-analyses, but only for adolescent samples (adolescen*, youth, teen*, youngst*, student*, emerging adult*, early adult*, and young adult*) and observational studies (observ*, code*, rater, tape*, task*, and record*). Studies were selected if they (a) included an adolescent sample, but did not include participants younger than 7 years or older than 25 years at the first measurement, (b) included longitudinal estimates for the cross-lagged parameters, and (c) assessed positive and/or negative interaction behavior from mother toward adolescent and/or from adolescent toward mother using observations. This search identified 275 studies, of which 11 were included (see **Figure 2**). Second, we searched all cross-sectional meta-analyses for studies that met the inclusion criteria and had estimates that were not included in the meta-analytic effect sizes. This resulted in an additional inclusion of 2 studies. Studies that failed to provide any or only partial standardized information were excluded (k = 8). The final inclusion yielded 47 effect sizes from 4 meta-analyses and 5 independent empirical studies (see **Supplementary Table 1** for all included studies per parameter and model).

Power Prior Weighting Scheme

To evaluate each previous study's contribution to our research question, we designed a scoring system that reflects each study's weight in the specification of prior distributions. Four experts on adolescent relationships and mental health (third, fifth, sixth, and seventh author) discussed and evaluated the importance of several methodological aspects, which were further quantified to represent one score (see **Table 1A**). For example, a longitudinal measurement most closely reflected our study design, and therefore received a higher score than a cross-sectional measurement. The final weighting scheme included ten categories: longitudinal associations, same time lag, controlling for earlier internalizing symptoms, mother-adolescent interaction behavior assessed solely observational, age range from early to mid-adolescence (12–16), included symptoms of depression and anxiety, or anxiety only, controlling for other partner's symptoms, controlling for other partner's interaction behavior, community sample, and meta-analysis. The ten categories were associated with 5–20 points depending on the importance of the criterion. Each included study received the allocated number of points per category depending on whether or not they fulfilled the criteria (see **Table 1B**). The final score for each study determined its associated weight, δ, in the power prior.

Specification of Prior Distributions

To be able to use previous information from studies with various measures, the data of the present study was standardized, and all prior distributions concerned standardized effects. Hence, only information from previous studies that presented or allowed to compute standardized effects *and* the associated standard errors was used[3]. The hyperparameters for the normally distributed prior distributions were a mean and standard deviation.

The longitudinal associations of maternal and adolescent internalizing symptoms with mother-adolescent positive (i.e., model A) and negative interaction behavior (i.e., model B)

[2] As starting year, we chose the date of the last updated search of the meta-analyses.

[3] If only the standard error of the unstandardized effect was present, we multiplied that standard error with the standard deviation of the independent variable and divided by the standard deviation of the dependent variable. If a t-statistic was present, the standard error was computed by dividing the standardized effect by t. If a confidence interval for the standardized effect was provided, the difference between upper and lower limit was divided by 2 and by 1.96.

TABLE 1A | Weighting scheme for informative priors.

Category	Points	Details
T1-T2 (longitudinal)	10	The estimates of longitudinal studies are usually smaller than those of cross-sectional studies. As our parameter are longitudinal estimates as well, longitudinal designs should receive most weight in relation other categories.
- controlling for symptoms at T1	20	Longitudinal studies that do not control for symptoms at T1 might have quite large estimates and cannot indicate change. As this is the most crucial aspect of longitudinal research, studies that also control for T1 symptoms should receive more weight. *Not applicable for T1 → T2 associations (deleted from final score)!*
- Same time lag *- (1 year)*	5	Studies that use the same time lag as we do are closer to our study design and thus deserve more weight.
Observation	15	The study list only includes empirical studies with observational assessments of the parent-adolescent interaction as these (multi-method) estimates are usually smaller than self-reports. However, meta-analyses often include a combination of observations and self-reports, which is difficult to disentangle. Therefore, estimates from "pure" observations should receive more weight than mixed studies (and most weight in relation to other categories as this is another main aspect of our study).
Early adolescence (12–16)	10	Some studies, and particularly the meta-analyses, used a broader age range than our study or even just adolescence (but all studies include adolescence). As our study focuses on early-mid adolescence, studies that included a similar age group should receive some more weight.
Internalizing symptoms include both anxiety and depression, or anxiety only	10	Most studies do not focus on a combination of depression and anxiety symptoms, but only include one of those symptoms (mostly depression). As we will use a combination of both, studies that include measures on internalizing symptoms or both depression and anxiety symptoms should receive more weight. *Most studies focus on mother or adolescent depression (rather than anxiety). To counterbalance that, we will also award 5 points if the study only focused on anxiety (i.e., either combined or anxiety only).*
Including covariates		If studies include other relevant covariates that might better reflect our study associations, such as parental symptoms (for T2-T3 parameters), they might receive additional weight.
- parental symptoms	5	
- other interaction behaviors	5	
Community sample (does not include clinical/diagnostic groups)	10	Many (older) studies include two subsamples, of which one is usually clinical. Therefore, the final sample includes participants who may have higher levels of internalizing symptoms than our participants. For these participants, the associations may be stronger. Thus, studies with a community sample which is closer to our sample should receive more weight.
Meta-analysis	10	Meta-analyses combine information from several studies and thus provide the most comprehensive evidence. Therefore they should receive somewhat more weight than individual studies.
10 categories (standard 5)	**100 (80)**	**Each study can score between 0 and 100 points (or between 0 and 80 points for T1 → T2 associations).**

describe the main parameters in the model (see **Figure 1**). We did not consider the datapoints from previous studies to be exchangeable with our current dataset, nor to be a previous sample from the same population (Spiegelhalter et al., 2004). The previous information was thus considered less relevant than the current data, and therefore, needed to be downweighed by power priors. The power prior weights δ were systematically determined through our weighting scheme (section "Power Prior Weighting Scheme"). Studies with lower relevance obtained lower scores for δ, which means that their variance was inflated. The larger variance (i.e., uncertainty) diminishes its impact on the posterior distribution.

When multiple studies contained information on one parameter, the information needed to be aggregated into one distribution. We evaluated three methods to aggregate previous information: (1) linear pooling, (2) logarithmic pooling, and (3) a normal distribution fitted to the linear pool. Additionally, we conducted sensitivity analyses with default priors from the statistical R package brms as a reference (Bürkner, 2017). The four posterior distributions were compared and evaluated based on estimation issues and interpretability to indicate the role of previous information. The defined informative priors for all longitudinal regression parameters are provided in **Table 2**. For all other parameters in the model, the following low-informative prior was used: $N(0,10)$.

The linear and logarithmic pool both used the study's normal prior distributions with σ/δ as input for the standard deviation. Subsequently, each of the distributions received an equal weight in the pooling procedure. The normal pool was programmed in Stan (see syntax in the **Supplementary Material**). The hyperparameters for the logarithmic pool of normal distributions were calculated according to de Carvalho et al. (2020). To obtain a normal distribution fitting to the linear pool, we first drew 5,000 random samples from each of the weighted normal prior distributions for one parameter. Subsequently, we fitted a normal distribution to these samples (i.e., fitted normal) by means of the fitdist function of the R-package fitdistrplus (Delignette-Muller and Dutang, 2015) using maximum likelihood estimation.

TABLE 1B | Final scoring of all included studies.

Study	T1-T2	lag	cT1	obs	Age	$M_{dep+anx\ (or\ anx)}$	$A_{dep+anx\ (or\ anx)}$	cov_s	cov_i	comm	MA	Score
Points	**10**	**5**	**20**	**15**	**10**		**10**	**5**	**5**	**10**	**10**	**100**
Lovejoy et al. (2000)				x							x	25
Simons et al. (1993)*	x				x							20
McCabe (2014)				x		x					x	35
Pinquart (2017)	x		x				x			x	x	60
Weymouth et al. (2016)							x			x	x	30
Allen et al. (2006)	x	x	x	x	x					x		70
Asbrand et al. (2017)	x			x			x					35
Dadds et al. (1992)				x								15
Dietz et al. (2008)				x		x						25
Griffith et al. (2019), (neg)	x		x	x				x		x		60
Griffith et al. (2019), (pos)	x		x	x						x		55
Hofer et al. (2013)	x		x	x	x		x		x	x		80
Jackson et al. (2011)				x								15
Milan and Carlone (2018), (only cs)				x				x	x	x		30
Milan and Carlone (2018)	x		x	x				x	x	x		60
Nelson et al. (2017)	x			x					x			30
Olino et al. (2016)	x		x	x				x				50
Schwartz et al. (2012)	x		x	x	x		x	x				70
Szwedo et al. (2017)	x		x	x						x		55
van Doorn et al. (2016)				x				x	x			25

Note. T1-T2 = longitudinal assessment, lag = same time lag used (for longitudinal studies), cT1, controlling for T1 symptoms (for longitudinal studies); obs, observational assessment of parent-adolescent interaction; age, age range early adolescence; N, sample size; M, maternal; A, adolescent; year, publication year; cov_s, controlling for parental symptoms; cov_i, controlling for other interaction behaviors; comm, community sample; MA, meta-analysis; x, indicates that the category is met, gray studies were excluded from the final analyses due to insufficient standardized information.
**Study included in aforementioned meta-analysis.*

The estimated mean and standard deviation associated with the best fit were used as hyperparameters for the priors in Stan.

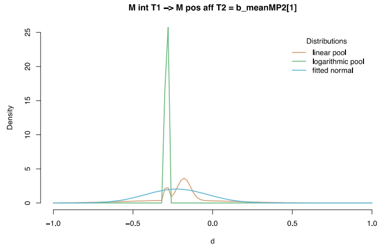
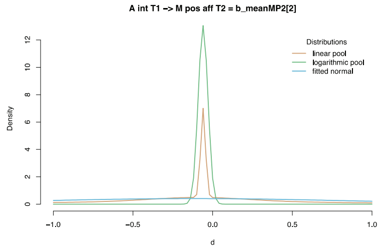
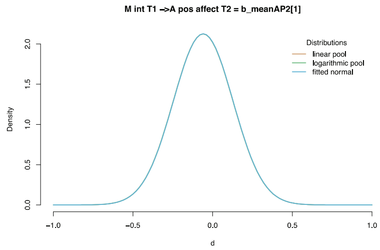
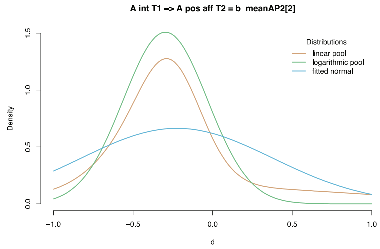
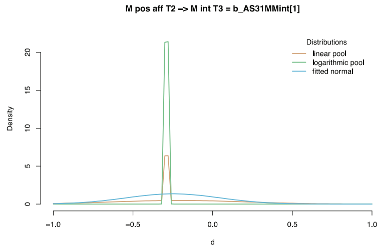
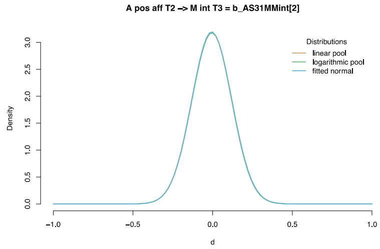
Statistical Analyses

Missing data was modest and ranged from 2-13% for most variables. Only at T1, 54% of the RADS-2, which is one of the two scales for internalizing problems, was missing because not all subscales of this questionnaire were administered to all participants. Based on Little's missing completely at random (MCAR) test that detected no systematic patterns of missingness, normed χ^2/df = 1.19, we inferred that missing data was not likely to bias our analyses. To handle the missing data, multiple imputation was conducted by means of the R-package mice (Van Buuren and Groothuis-Oudshoorn, 2011). All variables that had a correlation >0.10 with the variables to be imputed were included as predictors in the imputation model, except for the identification variable. As indicated by the imputation plots and absence of logged events, the 20 imputations were successful. The fraction of missing information (fmi) in all regression paths ranged from 0.07 to 0.38.

To evaluate the impact of the different prior distributions, we assessed convergence, conducted prior predictive checks, estimated the posterior distributions, and calculated posterior shrinkage. Convergence was assessed in randomly selected posteriors based on three imputed datasets to avoid false positives (Bürkner, 2020), using the potential scale reduction (PSR; Gelman and Rubin, 1992) and effective sample size (ESS). The PSR (or $\hat{R}$) compares the variance between and within chains. A PSR value near 1.0 indicates convergence. Originally, 1.05 was taken as an upper bound for convergence or even 1.10 with many model parameters, but more recently, smaller values like 1.01 and 1.001 have been recommended (e.g., Vehtari et al., 2019; Zitzmann and Hecht, 2019). The ESS quantifies the number of effectively independent draws from the posterior distribution, and is a measure of precision as it indicates how well an estimate is approximated. An ESS larger than 400 is recommended to get a stable estimate (e.g., Vehtari et al., 2019; Zitzmann and Hecht, 2019).

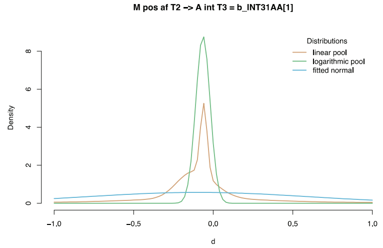
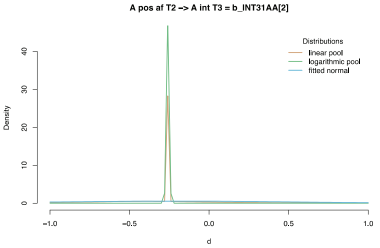
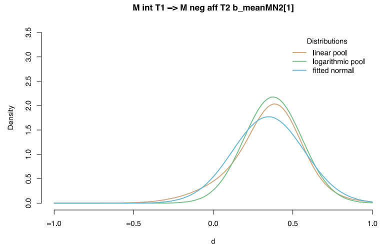
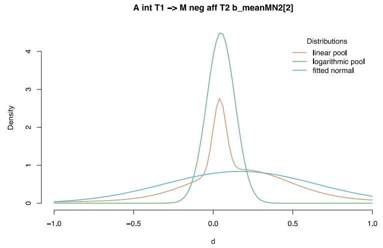
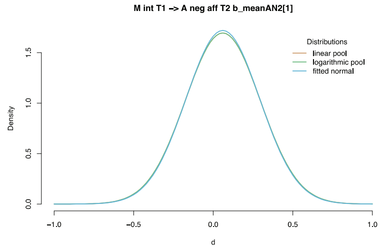
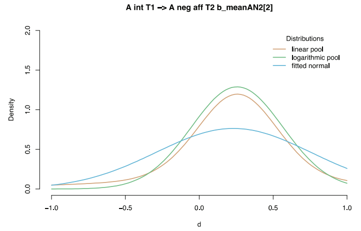
In a prior predictive check, samples are taken from the prior distribution to simulate new data based on the sampled parameter estimates. Together, the simulated datasets form the predictive distribution. The predictive distribution encompasses the data that can be expected given the multivariate prior distribution on the parameters. With a predictive distribution, the analyst can evaluate whether the (multivariate) prior relates to sensible data. Furthermore, the current observed data can be compared to the predictive distribution. In the present study, prior predictive distributions were evaluated for each of the four dependent variables and each of the four prior specifications in both models.

TABLE 2 | Informative priors for the regression parameters in Model A and Model B.

Parameter description and names	Linear pool	Logarithmic pool	Fitted normal	Image
Maternal internalizing symptoms T1 → Maternal positive interaction T2 *MPonMint* b_meanMP2[1]	$N(-0.18, 0.0179)^{0.4375} +$ $N(-0.21, 0.1040)^{0.3125} +$ $N(-0.29, 0.0015)^{0.3750}$	$N\,(-0.29, 0.01)$	$N\,(-0.23, 0.20)$	
Adolescent internalizing symptoms T1 → Maternal positive interaction T2 *MPonAint* b_meanMP2[2]	$N(-0.06, 0.0077)^{0.5000} +$ $N(-0.09, 0.0950)^{0.3125} +$ $N(-0.12, 0.1755)^{0.1875} +$ $N(-0.16, 0.6407)^{0.3750}$	$N\,(-0.06, 0.03)$	$N\,(-0.10, 0.98)$	
Maternal internalizing symptoms T1 → Adolescent positive interaction T2 *APonMint* b_meanAP2[1]	$N(-0.06, 0.0704)^{0.3750}$	$N\,(-0.06, 0.19)$	$N\,(-0.06, 0.19)$	
Adolescent internalizing symptoms T1 → Adolescent positive interaction T2 *APonAint* b_meanAP2[2]	$N(-0.01, 0.1768)^{0.1875} +$ $N(-0.41, 0.0871)^{0.3125} +$ $N(-0.26, 0.0697)^{0.3750}$	$N\,(-0.30, 0.26)$	$N\,(-0.22, 0.61)$	
Maternal positive interaction T2 → Maternal internalizing symptoms T3 *MintonMP* b_AS31MMInt[1]	$N(-0.21, 0.1040)^{0.2500} +$ $N(-0.29, 0.0015)^{0.3000}$	$N\,(-0.29, 0.01)$	$N\,(-0.25, 0.29)$	
Adolescent positive interaction T2 → Maternal internalizing symptoms T3 *MintonMP* b_AS31MMInt[2]	$N(-0.01, 0.0753)^{0.6000}$	$N\,(-0.01, 0.13)$	$N\,(-0.01, 0.13)$	

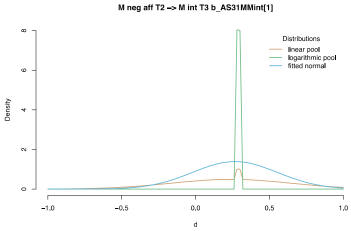
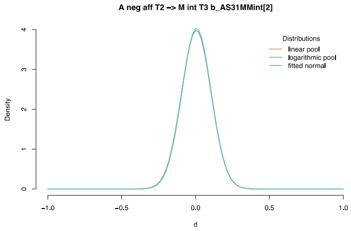
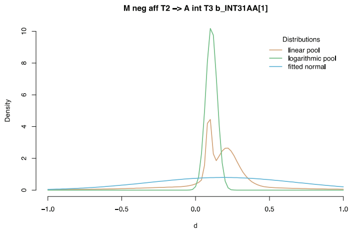
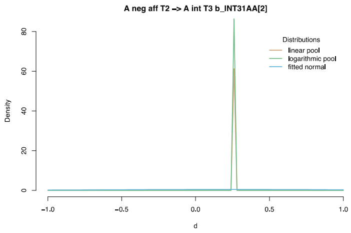
(Continued)

TABLE 2 | Continued

Parameter description and names	Linear pool	Logarithmic pool	Fitted normal	Image
Maternal positive interaction T2 → Adolescent internalizing symptoms T3 *AintonMP* b_INT31AA[1]	$N(-0.06, 0.0128)^{0.6000}+$ $N(-0.16, 0.0546)^{0.6000}+$ $N(-0.09, 0.0950)^{0.2500}+$ $N(-0.12, 0.2219)^{0.1500}+$ $N(-0.05, 0.0578)^{0.5500}$	$N\,(-0.06, 0.05)$	$N\,(-0.10, 0.68)$	
Adolescent positive interaction T2 → Adolescent internalizing symptoms T3 *AintonAP* b_INT31AA[2]	$N(-0.01, 0.1768)^{0.1500}+$ $N(-0.41, 0.0871)^{0.2500}+$ $N(-0.26, 0.0014)^{0.3000}$	$N\,(-0.26, 0.01)$	$N\,(-0.21, 0.73)$	
Maternal internalizing symptoms T1 → Maternal negative interaction T2 *MNonMint* b_meanMN2[1]	$N(0.40, 0.0459)^{0.3125}+$ $N(0.29, 0.1030)^{0.3750}$	$N\,(0.38, 0.18)$	$N\,(0.34, 0.22)$	
Adolescent internalizing symptoms T1 → Maternal negative interaction T2 *MNonAint* b_meanMN2[2]	$N(0.04, 0.0204)^{0.5000}+$ $N(0.10, 0.0948)^{0.3125}+$ $N(0.27, 0.1699)^{0.1875}+$ $N(0.16, 0.1020)^{0.3750}+$ $N(0.26, 0.1338)^{0.4375}$	$N\,(0.05, 0.09)$	$N\,(0.17, 0.47)$	
Maternal internalizing symptoms T1 → Adolescent negative interaction T2 *ANonMint* b_meanAN2[1]	$N(0.06, 0.09)^{0.3750}$	$N\,(0.06, 0.24)$	$N\,(0.06, 0.24)$	
Adolescent internalizing symptoms T1 → Adolescent negative interaction T2 *ANonAint* b_meanAN2[2]	$N(0.17, 0.1743)^{0.1875}+$ $N(0.28, 0.0916)^{0.3125}+$ $N(0.26, 0.0875)^{0.3750}+$ $N(0.23, 0.1348)^{0.4375}$	$N\,(0.26, 0.31)$	$N\,(0.23, 0.53)$	

(Continued)

TABLE 2 | Continued

Parameter description and names	Linear pool	Logarithmic pool	Fitted normal	Image
Maternal negative interaction T2 → Maternal internalizing symptoms T3 *MintonMN* b_AS31MMInt[1]	$N(0.24, 0.1017)^{0.2500}$ + $N(0.29, 0.0010)^{0.3000}$	N (0.29, 0.0046)	N (0.27, 0.29)	
Adolescent negative interaction T2 → Maternal internalizing symptoms T3 *MintonAN* b_AS31MMInt[2]	$N(0.01, 0.0601)^{0.6000}$	N (0.01, 0.10)	N (0.01, 0.10)	
Maternal negative interaction T2 → Adolescent internalizing symptoms T3 *AintonMN* b_INT31AA[1]	$N(0.09, 0.0102)^{0.6000}$ + $N(0.10, 0.0948)^{0.2500}$ + $N(0.27, 0.1699)^{0.1500}$ + $N(0.21, 0.0343)^{0.6000}$ + $N(0.26, 0.0260)^{0.3000}$ + $N(0.15, 0.0537)^{0.6000}$	N (0.11, 0.04)	N (0.19, 0.50)	
Adolescent negative interaction T2 → Adolescent internalizing symptoms T3 *AintonAN* b_INT31AA[2]	$N(0.17, 0.1743)^{0.1500}$ + $N(0.26, 0.0010)^{0.3000}$	N (0.25, 0.0046)	N (0.20, 0.82)	

M, maternal; A, adolescent; int, internalizing; P, positive interaction behavior; N, negative interaction behavior;on, describes the direction of regression (e.g., MPonMint indicates the association from maternal internalizing symptoms at T1 to maternal positive interaction behavior at T2). The hyperparameters of the normal distributions are a mean and a standard deviation.

Posterior shrinkage (or contraction) s describes the degree of reduction in uncertainty from the prior to the posterior distribution of a parameter:

$$s = 1 - \frac{\sigma^2_{\text{posterior}}}{\sigma^2_{\text{prior}}},$$

where $\sigma^2_{\text{posterior}}$ is the variance of the posterior distribution and σ^2_{prior} is the variance of the prior distribution. The inclusion of the likelihood of the data in the posterior tends to decrease the prior uncertainty, resulting in shrinkage. If the data is highly informative compared to the prior, the posterior shrinkage will be close to 1. If the data provides little additional information, the posterior shrinkage will be close to 0.

All Bayesian analyses were conducted in Stan by means of the rstan (Stan Development Team, 2020) and brms (Bürkner, 2017) R-packages in R 4.0 (R Core Team, 2020). We conducted our analyses with 3 chains, each running 8,000 iterations of which the first 3,000 were discarded. The software analyzed each of the 20 imputed datasets separately. Afterward, the separate posterior distributions were taken together to aggregate the results (Gelman et al., 2004, p. 520; Zhou and Reiter, 2010)[4]. We constructed two structural equation models (SEMs) to examine whether maternal and adolescent positive (see **Figure 1A**) and negative interaction behavior (see **Figure 1B**) mediated the association between maternal and adolescent internalizing symptoms across time. All models included 2-year autoregressive paths for adolescent and maternal internalizing

[4] brms includes a function that applies multiple imputation and the aggregation of results in one step, but we did not use it for reasons of comparability between methods.

symptoms. We further included correlations between maternal and adolescent interaction behavior. Finally, we calculated eight indirect effects to assess whether maternal and adolescent positive and negative interaction behavior mediated the associations from maternal to adolescent internalizing symptoms as well as from adolescent to maternal internalizing symptoms by multiplying the associations between internalizing symptoms and mother-adolescent interaction behavior from T1 to T2 and from T2 to T3.

RESULTS

Descriptive Statistics

Table 3 displays the means, standard deviations, and correlations among all study variables. Interaction behavior correlated moderately to strongly, both within mother and adolescent interaction behavior as well as between mother and adolescent interaction behavior. Maternal and adolescent interaction behavior correlated moderately with maternal and adolescent internalizing symptoms.

Convergence and Precision

PSR values were <1.01 and ESS >1,000 for all parameters in all viewed analyses for the analyses with logarithmic pooled priors, fitted normal priors, and default priors. However, the analyses with linear pooled priors also showed some insufficient results with respect to convergence and precision. In Model A, a PSR of 1.02 was observed for maternal internalizing symptoms at T1 predicting maternal positive interaction behavior at T2 (MPonMint), and a PSR of 1.04 for maternal positive interaction behavior at T2 predicting maternal internalizing symptoms at T3 (MintonMP). The ESS was <200 for maternal internalizing symptoms at T1 predicting maternal positive interaction behavior at T2 (MPonMint), maternal positive interaction behavior at T2 predicting maternal internalizing symptoms at T3 (MintonMP), and in some analyses also for adolescent positive interaction behavior at T2 predicting maternal internalizing symptoms at T3 (MintonAP). In model B, two PSR values >1.05 were observed: 1.07 for maternal negative interaction behavior at T2 predicting maternal internalizing symptoms at T3 (MintonMN), and 1.12 for adolescent negative interaction behavior at T2 predicting adolescent internalizing symptoms at T3 (AintonAN). The regression of adolescent internalizing problems on adolescent negative interaction behavior (AintonAN) was also repeatedly associated with a particularly low ESS (i.e., <50). For the purpose of this illustration, we will continue to evaluate the results as they are, without any further modifications to the estimation process.

Prior Predictive Check

We evaluated the predictive distributions of the four dependent model variables in both studies for each of the four methods (i.e., 32 predictive distributions). **Figure 3** displays a selection of four illustrative predictive distributions.

For each of the informative prior specifications, there was a considerable spread in predicted likelihoods and their associated means and standard deviations. The predicted means mostly ranged from −40 to +40, centered around the observed data mean of 0 (all variables were centered). The predictive distribution for the default brms priors, however, almost had an infinite range including many implausible predicted likelihoods. This behavior was expected, as default priors are not supposed to direct the estimation process, but it also demonstrates that default priors do not contain meaningful information.

Shrinkage

The posterior shrinkage for all parameters of interest and all prior specifications in both models can be found in **Table 4**. In all cases, the posterior shrinkage for the default brms priors was approaching 1.00, indicating that the data strongly diminished the posterior variance as compared to the prior variance. This finding was expected as default priors usually have an extremely wide variance to let the likelihood of the data predominate the posterior results. The logarithmic pool generally showed the lowest posterior shrinkage. In 9 out of 16 posterior parameter distributions, the posterior shrinkage for the logarithmic pooled prior was <0.20, and in 6 out of 16 it was even <0.05. In these cases, the logarithmic pooled prior greatly affected the posterior results. The shrinkage of the linear pooled prior and the fitted normal prior were relatively similar and varied between 0.27 and 0.90. It should be noted however, that the multimodality of the linear pooled prior and its associated posterior was not captured by our shrinkage measure that summarizes the distributions by their variances. Consequently, even though the shrinkage was larger than that of the fitted normal prior in 50% of the cases, we cannot interpret this outcome as if the likelihood had a larger impact on the posterior of the linear pooled prior than the fitted normal.

Indirect Pathways Through Maternal and Adolescent Interaction Behavior

Positive Interaction Behavior

The results for the positive interaction behavior model as analyzed with the three different prior settings are provided in **Table 5**. Based on the analysis with linear pooled priors, we found that only for the longitudinal associations from maternal and adolescent internalizing symptoms at T1 to maternal positive interaction behavior at T2 ($M_{maternal} = -0.24$, 95% HPD = [−0.30,−0.13], $M_{adolescent} = -0.15$, 95% HPD = [−0.35,−0.04]), the 95% highest posterior density (HPD) interval did not include 0 as probable value. The completely negative 95% HPD indicates that higher levels of maternal and adolescent internalizing symptoms predicted lower levels of subsequent maternal positive interaction behavior 1 year later. Although there was limited evidence that maternal and adolescent internalizing symptoms predicted adolescent positive interaction behavior as the 95% HPD included both negative and positive values, the values were mostly negative. This indicates that there was more probability toward such a negative effect, but still some probability that the effect was positive. For all other associations, negative as well as positive values were part of the 95% HPD. Hence, we are

TABLE 3 | Descriptives of all study variables.

Variable	*M*	*SD*	1	2	3	4	5	6	7
1 Adolescent internalizing T_1	−0.12	0.95							
2 Adolescent internalizing T_3	−0.02	0.88	0.601						
3 Maternal internalizing T_1	0.19	0.17	0.195	0.148					
4 Maternal internalizing T_3	0.19	0.18	0.105	0.195	0.669				
5 Maternal positive interaction T_2	3.50	0.79	−0.177	−0.293	−0.293	−0.366			
6 Maternal negative interaction T_2	1.48	0.72	−0.81	0.082	0.184	0.185	−0.574		
7 Adolescent positive interactionT_2	3.30	0.91	−0.185	−0.309	−0.232	−0.106	0.471	−0.260	
8 Adolescent negative interactionT_2	1.43	0.79	0.152	0.194	0.171	0.238	−0.279	0.223	−0.735

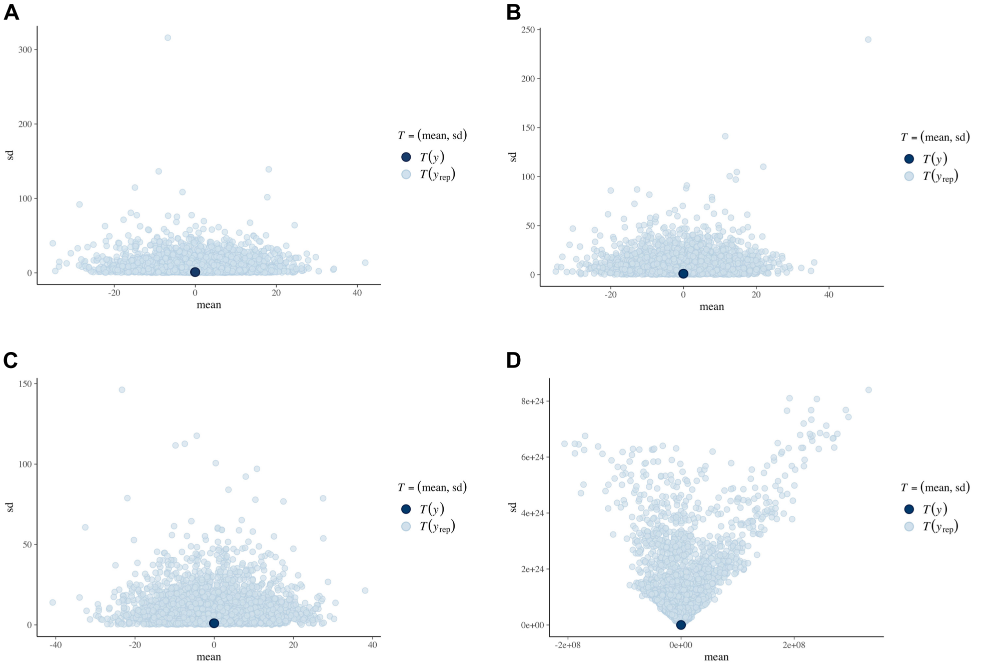


FIGURE 3 | Means (*x*-axis) and standard deviations (*y*-axis) in the prior predictive distribution for the four prior specifications. The dark-blue dots represent the means in the imputed observed datasets (centered at 0). **(A)** Linear pool. **(B)** Logarithmic pool. **(C)** Fitted normal distribution. **(D)** Default.

not certain if and how positive interaction behavior at T2 predicted maternal or adolescent internalizing symptoms at T3, 1 year later. Furthermore, the 95% HPD of the autoregressive paths from maternal and adolescent internalizing symptoms at T1 to their internalizing symptoms at T3 were completely positive ($M_{maternal}$ = 0.49, 95% HPD = [0.33,0.66], $M_{adolescent}$ = 0.44, 95% HPD = [0.27,0.61]), indicating that maternal and adolescent symptoms showed modest stability across time. All mediational paths included negative as well as positive values in their 95% HPD, indicating that there was no clear evidence on the existence and direction of the indirect effects from maternal to adolescent or adolescent to maternal internalizing symptoms through maternal or adolescent positive interaction behavior.

The analyses based on logarithmic pooled priors showed generally similar results. As for the analyses with the linear pooled priors, both higher levels of maternal and adolescent internalizing symptoms at T1 predicted lower levels of maternal positive interaction behavior at T2 ($M_{maternal}$ = −0.29, 95% HPD = [−0.30,−0.28], $M_{adolescent}$ = −0.07, 95% HPD = [−0.12,−0.02]). In contrast to the linear pooled priors, lower levels of maternal and adolescent positive interaction behavior at T2 predicted higher levels of their own, but not the other's internalizing symptoms at T3 ($M_{maternal}$ = −0.29, 95% HPD = [−0.30,−0.28]; $M_{adolescent}$ = −0.26, 95% HPD = [−0.27,−0.24]). For all other direct associations, the 95% HPD included both positive and negative values. Similar to the linear pooled priors, maternal and adolescent internalizing symptoms at T1 predicted their own respective symptoms at T2. Furthermore, maternal positive interaction behavior mediated the association between adolescent and maternal internalizing symptoms, as indicated by the 95% HPD of the indirect effect that was completely positive ($M_{indirect}$ = 0.02, 95% HPD = [0.01,0.04]). This suggests that higher levels of adolescent internalizing symptoms predicted higher levels of maternal internalizing symptoms 2 years later through decreased positive maternal interaction behavior. No other indirect effects were found.

Based on the analysis with normal distributions fitted to the linear pooled priors, we detected similar results as for the analysis using linear pooled priors. Comparable to the analyses with both linear and logarithmic pooled priors, maternal and adolescent internalizing symptoms at T1 predicted maternal positive interaction behavior at T2 ($M_{maternal} = -0.23$, 95% HPD = [−0.38,−0.07]; $M_{adolescent} = -0.20$, 95% HPD = [−0.38,−0.03]). However, we found no evidence for associations between maternal or adolescent interaction behavior and their subsequent internalizing symptoms, which is in line with the linear pooled priors, but only partially in line with the logarithmic pooled priors. As in the other analyses using linear and logarithmic pooled priors, maternal and adolescent internalizing symptoms at T1 predicted their own respective symptoms at T2. No indirect effects were found. Further sensitivity analyses with default priors, which relied on prespecified non-informative priors, yielded the same conclusions as the analysis with fitted normal priors.

Examining the posterior samples per parameter (see **Figure 4**) indicated that the linear pooled priors affected posterior samples for some parameters in such a manner that they became bimodal. For example, the posterior distribution of the association between maternal internalizing behavior and subsequent maternal positive interaction behavior (MPonMint) shows that there was some strong evidence from previous studies. This previous evidence supports an effect that is larger than what is found in the current data, as indicated by the shift in modes as compared to the analyses with default priors. On the other hand, for the association between adolescent internalizing symptoms and subsequent maternal positive interaction behavior (MPonAint), the posterior distribution still reflects some strong evidence from previous studies for an effect smaller than found in the data.

TABLE 4 | Shrinkage in model A and B.

	Linear pool	Logarithmic pool	Fitted normal	Default
MPonMint	0.67	−0.00	0.54	> 0.99
MPonAint	0.89	0.04	0.89	> 0.99
APonMint	0.51	0.50	0.51	> 0.99
APonAint	0.83	0.60	0.81	> 0.99
MintonMP	0.54	0.00	0.66	> 0.99
MintonAP	0.27	0.39	0.32	> 0.99
AintonMP	0.89	0.10	0.84	> 0.99
AintonAP	0.82	0.01	0.84	> 0.99
MNonMint	0.57	0.49	0.57	> 0.99
MNonAint	0.82	0.20	0.77	> 0.99
ANonMint	0.58	0.58	0.58	> 0.99
ANonAint	0.81	0.67	0.80	> 0.99
MintonMN	0.58	−0.00	0.69	> 0.99
MintonAN	0.28	0.30	0.30	> 0.99
AintonMN	0.86	0.07	0.81	> 0.99
AintonAN	0.87	−0.01	0.88	> 0.99

M, maternal; A, adolescent; int, internalizing; P, positive interaction behavior; N, negative interaction behavior;on, describes the direction of regression, indirect effects are reported in direction of the association (e.g., MintMPAint describes the indirect effect from maternal to adolescent internalizing symptoms via maternal positive interaction behavior).

Negative Interaction Behavior

The results for the negative interaction behavior model as analyzed with the three different prior settings are provided in **Table 6**. Based on the analysis with linear pooled priors, we found that higher levels of maternal, but not adolescent internalizing symptoms predicted higher levels of subsequent maternal, but not adolescent negative interaction behavior 1 year later (M = 0.23, 95% HPD = [0.06,0.39]). In turn, maternal negative interaction behavior at T2 predicted adolescent, but not maternal internalizing symptoms 1 year later at T3 (M = 0.11, 95% HPD = [0.01,0.23]). There was limited evidence that adolescent negative interaction behavior at T2 predicted their own or their mothers' internalizing symptoms at T3, 1 year later. Although for these associations the 95% HPD included both positive and negative values, the values were mostly positive. This indicates that there was more probability toward a positive effect, but still some probability that the effect was negative. Maternal negative interaction behavior mediated the association between maternal and adolescent internalizing symptoms, as indicated by the 95% HPD of the indirect effect that was completely positive ($M_{indirect}$ = 0.03, 95% HPD = [0,0.06]). This suggests that higher levels of maternal internalizing symptoms predicted higher levels of adolescent internalizing symptoms 2 years later through increased maternal negative interaction behavior. No other indirect effects were found.

The analyses with logarithmic pooled priors again demonstrated generally similar results. Higher levels of maternal, but not adolescent internalizing symptoms at T1 predicted higher levels of maternal negative interaction behavior at T2 (M = 0.23, 95% HPD = [0.08,0.39]). Maternal negative interaction behavior at T2 in turn predicted adolescent internalizing symptoms (M = 0.10, 95% HPD = [0.04,0.16]) and, in contrast to the linear pooled priors, also maternal internalizing symptoms at T3 (M = 0.29, 95% HPD = [0.28,0.30]). Higher levels of adolescent negative interaction behavior at T2 further predicted higher levels of subsequent adolescent internalizing symptoms at T3 as indicated by the completely positive 95% HPD (M = 0.26, 95% HPD = [0.25,0.27]), which contrasts with the analysis using linear pooled priors. For all other direct associations, the 95% HPD included both positive and negative values. Similar to the linear pooled priors, we detected evidence for an indirect effect from maternal to subsequent adolescent internalizing symptoms through increased maternal negative interaction behavior ($M_{indirect}$ = 0.02, 95% HPD = [0.00,0.05]).

Based on the analysis with fitted normal priors, we found slightly different results. While maternal internalizing symptoms at T1 also predicted maternal negative interaction behavior 1 year later at T2 (M = 0.22, 95% HPD = [0.06,0.38]), there was only limited evidence that maternal negative interaction behavior predicted subsequent adolescent internalizing

TABLE 5 | Parameter estimates using different prior settings for model A.

	Linear pool priors			Logarithmic pool priors			Normal fitted to linear pool priors			Default priors		
Parameter	**Mean**	**95% HPD**		**Mean**	**95% HPD**		**Mean**	**95% HPD**		**Mean**	**95% HPD**	
MPonMint	−0.24	−0.30	−0.13	−0.29	−0.30	−0.28	−0.23	−0.38	−0.07	−0.23	−0.40	−0.06
MPonAint	−0.15	−0.35	−0.04	−0.07	−0.12	−0.02	−0.20	−0.38	−0.03	−0.21	−0.39	−0.03
APonMint	−0.12	−0.28	0.03	−0.13	−0.28	0.03	−0.13	−0.28	0.03	−0.15	−0.32	0.03
APonAint	−0.16	−0.33	0.01	−0.16	−0.33	0.01	−0.14	−0.33	0.05	−0.14	−0.32	0.05
MintonMP	−0.12	−0.29	0.10	−0.29	−0.30	−0.28	−0.07	−0.24	0.09	−0.04	−0.24	0.15
MintonAP	0	−0.15	0.15	0.07	−0.06	0.19	−0.02	−0.16	0.12	−0.04	−0.24	0.16
MintonMint	0.49	0.33	0.66	0.46	0.30	0.62	0.50	0.35	0.66	0.51	0.35	0.67
AintonMP	−0.06	−0.20	0.07	−0.04	−0.11	0.03	−0.08	−0.27	0.10	−0.08	−0.27	0.11
AintonAP	−0.09	−0.26	0.12	−0.26	−0.27	−0.24	−0.03	−0.23	0.16	−0.03	−0.22	0.17
AintonAint	0.44	0.27	0.61	0.42	0.24	0.59	0.45	0.27	0.61	0.45	0.28	0.61
AintMPMint	0.02	−0.01	0.08	0.02	0.01	0.04	0.01	−0.02	0.06	0.01	−0.03	0.06
AintAPMint	0	−0.03	0.03	−0.01	−0.04	0.01	0	−0.02	0.03	0.01	−0.03	0.04
MintMPAint	0.02	−0.02	0.05	0.01	−0.01	0.03	0.02	−0.02	0.07	0.02	−0.02	0.08
MintAPAint	0.01	−0.02	0.05	0.03	−0.01	0.07	0	−0.02	0.04	0	−0.03	0.04

M, maternal; A, adolescent; int, internalizing; P, positive interaction behavior; N, negative interaction behavior; on, describes the direction of regression, indirect effects are reported in direction of the association (e.g., MintMPAint describes the indirect effect from maternal to adolescent internalizing symptoms via maternal positive interaction behavior).

symptoms at T3 as the posterior distribution was wider and the 95% HPD thus included positive and negative values (M = 0.11, 95% HPD = [−0.04,0.27]. However, adolescent negative interaction behavior predicted maternal internalizing symptoms 1 year later at T3 (M = 0.11, 95% HPD = [0.00,0.23]). No indirect effects were found in this analysis. Further sensitivity analyses using default priors again yielded the same conclusions as the normal priors fitted to the linear pool. For the association between adolescent negative interaction behavior at T2 and subsequent maternal internalizing symptoms at T3, the effect size doubled in size compared to the linear pool, logarithmic, and fitted normal priors. The 95% HPD was even further from 0 (M = 0.22, 95% HPD = [0.06,0.38]), indicating stronger evidence that negative behaviors of adolescents predicted internalizing symptoms in mothers.

Some deviations between the results above stand out. For example, even though the mode of the pooled priors is closer to zero than the data (see **Figure 5**; AintonMN) maternal negative interaction behavior at T2 predicted adolescent internalizing symptoms at T3 with both pooled priors, but not with the fitted normal and default priors. Apparently, the density in the region slightly above 0 was so high that 0 was excluded from the 95% HPD for the pooled priors. On the other hand, adolescent negative interaction behavior at T2 only predicted maternal internalizing symptoms at T3 with default priors, suggesting that the prior for this parameter had a higher probability in the region around zero than our data. The posterior distribution of the association between adolescent negative interaction behavior and subsequent adolescent internalizing symptoms seemed strongly affected by the prior distribution as well, as there was a small region with extremely high probability (i.e., a spike) in the posterior around 0.25 in the analyses using linear and logarithmic pooled priors (see **Figure 5**; AintonAN); the 95% HPD of the linear pooled results, however, still included 0.

DISCUSSION

The present study used Bayesian estimation with systematically obtained results from previous studies and systematically defined prior weights, following three prior aggregation methods. The illustrative empirical research question behind this analysis concerned the mediation of bidirectional associations between maternal and adolescent internalizing symptoms from early to mid-adolescence by mother-adolescent positive and negative interaction behavior. We retrieved 47 effect sizes from 9 studies that provided information on some of the relevant parameters of our model and were thus integrated into our analyses.

Empirical Discussion: The Mediating Role of Mother-Adolescent Interaction Behavior

Consistent with theoretical and empirical evidence that internalizing symptoms can lower maternal positive interaction behavior toward their children (Simons et al., 1993; Goodman and Gotlib, 1999; Lovejoy et al., 2000; McCabe, 2014), the distributions consistently showed that higher levels of maternal internalizing symptoms predicted lower levels of their own, but generally not adolescent positive and negative interaction behavior in the following year. Mothers with increased internalizing symptoms might be emotionally unavailable, easily irritated, and unable to sensitively respond to their children's needs, which can suppress encouraging or nurturing behaviors and exacerbate hostile, rejecting behaviors in subsequent interactions with their children (Lovejoy et al., 2000). As

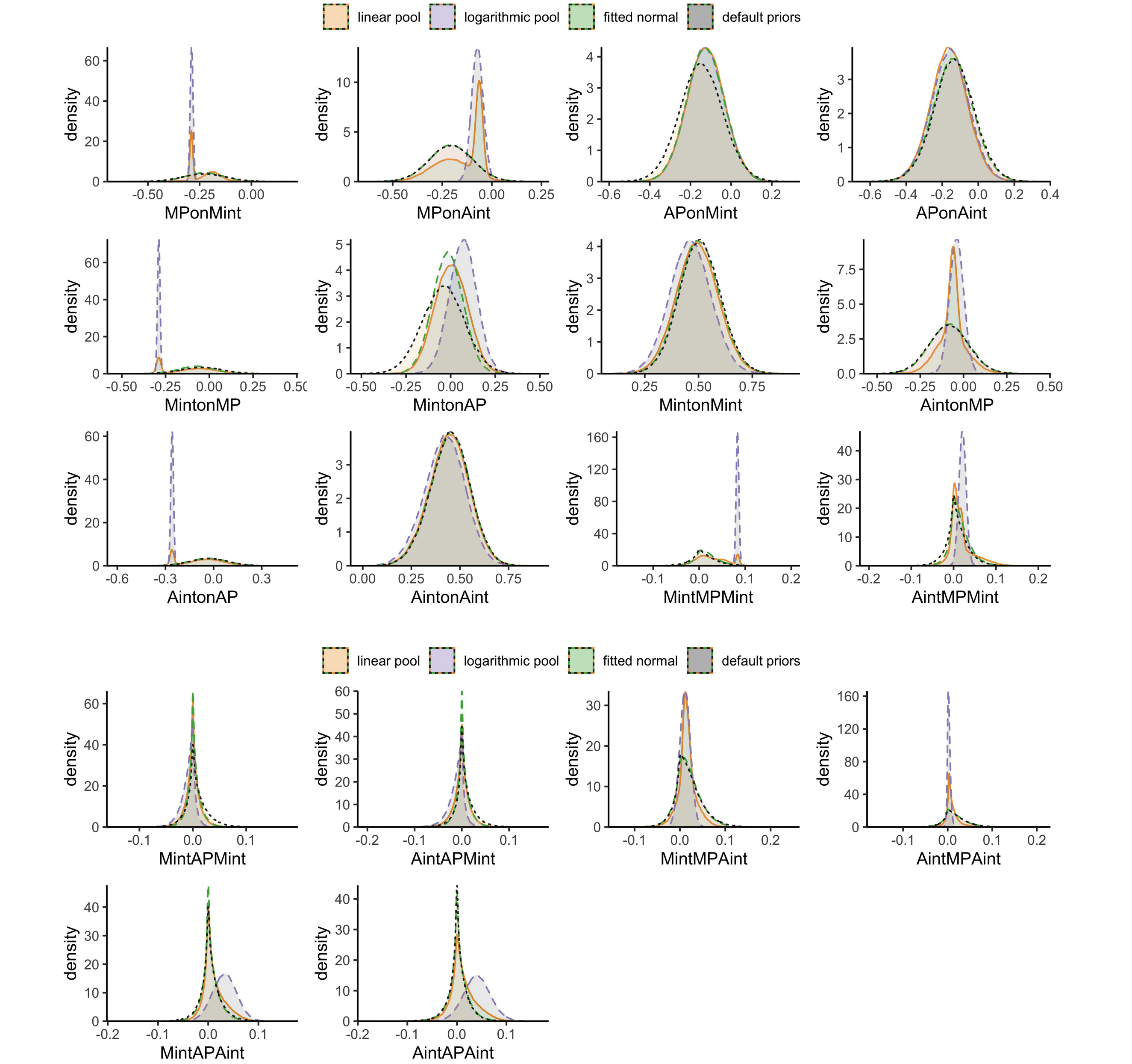


FIGURE 4 | Posterior distributions of the final results involving positive interaction behavior; linear pooled priors are displayed in orange, logarithmic pooled priors in light-purple, fitted normal priors in green, and default priors in gray.

maternal internalizing symptoms can disrupt interactional processes between mothers and adolescents, they are likely to drive relationship erosion in the long term (Coyne et al., 1991; Meeus, 2016). Interestingly, although the analyses using linear and logarithmic pooled priors suggested a clear negative association from adolescent internalizing symptoms to maternal positive interaction behavior as well, our findings generally provided only little evidence for the theoretical propositions that adolescent internalizing symptoms disrupt interactions in the family (Sheeber et al., 2001; Berg-Nielsen et al., 2002).

Despite theoretical propositions and empirical findings that less positive and more negative mother-adolescent interaction behavior predict adolescent internalizing symptoms (McLeod et al., 2007a,b; Yap et al., 2014; Pinquart, 2017), we found that maternal or adolescent internalizing symptoms predicted later mother-adolescent interaction behavior more often than that mother-adolescent interaction behavior predicted later internalizing symptoms. This is in line with one of the few mediation studies that found associations between maternal internalizing symptoms and observed maternal interaction behavior, but not between interaction behavior and

TABLE 6 | Parameter estimates using different prior settings for Model B.

	Linear pool priors			Logarithmic pool priors			Normal fitted to linear pool priors			Default priors		
Parameter	**Mean**	**95% HPD**		**Mean**	**95% HPD**		**Mean**	**95% HPD**		**Mean**	**95% HPD**	
MNonMint	0.23	0.06	0.39	0.23	0.08	0.39	0.22	0.06	0.38	0.19	0.01	0.37
MNonAint	0.04	−0.10	0.19	0.04	−0.08	0.16	0.04	−0.15	0.22	0.04	−0.15	0.23
ANonMint	0.10	−0.06	0.27	0.10	−0.06	0.27	0.10	−0.06	0.27	0.11	−0.07	0.30
ANonAint	0.13	−0.04	0.30	0.13	−0.04	0.30	0.12	−0.05	0.29	0.11	−0.06	0.29
MintonMN	0.11	−0.07	0.29	0.29	0.28	0.30	0.08	−0.07	0.23	0.03	−0.14	0.19
MintonAN	0.11	−0.01	0.23	0.07	−0.05	0.18	0.11	0	0.23	0.22	0.06	0.38
MintonMint	0.49	0.34	0.64	0.46	0.30	0.61	0.49	0.34	0.65	0.49	0.34	0.65
AintonMN	0.11	0.01	0.23	0.10	0.04	0.16	0.11	−0.04	0.27	0.11	−0.05	0.27
AintonAN	0.20	−0.02	0.26	0.26	0.25	0.27	0.10	−0.06	0.26	0.10	−0.06	0.26
AintonAint	0.44	0.27	0.60	0.43	0.26	0.59	0.45	0.28	0.61	0.45	0.28	0.61
AintMNMint	0	−0.01	0.03	0.01	−0.02	0.05	0	−0.02	0.03	0	−0.02	0.02
AintANMint	0.01	−0.01	0.04	0.01	−0.01	0.03	0.01	−0.01	0.04	0.02	−0.01	0.08
MintMNAint	0.03	0	0.06	0.02	0	0.05	0.02	−0.01	0.07	0.02	−0.01	0.07
MintANAint	0.02	−0.01	0.06	0.03	−0.02	0.07	0.01	−0.01	0.04	0.01	−0.01	0.05

M, maternal; A, adolescent; int, internalizing; P, positive interaction behavior; N, negative interaction behavior; on, describes the direction of regression; indirect effects are reported in direction of the associxation (e.g., MintMNAint describes the indirect effect from maternal to adolescent internalizing symptoms via maternal negative interaction behavior).

adolescent internalizing symptoms (van Doorn et al., 2016). One possible reason for this finding may be that mother-adolescent interaction behavior is more likely to influence immediate, short-term emotions in mothers or adolescents. While particularly maternal internalizing symptoms may have long-lasting effects, maladaptive interactions may exert their effects at a shorter time interval than we could detect with annual assessments. Alternatively, highly negative and less positive interactions between mothers and adolescents are quite common in early to mid-adolescence as mother-adolescent conflicts become more intense (Hadiwijaya et al., 2017). It is possible that because such behaviors are relatively typical during this time in adolescence, they are experienced as tied to that specific interaction and thus do not directly influence adolescent mood in the long term.

The limited evidence that we found for the associations between mother-adolescent interaction behavior and later internalizing symptoms concerned mainly negative interaction behavior in the analyses using linear and logarithmic pooled priors. This may be expected given that the impact of negative events and emotions is generally stronger than the impact of positive events or emotions (Baumeister et al., 2001). Although the effect sizes were generally comparable in all analyses, using different informative priors yielded somewhat different conclusions based on the distributions and credibility intervals. Together with the detected indirect effect that maternal negative interaction behavior mediated the associations between maternal internalizing and subsequent adolescent symptoms using linear pooled and logarithmic priors, however, they suggest that negative interaction behavior may play a role in the transmission of internalizing symptoms. Hostile behaviors might make interaction partners feel rejected and helpless, undermine their self-esteem, and elicit negative self-evaluations, which might in turn increase their risk for internalizing symptoms in the long-term (Gottman et al., 1997; Garber and Flynn, 2001). Interestingly, we also found that decreased maternal positive interaction behavior mediated the associations between adolescent internalizing symptoms and subsequent maternal internalizing symptoms, but this indirect effect was only evident using logarithmic pooled priors.

The different conclusions using different priors also warrant caution. Specifically, they suggest that our data contrasts with previous findings. In the linear pooled prior distribution, we indeed detected two spikes toward a positive distribution for the associations from maternal negative interaction behavior to later adolescent internalizing symptoms, whereas the logarithmic pooled priors suggested one extreme dense, narrow distribution closer to zero and the fitted normal priors indicated a similar, but flat distribution. The detected spikes in the linear and logarithmic pooled priors resulted from information found in previous studies, which drove these conclusions, whereas associations that we only detected with fitted normal and default priors suggested that our data provided stronger evidence than previous findings. Using different approaches to define informative priors allowed us to compare their impact on the results and evaluate the robustness of our conclusions. Once we updated the information collected in previous studies with our new data, the posterior distributions shifted to a varying degree depending on how we specified the priors. Differences between the posteriors were particularly pronounced when our data strongly diverged from previous studies. While the posteriors generated from logarithmic pooled priors were strongly influenced by previous data and thus only shifted little compared to the prior distributions, the linear pooled priors often resulted in bimodal distributions that reflected the discrepancy between previous and new data. These differences in priors and previous compared to new data emphasize that for some

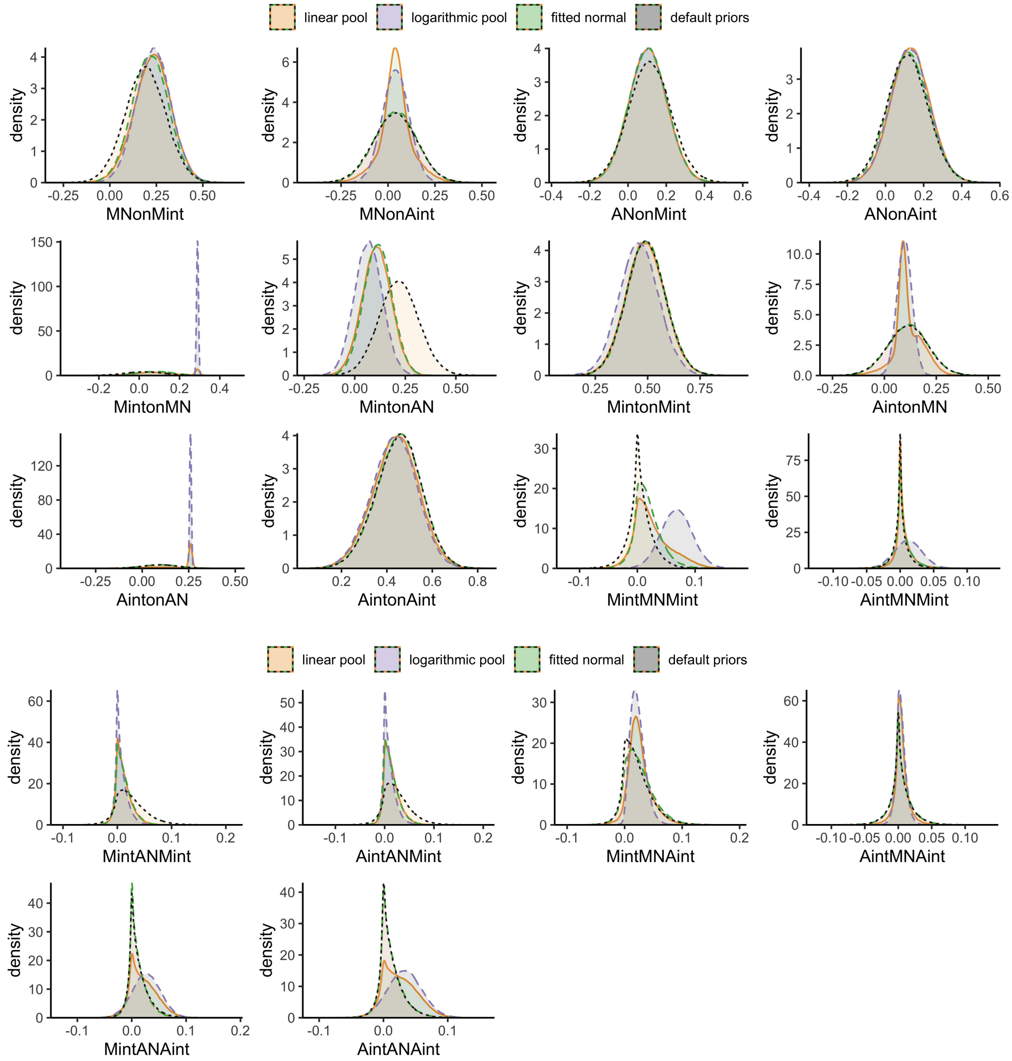


FIGURE 5 | Posterior distributions of the final results involving negative interaction behavior; linear pooled priors are displayed in orange, logarithmic pooled priors in light-purple, fitted normal priors in green, and default priors in gray.

associations, we may not yet have enough evidence to draw final conclusions.

The Role of Different Informative Priors

While we were able to include a range of findings relating to our model parameters, these studies reflected our own study's design to a varying degree and might thus introduce potential bias (Hobbs et al., 2011; Viele et al., 2014). Each included study provides a varying amount of relevant information and certainty, which is essential to take into account when specifying informative priors. How much a previous study contributes, depends on the focus and methodological considerations of the specific study. A weighting scheme therefore needs to be tailored to each new study's purpose and design. To avoid bias, such as increased subjectivity, it is important to engage content experts who can judge the relevance of weighting aspects and justify all decisions transparently in an accessible logbook (Zondervan-Zwijnenburg et al., 2017). Therefore, we involved content experts to design a weighting scheme and scoring system that allowed us to consider each study's specific contribution with respect to our data. Our illustrative example represented a longitudinal, multi-method design, which constituted the core of our weighting scheme. As cross-sectional studies cannot be used to disentangle the temporal order of associations, they provided only weak evidence for our parameters. Similarly, longitudinal studies that did not control for previous levels of psychopathological symptoms at an earlier point in time are not useful to measure change, and therefore received less weight as well. While a weighting scheme is an essential tool to combine findings from more or less comparable studies, it needs to be carefully constructed and reviewed to avoid inaccurate inferences and conclusions. In this study, we followed recommendations, such as including experts for the composition of the weighting scheme or the estimation of the weights (e.g., Zondervan-Zwijnenburg et al., 2017), which can further help to reduce subjectivity. Instead of weights based on the match between previous studies and the design of the study at hand, weights can also be based on optimality criteria (e.g., maximum entropy, minimum Kullback-Leibler divergence) or modeled by means of a prior on the weights (e.g., de Carvalho et al., 2020). These methods do not take the content of studies into account, which can be regarded their strength because of increased objectivity, but also their weakness because previous studies are not valued based on criteria that are considered important by experts.

Our statistical evaluation showed that analyses based on linear pooled priors may suffer from estimation problems (i.e., insufficient convergence and precision), where other prior specifications do not show the same issues. Furthermore, the prior predictive distributions were comparable across prior specification methods, except for the default prior, which does not produce a meaningful predictive distribution. Generally, we found that the posterior distributions based on the analyses with linear pooled priors displayed bimodal distributions and strong spikes in multiple occasions. The posteriors resulting from the logarithmic pooled priors were spiked and highly driven by the previous information as confirmed by the low associated shrinkage. In the current study, two studies (i.e., Pinquart, 2017; Milan and Carlone, 2018) caused all spikes. These studies reported estimates with extremely small (standardized) standard errors, thus strengthening the evidence for these effects. While Pinquart (2017) conducted a longitudinal meta-analysis on the associations between parental behaviors and adolescent internalizing symptoms with over 1,000 included studies, Milan and Carlone (2018) investigated actor and partner effects in how mother and adolescent internalizing symptoms predicted maternal and adolescent behaviors during an interaction task. Both studies provide important information for our analyses, but do not precisely reflect our study design. Specifically, Pinquart's meta-analysis also included (young) children and reported parental behaviors. Milan and Carlone, on the other hand, only sampled adolescent girls, who have been found to show higher levels of internalizing symptoms (Zahn-Waxler et al., 2008) and to be more sensitive to interpersonal experiences than adolescent boys (Flook, 2011). The design differences were taken into account by using power priors based on a systematic weighting scheme. In the current study, we did not lower study weights based on the specificity of the results. From a perspective of building cumulative knowledge, that would be a questionable practice. From a more pragmatic perspective, however, it may be sensible to downweigh information that appears unreasonably specific. For example, when expert elicitation is used to form prior distributions, it is suggested that the analyst decides to exclude an expert's distribution if their probability density is too narrow (de Carvalho et al., 2020).

Linear pooled priors integrate all available literature to its full avail and consider the influence of potentially differing previous findings. These distributions allow – or even demand – researchers to examine extreme or varying findings and discuss their data more specifically in relation to the literature. In this manner, the linear pooled prior and its associated posterior may also provide directions for future research. However, a multimodal posterior distribution may also render it difficult to interpret the findings directly. Furthermore, extra caution is warranted to establish sufficient convergence and precision.

Logarithmic pooled priors reflect an updating process of previous studies. As such, they are closely tied to the idea of building cumulative knowledge. In the current study, the specificity of some of the previous results overruled other previous findings and the current data in the posterior. However, this does not disqualify the logarithmic pooling procedure in general, nor in this case specifically. The posterior still represents our updated previous knowledge.

An alternative to downweighing previous results based on their extreme specificity, is to fit normal distributions to the linear pooled previous information. Similar to logarithmic pooled priors, fitted normal distributions behave well during Bayesian estimation and, similar to linear pooled priors, use previous information to inform the analyses. Particularly if previous research is scarce, contradictory or only few studies are sampled, fitted normal distributions are useful to specify informative priors without overemphasizing the effect of one individual study. Fitted normal priors are best suited when it can be assumed that the previous results are random samples from an underlying normal distribution, or when the analyst considers it a pragmatic midway between the more informative pooled and default priors.

In contrast to informative priors, default priors neglect previous knowledge about how mother-adolescent interaction behavior mediate the associations between maternal and adolescent internalizing symptoms. The predictive distribution clearly showed that default priors are highly unspecific with regards to expected future data. The associated shrinkage confirmed that the observed data completely overruled the unspecific previous information. For default priors, this behavior is desired. Previous studies, however, have shown that the use of default, non-informative priors may strongly bias the results and decrease estimation accuracy, particularly in small samples (Smid et al., 2019; Zitzmann et al., 2020).

Strengths, Limitations, and Implications

This study applied Bayesian estimation with informative priors to examine in an illustrative example whether observed mother-adolescent interaction behavior underlies the longitudinal associations between maternal and adolescent internalizing symptoms from early to mid-adolescence. Using a novel, comprehensive approach in which we first systematically quantified previous study findings in a meta-analytic design and then used this previous knowledge as input for the analyses allowed us to draw more precise conclusions about the potential mediating role of mother-adolescent interaction behavior. Such a strategy exceeds a pure meta-analytical approach, because it allowed us to incorporate existing information from a wide variety of studies that resemble our present study to varying degrees. Meta-analyses provide good starting points for new Bayesian analyses. Previous studies generate and raise new research questions, and Bayesian estimation with informative priors allows for a cumulative approach that does not ignore existing knowledge, but gradually updates it. This way, existing knowledge will be integrated into the empirical process. Particularly when previous research is scarce or when new studies are needed to address important limitations of previous research, including prior distributions can help to further cumulate knowledge. In our study, information was available on only some parameters, but not on the complete mediation model that we aimed to test. While the limited previous information was not sufficient to perform a meaningful meta-analysis, we were able to use the existing information to conduct new analyses that addressed previous limitations or remaining questions and integrated previous knowledge. By using three different priors, we were further able to show the robustness of our results across different approaches.

Despite these strengths, this study had some limitations with respect to the empirical mediation analysis. First, we only observed mother-adolescent interaction behavior at one time during early to mid-adolescence. While this approach allowed us to reduce the complexity of our model to fit our sample size, summary scores may not accurately reflect the processes that occur during the interactions between mothers and adolescents. It may be important to not only examine which average behaviors mothers and adolescents show during interactions, but also how these behaviors mutually influence each other on a moment-to-moment basis.

Second, a full longitudinal mediation approach would further require the assessment of all variables at each time point to account for the stability of not only internalizing symptoms across time, but also the stability of interaction behavior as well as concurrent associations between interaction behavior and internalizing symptoms. Due to the limited sample size in our data (N = 102 mother-adolescent dyads), we had insufficient information to inform a three-wave fully recursive model, which would have been ideal.

Third, longitudinal studies rarely employ the same time intervals between measurements, which renders comparing the findings from these studies difficult. Parameter estimates often depend on the time interval that was used (e.g., Gollob and Reichardt, 1987) as the underlying processes that measure change on a micro-scale, such as moment-to-moment or day-to-day, can differ from those on a macro-scale, such as year-to-year (Ebner-Priemer and Trull, 2009; Voelkle et al., 2012; Hollenstein et al., 2013). Consequently, studies with varying time scales might result in different conclusions that are not directly comparable. In our study, we tried to address time dependency by adding additional weight to studies that incorporated the same time interval as we did. However, we were only able to include few longitudinal studies, of which none received this additional weight. Future studies that aim to incorporate more, or exclusively, longitudinal studies might consider continuous rating options, such as continuous-time modeling or continuous-time meta-analytical procedures that allow to account for the effect of time more precisely (e.g., Van Montfort et al., 2018; Kuiper and Ryan, 2020). Another option could be to include a selection of varying weighting schemes and subsequently evaluate how different rating decisions affect the results. However, these approaches were beyond the scope of our empirical illustration.

In this study, we made use of differently composed informative priors to compare their effects on the posterior distributions. While our approach allowed us to systematically specify and use informative priors for the analyses of our data, quantifying, and weighing each previous study in such a systematic way requires a substantial amount of time and effort. If taken seriously, the task is equivalent to conducting a weighted meta-analysis with the additional benefit of including information from studies that resemble the present study to a varying extend. By allowing researchers to integrate new data and evaluate novel research questions using existing knowledge, this approach moves beyond where meta-analytical methods usually end and allows for knowledge to further cumulate over time.

As such, Bayesian estimation with informative priors can address important shortcomings of current empirical practices and serve the goal of empirical research to generate scientific growth of knowledge. Nevertheless, in such a systematic approach it is essential to effectively use previous information for Bayesian estimation. Knowing the literature and making informed decisions about relevant studies allows researchers to consider the most suitable approach to defining priors for their specific situation. This is important to avoid incorporating information from only one individual sample, while years of research already established well-grounded expectations. Focusing on the 95% HPD for hypothesis testing, our results did not detect many differences between the use of pooled or fitted normal priors.

How results from multiple previous studies on the same parameter should be included in the associated prior depends on theoretical considerations: Should the prior reflect the previous results as they are (linear pool), be an update of previous results (logarithmic pool), or be considered a set of random samples from an underlying normal distribution (normal fitted to the linear pool)? The differences between the approaches are emphasized when results diverge across previous studies: Are all results plausible and can they coexist in the prior distribution (linear pool), is only the consensual part plausible (logarithmic prior), or is there an underlying truth that is best resembled by a fitted normal distribution (fitted normal)? Additionally, pragmatic considerations can be taken into account. For example, the logarithmic pool is a theoretically sound (Bayesian) approach to aggregate multiple previous results that will emphasize consensual values, but extremely specific results from previous studies lead it to exclude large portions of the sample space. In the same situation, the posterior distributions based on the linear pooled priors do not exclude the observed values. However, the bimodal results that can result from diverging previous findings are difficult to interpret substantively. In these cases, a prior distribution like the fitted normal may be preferred, as it eliminates most of the impact of studies with high density when more studies contribute to the previous information.

CONCLUSION

Testing a comprehensive model that includes mediation effects requires a large sample size to detect small-to-moderate effects that are common in social science. Typically, studies including longitudinal, observational designs include only relatively small samples as they are time-consuming, costly, demand more of the participants, and face recruitment difficulties, such as dropout. Attempting to estimate complex models with traditional analytical techniques can result in estimation problems as well as inaccurate parameter estimates (e.g., van de Schoot et al., 2017), and thus limit the conclusions that can be drawn from such models. Furthermore, by using informative priors, we gain insight into how our data relate to the results from previous studies.

The findings of our study indicated that posterior distributions were generally stable across different prior distributions with differing levels of existing knowledge on the associations between mother-adolescent interaction behavior and internalizing symptoms. Specifically, we consistently found that even though mother-adolescent interaction behavior might play a relatively limited role in the transmission of internalizing symptoms from early to mid-adolescence, particularly negative interaction behavior might still be relevant. Nevertheless, the choice of prior aggregation did alter the results for some parameters and may well make a difference in other studies. Researchers should carefully consider how to aggregate previous results into one prior distribution, and always conduct sensitivity analyses to demonstrate if the results hold with different prior specifications. As illustrated by our example, using Bayesian estimation with informative priors offers a great opportunity to use accumulated knowledge to increase the precision of our outcomes. If conducted thoroughly, the approach equals and moves beyond where a weighted meta-analysis usually ends as it not only quantifies previous knowledge, but also integrates new data into a cumulative process. Such precision and accumulation of knowledge is important in moving empirical science forward, but also in informing therapeutic programs that aim to prevent or reduce adolescent internalizing symptoms by targeting often proposed risk factors, such as maladaptive interaction behavior between mothers and adolescents.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: Data Archiving and Networked Services (DANS): https://easy.dans.knaw.nl/ui/datasets/id/easy-dataset:113721; doi: 10.17026/dans-zrb-v5wp.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Medical Ethical Committee of the University Medical Center Utrecht (protocol number: 05/159-K). Written informed consent to participate in this study was provided by all adolescents and their parents.

AUTHOR CONTRIBUTIONS

SS conceived of the study and design, which was further refined by SN, AO, SB, and WM. SS drafted the manuscript, MZ-Z drafted the statistical sections. MZ-Z and DV verified the analytical methods and performed the statistical analyses. All authors discussed the results, critically revised the manuscript, and approved its final version.

FUNDING

The data of the RADAR study were used. RADAR has been financially supported by main grants from Netherlands Organization for Scientific Research (GB-MAGW 480-03-005 and GB-MAGW 480-08-006), and Stichting Achmea Slachtoffer en Samenleving (SASS), a grant from Netherlands Organization for Scientific Research to the Consortium Individual Development (CID; 024.001.003), and various other grants from Netherlands Organization for Scientific Research, the VU University Amsterdam, and Utrecht University. AO, MZ-Z, SB, SN, SS, and WM are members of the Consortium on Individual Development (CID). The CID is funded through the Gravitation program of the Dutch Ministry of Education, Culture, and Science and the Netherlands Organization for Scientific Research (NWO Gravitation 024-001-003).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.620802/full#supplementary-material

REFERENCES

Achenbach, T. M. (1966). The classification of children's psychiatric symptoms: a factor-analytic study. *Psychol. Monogr. Gen. Appl.* 80, 1–37. doi: 10.1037/h0093906

Achenbach, T. M., and Rescorla, L. A. (2003). *Manual for the ASEBA Adult Forms & Profiles*. Burlington, VT: University of Vermont.

Allen, J. P., Insabella, G., Porter, M. R., Smith, F. D., Land, D., and Phillips, N. (2006). A social-interactional model of the development of depressive symptoms in adolescence. *J. Consult. Clin. Psychol.* 74, 55–65. doi: 10.1037/0022-006X.74.1.55

Asbrand, J., Hudson, J., Schmitz, J., and Tuschen-Caffier, B. (2017). Maternal parenting and child behaviour: An observational study of childhood social anxiety disorder. *Cognit. Ther. Res.* 41, 562–575. doi: 10.1007/s10608-016-9828-3

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., and Vohs, K. D. (2001). Bad is stronger than good. *Rev. Gen. Psychol.* 5, 323–370. doi: 10.1037/1089-2680.5.4.323

Berg-Nielsen, T. S., Vikan, A., and Dahl, A. A. (2002). Parenting related to child and parental psychopathology: a descriptive review of the literature. *Clin. Child Psychol. Psychiatry* 7, 529–552. doi: 10.1177/1359104502007004006

Birmaher, B., Khetarpal, S., Brent, D., Cully, M., Balach, L., Kaufman, J., et al. (1997). The screen for child anxiety related emotional disorders (SCARED): scale construction and psychometric characteristics. *J. Am. Acad. Child Adolesc. Psychiatry* 36, 545–553. doi: 10.1097/00004583-199704000-00018

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

Bürkner, P. (2020). *Handle Missing Values With Brms.* The Comprehensive R Archive Network (CRAN). Available online at: https://cran.r-project.org/web/packages/brms/vignettes/brms_missings.html (accessed February 18, 2021).

Bürkner, P.-C. (2017). brms: an R Package for Bayesian multilevel models using stan. *J. Stat. Softw.* 80, 1–28. doi: 10.18637/jss.v080.i01

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., et al. (2017). Stan: a probabilistic programming language. *J. Stat. Softw.* 76, 1–32. doi: 10.18637/jss.v076.i01

Clayborne, Z. M., Varin, M., and Colman, I. (2019). Systematic review and meta-analysis: adolescent depression and long-term psychosocial outcomes. *J. Am. Acad. Child Adolesc. Psychiatry* 58, 72–79. doi: 10.1016/j.jaac.2018.07.896

Connell, A. M., and Goodman, S. H. (2002). The association between psychopathology in fathers versus mothers and children's internalizing and externalizing behavior problems: a meta-analysis. *Psychol. Bull.* 128, 746–773. doi: 10.1037/0033-2909.128.5.746

Coyne, J. C., Burchill, S. A. L., and Stiles, W. B. (1991). "An interactional perspective on depression," in *Handbook of Social and Clinical Psychology: The Health Perspective.* Pergamon General Psychology Series, eds C. R. Snyder and D. R. Forsyth (Oxford: Pergamon Press), 327–349.

Dadds, M. R., Sanders, M. R., Morrison, M., and Rebgetz, M. (1992). Childhood depression and conduct disorder: II. An analysis of family interaction patterns in the home. *J. Abnorm. Child Psychol.* 101, 505–513. doi: 10.1037//0021-843X.101.3.505

de Carvalho, L. M., Villela, D. A. M., Coelho, F. C., and Bastos, L. S. (2020). *Combining Probability Distributions: Extending the Logarithmic Pooling Approach.* 1–36. Available online at: http://arxiv.org/abs/1502.04206 (accessed February 3, 2021).

Delignette-Muller, M. L., and Dutang, C. (2015). fitdistrplus: an R Package for fitting distributions. *J. Stat. Softw.* 64, 1–34. doi: 10.18637/jss.v064.i04

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

Dietz, L. J., Birmaher, B., Williamson, D. E., Silk, J. S., Dahl, R. E., Axelson, D. A., et al. (2008). Mother-child interactions in depressed children and children at high risk and low risk for future depression. *J. Am. Acad. Child Adolesc. Psychiatry* 47, 574–582. doi: 10.1097/CHI.0b013e3181676595

Ebner-Priemer, U. W., and Trull, T. J. (2009). Ecological momentary assessment of mood disorders and mood dysregulation. *Psychol. Assess.* 21, 463–475. doi: 10.1037/a0017075

Eley, T. C., McAdams, T. A., Rijsdijk, F. V., Lichtenstein, P., Narusyte, J., Reiss, D., et al. (2015). The intergenerational transmission of anxiety: a children-of-twins study. *Am. J. Psychiatry* 172, 630–637. doi: 10.1176/appi.ajp.2015.14070818

Feinberg, M. E., Kan, M. L., and Hetherington, E. M. (2007). The longitudinal influence of coparenting conflict on parental negativity and adolescent maladjustment. *J. Marriage Fam.* 69, 687–702. doi: 10.1111/j.1741-3737.2007.00400.x

Feng, X., Shaw, D. S., Skuban, E. M., and Lane, T. (2007). Emotional exchange in mother-child dyads: stability, mutual influence, and associations with maternal depression and child problem behavior. *J. Fam. Psychol.* 21, 714–725. doi: 10.1037/0893-3200.21.4.714

Flook, L. (2011). Gender differences in adolescents' daily interpersonal events and well-being. *Child Dev.* 82, 454–461. doi: 10.1111/j.l467-8624.2010.01521.x

Fogel, A. (1993). "Two principles of communication: co-regulation and framing," in *New Perspectives in Early Communicative Development*, eds J. Nadel and L. Camaioni (London: Routledge), 9–22.

Garber, J., and Flynn, C. (2001). "Vulnerability to depression in childhood and adolescence," in *Vulnerability to Psychopathology: Risk Across the Lifespan*, eds R. E. Ingram and J. M. Price (New York, NY: Guilford), 175–225.

Gelman, A., Carlin, J., Stern, H. S., and Rubin, D. (2004). *Bayesian Data Analysis.* London: Chapman & Hall.

Gelman, A., and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Stat. Sci.* 7, 457–472.

Genest, C., and Zidek, J. V. (1986). Combining probability distributions: a critique and an annotated bibliography. *Stat. Sci.* 1, 147–148. doi: 10.1214/ss/1177013831

Gollob, H. F., and Reichardt, C. S. (1987). Taking account of time lags in causal models. *Child Dev.* 58, 80–92.

Goodman, S. H., and Gotlib, I. H. (1999). Risk for psychopathology in the children of depressed mothers: a developmental model for understanding mechanisms of transmission. *Psychol. Rev.* 106, 458–490. doi: 10.1037/0033-295X.106.3.458

Gottman, J. M., Katz, L. F., and Hooven, C. (1997). *Meta-Emotion: How Families Communicate Emotionally*. Mahwah, NJ: Lawrence Erlbaum Associates Inc.

Griffith, J. M., Crawford, C. M., Oppenheimer, C. W., Young, J. F., and Hankin, B. L. (2019). Parenting and youth onset of depression across three years: examining the influence of observed parenting on child and adolescent depressive outcomes. *J. Abnorm. Child Psychol.* 47, 1969–1980. doi: 10.1007/s10802-019-00564-z

Hadiwijaya, H., Klimstra, T. A., Vermunt, J. K., Branje, S. J. T., and Meeus, W. H. J. (2017). On the development of harmony, turbulence, and independence in parent–adolescent relationships: a five-wave longitudinal study. *J. Youth Adolesc.* 46, 1772–1788. doi: 10.1007/s10964-016-0627-7

Hald, T., Aspinall, W., Devleesschauwer, B., Cooke, R., Corrigan, T., Havelaar, A. H., et al. (2016). World Health Organization estimates of the relative contributions of food to the burden of disease due to selected foodborne hazards: a structured expert elicitation. *PLoS One* 11:e0145839. doi: 10.1371/journal.pone.0145839

Hobbs, B. P., Carlin, B. P., Mandrekar, S., and Sargent, D. J. (2011). Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials. *Biometrics* 67, 1047–1056. doi: 10.1111/j.1541-0420.2011.01564.x.Hierarchical

Hofer, C., Eisenberg, N., Spinrad, T. L., Morris, A. S., Gershoff, E., Valiente, C., et al. (2013). Mother-adolescent conflict: stability, change, and relations with externalizing and internalizing behavior problems. *Soc. Dev.* 22, 259–279. doi: 10.1111/sode.12012

Hollenstein, T., Lichtwarck-Aschoff, A., and Potworowski, G. (2013). A model of socioemotional flexibility at three time scales. *Emot. Rev.* 5, 397–405. doi: 10.1177/1754073913484181

Holtmann, J., Koch, T., Lochner, K., and Eid, M. (2016). A comparison of ML, WLSMV, and Bayesian methods for multilevel structural equation models in small samples: a simulation study. *Multivar. Behav. Res.* 51, 661–680. doi: 10.1080/00273171.2016.1208074

Hughes, E. K., and Gullone, E. (2010). Reciprocal relationships between parent and adolescent internalizing symptoms. *J. Fam. Psychol.* 24, 115–124. doi: 10.1037/a0018788

Ibrahim, J. G., and Chen, M. H. (2000). Power prior distributions for regression models. *Stat. Sci.* 15, 46–60.

Jackson, J., Kuppens, P., Sheeber, L. B., and Allen, N. B. (2011). Expression of anger in depressed adolescents: the role of the family environment. *J. Abnor. Child Psychol.* 39, 463–474. doi: 10.1007/s10802-010-9473-3

Kessler, R. C., Avenevoli, S., Costello, E. J., Georgiades, K., Green, J. G., Gruber, M. J., et al. (2012). Prevalence, persistence, and sociodemographic correlates of DSM-IV disorders in the national comorbidity survey replication adolescent supplement. *Arch. Gen. Psychiatry* 69, 372–380. doi: 10.1001/archgenpsychiatry.2011.160.Prevalence

König, C., and van de Schoot, R. (2018). Bayesian statistics in educational research: a look at the current state of affairs. *Educ. Rev.* 70, 486–509. doi: 10.1080/00131911.2017.1350636

Kuiper, R. M., and Ryan, O. (2020). Meta-analysis of lagged regression models: a continuous-time approach. *Struct. Equ. Modeling* 27, 396–413. doi: 10.1080/10705511.2019.1652613

Lahey, B. B., Krueger, R. F., Rathouz, P. J., Waldman, I. D., and Zald, D. H. (2017). A hierarchical causal taxonomy of psychopathology across the life span. *Psychol. Bull.* 143, 142–186. doi: 10.1037/bul0000069

Lovejoy, M. C., Graczyk, P. A., O'Hare, E., and Neuman, G. (2000). Maternal depression and parenting behavior: a meta-analytic review. *Clin. Psychol. Rev.* 20, 561–592. doi: 10.1016/S0272-7358(98)00100-7

McCabe, J. E. (2014). Maternal personality and psychopathology as determinants of parenting behavior: a quantitative integration of two parenting literatures. *Psychol. Bull.* 140, 722–750. doi: 10.1037/a0034835

McLeod, B. D., Weisz, J. R., and Wood, J. J. (2007a). Examining the association between parenting and childhood depression: a meta-analysis. *Clin. Psychol. Rev.* 27, 986–1003. doi: 10.1016/j.cpr.2007.03.001

McLeod, B. D., Wood, J. J., and Weisz, J. R. (2007b). Examining the association between parenting and childhood anxiety: a meta-analysis. *Clin. Psychol. Rev.* 27, 155–172. doi: 10.1016/j.cpr.2006.09.002

Meeus, W. (2016). Adolescent psychosocial development: a review of longitudinal models and research. *Dev. Psychol.* 52, 1969–1993. doi: 10.1037/dev0000243

Milan, S., and Carlone, C. (2018). A two-way street: mothers' and adolescent daughters' depression and PTSD symptoms jointly predict dyadic behaviors. *J. Fam. Psychol.* 32, 1097–1108. doi: 10.1037/fam0000467

Natsuaki, M. N., Shaw, D. S., Neiderhiser, J. M., Ganiban, J. M., Harold, G. T., Reiss, D., et al. (2014). Raised by depressed parents: is it an environmental risk? *Clin. Child Fam. Psychol. Rev.* 17, 357–367. doi: 10.1007/s10567-014-0169-z

Neelon, B., and O'Malley, A. J. (2010). Bayesian analysis using power priors with application to pediatric quality of care. *J. Biom. Biostat.* 1:103. doi: 10.4172/2155-6180.1000103

Nelemans, S. A., Hale, W. W., Branje, S. J. T., Hawk, S. T., and Meeus, W. H. J. (2014). Maternal criticism and adolescent depressive and generalized anxiety disorder symptoms: A 6-year longitudinal community study. *J. Abnorm. Child Psychol.* 42, 755–766. doi: 10.1007/s10802-013-9817-x

Nelson, B. W., Byrne, M. L., Sheeber, L., and Allen, N. B. (2017). Does context matter? A multi-method assessment of affect in adolescent depression across multiple affective interaction contexts. *Clin. Psychol. Sci.* 5, 239–258. doi: 10.1177/2167702616680061

Olino, T. M., McMakin, D. L., Nicely, T. A., Forbes, E. E., Dahl, R. E., and Silk, J. S. (2016). Maternal depression, parenting, and youth depressive symptoms: mediation and moderation in a short-term longitudinal study. *J. Clin. Child Adolesc. Psychol.* 45, 279–290. doi: 10.1080/15374416.2014.971456

Pinquart, M. (2017). Associations of parenting dimensions and styles with externalizing problems of children and adolescents: an updated meta-analysis. *Dev. Psychol.* 53, 873–932. doi: 10.1037/dev0000295

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Reynolds, W. M. (2000). *Reynolds Adolescent Depression Scale (RADS-2). Professional Manual*, 2nd Edn. Lutz, FL: Psychological Assessment Resources.

Sameroff, A.J. (Ed.) (2009). *The Transactional Model of Development: How Children and Contexts Shape Each Other*. Washington, DC: American Psychological Association.

Schwartz, O. S., Dudgeon, P., Sheeber, L. B., Yap, M. B. H., Simmons, J. G., and Allen, N. B. (2012). Parental behaviors during family interactions predict changes in depression and anxiety symptoms during adolescence. *J. Abnorm. Child Psychol.* 40, 59–71. doi: 10.1007/s10802-011-9542-2

Sheeber, L., Hops, H., and Davis, B. (2001). Family processes in adolescent depression. *Clin. Child Fam. Psychol. Rev.* 4, 19–35. doi: 10.1023/A:1009524626436

Simons, R. L., Lorenz, F. O., Wu, C. I., and Conger, R. D. (1993). Social network and marital support as mediators and moderators of the impact of stress and depression on parental behavior. *Dev. Psychol.* 29, 368–381. doi: 10.1037/0012-1649.29.2.368

Smid, S. C., McNeish, D., Miočević, M., and van de Schoot, R. (2019). Bayesian versus frequentist estimation for structural equation models in small sample contexts: a systematic review. *Struct. Equ. Modeling* 27, 131–161. doi: 10.1080/10705511.2019.1577140

Spiegelhalter, D. J., Abrams, K. R., and Myles, J. P. (2004). *Bayesian Approaches to Clinical Trials and Health-Care Evaluation, Vol. 13.* John Wiley & Sons

Stan Development Team (2020). *RStan: The R Interface to Stan. R Package Version 2.21.1*. Available online at: http://mc-stan.org/

Szwedo, D. E., Hessel, E. T., and Allen, J. P. (2017). Supportive romantic relationships as predictors of resilience against early adolescent maternal negativity. *J. Youth Adolesc.* 46, 454–465. doi: 10.1007/s10964-016-0507-1

Van Buuren, S., and Groothuis-Oudshoorn, K. (2011). mice: multivariate imputation by chained equations in R. *J. Stat. Softw.* 45, 1–67. doi: 10.18637/jss.v045.i03

van de Schoot, R., Depaoli, S., King, R., Kramer, B., Märtens, K., Tadesse, M. G., et al. (2021). Bayesian statistics and modelling. *Nat. Rev. Methods Prim.* 1:1. doi: 10.1038/s43586-020-00001-2

van de Schoot, R., Kaplan, D., Denissen, J., Asendorpf, J. B., Neyer, F. J., and van Aken, M. A. G. (2014). A gentle introduction to Bayesian analysis: applications to developmental research. *Child Dev.* 85, 842–860. doi: 10.1111/cdev.12169

van de Schoot, R., Sijbrandij, M., Depaoli, S., Winter, S. D., Olff, M., and van Loey, N. E. (2018). Bayesian PTSD-trajectory analysis with informed priors based on a systematic literature search and expert elicitation. *Multivar. Behav. Res.* 53, 267–291. doi: 10.1080/00273171.2017.1412293

van de Schoot, R., Winter, S. D., Ryan, O., Zondervan-Zwijnenburg, M., and Depaoli, S. (2017). A systematic review of Bayesian articles in psychology: the last 25 years. *Psychol. Methods* 22, 217–239. doi: 10.1037/met0000100

van Doorn, M. M. E. M., Kuijpers, R. C. W. M., Lichtwarck-Aschoff, A., Bodden, D., Jansen, M., and Granic, I. (2016). Does mother–child interaction mediate the relation between maternal depressive symptoms and children's mental health problems? *J. Child Fam. Stud.* 25, 1257–1268. doi: 10.1007/s10826-015-0309-1

van Erp, S. J., Mulder, J., and Oberski, D. L. (2018). Prior sensitivity analysis in default Bayesian structural equation modeling. *Psychol. Methods* 23, 363–388. doi: 10.1037/met0000162

Van Montfort, K., Oud, J. H. L., and Voelkle, M. C. (Eds.) (2018). *Continuous Time Modeling in the Behavioral and Related Sciences*. Cham: Springer.

Veen, D., Egberts, M. R., van Loey, N. E. E., and van de Schoot, R. (2020). Expert elicitation for latent growth curve models: the case of posttraumatic stress symptoms development in children with burn injuries. *Front. Psychol.* 11:1197. doi: 10.3389/fpsyg.2020.01197

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., and Bürkner, P. C. (2019). Rank-normalization, folding, and localization: an improved Rb for assessing convergence of MCMC. *ArXiv*[Preprint] 1–27. doi: 10.1214/20-ba1221

Viele, K., Berry, S., Neuenschwander, B., Amzal, B., Chen, F., Enas, N., et al. (2014). Use of historical control data for assessing treatment effects in clinical trials. *Pharm. Stat.* 13, 41–54. doi: 10.1002/pst.1589

Voelkle, M. C., Oud, J. H. L., Davidov, E., and Schmidt, P. (2012). An SEM Approach to continuous time modeling of panel data: relating authoritarianism and anomia. *Psychol. Methods* 17, 176–192. doi: 10.1037/a0027543

Weinfield, N. S., Egeland, B., Hennighausen, K., Lawrence, C., Carlson, E., Meijer, S., et al. (1999). *Jobs Wave 2 Middle Childhood Observational Manual Coding Scheme for Affective and Behavioral Quality of Mother-Child Interaction*. Minneapolis, MN: Institute of Child Development, University of Minnesota.

Weinfield, N. S., Ogawa, J. R., and Egeland, B. (2002). Predictability of observed mother-child interaction from preschool to middle childhood in a high-risk sample. *Child Dev.* 73, 528–543.

Weymouth, B. B., Buehler, C., Zhou, N., and Henson, R. A. (2016). A meta-analysis of parent-adolescent conflict: disagreement, hostility, and youth maladjustment. *J. Fam. Theory Rev.* 8, 95–112. doi: 10.1111/jftr.12126

Wilkinson, P. O., Harris, C., Kelvin, R., Dubicka, B., and Goodyer, I. M. (2013). Associations between adolescent depression and parental mental health, before and after treatment of adolescent depression. *Eur. Child Adolesc. Psychiatry* 22, 3–11. doi: 10.1007/s00787-012-0310-9

Yap, M. B. H., Pilkington, P. D., Ryan, S. M., and Jorm, A. F. (2014). Parental factors associated with depression and anxiety in young people: a systematic review and meta-analysis. *J. Affect. Disord.* 156, 8–23. doi: 10.1016/j.jad.2013.11.007

Zahn-Waxler, C., Shirtcliff, E. A., and Marceau, K. (2008). Disorders of childhood and adolescence: gender and psychopathology. *Annu. Rev. Clin. Psychol.* 4, 275–303. doi: 10.1146/annurev.clinpsy.3.022806.091358

Zhou, X., and Reiter, J. P. (2010). A note on Bayesian inference after multiple imputation. *Am. Stat.* 64, 159–163.

Zitzmann, S., and Hecht, M. (2019). Going beyond convergence in Bayesian estimation: why precision matters too and how to assess it. *Struct. Equ. Modeling* 26, 646–661. doi: 10.1080/10705511.2018.1545232

Zitzmann, S., Lüdtke, O., Robitzsch, A., and Hecht, M. (2020). On the performance of Bayesian approaches in small samples: a comment on Smid, McNeish, Miocevic, and van de Schoot (2020). *Struct. Equ. Modeling* 1–11. doi: 10.1080/10705511.2020.1752216

Zondervan-Zwijnenburg, M., Peeters, M., Depaoli, S., and Van de Schoot, R. (2017). Where do priors come from? Applying guidelines to construct informative priors in small sample research. *Res. Hum. Dev.* 14, 305–320. doi: 10.1080/15427609.2017.1370966

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 31 March 2021
doi: 10.3389/fpsyg.2021.624588

Assessing the Impact of Precision Parameter Prior in Bayesian Non-parametric Growth Curve Modeling

Xin Tong[1] *and Zijun Ke*[2]*

[1] *Department of Psychology, University of Virginia, Charlottesville, VA, United States,* [2] *Department of Psychology, Sun Yat-sen University, Guangzhou, China*

Bayesian non-parametric (BNP) modeling has been developed and proven to be a powerful tool to analyze messy data with complex structures. Despite the increasing popularity of BNP modeling, it also faces challenges. One challenge is the estimation of the precision parameter in the Dirichlet process mixtures. In this study, we focus on a BNP growth curve model and investigate how non-informative prior, weakly informative prior, accurate informative prior, and inaccurate informative prior affect the model convergence, parameter estimation, and computation time. A simulation study has been conducted. We conclude that the non-informative prior for the precision parameter is less preferred because it yields a much lower convergence rate, and growth curve parameter estimates are not sensitive to informative priors.

Keywords: non-parametric Bayesian, robust method, growth curve modeling, Dirichlet process mixture, prior, precision parameter

Edited by:
Haiyan Liu,
University of California, Merced,
United States

Reviewed by:
Sonja Winter,
University of California, Merced,
United States
Steffen Zitzmann,
University of Tübingen, Germany
Kenneth Wilcox,
University of Notre Dame,
United States

Correspondence:
Zijun Ke
keziyun@mail.sysu.edu.cn

Specialty section:
This article was submitted to
Quantitative Psychology and
Measurement,
a section of the journal
Frontiers in Psychology

Received: *31 October 2020*
Accepted: *09 February 2021*
Published: *31 March 2021*

Citation:
Tong X and Ke Z (2021) Assessing the Impact of Precision Parameter Prior in Bayesian Non-parametric Growth Curve Modeling.
Front. Psychol. 12:624588.
doi: 10.3389/fpsyg.2021.624588

1. INTRODUCTION

Bayesian non-parametric (BNP) modeling, also called semiparametric Bayesian modeling in the literature, has been recognized as a valuable data analytical technique due to its great flexibility and adaptivity (e.g., Müller and Mitra, 2004; Gershman and Blei, 2012). It is rapidly gaining popularity among methodologists and practitioners and has been applied to a variety of models including regressions, latent variable models with complex structures, sequential models, etc. BNP models are on an infinite dimensional parameter space and the complexity of the models adapts to the data. One of the most popular BNP models is Dirichlet process (DP) mixtures. Being able to adapt the number of latent classes to the complexity of the data, DP mixtures are powerful in modeling empirical data. However, they also face technical challenges. One challenge is the estimation of the precision parameter in the DP mixture. In this study, we focus on the prior of precision parameter and investigate how it affects model convergence, parameter estimation, and computation time in BNP growth curve modeling.

Growth curve models are broadly used in longitudinal research (e.g., Meredith and Tisak, 1990; McArdle and Nesselroade, 2014). Many popular longitudinal models in social and behavioral sciences, such as multilevel models, some mixed-effects models, and linear hierarchical models, can be written as a form of growth curve models. In growth curve models, dependent variables are repeatedly measured and explained as a function of time and possible control variables. The mean function between the dependent variables and time is the mean growth. Random effects and measurement errors cause the individual growth trajectories to deviate from the mean growth curve. Traditional growth curve modeling is typically based on the normality assumption. That is, both the random effects and measurement errors are assumed to follow normal distributions. However, empirical data often violate the normality assumption (Micceri, 1989; Cain et al., 2017). Non-normal population distributions and data contamination are two common causes of non-normality. Although standard errors and test statistics have been corrected to reduce the adverse effect of distributional assumption violation (e.g., Chou et al., 1991; Curran et al., 1996), normal-distribution-based maximum likelihood estimation may still yield inefficient or inaccurate parameter estimates, and thus misleading statistical inferences (e.g., Yuan and Bentler, 2001; Maronna et al., 2006). Therefore, researchers have developed robust methods to obtain accurate parameter estimation and statistical inference.

The ideas of robust methods can be divided into two types. For the first type, the key idea is to downweight extreme cases. To do so, this type of robust methods assigns a weight to each subject in a dataset according to its distance from the center of the majority of the data (e.g., Pendergast and Broffitt, 1985; Singer and Sen, 1986; Silvapulle, 1992; Yuan and Bentler, 1998; Zhong and Yuan, 2010). For the second type, the key idea is to use non-normal distributions that are mathematically tractable while building the statistical model. For example, latent variables and/or measurement errors are assumed to follow a *t* or skew-*t* distribution (Tong and Zhang, 2012; Zhang, 2016) or a mixture of certain distributions (Muthén and Shedden, 1999; Lu and Zhang, 2014). While being useful, these methods have limitations under certain conditions. For example, the downweighting method does not perform well when latent variables contain extreme scores (see Zhong and Yuan, 2011). Using a *t* distribution or a mixture of normal distributions still imposes restrictions on the shape of the data distribution.

The aforementioned issues are automatically resolved by BNP methods. BNP modeling relies on a building block, DP, to handle the non-normality issue. DP is a distribution over probability measures that can be used to estimate unknown distributions. Consequently, the non-normality issue can be addressed by directly estimating the unknown random distributions of latent variables or measurement errors (i.e., obtaining the posteriors of the distributions).

The advantages of using BNP methods with DP priors have been discussed in the literature (e.g., Ghosal et al., 1999; MacEachern, 1999; Hjort, 2003; Müller and Mitra, 2004; Fahrmeir and Raach, 2007; Hjort et al., 2010). They do not constrain models to a specific parametric form which may limit the scope and type of statistical inferences in many situations, especially when data are not normally distributed. Thus, a typical motivation of using BNP methods is that one is unwilling to make somewhat arbitrary and unverified assumptions for latent variables or error distributions as in the parametric modeling. Meanwhile, BNP methods can provide full probability models for the data-generating process and lead to analytically tractable posterior distributions.

BNP methods have been applied to complex models. For example, Bush and MacEachern (1996), Kleinman and Ibrahim (1998), and Brown and Ibrahim (2003) used DP mixtures to handle non-normal random effects. Burr and Doss (2005) used a conditional DP to handle heterogeneous effect sizes in the context of meta-analysis. Ansari and Iyengar (2006) included Dirichlet components to build a semiparametric recurrent choice model. Dunson (2006) used dynamic mixtures of DP to estimate the varied distributions of a latent variable, which change non-parametrically across groups. Si and Reiter (2013), Si et al. (2015) used DP mixtures of multinomial distributions for categorical data with missing values. BNP approach has also been adapted to structural equation modeling to relax the normality assumption of the latent variables (e.g., Lee et al., 2008; Yang and Dunson, 2010). Tong and Zhang (2019) directly used a DP mixture to model non-normal data in growth curve modeling.

Although the application of BNP modeling has increased dramatically since the theoretical properties of BNP methods were better understood and their computational hurdles were removed (e.g., Neal, 2000), BNP modeling is still unfamiliar to the majority of researchers in social and behavioral sciences. Additionally, there are technical issues that have not yet been fully addressed (Sharif-Razavian and Zollmann, 2009). The convergence issue is one of such unanswered questions. Non-convergence can occur when BNP method is applied to complex models. Tong and Zhang (2019) found that non-convergence was largely caused by the precision parameter of the mixing DP. The precision parameter is a critical hyperparameter that governs the expected number of mixture components. When a non-informative prior was used for the precision parameter, non-convergence occurred or a longer computation time was observed (Tong and Zhang, 2019). Informative priors may help solve this issue. However, only a few studies have noticed and discussed the effect of the precision parameter in DP mixtures (e.g., West, 1992; Ohlssen et al., 2007; Jara et al., 2011). Ishwaran (2000) was among the few that studied the informative prior for the precision parameter. Ishwaran (2000) suggested to use the $Gamma(2, 2)$ prior to encourage both small and large values of the precision parameter. In sum, despite its impact on the model convergence issue, no study has systematically investigated how the prior for the precision parameter should be specified.

Therefore, in this study, we evaluate and compare non-informative, weakly informative, accurate informative, and inaccurate informative priors for the precision parameter of DP mixtures. We study how these priors influence model convergence, model estimation, and computation time in BNP growth curve modeling. In the next section, we introduce BNP growth curve modeling. After providing the conditional posterior distribution of the precision parameter, we use a simulation study to assess the impact of four types of priors for the precision parameter. Recommendations are provided at the end of the article. We also provide a guideline about the implementation of BNP growth curve modeling using R (R Core Team, 2019) in the **Appendix**.

2. BAYESIAN NON-PARAMETRIC GROWTH CURVE MODELING

We now introduce a typical growth curve model and a BNP method based on this model. Consider a longitudinal dataset with N subjects and T measurement occasions. Let $\mathbf{y}_i = (y_{i1}, \ldots, y_{iT})'$

be a $T \times 1$ random vector with y_{ij} being a measurement from individual i at time j $(i = 1, \ldots, N; j = 1, \ldots, T)$. A growth curve model without covariates can be written as

$$\begin{aligned} \mathbf{y}_i &= \Lambda \mathbf{b}_i + \mathbf{e}_i, \\ \mathbf{b}_i &= \boldsymbol{\beta} + \mathbf{u}_i, \end{aligned}$$

where Λ is a $T \times q$ factor loading matrix that determines the growth curves, $\mathbf{b}_i$ is a $q \times 1$ vector of random effects, and $\mathbf{e}_i$ is a vector of measurement errors. The vector of random effects $\mathbf{b}_i$ varies around its mean $\boldsymbol{\beta}$. The residual vector $\mathbf{u}_i$ represents the deviation of $\mathbf{b}_i$ from $\boldsymbol{\beta}$. When

$$\boldsymbol{\Lambda} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & T-1 \end{pmatrix}, \mathbf{b}_i = \begin{pmatrix} L_i \\ S_i \end{pmatrix}, \ and\ \boldsymbol{\beta} = \begin{pmatrix} \beta_L \\ \beta_S \end{pmatrix},$$

the model is reduced to a linear growth curve model with random intercept L_i and random slope S_i. The mean intercept and slope are denoted as β_L and β_S, respectively.

Traditionally, $\mathbf{e}_i$ and $\mathbf{u}_i$ are assumed to follow multivariate normal distributions with mean vectors of zero and covariance matrices $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$, respectively, so $\mathbf{e}_i \sim MN_T(\mathbf{0}, \boldsymbol{\Phi})$ and $\mathbf{u}_i \sim MN_q(\mathbf{0}, \boldsymbol{\Psi})$. Here, MN denotes a multivariate normal distribution and its subscript indicates its dimension. Measurement errors are often assumed to be uncorrelated with each other and have equal variances across time. Statistically, this simplification means the covariance matrix of measurement error $\boldsymbol{\Phi}$ is reduced to $\boldsymbol{\Phi} = \sigma_e^2 \mathbf{I}$ where σ_e^2 is a scale parameter. In linear growth curve models, $\mathbf{u}_i = (u_{Li}, u_{Si})'$. Its covariance matrix is then $\boldsymbol{\Psi} = cov(\mathbf{u}_i) = \begin{pmatrix} \sigma_L^2 & \sigma_{LS} \\ \sigma_{LS} & \sigma_S^2 \end{pmatrix}$. Here, σ_L^2 and σ_S^2 represent the variances of the random intercept and slope across individuals, respectively, and σ_{LS} represents the covariance between the random intercept and slope.

BNP methods do not make arbitrary distributional assumptions as in the parametric modeling and thus are more flexible in handling non-normal data (e.g., Lee et al., 2008; Tong and Zhang, 2019). Unlike conventional non-parametric methods such as permutation tests, BNP methods use full probability models to describe the data-generating process and thus can derive posterior distributions for model parameters.

Within the BNP modeling scope, the parametric distributions of latent variables and measurement errors in traditional methods are replaced by unknown random distributions. To estimate these unknown distributions, DP is frequently used as the prior (Ferguson, 1973, 1974). Specifically, a random "sample" from a DP is a random distribution. Here, we denote it as G. A DP has two hyperparameters, α and G_0. The base distribution, G_0, represents the central tendency or "mean" distribution in the distribution space. The precision parameter, α, quantifies how far away realizations of G deviate from G_0. According to Ferguson (1973), DP is a conjugate prior that has two desirable properties: (1) a sufficiently large support, and (2) analytically manageable posterior distributions. Ferguson further derived the posterior of G, $DP(\tilde{\alpha}, \tilde{G_0})$. Here, $\tilde{\alpha} = \alpha + N$ and

$$\tilde{G_0} = \frac{\alpha}{\alpha + N} G_0 + \frac{N}{\alpha + N} G_N$$

with G_N being the empirical distribution of the data. Notably, the posterior point estimate of G, $E(G|data) = \tilde{G_0}$, is a weighted average of the base distribution or prior mean G_0 and the empirical distribution or data G_N. When $\alpha = 0$, the posterior point estimate is reduced to the empirical distribution G_N, which is pure non-parametric. When α approaches to infinity, the posterior point estimate gradually approximates G_0, which is parametric. A common practice is to specify a gamma prior for α, which would yield a posterior estimate that is neither 0 nor infinity.

In BNP growth curve modeling, latent variables and/or measurement errors can be modeled non-parametrically. In this article, we focus on the distributional assumption of measurement errors. When the normality of measurement errors is suspected, we assume that $\mathbf{e}_i \sim G_e$ where G_e is an unknown random distribution that is determined by the data. In the BNP framework, DP is typically adopted to specify G_e. Because the distribution of $\mathbf{e}_i$ is continuous but DP is essentially discrete, a DP mixture (DPM) can be used to model the measurement errors such that

$$G_e = \begin{cases} D(\boldsymbol{\mu}_e^{(1)}, \boldsymbol{\Phi}^{(1)}), & with\ p = p_1 \\ D(\boldsymbol{\mu}_e^{(2)}, \boldsymbol{\Phi}^{(2)}), & with\ p = p_2 \\ \vdots & \vdots \\ D(\boldsymbol{\mu}_e^{(k)}, \boldsymbol{\Phi}^{(k)}), & with\ p = p_k \\ \vdots & \vdots \end{cases},$$

where D represents a predetermined multivariate distribution (e.g., multivariate normal, t, multinomial, etc.), and $\boldsymbol{\mu}_e^{(k)}$ and $\boldsymbol{\Phi}^{(k)}, k = 1, \ldots, \infty$ are means and covariances of the multivariate distribution in the kth component with probability p_k. Theoretically, given an arbitrary distributional shape, there could be infinite number of mixture components as k goes to infinity. In practice, a finite number of mixture components often can describe a distribution well and the number of mixture components is determined by the DP precision parameter α. Smaller α yields a smaller number of mixture components. If α approaches infinity, there would be N mixture components, one associated with each subject. Namely, the precision parameter α is an important parameter that can determine the complexity of the model and how well the model fits the data, and thus may affect the convergence of the model. For the intraindividual measurement errors in the typical linear growth curve model, Tong and Zhang (2019) proposed that

$$\begin{aligned} \mathbf{e}_i|\boldsymbol{\Phi}_i &\sim MN_T(\mathbf{0}, \boldsymbol{\Phi}_i), \\ \boldsymbol{\Phi}_i|G &\sim G, \\ G &\sim DP(\alpha, G_0). \end{aligned}$$

That is, the unknown distribution G_e is approximated by a mixture of multivariate normal distributions where the mixing measure has a DP prior, $G_e \sim DPM$. The DP prior $DP(\alpha, G_0)$ can be obtained using the truncated stick-breaking construction (e.g., Sethuraman, 1994; Lunn et al., 2013). Specifically, $DP(\cdot) = \sum_{j=1}^{C} p_j \delta_{z_j}(\cdot), 1 \leq C < \infty$, where C ($1 \leq C \leq N$,often set at a large number) is a possible maximum number of mixture components, $\delta_{z_j}(\cdot)$ denotes a point mass at z_j and $z_j \sim G_0$ independently. The random weights p_j can be generated through the following procedure. With $q_1, q_2, \ldots, q_C \sim Beta(1, \alpha)$, define

$$p_j^{'} = q_j \prod_{k=1}^{j-1} (1 - q_k), j = 1, \ldots, C.$$

Then, p_j is obtained by

$$p_j = \frac{p_j^{'}}{\sum_{k=1}^{C} p_k^{'}}, \tag{1}$$

to satisfy that $\sum_{j=1}^{C} p_j = 1$. In practice, the updating of $\mathbf{e}_i$ can proceed as in a typical DP mixture model and its distribution is an infinite mixture distribution[1].

In general, the distribution of $\mathbf{e}_i$ through the truncated stick-breaking construction is

$$G_e = \begin{cases} D(\boldsymbol{\mu}_e^{(1)}, \boldsymbol{\Phi}^{(1)}), & with\ p = p_1 \\ D(\boldsymbol{\mu}_e^{(2)}, \boldsymbol{\Phi}^{(2)}), & with\ p = p_2 \\ \vdots & \vdots \\ D(\boldsymbol{\mu}_e^{(C)}, \boldsymbol{\Phi}^{(C)}), & with\ p = p_C \end{cases},$$

where D represents a predetermined multivariate distribution, $\boldsymbol{\mu}_e^{(j)}$ and $\boldsymbol{\Phi}^{(j)}, j = 1, \ldots, C$ are means and covariances of the multivariate distribution in the jth component, and p_j is obtained using Equation (1). Given that the mean of $\mathbf{e}_i$ is $\mathbf{0}$, we constrain $\sum_{j=1}^{C} p_j \boldsymbol{\mu}_e^{(j)} = \mathbf{0}$. For simplicity, in this study, we follow Tong and Zhang (2019) and use multivariate normal distributions for the mixing components and constrain $\boldsymbol{\mu}_e^{(j)}$ to be 0. We use inverse Wishart priors $p(\boldsymbol{\Phi}^{(j)}) = IW(n_0, W_0)$ for the covariance matrices of the mixture components, $\boldsymbol{\Phi}^{(j)}$, $j = 1, \ldots, C$. Following Lunn et al. (2013, p. 294), we fix the shape parameter n_0 at a specific number and assign an inverse Wishart prior to the scale matrix W_0. With such a specification, the measurement error for individual i, $\mathbf{e}_i$, has a p_j probability of coming from the mixing component $MN(\mathbf{0}, \boldsymbol{\Phi}^{(j)})$. The measurement errors for other individuals may also come from the same mixing component. Let K denotes the number of mixing components or $MN(\mathbf{0}, \boldsymbol{\Phi}^{(j)})$ with $j = 1, \ldots, C$. In other words, K is the number of latent classes for $\mathbf{e}_i$ and K can be smaller than C, $K \leq C$. Within each class, $\mathbf{e}_i$s come from the same distribution.

We would like to note that a similar approach to BNP modeling is finite mixture modeling (FMM). FMM estimates or equivalently approximates an unknown distribution using a mixture of known distributions. A key difference between FMM and BNP modeling is that the number of mixture components is treated as known in FMM, whereas this number is treated as unknown and is freely estimated in BNP modeling. As a result, when FMM is used to handle non-normality, additional analyses such as model comparison are needed to determine the unknown number of mixture components. BNP modeling therefore is believed to have the advantage of being more objective and data-driven, given that additional analyses such as model comparison that may be vulnerable to subjectivity are avoided.

Bayesian methods are applied to estimate BNP growth curve models. Bayesian methods are becoming increasingly popular in recent years because of their flexibility and powerfulness in estimating models with complex structures (e.g., Lee and Shi, 2000; Lee and Song, 2004; Zhang et al., 2007; Lee and Xia, 2008; Tong and Zhang, 2012; Serang et al., 2015). The key idea of Bayesian methods is to compute the posterior distributions for model parameters by combining the likelihood function and the priors. As introduced previously, $\boldsymbol{\beta}, \boldsymbol{\Phi}$, and $\boldsymbol{\Psi}$ are the model parameters in traditional growth curve model. In a BNP growth curve model, $\boldsymbol{\beta}$ and $\boldsymbol{\Psi}$ remain model parameters. In contrast, the measurement error covariance matrix $\boldsymbol{\Psi}$ is not directly estimated. Instead, we obtain $\mathbf{e}_i$ based on which we can get $\boldsymbol{\Phi}$. Another important parameter in BNP growth curve modeling is the precision parameter α. Let $p(\boldsymbol{\beta}, \boldsymbol{\Psi}, \alpha)$ be the joint prior distribution of model parameters, and let L be the likelihood function. The joint posterior distribution of model parameters is

$$p(\boldsymbol{\beta}, \boldsymbol{\Psi}, \alpha | \mathbf{y}_i) \propto \int p(\boldsymbol{\beta}, \boldsymbol{\Psi}, \alpha) \times L\, d\mathbf{b},$$

where $\mathbf{b} = (\mathbf{b}_1^{'}, \ldots, \mathbf{b}_N^{'})^{'}$. It is difficult to solve for this integral in practice. Instead, Markov chain Monte Carlo (MCMC) methods (e.g., Gibbs sampling; Robert and Casella, 2004) are often used to obtain parameter estimates and statistical inferences. Specifically, we first derive the conditional posterior distribution for each of the parameters. We then iteratively draw samples from the derived conditional posteriors to obtain empirical marginal distributions of the model parameters. Finally, statistical inferences are made based on the empirical marginal distributions (Geman and Geman, 1984).

3. PRECISION PARAMETER IN BNP MODELS

The convergence issue in BNP growth curve modeling is likely related to the precision parameter (Tong and Zhang, 2019). Here, we provide a theoretical discussion on how the prior of the precision parameter can influence the number of latent classes for $\mathbf{e}_i$.

The DP precision parameter α is the key to govern the expected number of latent classes. It directly determines the distribution of K, the number of latent classes of $\mathbf{e}_i$. With a larger K, measurement errors of different individuals are more likely to have different distributions. West (1992) found that K asymptotically follows a Poisson distribution

[1] In practice, infinite-dimension means finite but unbounded dimension.

TABLE 1 | Different percentiles (5, 50, 95%) of the distribution of the number of clusters K, given different values of precision parameter α, and sample size N.

	$\alpha = 0.1$			$\alpha = 1$			$\alpha = 2$		
	5%	50%	95%	5%	50%	95%	5%	50%	95%
$N = 200$	1	1	3	3	7	11	7	13	19
$N = 600$	1	2	3	4	8	13	9	15	21
$N = 1,000$	1	2	3	4	8	13	10	16	23

$$K = 1 + x,\ x \sim Poisson\left(\alpha\left(\gamma + logN\right)\right) \tag{2}$$

where γ is Euler's constant. Several percentiles of the distribution of K are given in **Table 1**. As shown in the table, K increases as α and N increases.

As discussed previously, a gamma prior $Gamma(a_1, a_2)$ is often used for the hyperparameter α. Given such a prior, West (1992) derived the posterior of α as a mixture of two gamma densities

$$\begin{aligned}\alpha|\cdot &\sim \pi_x Gamma(a_1 + K, a_2 - logx)\\ &+(1-\pi_x)Gamma(a_1 + K - 1, a_2 - logx),\end{aligned}$$

where x is an augmented variable $x|\cdot \sim Beta(\alpha + 1, N)$ and the weights π_x is defined by $\pi_x/(1-\pi_x) = \dfrac{a_1 + K - 1}{N(a_2 - logx)}$. Although West (1992) also provided an approximation to the posterior of α, $p(\alpha|\cdot) \approx Gamma(a_1 + K - 1, a_2 + \gamma + logN)$, how good the approximation was has not been investigated.

A non-informative prior for α seems to be reasonable, especially when the information about number of latent classes are not available. However, a non-informative prior may cause non-convergence of Markov chains. Therefore, it is worth evaluating different priors for the precision parameter.

4. A SIMULATION STUDY

We now present a simulation study to evaluate the influence of the prior for the precision parameter in BNP growth curve modeling when data are normally distributed and contain outliers[2]. The linear growth curve model in the previous section is used. Measurement errors are modeled non-parametrically to address the non-normality. Based on the results of previous studies, the number of times points (T), the covariance between the random intercept and slope (σ_{LS}), and the measurement error variance (σ_e^2) have trivial effects on the performance of BNP growth curve modeling (e.g., Tong and Zhang, 2019). Therefore, we only consider a set of values for these parameters in this study. We follow the empirical data analysis results in Tong and Zhang (2019) to select the population parameter values: the fixed effects are fixed at $\boldsymbol{\beta} = (\beta_L, \beta_S)' = (6.2, 0.3)'$; the number of measurement occasion is $T = 4$; measurement error variance $\sigma_e^2 = 0.5$; variances of the random intercept and slope are 1 and 0.1, respectively, and the covariance between the random intercept and slope $\sigma_{LS} = 0$.

Three potentially influential factors are manipulated in the simulation study, including sample size, data distribution, and precision parameter prior. First, two sample sizes are considered, N =200 or 600, representing small and large sample sizes. Second, data are either normal or containing outliers. When generating outliers, three proportions of outliers are considered, $r\%$ =5, 10, or 20%. To generate outliers, we randomly select $r\%$ observations at each measurement occasion and replace them by extreme values. The extreme values are generated from 10 different distributions with a large mean of $L_i + S_i(j-1) + m\sigma_e$ where $m \geq 5$ is generated from a truncated Poisson distribution, and a variance of σ_e^2 which is the same as that of the normal data. As a result, the true distribution of the data is a mixture of 11 distributions. Outliers generated in this way conform to the definition of outliers (Yuan and Zhong, 2008; Tong and Zhang, 2017). See **Supplementary Figures 1, 2** to aid the understanding of the shape of generated normal data and data with outliers. Third, four priors for the precision parameter are investigated (see **Figure 1**): a diffuse prior $Gamma(0.001, 0.001)$, a weakly informative prior $Gamma(2, 2)$ suggested by Ishwaran (2000), an accurate informative prior $Gamma(100, 100)$, and an inaccurate informative prior $Gamma(10, 100)$. Gamma(10,100) is an inaccurate informative prior because its mean is 0.1 and its variance is as small as 0.001. According to **Table 1**, the resulting number of latent classes ranges from 1 to 3, whereas the true number of mixed underlying distribution is 11. For all the other model parameters, conventional non-informative priors such as those in Zhang et al. (2013) are used. Specifically, fixed effects $\boldsymbol{\beta}$ have non-informative diffuse priors $N(0, 10^6)$. The covariance matrix of the random intercept and slope $\boldsymbol{\Psi}$ has an inverse-Wishart prior with an identity scale matrix and degrees of freedom being 2.

In each simulation condition, 500 datasets are generated. BNP growth curve modeling is applied for each dataset using JAGS with the rjags package in software R (Plummer, 2017; R Core Team, 2019). The total length of Markov chains is set at 50,000 and the first half of iterations is the burn-in period[3]. We assess how different priors affect model convergence rate, parameter estimation, and computation time.

Geweke tests (Geweke, 1991) are used to perform the convergence diagnostics. After the burn-in period, if parameter values are sampled from the stationary distribution of the chain,

[2] Note that non-normal data may be caused by non-normal population distributions or data contaminations. We work with outliers in this simulation study because BNP methods are essentially infinite mixture modeling procedures. Generating and dealing with outliers from multiple different distributions are more manageable as we easily know the true number of underlying classes. It is worth verifying the conclusions of this paper for non-normal population distributions in the future.

[3] Multiple lengths of Markov chains were tested before the current setting was selected. The convergence results with 50,000 iterations were about the same as those for longer chains.

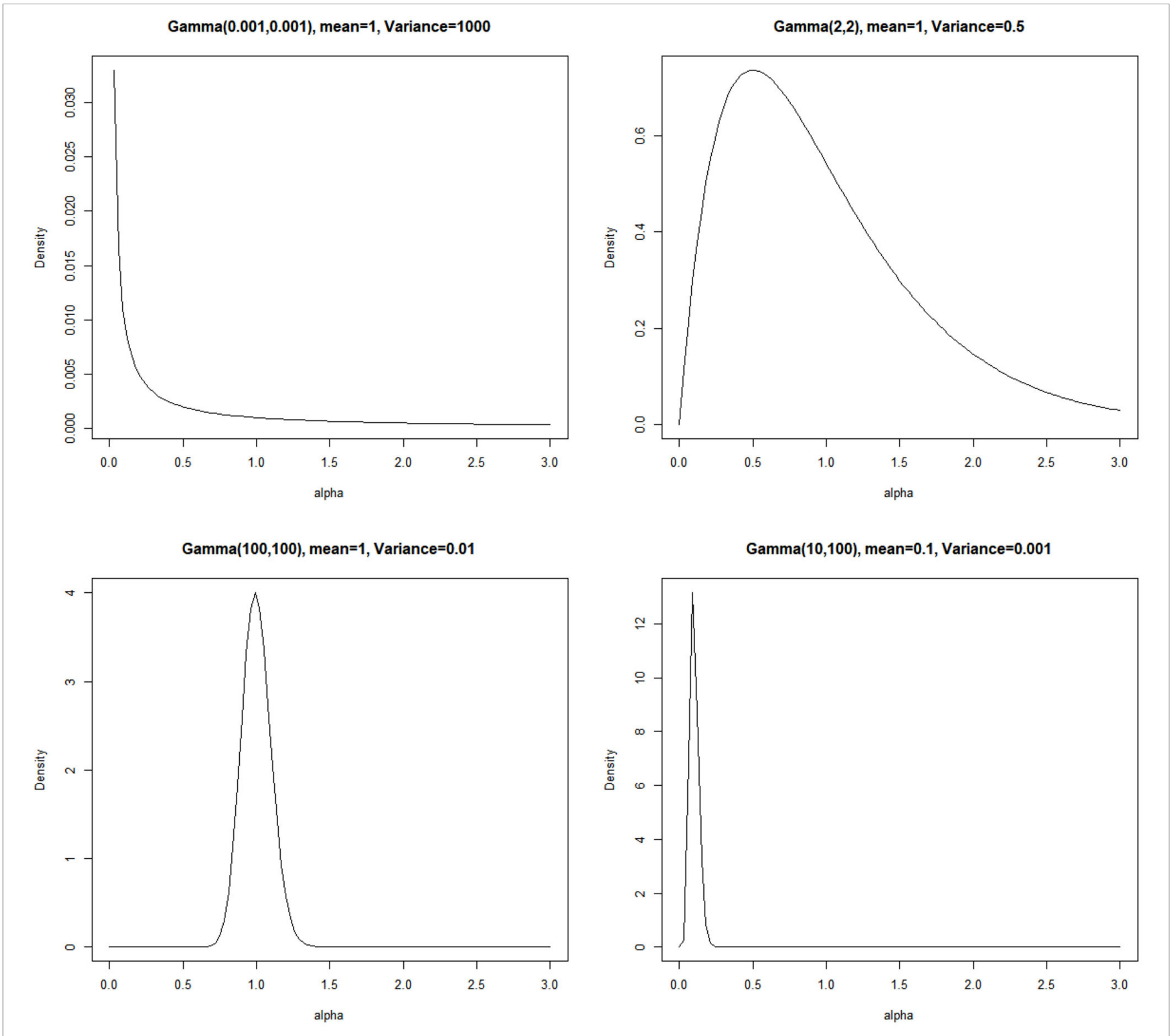


FIGURE 1 | Density curves for the four precision parameter priors used in the simulation study.

the means of the first and last parts of the Markov chain (by default the first 10% and the last 50%) should be equal and Geweke's statistic asymptotically follows a standard normal distribution. A Markov chain converges when the Geweke's statistic is between –1.96 and 1.96. If none of the convergence diagnostics (i.e., Geweke tests) for all model parameters suggest non-convergence, the model is said to have converged. In each simulation condition, the convergence rate is defined as the proportion of converged models out of the total 500 generated replications.

For the assessment of model estimation, we obtain the parameter estimate bias, average standard error (ASE), empirical standard error (ESE), mean squared error (MSE), and coverage probability (CP) of the 95% highest posterior density (HPD) credible intervals for each parameter based on converged simulation replications[4].

In addition, the estimation time (in seconds) is recorded for each replication. The average estimation time (AET) is the average of the estimation time for all the converged replications.

[4] ASE is the mean estimated standard error across replications. ESE is the standard deviation of the parameter estimates from all replications. MSE is computed as squared bias plus squared ESE. Posterior credible interval, also called credible interval, is the Bayesian counterpart of the frequentist confidence interval. A HPD interval is essentially the narrowest interval on a posterior that covers a given proportion of the probable posterior values.

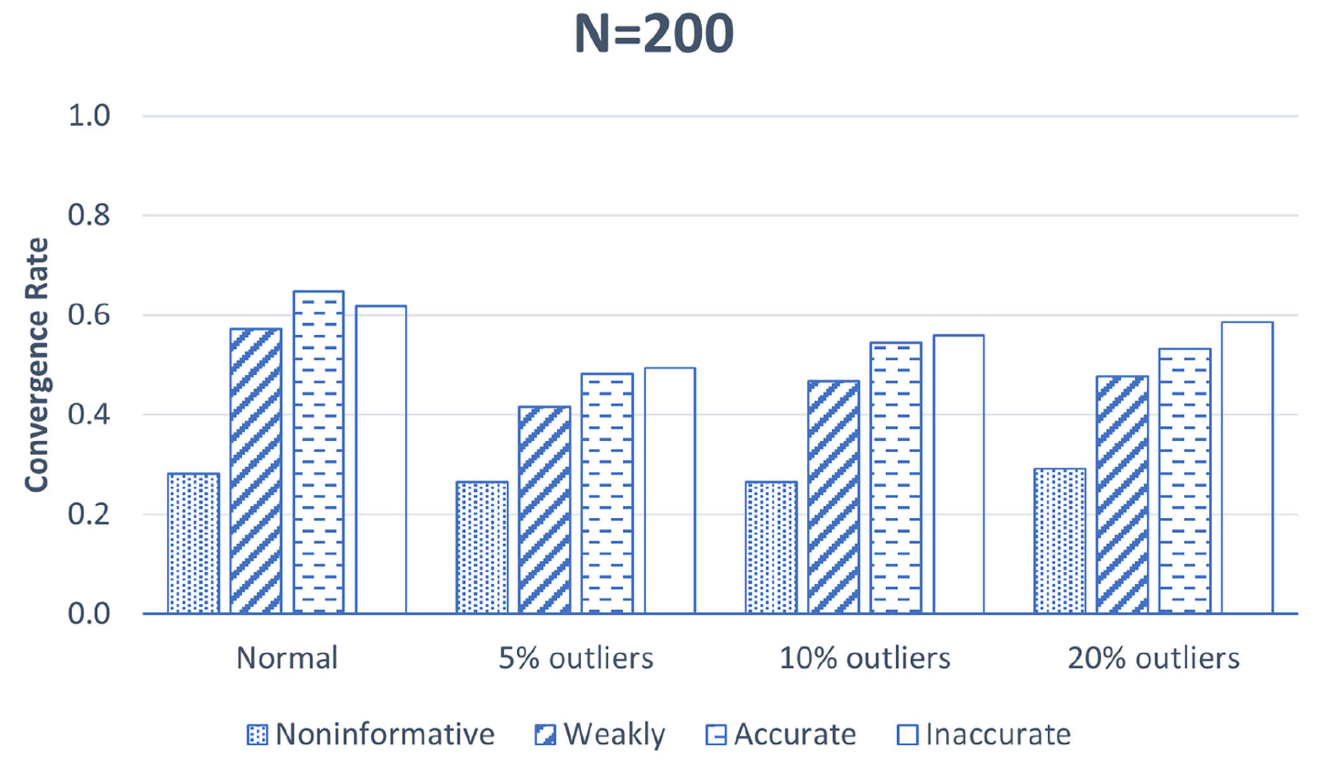


FIGURE 2 | Convergence rate for different priors when $N = 200$.

All program code and detailed results for the simulation study are available on our GitHub site: https://github.com/CynthiaXinTong/PrecisionParPrior_BNP_GCM.

4.1. Main Results

Figure 2 shows the convergence rate for BNP growth curve modeling with different precision parameter priors when sample size is 200. This figure clearly shows that outliers harm model convergence. Note that the convergence rate for data with 5% outliers is the lowest. This may be because a small proportion of outliers (e.g., 5%) creates a steep and high-curvature region for the Markov chain to enter and thus more difficult to converge. As the outlier proportion increases, the curvature becomes smoother so the convergence rate is higher. Among the four studied priors, the non-informative prior for the precision parameter always leads to the lowest convergence rate, i.e., less than 30% across all the simulation conditions. Informative priors substantially increase the model convergence rate. Specifically, the convergence rate doubles when we switch from the non-informative prior to the weakly informative prior suggested by Ishwaran (2000) in the condition with normal data. The incremental amount is about 30% of the original convergence rate in the conditions with outliers. Both accurate informative priors and inaccurate informative priors lead to higher convergence rates. The importance of using informative priors is more salient when data are not normal. Note that inaccurate informative priors yield slightly higher convergence rates than accurate informative priors because the variance of the inaccurate prior is lower and thus its precision is higher. When $N = 600$, model convergence results for BNP growth curve models follow the same pattern, and thus are not reported here.

For converged replications, we evaluate the impact of precision parameter priors on parameter estimation and computation time. Results for $N = 200$ are summarized in **Tables 2–5**. The relative performance of the four priors in conditions with a larger sample size ($N = 600$) has a similar pattern. Detailed results for $N = 600$ are available in the **Supplementary Document**.

From **Tables 2–5**, we obtain the following findings. First, the estimates of growth curve parameters $(\beta_L, \beta_S, \sigma_L^2, \sigma_S^2, \sigma_{LS}, \sigma_e^2)$ are not affected by different priors. Estimation bias, standard errors, MSE, and coverage probability of the 95% HPD credible interval across different precision parameter prior conditions are very close to each other, respectively. Note that when outliers exist (see **Tables 3–5**), the true population parameter value of the measurement error variance σ_e^2 is unknown. So, bias, MSE, and CP for this parameter cannot be calculated.

Second, the estimation of the hyperparameter α is greatly affected by different priors. When the non-informative prior is used, the estimated α can be very large (e.g., 28.284 in

TABLE 2 | Model estimation for BNP growth curve modeling with different precision parameter priors when data are normal and $N = 200$.

Prior		Est.	Bias	ASE	ESE	MSE	CP	AET
Gamma(0.001, 0.001)	β_L	6.204	0.004	0.082	0.084	0.007	0.957	539.332
	β_S	0.301	0.001	0.033	0.032	0.001	0.957	539.332
	σ_L^2	0.999	–0.001	0.138	0.142	0.020	0.936	539.332
	σ_S^2	0.118	0.018	0.021	0.018	0.001	0.922	539.332
	σ_{LS}	–0.010	–0.010	0.040	0.034	0.001	0.993	539.332
	σ_e^2	0.497	–0.003	0.024	0.036	0.001	0.816	539.332
	K	2.113	–	0.803	2.331	–	–	539.332
	α	11.134	–	18.109	104.805	–	–	539.332
Gamma(2, 2)	β_L	6.198	–0.002	0.082	0.080	0.006	0.958	740.331
	β_S	0.302	0.002	0.033	0.031	0.001	0.965	740.331
	σ_L^2	1.008	0.008	0.139	0.131	0.017	0.965	740.331
	σ_S^2	0.118	0.018	0.021	0.018	0.001	0.927	740.331
	σ_{LS}	–0.010	–0.010	0.040	0.034	0.001	0.983	740.331
	σ_e^2	0.499	–0.001	0.024	0.034	0.001	0.823	740.331
	K	4.106	–	2.415	0.776	–	–	740.331
	α	0.732	–	0.526	0.126	–	–	740.331
Gamma(100, 100)	β_L	6.200	0.000	0.083	0.083	0.007	0.948	1024.509
	β_S	0.299	-0.001	0.033	0.032	0.001	0.958	1024.509
	σ_L^2	1.014	0.014	0.139	0.133	0.018	0.967	1024.509
	σ_S^2	0.117	0.017	0.021	0.018	0.001	0.942	1024.509
	σ_{LS}	–0.010	–0.010	0.040	0.036	0.001	0.976	1024.509
	σ_e^2	0.499	–0.001	0.024	0.036	0.001	0.827	1024.509
	K	5.037	–	1.924	0.407	–	–	1024.509
	α	0.992	–	0.099	0.004	–	–	1024.509
Gamma(10, 100)	β_L	6.202	0.002	0.082	0.082	0.007	0.945	370.307
	β_S	0.301	0.001	0.033	0.031	0.001	0.971	370.307
	σ_L^2	1.001	0.001	0.138	0.129	0.017	0.971	370.307
	σ_S^2	0.117	0.017	0.021	0.018	0.001	0.942	370.307
	σ_{LS}	–0.012	–0.012	0.040	0.037	0.001	0.974	370.307
	σ_e^2	0.498	–0.002	0.024	0.035	0.001	0.835	370.307
	K	1.981	–	0.874	0.199	–	–	370.307
	α	0.099	-	0.031	0.001	–	–	370.307

Est, estimate; ASE, average standard error; ESE, empirical standard error; MSE, mean squared error; CP, coverage probability of the 95% HPD credible interval; AET, average estimation time.

Table 3) or small (e.g., 0.019 in **Table 5**), associating with a large standard error. When *Gamma*(2, 2) or *Gamma*(100, 100) is used, estimated α is almost always close to 1. When *Gamma*(10, 100) is used, estimated α is around 0.1. Different α values indicate a different total number of classes K. In general, a larger α value may yield a larger number of latent classes. Since the estimated α has a large standard error when the non-informative diffuse prior is used, the corresponding estimated K can also be large or small. For the weakly informative and accurate informative priors, the estimated number of latent classes ranges from 4 to 6 for different data conditions, whereas for the inaccurate informative prior, the estimated number of latent classes is about 2 or 3. It is interesting to see that although distinctively different hyperparameter estimates are obtained leading to different number of latent classes, the estimated growth curve parameters are essentially similar. This is because although outliers are generated from 10 different distributions, the 10 different distributions are not separated far apart. With a low class separation, one distribution may be enough to describe several outliers generated from different distributions. Thus, even the inaccurate informative prior can yield a precision parameter that is adequate to model the measurement errors.

Third, BNP growth curve modeling with the inaccurate informative prior *Gamma*(10, 100) requires the shortest computation time. This is because the inaccurate informative prior here has the smallest variance and thus is most "informative" among the four priors.

Fourth, outliers affect the performance of BNP growth curve modeling. When data contain a large proportion of outliers (e.g., 20%), estimation bias for the average of random intercepts β_L

TABLE 3 | Model estimation for BNP growth curve modeling with different precision parameter priors when data contain 5% of outliers and $N = 200$.

Prior		Est.	Bias	ASE	ESE	MSE	CP	AET
Gamma(0.001, 0.001)	β_L	6.300	0.100	0.092	0.083	0.017	0.793	841.706
	β_S	0.313	0.013	0.037	0.036	0.001	0.948	841.706
	σ_L^2	1.006	0.006	0.160	0.145	0.021	0.956	841.706
	σ_S^2	0.118	0.018	0.024	0.017	0.001	0.985	841.706
	σ_{LS}	–0.009	–0.009	0.046	0.044	0.002	0.985	841.706
	σ_e^2	3.133	–	0.125	0.138	–	–	841.706
	K	5.184	–	1.189	5.433	–	–	841.706
	α	28.284	–	51.922	75.124	–	–	841.706
Gamma(2, 2)	β_L	6.311	0.111	0.092	0.088	0.020	0.763	971.782
	β_S	0.311	0.011	0.037	0.034	0.001	0.957	971.782
	σ_L^2	1.007	0.007	0.161	0.146	0.021	0.967	971.782
	σ_S^2	0.117	0.017	0.024	0.018	0.001	0.976	971.782
	σ_{LS}	–0.008	–0.008	0.046	0.041	0.002	0.976	971.782
	σ_e^2	3.119	–	0.124	0.136	–	–	971.782
	K	6.515	–	2.905	0.981	- -	–	971.782
	α	1.126	–	0.676	0.171	–	–	971.782
Gamma(100, 100)	β_L	6.298	0.098	0.091	0.090	0.018	0.794	1088.448
	β_S	0.314	0.014	0.037	0.034	0.001	0.944	1088.448
	σ_L^2	0.987	–0.013	0.158	0.134	0.018	0.964	1088.448
	σ_S^2	0.117	0.017	0.024	0.018	0.001	0.976	1088.448
	σ_{LS}	–0.004	–0.004	0.045	0.041	0.002	0.992	1088.448
	σ_e^2	3.133	–	0.124	0.130	–	–	1088.448
	K	6.161	–	1.930	0.467	–	–	1088.448
	α	1.003	–	0.100	0.005	–	–	1088.448
Gamma(10, 100)	β_L	6.311	0.111	0.091	0.090	0.020	0.767	561.074
	β_S	0.311	0.011	0.037	0.034	0.001	0.952	561.074
	σ_L^2	0.985	–0.015	0.158	0.144	0.021	0.960	561.074
	σ_S^2	0.119	0.019	0.024	0.018	0.001	0.964	561.074
	σ_{LS}	–0.009	–0.009	0.046	0.042	0.002	0.968	561.074
	σ_e^2	3.118	–	0.124	0.136	–	–	561.074
	K	2.903	–	0.872	0.240	–	–	561.074
	α	0.103	–	0.032	0.001	–	–	561.074

Est, estimate; ASE, average standard error; ESE, empirical standard error; MSE, mean squared error; CP, coverage probability of the 95% HPD credible interval; AET, average estimation time.

and variance of random intercepts σ_L^2 are much larger than those when outlier proportion is low. In addition, outliers influence computation time. It is worth mentioning that it is most time consuming when the outlier proportion is 5%. A possible reason is that a small proportion of outliers creates a steep and high-curvature region for Markov chains to enter and thus takes longer time to converge. With more outliers, the curvature is smoother so the computation is faster.

5. DISCUSSION

Restricting to a parametric probability family can delude investigators and falsely make an illusion of posterior certainty (Müller and Mitra, 2004). On the contrary, BNP methods are adaptive and powerful to discover complex patterns in real data. Although BNP growth curve modeling has been proposed, the effect of the precision parameter was not fully studied. In this article, we have conducted a simulation study to investigate how different types of precision parameter priors impact the convergence rate, model estimation, and computation time in BNP growth curve modeling. We found that the non-informative prior suffered from the lowest convergence rates while the inaccurate informative prior with the smallest prior variance yielded the highest convergence rates and the fastest computations. Furthermore, we found that the estimation of growth curve parameters was not affected by the prior of the precision parameter. Based on these results, we recommend to use informative priors with high precision in practice.

We would like to note that although it seems counterintuitive that the inaccurate informative prior for the precision parameter

TABLE 4 | Model estimation for BNP growth curve modeling with different precision parameter priors when data contain 10% of outliers and $N = 200$.

Prior		Est.	Bias	ASE	ESE	MSE	CP	AET
Gamma(0.001, 0.001)	β_L	6.437	0.237	0.103	0.102	0.066	0.348	591.282
	β_S	0.335	0.035	0.043	0.039	0.003	0.917	591.282
	σ_L^2	1.018	0.018	0.187	0.173	0.030	0.977	591.282
	σ_S^2	0.126	0.026	0.028	0.022	0.001	0.917	591.282
	σ_{LS}	–0.007	–0.007	0.053	0.047	0.002	0.977	591.282
	σ_e^2	5.464	–	0.173	0.180	–	–	591.282
	K	3.112	–	1.106	1.728	–	–	591.282
	α	1.041	–	2.885	7.422	–	–	591.282
Gamma(2, 2)	β_L	6.424	0.224	0.103	0.105	0.061	0.426	938.496
	β_S	0.336	0.036	0.043	0.038	0.003	0.886	938.496
	σ_L^2	1.020	0.020	0.187	0.174	0.031	0.966	938.496
	σ_S^2	0.121	0.021	0.027	0.020	0.001	0.970	938.496
	σ_{LS}	–0.008	–0.008	0.053	0.044	0.002	0.979	938.496
	σ_e^2	5.448	–	0.172	0.171	–	–	938.496
	K	6.314	–	2.798	0.942	–	–	938.496
	α	1.090	–	0.652	0.163	–	–	938.496
Gamma(100, 100)	β_L	6.428	0.228	0.104	0.100	0.062	0.398	1045.439
	β_S	0.332	0.032	0.043	0.040	0.003	0.903	1045.439
	σ_L^2	1.020	0.020	0.188	0.172	0.030	0.964	1045.439
	σ_S^2	0.123	0.023	0.027	0.021	0.001	0.961	1045.439
	σ_{LS}	–0.009	–0.009	0.053	0.043	0.002	0.982	1045.439
	σ_e^2	5.459	–	0.174	0.175	–	–	1045.439
	K	6.091	–	1.911	0.409	–	–	1045.439
	α	1.002	–	0.100	0.004	–	–	1045.439
Gamma(10, 100)	β_L	6.426	0.226	0.103	0.102	0.062	0.395	389.282
	β_S	0.333	0.033	0.043	0.041	0.003	0.897	389.282
	σ_L^2	1.011	0.011	0.185	0.177	0.032	0.957	389.282
	σ_S^2	0.123	0.023	0.027	0.021	0.001	0.943	389.282
	σ_{LS}	–0.007	–0.007	0.052	0.045	0.002	0.975	389.282
	σ_e^2	5.457	–	0.172	0.169	–	–	389.282
	K	2.935	–	0.878	0.206	–	–	389.282
	α	0.103	–	0.032	0.001	–	–	389.282

Est, estimate; ASE, average standard error; ESE, empirical standard error; MSE, mean squared error; CP, coverage probability of the 95% HPD credible interval; AET, average estimation time.

performed the best, such findings have been observed in the literature. For example, Finch and Miller (2019) found that slightly informative priors can be advantageous in small samples even when these priors are incorrect. Depaoli (2013) showed that growth mixture model estimations obtained with inaccurate priors were still more accurate than maximum likelihood or Bayesian estimation with diffuse priors. Zitzmann et al. (2020) explicitly discussed this issue for small samples. Our simulation results also supported the argument that the amount of information in the prior can be more important than the accuracy of the prior under certain circumstances.

We also want to point out that the estimation bias was relatively large in our simulation study, when compared to that in previous studies (Tong and Zhang, 2019). This is because we consider much higher outlier proportions. When the outlier proportion is low (i.e., 5%), parameter estimates are very close to the true population values. As the outlier proportion increases, the bias increases. One possible way to improve the performance of BNP growth curve modeling when the outlier proportion is high is to use a non-normal base distribution. In our simulation study, for simplicity, we used normal distributions with zero mean as the mixing components of BNP modeling. This cannot handle asymmetric non-normal distributions, which may partly explain the less satisfactory performance of BNP modeling in the conditions with high outlier proportions. But BNP methods in general are very flexible. A non-normal base distribution may overcome this limitation. While future studies may continue along this path, we want to emphasize that BNP modeling as in our study still outperforms traditional growth curve modeling and is recommended to use in general when data are suspected

TABLE 5 | Model estimation for BNP growth curve modeling with different precision parameter priors when data contain 20% of outliers and $N = 200$.

Prior		Est.	Bias	ASE	ESE	MSE	CP	AET
Gamma(0.001, 0.001)	β_L	6.890	0.690	0.149	0.120	0.490	0.000	460.170
	β_S	0.385	0.085	0.061	0.054	0.010	0.735	460.170
	σ_L^2	1.321	0.321	0.315	0.284	0.183	0.884	460.170
	σ_S^2	0.141	0.041	0.038	0.027	0.002	0.952	460.170
	σ_{LS}	0.019	0.019	0.080	0.062	0.004	0.980	460.170
	σ_e^2	9.238	–	0.242	0.258	–	–	460.170
	K	2.713	–	0.810	0.307	–	–	460.170
	α	0.019	–	0.047	0.089	–	–	460.170
Gamma(2, 2)	β_L	6.890	0.690	0.150	0.120	0.490	0.000	949.186
	β_S	0.381	0.081	0.061	0.052	0.009	0.787	949.186
	σ_L^2	1.358	0.358	0.321	0.279	0.206	0.879	949.186
	σ_S^2	0.143	0.043	0.038	0.024	0.002	0.962	949.186
	σ_{LS}	0.011	0.011	0.082	0.064	0.004	0.983	949.186
	σ_e^2	9.167	–	0.245	0.265	–	–	949.186
	K	5.458	–	2.392	0.564	–	–	949.186
	α	0.941	–	0.566	0.095	–	–	949.186
Gamma(100, 100)	β_L	6.882	0.682	0.149	0.121	0.480	0.000	1056.953
	β_S	0.381	0.081	0.061	0.054	0.010	0.774	1056.953
	σ_L^2	1.323	0.323	0.314	0.284	0.185	0.878	1056.953
	σ_S^2	0.143	0.043	0.038	0.026	0.003	0.944	1056.953
	σ_{LS}	0.010	0.010	0.081	0.062	0.004	0.981	1056.953
	σ_e^2	9.172	–	0.243	0.256	–	–	1056.953
	K	5.695	–	1.811	0.321	–	–	1056.953
	α	0.998	–	0.099	0.003	–	–	1056.953
Gamma(10, 100)	β_L	6.897	0.697	0.150	0.116	0.499	0.000	391.429
	β_S	0.379	0.079	0.061	0.052	0.009	0.803	391.429
	σ_L^2	1.354	0.354	0.319	0.280	0.204	0.861	391.429
	σ_S^2	0.141	0.041	0.038	0.026	0.002	0.956	391.429
	σ_{LS}	0.014	0.014	0.081	0.064	0.004	0.980	391.429
	σ_e^2	9.166	–	0.242	0.255	–	–	391.429
	K	2.880	–	0.855	0.151	–	–	391.429
	α	0.103	-	0.032	0.001	–	–	391.429

Est, estimate; ASE, average standard error; ESE, empirical standard error; MSE, mean squared error; CP, coverage probability of the 95% HPD credible interval; AET, average estimation time.

to be non-normal (Tong and Zhang, 2019) no matter the non-normality is caused by non-normal population distribution or data contamination.

The convergence rate of BNP growth curve modeling was found to be higher in previous studies, i.e., close to one (Tong and Zhang, 2019). We would like to note that the difference is likely due to the list of parameters counted during convergence assessment. In Tong and Zhang (2019), the convergence rate was computed only for growth curve parameters. When only growth curve parameters are considered, non-convergence rarely occurred in our study. The major problem is the precision parameter. As shown in the simulation study, non-convergence frequently arose for this parameter (detailed Geweke tests results for each parameter are available on our GitHub site: https://github.com/CynthiaXinTong/PrecisionParPrior_BNP_GCM). Another possible reason why convergence rates were relatively low (below 70%) in our simulation is that Geweke tests often yield lower rates of convergence than other diagnostic methods (e.g., Jang and Cohen, 2020). However, as pointed out in Jang and Cohen, the pattern of convergence rates for model comparison was similar for different diagnostic tests. Namely, our conclusions about which precision parameter priors to use in BNP growth curve modeling will not be affected by the diagnostic tests. We further discuss the use of Geweke tests in the next paragraph. Notably, although the non-convergence for the precision parameter seemed not to impact parameter estimates for the growth curve parameters, such issue may mislead model fit assessment. Although model assessment and model comparison methods have been proposed for various models, samples of different sizes, and data structures (e.g., Celeux et al., 2006), their performance in BNP analysis has not been studied. Therefore, future studies on how different precision parameter priors affect model fit assessment are encouraged.

In our study, model convergence diagnostics were conducted using Geweke tests. Although Geweke tests are commonly used in the Bayesian literature, it is impossible to say with certainty that a finite sample from an MCMC algorithm is representative of an underlying stationary distribution and a combination of strategies aiming at evaluating and accelerating MCMC sampler convergence is recommended (Cowles and Carlin, 1996). For our simulation study, Geweke tests were relatively easy to systematically implement. In empirical studies, we recommend using multiple strategies (e.g., trace plots, multiple chains) to check model convergence. In addition, since Zitzmann and Hecht (2019) pointed out that it is possible that the approximation of the Bayesian estimates is still not optimal even when a chain converges, we recommend substantive researchers conducting sensitivity analysis and evaluating how the length of the Markov chains affects the model estimation results.

Our study echoed the previous literature in that using informative priors may help reduce computation time in Bayesian modeling. We would like to note that there are other approaches that can be used to further increase the computation efficiency. For example, Berger et al. (2020) and Daniels and Kass (1999) proposed shrinkage priors, and Hecht et al. (2020) proposed a model reformulation approach in which the sample covariance matrix was modeled instead of individual observations. This latter approach has been applied to the Bayesian continuous-time model (Hecht and Zitzmann, 2020) as well as the Bayesian STARTS model (Ludtke et al., 2018). Future research on BNP growth curve modeling could incorporate this approach and other potentially efficient approaches to reduce computation time.

The employment of BNP growth curve modeling is a field still in its early stage. New DP variants and generalizations are being proposed every year to cater to specific applications. BNP modeling was only used to handle the non-normality in intraindividual measurement errors in our study. The similar strategy can be used for random effects, such as random intercepts and slopes. Also, although we worked with balanced data, BNP growth curve modeling should be able to handle unbalanced data (e.g., individually varying time points). However, as implied by previous studies (Tong, 2014), the convergence issue may be more challenging, thereby awaiting future studies.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article are available by running the data generation R code on our GitHub site: https://github.com/CynthiaXinTong/PrecisionParPrior_BNP_GCM.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

FUNDING

This paper is based upon work supported by the National Science Foundation under grant no. SES-1951038.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.624588/full#supplementary-material

REFERENCES

Ansari, A., and Iyengar, R. (2006). Semiparametric Thurstonian models for recurrent choices: A Bayesian analysis. *Psychometrika* 71, 631–657. doi: 10.1007/s11336-006-1233-5

Berger, J. O., Sun, D., and Song, C. (2020). Bayesian analysis of the covariance matrix of a multivariate normal distribution with a new class of priors. *Ann. Stat.* 48, 2381–2403. doi: 10.1214/19-AOS1891

Brown, E. R., and Ibrahim, J. G. (2003). A Bayesian semiparametric joint hierarchical model for longitudinal and survival data. *Biometrics* 59, 221–228. doi: 10.1111/1541-0420.00028

Burr, D., and Doss, H. (2005). A Bayesian semiparametric model for random-effects meta-analysis. *J. Am. Stat. Assoc.* 100, 242–251. doi: 10.1198/016214504000001024

Bush, C. A., and MacEachern, S. N. (1996). A semiparametric Bayesian model for randomised block designs. *Biometrika* 83, 275–285. doi: 10.1093/biomet/83.2.275

Cain, M. K., Zhang, Z., and Yuan, K.-H. (2017). Univariate and multivariate skewness and kurtosis for measuring nonnormality: prevalence, influence and estimation. *Behav. Res. Methods* 49, 1716–1735. doi: 10.3758/s13428-016-0814-1

Celeux, G., Forbes, F., Robert, C. P., and Titterington, D. M. (2006). Deviance information criteria for missing data models. *Bayesian Anal.* 1, 651–673. doi: 10.1214/06-ba122

Chou, C.-P., Bentler, P. M., and Satorra, A. (1991). Scaled test statistics and robust standard errors for non-normal data in covariance structure analysis: a monte carlo study. *Br. J. Math. Stat. Psychol.* 44, 347–357. doi: 10.1111/j.2044-8317.1991.tb00966.x

Cowles, M. K., and Carlin, B. P. (1996). Markov chain monte carlo convergence diagnostics: a comparative review. *J. Am. Stat. Assoc.* 91, 883–904.

Curran, P. J., West, S. G., and Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychol. Methods* 1, 16–29. doi: 10.1037/1082-989x.1.1.16

Daniels, M. J., and Kass, R. E. (1999). Nonconjugate Bayesian estimation of covariance matrices and its use in hierarchical models. *J. Am. Stat. Assoc.* 94, 1254–1263. doi: 10.2307/2669939

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

Dunson, D. B. (2006). Bayesian dynamic modeling of latent trait distributions. *Biostatistics* 7, 551–568. doi: 10.1093/biostatistics/kxj025

Fahrmeir, L., and Raach, A. (2007). A Bayesian semiparametric latent variable model for mixed responses. *Psychometrika* 72, 327–346. doi: 10.1007/s11336-007-9010-7

Ferguson, T. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Stat.* 1, 209–230. doi: 10.1214/aos/1176342360

Ferguson, T. (1974). Prior distributions on spaces of probability measures. *Ann. Stat.* 2, 615–629. doi: 10.1214/aos/1176342752

Finch, W. H., and Miller, J. E. (2019). The use of incorrect informative priors in the estimation of MIMIC model parameters with small sample sizes. *Struct. Equat. Model.* 26, 497–508. doi: 10.1080/10705511.2018.1553111

Geman, S., and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.* 6, 721–741. doi: 10.1016/b978-0-08-051581-6.50057-x

Gershman, S. J., and Blei, D. M. (2012). A tutorial on Bayesian nonparametric models. *J. Math. Psychol.* 56, 1–12. doi: 10.1016/j.jmp.2011.08.004

Geweke, J. (1991). "Evaluating the accuracy of sampling-based approaches to calculating posterior moments," in *Bayesian Statistics 4*, eds J. M. Bernado, J. O. Berger, A. P. Dawid, and A. F. M. Smith (Oxford: Clarendon Press), 169–193. doi: 10.21034/sr.148

Ghosal, S., Ghosh, J., and Ramamoorthi, R. (1999). Posterior consistency of Dirichlet mixtures in density estimation. *Ann. Stat.* 27, 143–158. doi: 10.1214/aos/1018031105

Hecht, M., Gische, C., Vogel, D., and Zitzmann, S. (2020). Integrating out nuisance parameters for computationally more efficient Bayesian estimation - An illustration and tutorial. *Struct. Equat. Model.* 27, 483–493. doi: 10.1080/10705511.2019.1647432

Hecht, M., and Zitzmann, S. (2020). A computationally more efficient Bayesian approach for estimating continuous-time models. *Struct. Equat. Model.* 27, 829–840. doi: 10.1080/10705511.2020.1719107

Hjort, N. L. (2003). "Topics in nonparametric Bayesian statistics," in *Highly Structured Stochastic Systems*, eds P. Green, N. L. Hjort, and S. Richardson (Oxford: Oxford University Press), 455–487.

Hjort, N. L., Holmes, C., Müller, P., and Walker, S. G. (2010). *Bayesian Nonparametrics*. Cambridge: Cambridge University Press. doi: 10.1093/acprof:oso/9780199695607.003.0013

Ishwaran, H. (2000). "Inference for the random effects in Bayesian generalized linear mixed models," in *ASA Proceedings of the Bayesian Statistical Science Section*, ed A. S. Association (London), 1–10.

Jang, Y., and Cohen, A. S. (2020). The impact of Markov chain covergence on estimation of mixture IRT model parameters. *Educ. Psychol. Meas.* 80, 975–994. doi: 10.1177/0013164419898228

Jara, A., Hanson, T., Quintana, F., Müller, P., and Rosner, G. (2011). DP-package: Bayesian semi and nonparametric modeling in R. *J. Stat. Softw.* 40, 1–30. doi: 10.18637/jss.v040.i05

Kleinman, K. P., and Ibrahim, J. G. (1998). A semiparametric Bayesian approach to the random effects model. *Biometrics* 54, 921–938. doi: 10.2307/2533846

Lee, S. Y., Lu, B., and Song, X. Y. (2008). Semiparametric Bayesian analysis of structural equation models with fixed covariates. *Stat. Med.* 27, 2341–2360. doi: 10.1002/sim.3098

Lee, S. Y., and Shi, J. Q. (2000). Joint Bayesian analysis of factor scores and structural parameters in the factor analysis model. *Ann. Inst. Stat. Math.* 52, 722–736. doi: 10.1023/a:1017529427433

Lee, S. Y., and Song, X. Y. (2004). Bayesian model comparison of nonlinear structural equation models with missing continuous and ordinal categorical data. *Br. J. Math. Stat. Psychol.* 57, 131–150. doi: 10.1348/000711004849204

Lee, S. Y., and Xia, Y. M. (2008). A robust bayesian approach for structural equation models with missing data. *Psychometrika* 73, 343–364. doi: 10.1007/s11336-008-9060-5

Lu, Z., and Zhang, Z. (2014). Robust growth mixture models with non-ignorable missingness: models, estimation, selection, and application. *Comput. Stat. Data Anal.* 71, 220–240. doi: 10.1016/j.csda.2013.07.036

Ludtke, O., Robitzsch, A., and Wagner, J. (2018). More stable estimation of the STARTS model: a Bayesian approach using Markov Chain Monte Carlo techniques. *Psychol. Methods* 23, 570–593. doi: 10.1037/met0000155

Lunn, D., Jackson, C., Best, N., Thomas, A., and Spiegelhalter, D. (2013). *The BUGS Book: A Practical Introduction to Bayesian Analysis*. Boca Raton, FL: CRC Press.

MacEachern, S. (1999). "Dependent nonparametric processes," in *ASA Proceedings of the Section on Bayesian Statistical Science*, ed A. S. Association (Alexandria, VA).

Maronna, R. A., Martin, R. D., and Yohai, V. J. (2006). *Robust Statistics: Theory and Methods*. New York, NY: John Wiley & Sons, Inc. doi: 10.1002/0470010940

McArdle, J. J., and Nesselroade, J. R. (2014). *Longitudinal Data Analysis Using Structural Equation Models*. Washington, DC: American Psychological Association. doi: 10.1037/14440-000

Meredith, W., and Tisak, J. (1990). Latent curve analysis. *Psychometrika* 55, 107–122. doi: 10.1007/bf02294746

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychol. Bull.* 105, 156–166. doi: 10.1037/0033-2909.105.1.156

Müller, P., and Mitra, R. (2004). Bayesian nonparametric inference - why and how. *Bayesian Anal.* 1, 1–33. doi: 10.1214/13-ba811

Muthén, B., and Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics* 55, 463–469. doi: 10.1111/j.0006-341X.1999.00463.x

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *J. Comput. Graph. Stat.* 9, 249–265. doi: 10.2307/1390653

Ohlssen, D. I., Sharples, L. D., and Spiegelhalter, D. (2007). Flexible random-effects models using Bayesian semi-parametric models: applications to institutional comparisons. *Stat. Med.* 26, 2088–2112. doi: 10.1002/sim.2666

Pendergast, J. F., and Broffitt, J. D. (1985). Robust estimation in growth curve models. *Commun. Stat. Theor. Methods* 14, 1919–1939. doi: 10.1080/03610928508829021

Plummer, M. (2017). *Jags Version 4.3. 0 User Manual*.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Robert, C. P., and Casella, G. (2004). *Monte Carlo Statistical Methods*. New York, NY: Springer. doi: 10.1007/978-1-4757-4145-2

Serang, S., Zhang, Z., Helm, J., Steele, J. S., and Grimm, K. J. (2015). Evaluation of a Bayesian approach to estimating nonlinear mixed-effects mixture models. *Struct. Equat. Model.* 22, 202–215. doi: 10.1080/10705511.2014.937322

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Stat. Sin.* 4, 639–650.

Sharif-Razavian, N., and Zollmann, A. (2009). *An Overview of Nonparametric Bayesian Models and Applications to Natural Language Processing*. Available online at: http://www.cs.cmu.edu/?simzollmann/publications/nonparametric.pdf

Si, Y., and Reiter, J. P. (2013). Nonparametric Bayesian multiple imputation for incomplete categorical variables in large-scale assessment surveys. *J. Educ. Behav. Stat.* 38, 499–521. doi: 10.3102/1076998613480394

Si, Y., Reiter, J. P., and Hillygus, D. S. (2015). Semi-parametric selection models for potentially non-ignorable attrition in panel studies with refreshment samples. *Polit. Anal.* 23, 92–112. doi: 10.1093/pan/mpu009

Silvapulle, M. J. (1992). On M-methods in growth curve analysis with asymmetric errors. *J. Stat. Plan. Inference* 32, 303–309. doi: 10.1016/0378-3758(92)90013-i

Singer, J. M., and Sen, P. K. (1986). M-methods in growth curve analysis. *J. Stat. Plan. Inference* 13, 251–261. doi: 10.1016/0378-3758(86)90137-0

Tong, X. (2014). *Robust semiparametric bayesian methods in growth curve modeling* (Unpublished doctoral dissertation). University of Notre Dame, Notre Dame, IN, United States.

Tong, X., and Zhang, Z. (2012). Diagnostics of robust growth curve modeling using Student's t distribution. *Multivariate Behav. Res.* 47, 493–518. doi: 10.1080/00273171.2012.692614

Tong, X., and Zhang, Z. (2017). Outlying observation diagnostics in growth curve modeling. *Multivariate Behav. Res.* 52, 768–788. doi: 10.1080/00273171.2017.1374824

Tong, X., and Zhang, Z. (2019). Robust Bayesian approaches in growth curve modeling: using Student's t distributions versus a semiparametric method. *Struct. Equat. Model.* 27, 544–560. doi: 10.1080/10705511.2019.1683014

West, M. (1992). *Hyperparameter Estimation in Dirichlet Process Mixture Models*. Technical Report 92-A03, Duke University, ISDS.

Yang, M., and Dunson, D. B. (2010). Bayesian semiparametric structural equation models with latent variables. *Psychometrika* 75, 675–693. doi: 10.1007/s11336-010-9174-4

Yuan, K.-H., and Bentler, P. M. (1998). Structural equation modeling with robust covariances. *Sociol. Methodol.* 28, 363–396. doi: 10.1111/0081-1750.00052

Yuan, K.-H., and Bentler, P. M. (2001). Effect of outliers on estimators and tests in covariance structure analysis. *Br. J. Math. Stat. Psychol.* 54, 161–175. doi: 10.1348/000711001159366

Yuan, K.-H., and Zhong, X. (2008). Outliers, high-leverage observations and influential cases in factor analysis: Minimizing their effect using robust procedures. *Sociol. Methodol.* 38, 329–368. doi: 10.1111/j.1467-9531.2008.00198.x

Zhang, Z. (2016). Modeling error distributions of growth curve models through Bayesian methods. *Behav. Res. Methods* 48, 427–444. doi: 10.3758/s13428-015-0589-9

Zhang, Z., Hamagami, F., Wang, L., Grimm, K. J., and Nesselroade, J. R. (2007). Bayesian analysis of longitudinal data using growth curve models. *Int. J. Behav. Dev*. 31, 374–383. doi: 10.1177/0165025407077764

Zhang, Z., Lai, K., Lu, Z., and Tong, X. (2013). Bayesian inference and application of robust growth curve models using Student's t distribution. *Struct. Equat. Model.* 20, 47–78. doi: 10.1080/10705511.2013.742382

Zhong, X., and Yuan, K.-H. (2010). "Weights," in *Encyclopedia of Research Design*, ed N. J. Salkind (Thousand Oaks, CA: Sage), 1617–1620. doi: 10.4135/9781412961288

Zhong, X., and Yuan, K.-H. (2011). Bias and efficiency in structural equation modeling: maximum likelihood versus robust methods. *Multivariate Behav. Res.* 46, 229–265. doi: 10.1080/00273171.2011.558736

Zitzmann, S., and Hecht, M. (2019). Going beyond convergence in Bayesian estimation: why precision matters too and how to assess it. *Struct. Equat. Model.* 26, 646–661. doi: 10.1080/10705511.2018.1545232

Zitzmann, S., Ludtke, O., Robitzsch, A., and Hecht, M. (2020). On the performance of Bayesian approaches in small samples: a comment on Smid, McNeish, Miocevic, and van de Schoot. *Struct. Equat. Model.* 28, 40–50. doi: 10.1080/10705511.2020.1752216

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 29 April 2021
doi: 10.3389/fpsyg.2021.615162

A Comparison of Penalized Maximum Likelihood Estimation and Markov Chain Monte Carlo Techniques for Estimating Confirmatory Factor Analysis Models With Small Sample Sizes

Oliver Lüdtke[1,2]*, *Esther Ulitzsch*[1] *and Alexander Robitzsch*[1,2]

[1] *IPN – Leibniz Institute for Science and Mathematics Education, Kiel, Germany,* [2] *Centre for International Student Assessment, Kiel, Germany*

With small to modest sample sizes and complex models, maximum likelihood (ML) estimation of confirmatory factor analysis (CFA) models can show serious estimation problems such as non-convergence or parameter estimates outside the admissible parameter space. In this article, we distinguish different Bayesian estimators that can be used to stabilize the parameter estimates of a CFA: the mode of the joint posterior distribution that is obtained from penalized maximum likelihood (PML) estimation, and the mean (EAP), median (Med), or mode (MAP) of the marginal posterior distribution that are calculated by using Markov Chain Monte Carlo (MCMC) methods. In two simulation studies, we evaluated the performance of the Bayesian estimators from a frequentist point of view. The results show that the EAP produced more accurate estimates of the latent correlation in many conditions and outperformed the other Bayesian estimators in terms of root mean squared error (RMSE). We also argue that it is often advantageous to choose a parameterization in which the main parameters of interest are bounded, and we suggest the four-parameter beta distribution as a prior distribution for loadings and correlations. Using simulated data, we show that selecting weakly informative four-parameter beta priors can further stabilize parameter estimates, even in cases when the priors were mildly misspecified. Finally, we derive recommendations and propose directions for further research.

Keywords: measurement error, latent variable models, Bayesian methods, prior distribution, Markov Chain Monte Carlo, penalized maximum likelihood estimation, constrained maximum likelihood estimation, confirmatory factor analysis

OPEN ACCESS

Edited by:
Rens Van De Schoot,
Utrecht University, Netherlands

Reviewed by:
Sara Van Erp,
Utrecht University, Netherlands
Shaobo Jin,
Uppsala University, Sweden

****Correspondence:***
Oliver Lüdtke
oluedtke@ipn.uni-kiel.de

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *08 October 2020*
Accepted: *29 March 2021*
Published: *29 April 2021*

Citation:
Lüdtke O, Ulitzsch E and Robitzsch A (2021) A Comparison of Penalized Maximum Likelihood Estimation and Markov Chain Monte Carlo Techniques for Estimating Confirmatory Factor Analysis Models With Small Sample Sizes. Front. Psychol. 12:615162. doi: 10.3389/fpsyg.2021.615162

INTRODUCTION

In the social and behavioral sciences, constructs (e.g., intelligence, extraversion) are often conceptualized as latent variables that are measured by error-prone observed indicators (e.g., items). Structural equation modeling (SEM) is a very prominent approach that is used to correct for measurement error when assessing multivariate relationships among latent constructs (Bollen, 1989; Hoyle, 2012). In the SEM approach, a measurement part is distinguished from a structural part. In the measurement part, measurement models are specified to allow for an error-free estimation of the relations in the structural model. In research practice, maximum likelihood (ML) estimation is routinely used to obtain parameter estimates for structural equation models. However, one major limitation of ML estimation is that it needs large sample sizes to reveal its optimal properties (e.g., unbiasedness, efficiency). With small to modest sample sizes and complex models, ML estimation can show serious estimation problems such as non-convergence or parameter estimates that are outside the admissible parameter space (e.g., negative variances; see Anderson and Gerbing, 1984; Boomsma, 1985; Hoogland and Boomsma, 1998; Chen et al., 2001; Gagné and Hancock, 2006; Wolf et al., 2013; Smid and Rosseel, 2020).

In the last decades, several researchers have shown that the Bayesian approach has the potential to solve some of the estimation problems that occur in small sample applications of SEM (e.g., Lee, 2007; Song and Lee, 2009; Kaplan and Depaoli, 2012; Muthén and Asparouhov, 2012). First, if appropriate prior distributions are used, the Bayesian approach guarantees that parameter estimates will be within the admissible range, and estimation problems can usually be avoided. Second, Bayesian methods allow for the stabilization of parameter estimates by specifying weakly informative prior distributions for the SEM parameters (Lee and Song, 2004; Chen et al., 2014; Can et al., 2015; Depaoli and Clifton, 2015; McNeish, 2016; Lüdtke et al., 2018; van Erp et al., 2018; Miocevic et al., 2020). The basic idea is that incorporating even a small amount of information into the prior distribution of the SEM parameters provides some direction for their estimation, while inferences can still be driven by the data (Gelman et al., 2014).

In this article, we focus on the estimation of confirmatory factor analysis (CFA) models in which several latent factors are measured by a set of observed variables (Bollen, 1989). We investigate two critical issues in the Bayesian estimation of CFA models with small sample sizes. First, we discuss different Bayesian point estimators that can be used as estimates for CFA model parameters: the mode of the joint posterior, and the mode, mean, or median of the marginal posterior. Furthermore, we clarify that two popular methods for calculating Bayesian point estimates, penalized maximum likelihood (PML) estimation and Markov Chain Monte Carlo (MCMC) methods, produce different Bayesian point estimates and compare the performance of the different Bayesian point estimators to the traditional ML estimation of CFA models. Second, we discuss the specification of prior distributions in the Bayesian approach and argue that it can be advantageous to choose a parameterization in which the model parameters (i.e., standardized loadings, latent correlations) are bounded. More specifically, we suggest the four-parameter beta distribution as a prior distribution for bounded parameters (see also Muthén and Asparouhov, 2012; Merkle and Rosseel, 2018) and investigate in a simulation study how the specification of weakly informative prior distributions can help to stabilize parameter estimates in small sample size conditions.

The article is organized as follows. We start by describing how a basic CFA model is estimated with traditional ML estimation. We then discuss the specification of CFA models in the Bayesian approach and describe different Bayesian estimators that can be used to estimate CFA model parameters. In the context of a CFA model with two latent factors, we discuss issues of parameterization and the specification of prior distributions, and we illustrate conditions under which the different Bayesian estimators produce different results. We then present the results of two simulation studies in which we compare traditional ML estimation with the Bayesian approach. In the first simulation study, we evaluate the influence of correctly and misspecified prior distributions on the quality of parameter estimates in small sample size conditions. In the second simulation study, we investigate the robustness of the Bayesian approach against distributional misspecifications (i.e., non-normality). Finally, we derive recommendations and propose directions for further research.

CONFIRMATORY FACTOR ANALYSIS

Let $\mathbf{x}$ denote a vector of p observed variables. Then, a CFA model with m latent factors is represented as follows:

$$\mathbf{x} = \boldsymbol{\nu} + \boldsymbol{\Lambda}\boldsymbol{\eta} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\boldsymbol{\nu}$ is a $p \times 1$ vector containing intercepts, $\boldsymbol{\Lambda}$ is a $p \times m$ matrix of factor loadings, $\boldsymbol{\eta}$ is an $m \times 1$ vector of latent factors, and $\boldsymbol{\varepsilon}$ denotes the vector of multivariate normally distributed residuals with zero mean vector and covariance matrix $\boldsymbol{\Omega}$. In the following, we assume that the mean structure is saturated and completely reflected in the intercepts, that is, $\mathrm{E}(\mathbf{x}) = \boldsymbol{\nu}$. Thus, the focus is on modeling the covariance structure $\boldsymbol{\Sigma}$ of the observed variables.

The covariance matrix of the observed variables $\boldsymbol{\Sigma}$ can be written as a function of the model parameters of the CFA model:

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}^{\mathrm{T}} + \Omega, \tag{2}$$

Where $\boldsymbol{\Phi}$ is the $m \times m$ covariance matrix of the latent factors, $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_q)$ is a $q \times 1$ vector that contains the q non-redundant parameters in $\boldsymbol{\Lambda}$, $\boldsymbol{\Phi}$, and $\boldsymbol{\Omega}$ that are estimated; and $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is the model-implied covariance matrix. Thus, the covariance of the observed variables can be decomposed into a part due to the covariance structure of the latent factors and a part that is due to measurement error.

Maximum Likelihood Estimation

Maximum Likelihood estimation is routinely used to obtain parameter estimates of CFA models (Jackson et al., 2009). Let $\mathbf{x}_1,\ldots,\mathbf{x}_n$ denote a set of independently and identically distributed $p \times 1$ vectors of observed variables that are multivariate normally distributed. Then, for an observed data set, $\mathbf{X} = \{\mathbf{x}_i\}_{i=1,\ldots,n}$, the likelihood function is written as:

$$L(\boldsymbol{\theta}|\mathbf{X}) = \prod_{i=1}^{n} f(\mathbf{x}_i;\, \boldsymbol{\nu},\, \boldsymbol{\Sigma}(\boldsymbol{\theta})), \tag{3}$$

where $f(\mathbf{x};\, \boldsymbol{\mu},\, \boldsymbol{\Sigma})$ denotes the multivariate normal density with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. It is known that the sample-based covariance matrix $\mathbf{S} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'$ is a sufficient statistic for $\boldsymbol{\Sigma}$, and hence for $\boldsymbol{\Sigma}(\boldsymbol{\theta})$, which also implies sufficiency for $\boldsymbol{\theta}$. Thus, the likelihood can be written as $L(\boldsymbol{\theta}|\,\mathbf{X}) = L(\boldsymbol{\theta}|\mathbf{S})$, and the sample covariance matrix $\mathbf{S}$ of the p observed variables can be used as input in the SEM framework. The log-likelihood can be simplified as (Bollen, 1989):

$$\begin{aligned} l(\boldsymbol{\theta}) &= \log\ L(\boldsymbol{\theta}|\mathbf{S}) \\ &= -\frac{n}{2}[p\cdot\log(2\pi)\ + \log|\boldsymbol{\Sigma}(\boldsymbol{\theta})| + \mathrm{tr}\left(\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\,\mathbf{S}\right)], \end{aligned} \quad (4)$$

where tr is the trace operator, that is, the sum of the diagonal elements of a square matrix. The value $\widehat{\boldsymbol{\theta}}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\theta}} l(\boldsymbol{\theta})$ that maximizes $l(\boldsymbol{\theta})$ is the ML estimate. It should be emphasized that the latent variables $\boldsymbol{\eta}$ do not appear in the likelihood in Equation 4. Therefore, it has also been referred to as the marginal likelihood where the latent variables are integrated out (Fox et al., 2017; Merkle et al., 2019).

Statistical inference in ML estimation is based on the asymptotic covariance matrix of the ML estimator $\widehat{\boldsymbol{\theta}}_{\mathrm{ML}}$ which is obtained from the negative second partial derivatives of the log-likelihood function with respect to the model parameters:

$$\mathrm{ACOV}(\widehat{\boldsymbol{\theta}}_{\mathrm{ML}}) = \left\{ -\left[\frac{\partial^2 l(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right]\Bigg|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_{\mathrm{ML}}} \right\}^{-1}, \quad (5)$$

where the diagonal elements of the $q \times q$ matrix are used as estimates of standard errors. The term in brackets is also known as observed information matrix (with $\widehat{\boldsymbol{\theta}}_{\mathrm{ML}}$ plugged into the matrix of the second partial derivatives of $l(\boldsymbol{\theta})$; see Held and Bové, 2014). In research practice, robust standard error estimates are often used for statistical inference in SEM (Savalei, 2014; Maydeu-Olivares, 2017).

The desirable properties of ML estimation (e.g., most efficient estimates) are based on asymptotic theory and are only guaranteed to hold with large sample sizes (Yuan and Bentler, 2007). In small samples and complex models, ML estimation is prone to serious estimation problems such as failure to converge or inadmissible solutions (e.g., negative variance estimates or correlations that are larger than one; Anderson and Gerbing, 1984; Wothke, 1993; Chen et al., 2001; Yuan and Chan, 2008). Furthermore, in small to medium samples, SEMs that correct for measurement error, even though approximately unbiased, can produce much more variable estimates of structural relationships (i.e., larger empirical sampling variance) than biased manifest approaches that ignore measurement error and use manifest scale scores (Hoyle and Kenny, 1999; Ledgerwood and Shrout, 2011; Savalei, 2019; see also Li and Beretvas, 2013; Zitzmann et al., 2016).

Constrained Maximum Likelihood Estimation

As mentioned above, in standard ML estimation, parameter estimates are not constrained to any specific interval, and nothing prevents, for example, variance estimates from becoming negative (Savalei and Kolenikov, 2008; Held and Bové, 2014). *Constrained* ML estimation can mitigate estimation problems and avoid parameter estimates outside the admissible parameter space. For example, Lüdtke et al. (2018) showed in simulation studies that constrained ML estimation of the trait-state-error model for multi-wave data (Kenny and Zautra, 1995) outperformed unconstrained ML estimation in terms of the frequency of estimation problems and the accuracy of the parameter estimates (see also Gerbing and Anderson, 1987; Chen et al., 2001).

In constrained estimation, the parameter space over which optimization is performed is restricted to admissible values (e.g., variances are constrained to be positive; Schoenberg, 1997). To this end, inequality constraints that restrict parameter estimates to lower and upper bounds must be specified (see Savalei and Kolenikov, 2008). More specifically, in the constrained estimation approach, a multivalued function $\mathbf{h}$ is specified on the vector of SEM parameters, that is, $\mathbf{h}(\boldsymbol{\theta}) \geq\ \mathbf{0}$. For example, if a parameter θ (e.g., correlation) is supposed to be bounded by a lower bound l and an upper bound u, that is, $l \leq \theta \leq u$, the constraints would be given as follows: $\mathbf{h}(\theta) = (\theta - l,\ u - \theta) \geq (0, 0)$. Further possible constraints include restricting factor loadings or residual variances to positive values. Note that the constrained ML estimator $\widehat{\boldsymbol{\theta}}_{\mathrm{CML}}$ is the parameter vector $\boldsymbol{\theta}$ that maximizes the log-likelihood $l(\boldsymbol{\theta})$ in Equation 4 and fulfills the constraints that are imposed in $\mathbf{h}$. Statistical inference can be based on the asymptotic covariance matrix that is obtained from plugging $\widehat{\boldsymbol{\theta}}_{\mathrm{CML}}$ into the matrix of second derivatives of $l(\boldsymbol{\theta})$:

$$\mathrm{ACOV}(\widehat{\boldsymbol{\theta}}_{\mathrm{CML}}) = \left\{ -\left[\frac{\partial^2 l(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right]\Bigg|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_{\mathrm{CML}}} \right\}^{-1}, \quad (6)$$

where the diagonal elements of the $q \times q$ matrix are again used as estimates of standard errors (Dolan and Molenaar, 1991; but see Schoenberg, 1997, for alternative standard error estimation methods). The asymptotic covariance in Equation 6 can be enforced to be positive definite in empirical data if the parameter estimates are slightly pulled away from the boundary (e.g., by constraining correlations to the interval $[-1+\varepsilon, 1-\varepsilon]$).

In most SEM programs such as M*plus* (Muthén and Muthén, 2012) and lavaan (Rosseel, 2012), unconstrained ML estimation that does not impose any restrictions on the admissible parameter space is used as the default (see Kline, 2016). In the present article, we compare the performance of constrained and unconstrained ML estimation of CFA models with different Bayesian estimators. These are discussed in the next section.

Bayesian Approach to Confirmatory Factor Analysis

In the Bayesian approach, statistical inference is based on the posterior distribution, which is determined by the likelihood function and the prior distribution $\pi(\boldsymbol{\theta})$ of model parameters (for a general introduction to the Bayesian approach, see Jackman, 2009; Gelman et al., 2014; van de Schoot et al., 2021). Using the observed data $\mathbf{X}$ (or the sufficient statistic $\mathbf{S}$) and the prior distributions, the joint posterior distribution $p(\boldsymbol{\theta}|\mathbf{X})$ of the parameters is determined by multiplying the likelihood with the prior:

$$p\left(\boldsymbol{\theta}\,|\,\mathbf{X}\right) = \frac{L(\boldsymbol{\theta}|\mathbf{X})\pi\left(\boldsymbol{\theta}\right)}{\int L(\boldsymbol{\theta}|\mathbf{X})\pi\left(\boldsymbol{\theta}\right)\mathrm{d}\boldsymbol{\theta}} = L\left(\boldsymbol{\theta}\,|\,\mathbf{X}\right)\pi\left(\boldsymbol{\theta}\right)C \propto L\left(\boldsymbol{\theta}\,|\,\mathbf{X}\right)\pi\left(\boldsymbol{\theta}\right), \tag{7}$$

where $C = 1/\int L\left(\boldsymbol{\theta} \mid \mathbf{X}\right)\pi\left(\boldsymbol{\theta}\right)\mathrm{d}\boldsymbol{\theta}$ is a normalizing constant. As can be seen, the posterior distribution is proportional to the product of the likelihood and the prior. If a researcher does not want to make assumptions about a parameter, non-informative (diffuse) prior distributions that are intended to have only a minimal influence on the results are selected (van de Schoot et al., 2021). Moreover, the Bayesian approach offers the opportunity to stabilize parameter estimates by specifying a weakly informative prior distribution $\pi(\boldsymbol{\theta})$ "which contains some information – enough to 'regularize' the posterior distribution, that is, to keep it roughly within reasonable bounds – but without attempting to fully capture one's scientific knowledge about the underlying parameter" (Gelman et al., 2014, pp. 51). Thus, the idea is to incorporate a small amount of information into $\pi(\boldsymbol{\theta})$ that provides some direction for the estimation of model parameters but, at the same time, still allows the inferences to be driven by the likelihood (Baldwin and Fellingham, 2013; Chung et al., 2013; Lüdtke et al., 2013; Depaoli and Clifton, 2015). In the following, we discuss different Bayesian point estimates that are obtained from the posterior distribution $p\left(\boldsymbol{\theta}\,|\,X\right)$.

Bayesian Point Estimates

Point estimates in the Bayesian approach are usually calculated by summarizing the center of the marginal posterior distribution of the particular parameters of interest (e.g., latent correlation). More formally, let $\boldsymbol{\theta}_{(-d)} = (\theta_1,\ldots,\theta_{d-1}, \theta_{d+1},\ldots, \theta_q)$ denote the vector of parameters in which the dth entry of $\boldsymbol{\theta}$ has been omitted. The univariate marginal posterior distribution of θ_d, in which all other components of $\boldsymbol{\theta}$ are integrated out, is given by:

$$p_d\left(\theta_d \mid \mathbf{X}\right) = \int p(\boldsymbol{\theta}|\mathbf{X})\mathrm{d}\boldsymbol{\theta}_{(-d)} = C\int L(\boldsymbol{\theta}|\mathbf{X})\pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}_{(-d)}. \tag{8}$$

Bayesian point estimates of θ_d are obtained from location parameters (i.e., mean, median, mode) of the marginal posterior distribution. The posterior mean $\hat{\theta}_{d,\mathrm{EAP}}$ (d = 1,. . .,q) is given by the expectation of the posterior distribution:

$$\hat{\theta}_{d,\mathrm{EAP}} = \int \theta_d p_d\left(\theta_d \mid \mathbf{X}\right)\mathrm{d}\theta_d = \int \theta_d p\left(\boldsymbol{\theta} \mid \mathbf{X}\right)\mathrm{d}\boldsymbol{\theta}. \tag{9}$$

The posterior median (Med) $\hat{\theta}_{d,\mathrm{Med}}$ is the median of the marginal posterior distribution

$$\int_{-\infty}^{\hat{\theta}_{d,\mathrm{Med}}} p_d\left(\theta_d \mid \mathbf{X}\right)\mathrm{d}\theta_d = 0.5. \tag{10}$$

The posterior mode $\hat{\theta}_{d,\mathrm{MAP}}$ is given by the value that maximizes the marginal posterior distribution (maximum-a-posteriori; MAP):

$$\hat{\theta}_{d,\mathrm{MAP}} = \arg\max_{\theta_d}\; p_d\left(\theta_d|\mathbf{X}\right) \tag{11}$$

Note that all three Bayesian point estimates $\hat{\theta}_{d,\mathrm{EAP}}$, $\hat{\theta}_{d,\mathrm{Med}}$ and $\hat{\theta}_{d,\mathrm{MAP}}$ are functionals of the joint posterior distribution and involve high-dimensional integration to obtain the marginal posterior distribution. In practice, simulation-based methods such as MCMC are often used to evaluate these high-dimensional integrals. As another option for a low number of parameters, numerical integration techniques can be employed (Held and Bové, 2014).

Alternatively, the mode of the joint posterior distribution $p\left(\boldsymbol{\theta}|\mathbf{X}\right)$ can be used as a Bayesian point estimate:

$$\widehat{\boldsymbol{\theta}}_{\mathrm{PML}} = \arg\max_{\boldsymbol{\theta}} p\left(\boldsymbol{\theta}|\mathbf{X}\right) = \arg\max_{\boldsymbol{\theta}}\left[\log L\left(\boldsymbol{\theta}|\mathbf{X}\right) + \log\pi(\boldsymbol{\theta})\right]. \tag{12}$$

Note that for the computation of $\widehat{\boldsymbol{\theta}}_{\mathrm{PML}}$ (penalized maximum likelihood estimate; PML estimate; see Section "Penalized ML Estimation") it is not required to evaluate the normalization constant of the posterior distribution. Three points need to be made about the mode of the joint posterior. First, with a diffuse prior (i.e., π is a constant function with respect to $\boldsymbol{\theta}$), the likelihood is proportional to the posterior distribution (see Equation 12), and the mode of the joint posterior coincides with the ML estimator. Second, it needs to be emphasized that the univariate modes $\hat{\theta}_{d,\mathrm{MAP}}$ (d = 1,. . .,q) of the marginal posterior distributions may not equal the components of the mode $\widehat{\boldsymbol{\theta}}_{\mathrm{PML}}$ of the joint posterior distribution (Held and Bové, 2014). Note that the EAP has (in contrast to MAP and Med) the desirable property that it is invariant with respect to marginalization (see Equation 9); that is, the EAP for θ_d of the univariate posterior distribution p_d equals the EAP of the multivariate posterior p (see Fox, 2010, p. 69). Third, for a multivariate normally distributed estimate $\widehat{\boldsymbol{\theta}}_{\mathrm{ML}} \sim \mathrm{MVN}(\boldsymbol{\theta}, n^{-1}\mathbf{V}_1)$ and a multivariate normal prior distribution (i.e., $\pi\left(\boldsymbol{\theta}\right) \equiv \mathrm{MVN}(\boldsymbol{\theta}_0, \mathbf{V}_0)$), it is well known that the posterior distribution is also multivariate normal (Gelman et al., 2014), that is

$$p\left(\boldsymbol{\theta}|\mathbf{X}\right) \equiv \mathrm{MVN}\left(\left(\mathbf{V}_1^{-1} + n^{-1}\mathbf{V}_0^{-1}\right)^{-1}\left(\mathbf{V}_1^{-1}\widehat{\boldsymbol{\theta}}_{\mathrm{ML}} + n^{-1}\mathbf{V}_0^{-1}\boldsymbol{\theta}_0\right),\; n^{-1}\left(\mathbf{V}_1^{-1} + n^{-1}\mathbf{V}_0^{-1}\right)^{-1}\right). \tag{13}$$

In this case, all estimators $\widehat{\boldsymbol{\theta}}_{\mathrm{PML}}$, $\hat{\theta}_{d,\mathrm{MAP}}$, $\hat{\theta}_{d,\mathrm{Med}}$, and $\hat{\theta}_{d,\mathrm{EAP}}$ coincide. However, as ML estimates are only asymptotically normally distributed and often priors different from the normal distribution are used, it is not guaranteed that the different Bayesian estimators perform similarly, particularly in small samples. This fact is essential as different estimation methods produce different Bayesian point estimators. In the following, we distinguish between PML estimation and MCMC methods.

Penalized ML Estimation

Penalized ML estimation maximizes the log-posterior function:

$$w\left(\boldsymbol{\theta}\right) = l\left(\boldsymbol{\theta}\right) + \log\pi(\boldsymbol{\theta}). \tag{14}$$

The log-posterior is a function of the log-likelihood $l(\boldsymbol{\theta})$ = log $L(\boldsymbol{\theta}|\mathbf{X})$ and additional information given by the prior log $\pi(\boldsymbol{\theta})$. The maximizer $\widehat{\boldsymbol{\theta}}_{\mathrm{PML}}$ is also referred to as the maximum a posteriori (MAP) estimator (Gelman et al., 2014). Alternatively, the logarithm of the prior distribution $\log\pi\left(\boldsymbol{\theta}\right)$ can be interpreted as a penalty term that is added to the log-likelihood function, which motivates the label "penalized" ML (Cole et al., 2014; see also Cousineau and Helie, 2013). It should also be emphasized that constrained ML estimation can be regarded as a variant of PML estimation when uniform prior distributions are imposed on the admissible parameter space (e.g., Rindskopf, 2012). In this case, it holds that $\widehat{\boldsymbol{\theta}}_{\text{PML}} = \widehat{\boldsymbol{\theta}}_{\text{CML}}$.

Statistical inference in PML estimation can be obtained by plugging in the PML estimate $\widehat{\boldsymbol{\theta}}_{\text{PML}}$ into the matrix of second derivatives of the log-likelihood $l(\boldsymbol{\theta})$:

$$\text{ACOV}\left(\widehat{\boldsymbol{\theta}}_{\text{PML}}\right) = \left\{ -\left[\frac{\partial^2 l\left(\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{'}}\right]\Bigg|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_{\text{PML}}} \right\}^{-1}, \quad (15)$$

where the diagonal elements are used as estimates of standard errors. Note that the standard error estimates only rely on the log-likelihood $l(\boldsymbol{\theta})$ part and that the part of the prior distribution $\log\,\pi(\boldsymbol{\theta})$ is ignored. The motivation for pursuing this strategy is that the prior is only used to stabilize the estimation. Based on experience from our simulations, it is vital to ignore the prior part in the computation of uncertainty for obtaining valid frequentist statistical inference because it would otherwise result in undercoverage.

Penalized maximum likelihood estimation has been shown to circumvent estimation problems, particularly when the likelihood is flat, and has been successfully applied to stabilize parameter estimates in a wide range of models such as logistic regression models (Firth, 1993; Heinze and Schemper, 2002), latent class models (Galindo-Garre and Vermunt, 2006; DeCarlo et al., 2011), item response theory models (Mislevy, 1986; Harwell and Baker, 1991), and multilevel models (Chung et al., 2013). It should also be emphasized that in the pre-MCMC era, PML estimation was the standard approach for obtaining estimates for Bayesian factor analysis models (e.g., Martin and McDonald, 1975; Press and Shigemasu, 1989) and Bayesian SEM models (e.g., Lee, 1981; Lee, 1992; Poon, 1999). Furthermore, PML estimation bears strong similarities to regularized ML estimation, in which, too, penalty functions are added to the log-likelihood (Jacobucci and Grimm, 2018; van Erp et al., 2019; Fan et al., 2020). Note that regularized estimation is often applied for effect selection, such as the determination of non-vanishing item loadings in factor analysis (Jin et al., 2018) or the allowance of non-invariant item parameters in multiple-group factor analysis (Huang, 2018).

MCMC Estimation

Another strategy that is used to obtain Bayesian estimates is to apply simulation-based techniques. This is motivated by the fact that in practice, the joint posterior distribution of the parameters is often difficult to evaluate because high-dimensional integration is required to compute the normalization constant C (see Equation 7; Held and Bové, 2014, Ch. 8). Simulation-based techniques – implemented in general-purpose Bayesian software such as WinBUGS (Spiegelhalter et al., 2003), JAGS (Plummer, 2003), NIMBLE (de Valpine et al., 2017), or Stan (Carpenter et al., 2017) – use MCMC algorithms to approximate the posterior distribution by iteratively sampling from conditional distributions. The most prominent MCMC methods are Gibbs sampling, Metropolis-Hastings sampling, and the no-U-turn sampler (Gelman et al., 2014; Junker et al., 2016).

In the present study, we implemented a Metropolis-Hastings step within a Gibbs sampling algorithm to estimate the parameters of the CFA model. The Metropolis-within-Gibbs algorithm uses the following sampling steps to generate observations from the conditional distributions. At the $(t+1)$th iteration with current values $\left(\theta_1^{(t)}, \ldots, \theta_q^{(t)}\right)$ sample:

$$\begin{aligned}
&\theta_1^{(t+1)} \quad \text{from} \quad p(\theta_1|\mathbf{X},\theta_2^{(t)}, \theta_3^{(t)}, ..., \theta_q^{(t)}) \\
&\theta_2^{(t+1)} \quad \text{from} \quad p(\theta_2|\mathbf{X},\theta_1^{(t+1)}, \theta_3^{(t)}, ..., \theta_q^{(t)}) \\
&\vdots \\
&\theta_q^{(t+1)} \quad \text{from} \quad p(\theta_q|\mathbf{X},\theta_1^{(t+1)}, \theta_2^{(t+1)}, ..., \theta_{q-1}^{(t+1)}) \qquad (16)
\end{aligned}$$

There are q steps in the $(t+1)$th iteration. All conditional distributions are unidimensional, and parameters are updated conditional on the latest value of the other parameters.

We now show how one component of the parameter vector $\boldsymbol{\theta}$, say θ_d, is updated in the Metropolis-within-Gibbs algorithm. To generate a sample from the conditional distribution of θ_d given the most recent values of the other parameters, we rewrite the conditional distribution using Bayes theorem:

$$\begin{aligned}
&p(\theta_d|\mathbf{X},\theta_1^{(t+1)}, \theta_2^{(t+1)}, ..., \theta_{d-1}^{(t+1)}, \theta_{d+1}^{(t)}, ..., \theta_q^{(t)}) \propto \\
&L(\theta_1^{(t+1)}, \theta_2^{(t+1)}, ..., \theta_{d-1}^{(t+1)}, \theta_d, \theta_{d+1}^{(t)}, ..., \theta_q^{(t)}|\mathbf{X}) \cdot \pi(\theta_d) \quad (17)
\end{aligned}$$

The conditional distribution is proportional to the product of the likelihood and the prior distribution for θ_d. A new value is sampled from a proposal distribution $\text{N}(\theta_d^{(t)}, \tau_{\theta_d}^{2(t)})$ where $\theta_d^{(t)}$ is the value of θ_d from the previous iteration and $\tau_{\theta_d}^2$ is the proposal distribution standard deviation, which is adapted in the MCMC algorithm (see Section "Analysis Models and Outcomes"). Negative proposed values are not accepted, and the value from the previous iteration is used. Then the Metropolis-Hastings ratio is calculated as follows:

$$\begin{aligned}
&M\left(\theta_d^{(*)}, \theta_d^{(t)}\right) \\
&= \frac{L(\theta_1^{(t+1)}, \theta_2^{(t+1)}, \theta_{d-1}^{(t+1)}, \theta_d^{(*)}, \theta_{d+1}^{(t)}, ..., \theta_q^{(t)}|\mathbf{X}) \cdot \pi(\theta_d^{(*)})}{L(\theta_1^{(t+1)}, \theta_2^{(t+1)}, \theta_{d-1}^{(t+1)}, \theta_d^{(t)}, \theta_{d+1}^{(t)}, ..., \theta_q^{(t)}|\mathbf{X}) \cdot \pi(\theta_d^{(t)})},
\end{aligned} \quad (18)$$

where $M = M\left(\theta_d^{(*)}, \theta_d^{(t)}\right)$ is the Metropolis-Hastings ratio as a function of the proposed value $\theta_d^{(*)}$ and the previous value $\theta_d^{(t)}$. The proposed value $\theta_d^{(*)}$ is then accepted and set to $\theta_d^{(t+1)}$with probability min(1, M). Acceptance rates of roughly between 0.40 and 0.50 are considered optimal in the literature (Hoff, 2009; Gelman et al., 2014) to obtain an MCMC chain that has relatively low autocorrelation and mixes well (i.e., moves around the sample space in a seemingly random fashion without any long-term trends).

When the chain converges, the draws $\boldsymbol{\theta}^{(t)} = \left(\theta_1^{(t)}, ..., \theta_d^{(t)}, ..., \theta_q^{(t)}\right)$ can be seen as samples from the joint posterior distribution of the CFA model parameters (for a detailed discussion of assessing convergence in MCMC, see Cowles and Carlin, 1996; Gill, 2007; Jackman, 2009). Usually, the initial draws are discarded (burn-in phase) because the initial draws are affected by the starting values of the chain (Gelman et al., 2014). Bayesian point estimators are constructed from the samples of the posterior distribution. The expected a posteriori (EAP) estimator for a parameter θ_d, $\hat{\theta}_{d,\text{EAP}}$, is obtained by averaging across the T iterations, that is,

$$\hat{\theta}_{d,\text{EAP}} = T^{-1} \sum_{t=1}^{T} \theta_d^{(t)}. \tag{19}$$

The median $\hat{\theta}_{d,\text{Med}}$ is estimated by computing the sample median of the draws $\theta_d^{(t)}$ (t = 1, . . ., T). The mode $\hat{\theta}_{d,\text{MAP}}$ can be defined as the univariate mode of the kernel density estimate (Silverman, 1998) of the univariate density for the sample of $\theta_d^{(t)}$ (t = 1, . . ., T) (see Johnson and Sinharay, 2016). Notably, it has been proposed that the multivariate mode (PML) could also be estimated by choosing the sampled parameter that maximizes the posterior distribution (see the discussion in the Stan users group[1]):

$$\widehat{\boldsymbol{\theta}}_{\text{PML}-\text{MCMC}} = \underset{t=1,...,T}{\arg\max}\, p\left(\boldsymbol{\theta}^{(t)}|\mathbf{X}\right). \tag{20}$$

However, if the PML is of primary interest, deterministic optimization using the Newton approach (see Section "MCMC Estimation" and Equation 12) is generally preferable.

The standard deviation of the posterior distribution can be used as a measure of uncertainty (Gelman et al., 2014). Comparable to a confidence interval in the frequentist approach, it is possible to calculate a Bayesian credibility interval (BCI). The BCI is based on percentile points of the posterior distribution and describes the probability that the interval covers the true value of the parameter after observing the data. Note that in contrast to the confidence interval in the frequentist approach, no assumptions about the sampling distribution (e.g., symmetry, normality) need to be made for the BCI.

Finally, it should also be emphasized that in the presented MCMC approach, the Bayesian point estimates are based on the marginal likelihood $L(\boldsymbol{\theta}|\mathbf{X})$ – or $L(\boldsymbol{\theta}|\mathbf{S})$ if the sufficient statistic $\mathbf{S}$ is used – in which the latent variables $\boldsymbol{\eta}$ are integrated out (Equation 4). However, in many applications of MCMC-based SEM, a joint estimation approach is used that relies on the joint likelihood $L(\boldsymbol{\theta}, \boldsymbol{\eta}|\mathbf{X})$, which also includes the latent variables $\boldsymbol{\eta}$ in the likelihood (Lee, 2007; Muthén and Asparouhov, 2012). In the joint estimation approach, the MCMC method[2] generates samples for $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$. When individual factor scores are not of interest, however, the latent variables $\boldsymbol{\eta}$ are nuisance parameters that can reduce the computational efficiency of the MCMC algorithm (Hayashi and Arav, 2006; Choi and Levy, 2017; Lüdtke et al., 2018; Hecht et al., 2020; Merkle et al., 2020).

Previous Research Comparing Bayesian Estimation Approaches

The different Bayesian point estimators, that is, $\widehat{\boldsymbol{\theta}}_{\text{PML}}$, $\hat{\theta}_{d,\text{EAP}}$, $\hat{\theta}_{d,\text{MAP}}$, and $\hat{\theta}_{d,\text{Med}}$, can be evaluated from a frequentist point of view – population parameters $\boldsymbol{\theta}$ are treated as fixed but unknown constants, and the distribution of the Bayesian estimators is evaluated across all possible samples from the population (Stark, 2015). For simple univariate quantities (e.g., proportions, means), Bolstad and Curran (2017) compared frequentist properties (i.e., bias and RMSE) of mode, median, and mean using analytical derivations (see also Carlin and Louis, 2009; Efron, 2015). For more complex statistical models, several studies used simulated data to compare the performance of Bayesian estimators for different model parameters. Hoeschele and Tier (1995) compared the MAP and EAP (obtained from MCMC methods) for estimating variance components in multilevel models (see also Browne and Draper, 2006). Like the present study, Choi et al. (2011) evaluated two Bayesian estimators (MAP and EAP obtained from numerical integration) to estimate a polychoric correlation. The EAP estimates were biased and pulled toward the prior distribution (i.e., shrinkage effect), but less variable than the MAP estimates. In the context of IRT models, Azevedo et al. (2012; see also Azevedo and Andrade, 2013; Waller and Feuerstahler, 2017) and Kieftenbeld and Natesan (2012) compared PML and EAP estimates for a multiple-group 2PL model and the graded response model, respectively. In both studies, it turned out that the EAP estimates slightly outperformed PML in terms of RMSE (see also Bürkner, 2020). Yao (2014) and Johnson and Kuhn (2015) compared MAP and EAP estimation for person parameter estimation in unidimensional IRT models. Again, EAP estimates were biased (i.e., shrinkage effects) but were also less variable than MAP estimates (see also Johnson and Kuhn, 2015). For log-linear models, Galindo-Garre et al. (2004) found that the MAP outperformed the EAP for estimating main and interaction effects.

In the context of SEM and CFA models, systematic comparisons of the frequentist performance of Bayesian estimators are scarce. Simulation studies that evaluated the performance of different Bayesian estimators for estimating SEMs focused on either the Med (e.g., Hox et al., 2012, 2014; Depaoli and Clifton, 2015; Holtmann et al., 2016), the EAP (e.g., Lee and Song, 2004; Natesan, 2015) or the MAP (e.g., Zitzmann et al., 2016). One notable exception is the study by Miocevic et al. (2020) that evaluated the EAP, MAP, and Med for estimating an indirect effect (i.e., the product of two path coefficients) in a latent mediation model using MCMC methods. The relative performance of the different estimators in terms of RMSE depended on the specification of the prior distribution (accurate vs. inaccurate) and the size of the true indirect effect, with a slight disadvantage for the MAP when accurate priors were

[1] https://shortly.cc/ArTE6

[2] In the joint estimation approach, the joint likelihood $L(\boldsymbol{\theta}, \boldsymbol{\eta}|\mathbf{X}) = \prod_{i=1}^{n} f(\mathbf{x}_i; \boldsymbol{\nu} + \boldsymbol{\Lambda}\boldsymbol{\eta}_i, \boldsymbol{\Omega}) \cdot f(\boldsymbol{\eta}_i; \mathbf{0}, \boldsymbol{\Phi})$ is considered, and the MCMC method generates samples for $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ (t = 1,. . .,T): $\boldsymbol{\theta}^{(t)}$, $\boldsymbol{\eta}^{(t)} p(\boldsymbol{\theta}, \boldsymbol{\eta} \mid \mathbf{X}) \propto L(\boldsymbol{\theta}, \boldsymbol{\eta} \mid \mathbf{X}) \cdot \pi(\boldsymbol{\theta})$. Note that the joint estimation approach needs the raw data $\mathbf{X}$ instead of the sufficient statistic $\mathbf{S}$ (see Choi and Levy, 2017).

specified. However, it is unclear whether these findings generalize to other SEM parameters (e.g., loadings, latent correlations), making it difficult for applied researchers to choose between different Bayesian estimators. This lack of guidance is also reflected in the fact that popular software packages for SEM use different Bayesian estimators as default. The commercial software packages Mplus (Muthén, 2010) and Amos (Arbuckle, 2017) provide the Med, whereas the R package blavaan (Merkle and Rosseel, 2018) uses the EAP (see Taylor, 2019). In the present study, we evaluate the performance of four different Bayesian estimators for latent correlations and loadings in CFA models.

CFA WITH TWO FACTORS: PARAMETERIZATION, PRIOR DISTRIBUTIONS, AND ESTIMATION METHODS

In the following, we consider a CFA model with two latent variables, which are each measured by three observed indicator variables (see **Figure 1**). We use this model to discuss three relevant issues in the practical application of the Bayesian approach: model parameterization, specification of prior distributions, and different Bayesian estimators (mode of the joint posterior, and mean, median, and mode of the marginal posterior). Although this is a very simple model, it is a building block for many more complicated SEM models, such as latent mediation models or multilevel SEMs (Hoyle, 2012; Kline, 2016).

Parameterization

To mitigate estimation problems, it can be advantageous to choose a parameterization of the CFA model that transforms the optimization problem of estimating unbounded parameters into an optimization involving bounded parameters. We start with an unstandardized parameterization in which the two latent variables η_1 and η_2 are measured by indicators x_j (j = 1,. . ., 6):

$$x_j = \lambda_j^* \eta_{\mathrm{m}[j]} + \varepsilon_j^*, \tag{21}$$

Where m[·] is a function that maps the item index j to the corresponding index m[j] of the latent variable, λ_j^* are the unstandardized non-negative loadings, and ε_j^* are normally distributed residuals with $\mathrm{Var}\left(\varepsilon_j^*\right) = \omega_{jj}^*$. Two strategies for identifying the metric of the latent factors are often used (Kline, 2016; see also Gonzalez and Griffin, 2001). In the first strategy (reference variable method) the first loading of each latent factor is set to one, and the variances and covariance of the latent factors are freely estimated. In the second strategy, the variances of the latent variables are set to one, that is, $\mathrm{Var}(\eta_1) = \mathrm{Var}(\eta_2) = 1$, and the latent correlation between the factors is directly estimated. In the Bayesian framework, the reference variable method has been the most common choice (see Lee, 1981; Erosheva and Curtis, 2017; Merkle and Rosseel, 2018; Miocevic et al., 2020). This may be explained by the fact that the second strategy is not easily applicable to general SEM models because the variances of endogenous latent variables are not free parameters in standard SEM specifications (Kline, 2016; see also Kaplan and Depaoli, 2012; van Erp et al., 2018).

In this paper, we suggest a parameterization of the CFA model in which the parameters of interest are bounded, and the standardized loadings and the latent correlation are directly estimated (Little, 2013). Using a parameterization with bounded or standardized parameters has the advantage that it is straightforward to restrict correlations to admissible values between −1 and 1. This is more difficult to accomplish when the correlation is derived from the variances and covariance of the latent variables (e.g., very small variance estimates; Rindskopf, 1984)[3]. Furthermore, a parameterization with bounded parameters is often more convenient for applied researchers when specifying thoughtful prior distributions (Smid et al., 2020; Zitzmann et al., 2021). Let σ_j denote the standard deviation of the observed indicator x_j, then Equation 21 can be rewritten as:

$$x_j = \sigma_j(\lambda_j \eta_{\mathrm{m}[j]} + \varepsilon_j), \tag{22}$$

where λ_j (j = 1,. . .,6) are the standardized loadings, and ε_j are the residuals of the standardized solution. It can be shown that the parameterizations in Equations 21 and 22 are equivalent. It holds that $\sigma_j^2 = \left(\lambda_j^*\right)^2 + \omega_{jj}^*$, $\lambda_j = \lambda_j^* / \sqrt{\omega_{jj}^*}$, and $\mathrm{Var}(\varepsilon_j) = 1 - \lambda_j^2$. Thus, the standardized error variance is positive if the standardized loadings are restricted to be positive. In many applications, especially with established scales, restricting loadings to positive values seems plausible

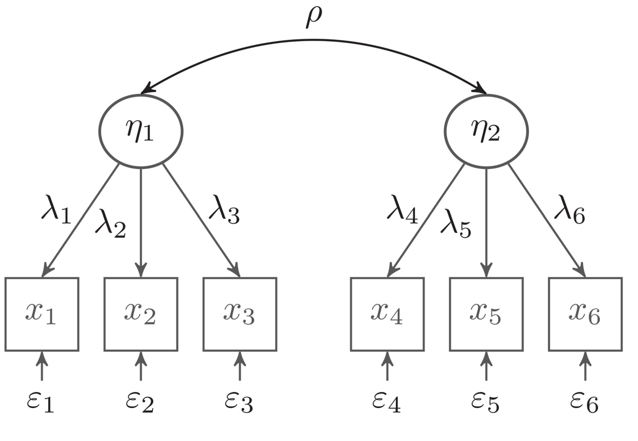


FIGURE 1 | Confirmatory factor analysis (CFA) model with two latent factors. The variances of the latent factors are set to 1.0.

[3] This was confirmed by preliminary simulation studies in which we also included unconstrained ML estimation with the reference variable method (i.e., first loading fixed to 1). The parameterization with bounded parameters clearly outperformed the parameterization with freely estimated variances and covariance of the latent variables in terms of estimation problems (e.g., convergence) and quality of parameter estimates (e.g., RMSE).

because researchers commonly have strong presumptions on the direction of relationships between the observed and latent variables. Technically, the parameterization in Equation 22 can be implemented in the SEM framework by introducing an intermediate layer of latent variables (phantom variables; see Rindskopf, 1984) and non-linear constraints for the measurement error variances.[4]

In the present study, our main focus is on estimating the correlation between the latent variables, that is, $\mathrm{Cov}(\eta_1, \eta_2) = \rho$. As we are interested in estimating the latent correlation that corrects for the unreliability of the scale scores, it is instructive to see how the reliability of the manifest sum score is related to the standardized loadings of the measurement model. In the data-generating model, we assume that the standardized loadings of the indicators for a latent factor are equal and set to λ (tau-congeneric measurement model; Traub, 1994). Then, the indicator-specific reliability is given by $\mathrm{Rel}_1 = \lambda^2$, and the reliability of the sum score of I items is:

$$\mathrm{Rel}_I = \frac{\lambda^2}{\lambda^2 + (1-\lambda^2)/I} = \frac{\mathrm{Rel}_1}{\mathrm{Rel}_1 + (1-\mathrm{Rel}_1)/I}. \tag{23}$$

As can be seen, the reliability of the sum score Rel_I is a function of the indicator-specific reliability Rel_1 and the number of items. Thus, by rearranging terms, the reliability of an indicator can be written as:

$$\mathrm{Rel}_1 = \frac{\mathrm{Rel}_I}{1 + I(1-\mathrm{Rel}_I)}. \tag{24}$$

For example, with $I = 3$, a standardized loading of $\lambda = 0.58$ translates into a reliability of 0.60 for the sum score (see **Table 1**). This relationship between the standardized loading and the reliability of the sum score is helpful when specifying the prior distributions for the loadings because, in most cases, it is easier to make plausible assumptions about the overall reliability of a scale than about every single item (Smid et al., 2020).

TABLE 1 | Relationship between standardized loading, indicator-specific reliability, and reliability of sum score for three indicators.

λ	Rel_1	Rel_I
0.35	0.13	0.30
0.43	0.18	0.40
0.50	0.25	0.50
0.58	0.33	0.60
0.66	0.44	0.70
0.76	0.57	0.80
0.87	0.75	0.90

λ = standardized loading; Rel_1 = indicator-specific reliability; Rel_I = reliability of sum score.

Specification of Prior Distributions

In the CFA model, the standardized loadings and latent correlations are bounded parameters. For bounded parameters, the beta distribution is a natural choice. The density f of the beta distribution $X \sim \mathrm{Beta}(a, b)$ on the interval [0, 1] is given as:

$$f(x) = B(a,b)^{-1} x^{a-1}(1-x)^{b-1}, x \in [0,\ 1] \tag{25}$$

where B is the Beta function. The mean and the variance can be computed as:

$$\mathrm{E}(X) = \frac{a}{a+b} \text{ and } \mathrm{Var}(X) = \frac{ab}{(a+b)^2(a+b+1)}. \tag{26}$$

Alternatively, the beta distribution can be parameterized as a function of a mean μ and a prior sample size ν, that is, $X \sim \mathrm{Beta}(\mu, \nu)$, where $\mu = a(a+b)^{-1}$ and $\nu = a + b - 2$ (Hoff, 2009). The prior sample size is explained by the fact that the uniform distribution, which reflects complete ignorance about a parameter, is given by setting $a = b = 1$. Thus, a prior sample size of $\nu = 1 + 1 - 2 = 0$ corresponds to the uniform prior on [0, 1]. The variance of the beta distribution is given as $\mathrm{Var}(X) = \mu(1-\mu)(\nu+3)^{-1}$. For the given μ and ν, the original a and b parameters are determined by $a = (\nu + 2)\mu$ and $b = (\nu + 2)(1 - \mu)$, respectively.

However, the beta distribution is only appropriate for parameters with a parameter space that equals [0, 1]. The four-parameter beta distribution (also known as a scaled, stretched, or generalized beta distribution) extends the support of the beta distribution to arbitrary bounded intervals and allows a more flexible specification of prior distributions (Johnson et al., 1994). The four-parameter beta distribution $Y \sim \mathrm{Beta4}(a, b, l, u)$ can be obtained by shifting a beta-distributed random variable $X \sim \mathrm{Beta}(a, b)$ by lower (l) and upper (u) bounds: $Y = l + (u - l)X$. The density of Y is then given as:

$$f(x) = (u-l)^{-1} B(a,b)^{-1} \left(\frac{x-l}{u-l}\right)^{a-1} \left(\frac{u-x}{u-l}\right)^{b-1}, x \in [l,\ u] \tag{27}$$

Again, the four-parameter beta distribution can be reparameterized as $Y \sim \mathrm{Beta4}(\mu, \nu, l, u)$ with a prior guess of $\mu = a(a+b)^{-1}$ and a prior sample size of $\nu = a + b - 2$. The parameters of the original specification $Y \sim \mathrm{Beta4}(a, b, l, u)$ can be obtained as $a = (\nu + 2)(\mu - l)(u - l)^{-1}$ and $b = (\nu + 2)(u - \mu)(u - l)^{-1}$. In previous research, the four-parameter beta distribution has been applied as prior distribution for item parameters in three-parameter logistic models (Zeng, 1997; Gao and Chen, 2005), and for correlations between observed scores (Gokhale and Press, 1982; O'Hagan et al., 2006) or latent variables in factor models (Muthén and Asparouhov, 2012; Merkle and Rosseel, 2018). However, to the best of our

[4] For constrained ML estimation, it can be shown that this parameterization of the CFA model with standard deviations that are constrained to be positive and loadings that are restricted to the interval [0, 1] is (analytically) equivalent to a parameterization in which the loadings and error variances of the unstandardized solution are constrained to be positive and the variances of the latent factors are set to one. Thus, using the parameterization of the CFA model in Equation 21 for constrained ML estimation should – in theory – provide the same results as a CFA model with the corresponding inequality constraints on the loadings and error variances.

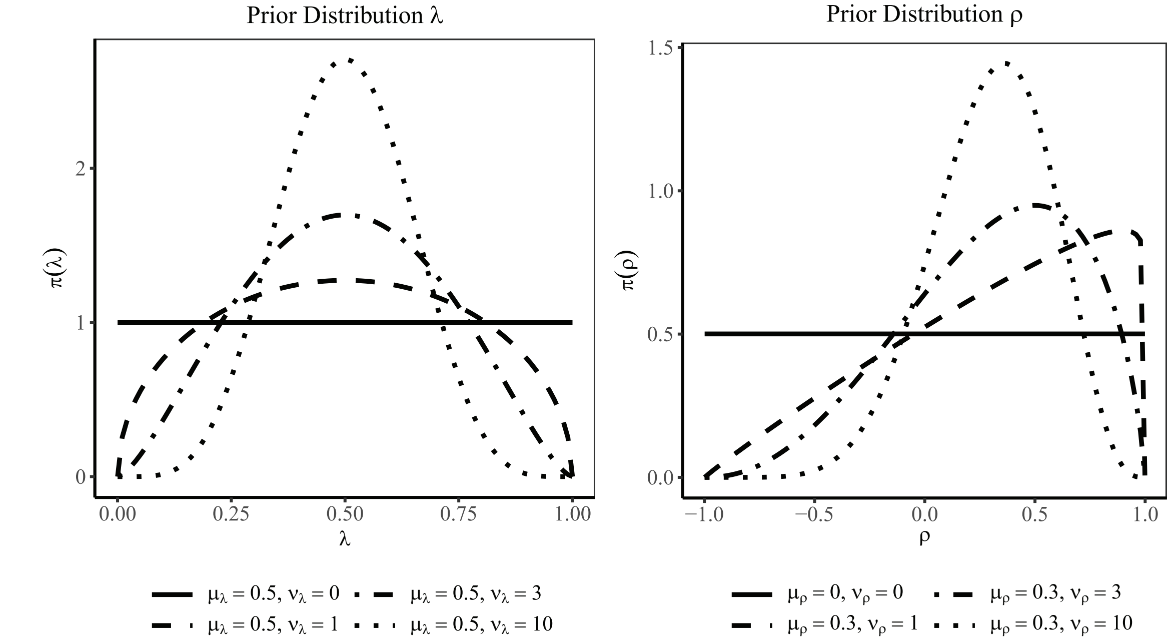


FIGURE 2 | Four-parameter beta distributions for a standardized loading λ **(left panel)** and the correlation ρ **(right panel)**. With a larger prior sample size (ν_λ or ν_ρ), the distribution puts more probability mass around the prior guess (μ_λ or μ_ρ).

knowledge, we are not aware of any applications of the four-parameter beta distribution as prior distribution for loadings in the SEM framework (see Table 1 in van Erp et al., 2018, for an overview of prior distributions in the SEM context).

When specifying the lower and upper bounds of the four-parameter beta distribution, the numerical stability of parameter estimates can be improved if parameters are coerced further away from the boundaries by a small value ε (e.g., ε = 0.001). For a standardized loading λ that can be assumed to be bounded between 0 and 1, we suggest a Beta4(μ_λ, ν_λ, ε, 1 − ε) prior distribution and interpret μ_λ as a prior guess for the standardized loading and ν_λ as the sample size of prior observations on which the prior guess is based (Lüdtke et al., 2018)[5]. If only little information is available about the standardized loading or the reliability of a scale, a small value for ν_λ is chosen so that the prior distribution is only weakly centered around the prior guess μ_λ. **Figure 2** (left panel) shows for a prior guess of $\mu_\lambda = 0.50$ (i.e., standardized loading of 0.50) how increasing ν_λ (i.e., prior sample sizes of 1, 3, and 10) changes the shape of the four-parameter beta distribution. Note that with $\mu_\lambda = 0.50$ and $\nu_\lambda = 0$, the four-parameter beta distribution corresponds to a uniform distribution on the interval [ε, 1 − ε], which reflects complete ignorance about the size of the loading.

For the latent correlation that is restricted to the interval [−1, 1], we suggest a Beta4(μ_ρ, ν_ρ, −1 + ε, 1 − ε) distribution where μ_ρ is the prior guess for the correlation and ν_ρ is again the prior sample size on which the prior guess is based (see for a similar approach Muthén and Asparouhov, 2012; Merkle and Rosseel, 2018). **Figure 2** (right panel) illustrates the influence of the prior sample size ν_ρ (1, 3, and 10) on the shape of the Beta4(μ_ρ, ν_ρ, − 1 + ε, 1 − ε) with a prior guess of $\mu_\rho = 0.30$. Setting $\mu_\rho = 0$ and $\nu_\rho = 0$ gives the uniform distribution on [−1 + ε, 1 − ε].

Illustrative Comparison of Different Bayesian Point Estimates

To illustrate how the different Bayesian point estimates can produce different estimates of the latent correlation, we further simplify the two-factor model and assume that all loadings are equal. Thus, we need to estimate only two parameters[6]: the correlation ρ (ranging between −1 and 1) and the standardized loading λ (ranging between 0 and 1). Thus, for this simplified model, the likelihood function $L(\rho, \lambda|\mathbf{S})$ is only a function of ρ and λ, given the sufficient statistic **S**. Furthermore, we assume uniform priors for both parameters (i.e., constant functions with respect to ρ and λ), which results in a joint posterior $p\,(\rho, \lambda \mid \mathbf{S})$

[5] Alternatively, if a researcher is not willing to make assumptions about the sign of the loading (i.e., loadings are assumed to be positive), the loadings can be restricted to [−1, 1]. However, if this specification is chosen, researchers need to define one marker item for which loadings must be restricted to [0, 1]. Otherwise, sign switching issues can occur in the MCMC algorithm. In preliminary simulations, we investigated the effect of both restrictions. If the loadings of the data-generating model were all positive, the performance for the two variants of restrictions was very similar.

[6] For this illustration, we further reduced the number of estimated parameters, assumed that the indicators had a variance of one, and also fixed the variances of the indicators in the analysis model to one. The main purpose of the illustration was to demonstrate the differences between the mode from the joint posterior and the EAP that is obtained from the marginal posterior distribution. A more systematic and realistic evaluation of the different estimators is provided in the two main simulations.

that is proportional to the likelihood. In PML estimation, the mode of the joint posterior distribution is calculated as:

$$(\hat{\rho}_{\mathrm{PML}}, \hat{\lambda}_{\mathrm{PML}}) = \underset{(\rho,\lambda)}{\arg\max}\, p\,(\rho, \lambda|\mathbf{S}) = \underset{(\rho,\lambda)}{\arg\max}\, L(\rho, \lambda|\mathbf{S}) \quad (28)$$

and $\hat{\rho}_{\mathrm{PML}}$ is used as a point estimate of ρ. Note that this is the first component of the multivariate mode, which is calculated by directly maximizing the density of the joint posterior distribution. It becomes clear from Equation 28 that the PML estimate is also the constrained ML estimate because the likelihood function is maximized, that is $\hat{\boldsymbol{\theta}}_{\mathrm{PML}} = \hat{\boldsymbol{\theta}}_{\mathrm{CML}}$.

In contrast, when using MCMC methods, the univariate mode (MAP), median (Med), and mean (EAP) are often used as point estimates for ρ. The marginal posterior distribution p_ρ of ρ is obtained by integrating the joint posterior $p\,(\rho, \lambda|\mathbf{S})$ with respect to λ:

$$p_\rho\,(\rho|\mathbf{S}) = \int p\,(\rho, \lambda|\mathbf{S})\, d\lambda = C \int L(\rho, \lambda, |\mathbf{S}) d\lambda, \quad (29)$$

where $C = 1/\iint L\,(\rho, \lambda|\mathbf{S})\, d\rho d\lambda$ is the normalizing constant. The corresponding marginal location parameters are given as follows:

$$\hat{\rho}_{\mathrm{MAP}} = \underset{\rho}{\arg\max}\; p_\rho\,(\rho|\mathbf{S}) = \underset{\rho}{\arg\max} \int L(\rho, \lambda, |\mathbf{S}) d\lambda, \quad (30)$$

$$\int_{-\infty}^{\hat{\rho}_{\mathrm{Med}}} p_\rho\,(\rho|\mathbf{S})\, d\rho = 0.5, \text{ and} \quad (31)$$

$$\hat{\rho}_{\mathrm{EAP}} = \int \rho p_\rho\,(\rho|\mathbf{S})\, d\rho = \frac{\iint \rho L(\rho, \lambda|\mathbf{S}) d\rho d\lambda}{\iint L(\rho, \lambda|\mathbf{S}) d\rho d\lambda}. \quad (32)$$

In our simple bivariate case, the univariate EAP $\hat{\rho}_{\mathrm{EAP}}$ can be calculated by numerically evaluating the posterior on a bivariate discrete grid of values ρ and λ. The integrals in Equation 32 can be obtained by numerical integration using a rectangle rule (see also Choi et al., 2011). Similarly, the median $\hat{\rho}_{\mathrm{Med}}$ and the univariate MAP $\hat{\rho}_{\mathrm{MAP}}$ can be obtained by a numerical evaluation of the integrals in Equations 30 and 31. However, this would not be practical with a larger number of parameters, and simulation-based MCMC techniques are needed to evaluate high-dimensional integrals (Held and Bové, 2014).

We now employ an idealized scenario to illustrate the difference between the different Bayesian estimates of the latent correlation. In this case, the empirical covariance matrix **S** (i.e., the sufficient statistic) obtained from the data was set to be equal to the true covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta})$. Hence, the likelihood estimates (i.e., the constrained ML and the PML estimates) coincided with the data-generating parameters. The true correlation was $\rho = 0.70$, and the standardized loading was $\lambda = 0.50$. **Figure 3** shows, for a small sample size of $N = 30$, a contour plot of the joint posterior distribution for ρ and λ (upper left panel) and the marginal posterior distribution of ρ (lower left panel). As can be seen, the mode of the joint posterior ($\hat{\rho}_{\mathrm{PML}} = 0.700$) provides a different Bayesian estimate of the correlation than the mode ($\hat{\rho}_{\mathrm{MAP}} = 0.710$), mean ($\hat{\rho}_{\mathrm{EAP}} = 0.568$) or median ($\hat{\rho}_{\mathrm{Med}} = 0.610$) of the marginal posterior. This can be explained by the fact that the marginal posterior is negatively skewed, and the mean and—to a slightly lesser extent—the median are pulled toward zero (i.e., shrinkage effect; see also Choi et al., 2011). However, with a larger sample of $N = 100$, the Bayesian estimates from the joint posterior (upper right panel) and the marginal posterior (lower right panel) agree more closely ($\hat{\rho}_{\mathrm{PML}} = 0.700$, $\hat{\rho}_{\mathrm{MAP}} = 0.704$, $\hat{\rho}_{\mathrm{EAP}} = 0.675$, and $\hat{\rho}_{\mathrm{Med}} = 0.686$), and the marginal posterior distribution of ρ is more symmetrically centered around the true value of 0.70.

Illustrative Simulation Study

To further explore how these differences between the Bayesian point estimates affect their frequentist properties, we ran a small simulation study in which we manipulated the sample size (N = 30, 50, 100, and 1000) and the magnitude of the true correlation (ρ = 0.10, 0.30, 0.50, 0.70, and 0.90). The standardized loading was set to 0.50. We generated 1000 replications for each condition and calculated the bias, variability (i.e., the standard deviation of the empirical sampling distribution), and the RMSE (which combines bias and variability into a measure of accuracy) for the different point estimates (PML, MAP, EAP, and Med) of the latent correlation ρ.

The results are shown in **Table 2** and confirm the findings from the illustration that the mode from the joint posterior (PML) and the mode from the marginal posterior (MAP) perform very similarly. Both produced approximately unbiased estimates of the latent correlation, except for the condition with a very large correlation (ρ = 0.90) and a small sample size (N = 30). By contrast, the mean (EAP) and the median (Med) of the marginal posterior provided negatively biased estimates, particularly in conditions with small sample sizes. However, the EAP and Med were also less variable (i.e., smaller empirical sampling variability) than the estimates produced by both the PML estimate and the MAP, resulting in overall more accurate estimates in terms of the RMSE, which combines bias and variability. The results also show that there is a turning point at which, with a larger true correlation, the bias introduced by the EAP outweighs the gains in efficiency (i.e., less variable estimates of the EAP). Thus, the EAP seems to be most beneficial with a small to moderate true correlation (i.e., $\rho \leq 0.50$) and does not generally result in more accurate estimates of the latent correlation. A very similar pattern holds true for the Med. However, in almost all conditions, the Med was outperformed by either the EAP or the MAP in terms of RMSE. Notably, the multivariate mode (PML) performed similarly to the univariate mode (MAP). However, in other models, particularly with strongly correlated parameter estimates, the multivariate and univariate modes can provide substantially different point estimates.[7] Finally, the findings confirm that with large samples, the

[7] For example, in the context of state-trait models, Lüdtke et al. (2018) found in simulation studies that PML (obtained from constrained ML estimation) was clearly outperformed by the MAP (obtained from MCMC) in terms of the accuracy of the estimated variance components (e.g., stable trait variance, state variance). This can be explained by the fact that using marginal distributions stabilizes point estimates if model parameters are substantially correlated.

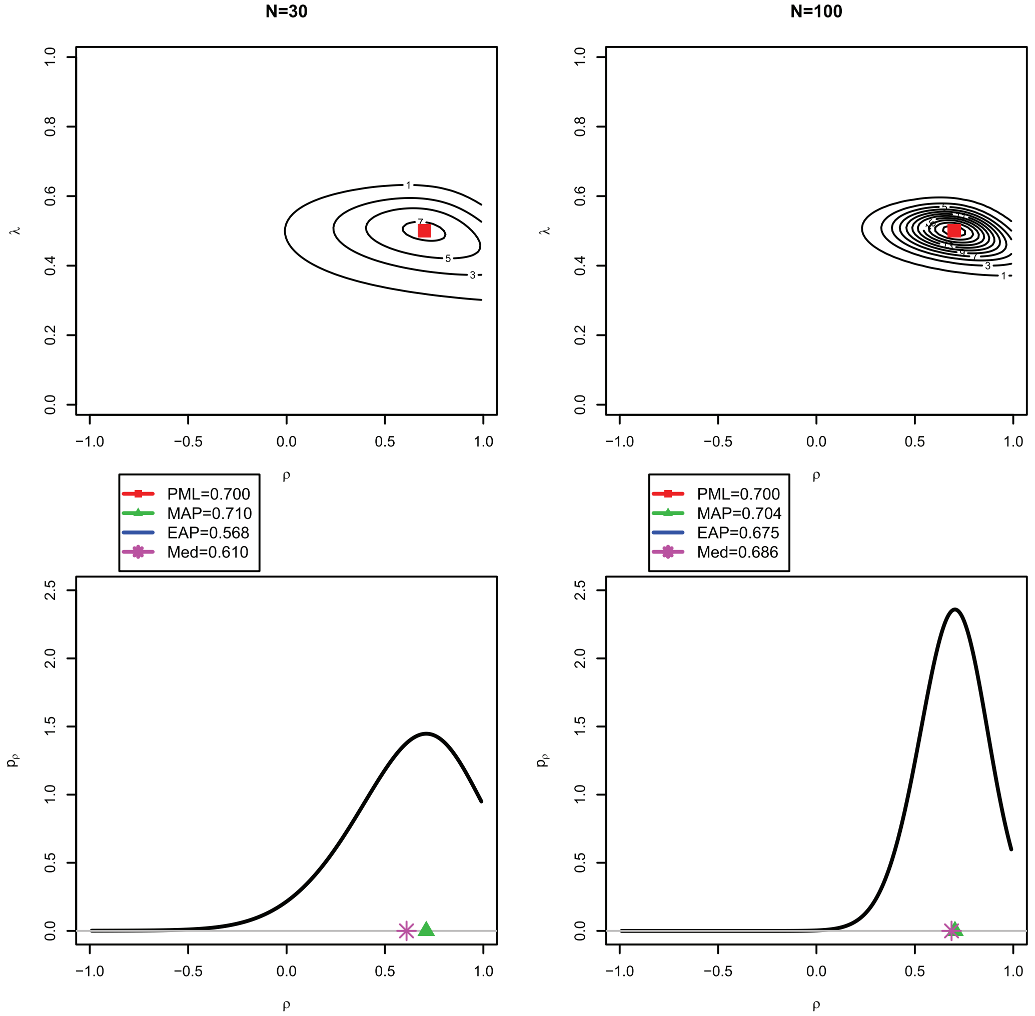


FIGURE 3 | Illustrating the difference between the joint and marginal posterior distribution: the red square in the first row indicates the (multivariate) mode of the joint posterior distribution for $N = 30$ **(upper left panel)** and $N = 100$ **(upper right panel)**; the green triangle, blue circle, and purble star in the second row indicate the (univariate) mode, mean, and median of the marginal posterior distribution for the correlation ρ for $N = 30$ **(lower left panel)** and $N = 100$ **(lower right panel)**. Note that with $N = 30$, the mode of the joint posterior (PML) for ρ strongly deviates from the mean of the marginal posterior (EAP).

different Bayesian point estimates converge and produce almost identical results.

We also investigated the performance of the different Bayesian point estimates for the loading parameter λ (see for the detailed results **Supplementary 1** at https://doi.org/fwr7). Across all conditions (i.e., true correlations and sample sizes) the biases for the four estimators were relatively small (PML: $M = -0.001$, range = −0.010 to 0.005; MAP: $M = -0.004$, range = −0.019 to 0.006; Med: $M = -0.013$, range = −0.040 to 0.003; EAP: $M = -0.017$, range = −0.050 to 0.001). In addition, the PML provided slightly more accurate estimates in terms of RMSE than the three Bayesian estimates that were based on the marginal posterior.

In the following, we report the results of two simulation studies that provide a more comprehensive comparison of the different Bayesian point estimates. In these simulations, MCMC methods are used to evaluate the high-dimensional integrals that are needed for computing the MAP, EAP, and Med from the marginal posterior distribution.

SIMULATION STUDY 1

Simulation study 1 had two main goals. First, we evaluated the performance of the different Bayesian estimators and compared them with unconstrained ML estimation. For small sample sizes, we expected unconstrained ML estimation to show serious estimation problems (i.e., non-convergence or inadmissible parameter estimates). In addition, based on our illustration, we assumed that using the EAP (obtained from MCMC) as a point estimate would produce more stable estimates of latent correlations than the multivariate mode from PML estimation, particularly in conditions with small sample sizes and small to moderate correlations. Second, we

TABLE 2 | Illustrating differences between the mode of the joint posterior (PML) and the mode (MAP), median (Med), and mean (EAP) of the marginal posterior as point estimators of the correlation: bias, standard deviation, and RMSE as a function of the true correlation (ρ) and sample size (N).

		Bias				SD				RMSE			
ρ	*N*	PML	MAP	Med	EAP	PML	MAP	Med	EAP	PML	MAP	Med	EAP
0.1	30	0.000	0.001	−0.017	−0.024	0.417	0.416	0.322	0.293	0.417	0.416	0.322	0.294
	50	0.005	0.007	−0.003	−0.007	0.301	0.305	0.273	0.259	0.301	0.305	0.272	0.259
	100	−0.011	−0.009	−0.011	−0.012	0.210	0.213	0.208	0.205	0.211	0.213	0.208	0.206
	1000	0.001	0.001	0.001	0.001	0.061	0.061	0.060	0.060	0.061	0.061	0.060	0.060
0.3	30	−0.001	0.002	**−0.053**	**−0.073**	0.410	0.406	0.308	0.281	0.410	0.405	0.312	0.290
	50	0.007	0.010	−0.019	−0.031	0.294	0.295	0.258	0.246	0.294	0.295	0.259	0.248
	100	0.002	0.006	−0.002	−0.006	0.196	0.198	0.192	0.189	0.196	0.198	0.192	0.189
	1000	−0.002	−0.002	−0.003	−0.003	0.061	0.061	0.061	0.061	0.061	0.061	0.061	0.061
0.5	30	0.005	0.006	**−0.093**	**−0.126**	0.345	0.340	0.262	0.244	0.345	0.340	0.278	0.274
	50	0.006	0.010	−0.040	**−0.060**	0.278	0.276	0.237	0.226	0.278	0.276	0.241	0.234
	100	0.004	0.008	−0.008	−0.015	0.186	0.185	0.174	0.169	0.186	0.185	0.174	0.170
	1000	0.001	0.002	0.001	0.001	0.060	0.060	0.060	0.060	0.060	0.060	0.060	0.060
0.7	30	−0.026	−0.024	**−0.150**	**−0.192**	0.299	0.292	0.231	0.221	0.300	0.293	0.275	0.292
	50	−0.004	−0.002	**−0.081**	**−0.108**	0.234	0.229	0.190	0.183	0.234	0.229	0.206	0.212
	100	0.001	0.003	−0.030	−0.043	0.169	0.166	0.142	0.137	0.169	0.166	0.145	0.143
	1000	−0.001	0.000	−0.001	−0.001	0.054	0.054	0.055	0.055	0.054	0.054	0.055	0.055
0.9	30	**−0.069**	**−0.070**	**−0.227**	**−0.277**	0.218	0.213	0.182	0.181	0.228	0.224	0.290	0.331
	50	−0.039	−0.039	**−0.145**	**−0.178**	0.164	0.160	0.132	0.131	0.169	0.165	0.196	0.221
	100	−0.021	−0.022	**−0.083**	**−0.102**	0.124	0.122	0.095	0.091	0.126	0.124	0.126	0.137
	1000	0.001	0.001	−0.006	−0.009	0.052	0.052	0.045	0.043	0.052	0.052	0.045	0.044

SD = standard deviation of empirical sampling distribution; RMSE = root mean square error; ρ = true latent correlation; N = sample size; PML = mode of joint posterior (obtained from maximizing the joint posterior); MAP = mode of marginal posterior (obtained from maximizing the marginal posterior); Med = median of marginal posterior (obtained from numerical integration); EAP = mean of marginal posterior (obtained from numerical integration). Biases smaller than −0.05 or larger than 0.05 are printed in bold.

evaluated the extent to which the parameter estimates of the Bayesian approach are sensitive to different specifications of the prior distributions for the standardized loadings and the latent correlation. We assumed that by choosing weakly informative and correctly specified prior distributions (i.e., four-parameter beta distributions), the estimates of the latent correlations could be stabilized. Furthermore, we explored whether the Bayesian approach produces more accurate estimates, even with mildly misspecified prior distributions. Overall, we expected the impact of choosing different prior distributions to be more pronounced with small sample sizes.

Simulation Model and Conditions

The data-generating model was a two-factor CFA model, as given by **Figure 1**. Each factor was measured by three mean-centered and normally distributed indicators. The indicators were assumed to be parallel, with standardized loadings of 0.50 and a variance of one. This resulted in a reliability of $\text{Rel}_I = 0.50$ for each scale (i.e., sum score of the three items) and a reliability of $\text{Rel}_1 = 0.25$ for a single indicator. We manipulated the latent correlation between the two factors (ρ = 0.30, 0.50, and 0.70) and the sample size (N = 30, 50, and 100). For each of the 3 × 3 = 9 conditions, we generated 1,000 simulated data sets.

Analysis Models and Outcomes

Each of the simulated data sets was analyzed with a two-factor CFA model in which the loadings were freely estimated, and the variances of the two factors were each fixed to one. The model had 21 − 13 = 8 degrees of freedom (the mean structure was assumed to be saturated). Two ML estimation approaches were used. In unconstrained ML estimation, we imposed no constraints on the parameter estimates (loadings, residual variances, and the latent correlation). In constrained ML estimation, we used the parameterization in which standard deviations of the indicators are constrained to be positive, the standardized loadings are restricted to the interval [0, 1], and the latent correlation is restricted to the interval [−1, 1][8]. In the Bayesian approach, we used PML estimation and MCMC methods, and varied the prior distribution for the standardized loadings and the latent correlation. PML estimation was implemented using a quasi-Newton optimization (employing the `nlminb` optimizer in the R package stats) using the wrapper function pmle from the R package LAM (Robitzsch, 2020). The standard errors were calculated based on the second derivatives of the observed log-likelihood function (see Equation 15). The estimated standard errors were used to calculate 95% confidence intervals.

[8] For constrained ML estimation, we also included the equivalent parameterization (see Equation 20) in which the loadings and residual variances were constrained to be positive and the latent correlation was restricted to the interval [−1, 1]. As expected, the results for both parameterizations were virtually identical in every replication (results were numerically identical in 95.5% of the replications for N = 30, in 98.4% for N = 50, and in 99.9% for N = 100). Moreover, the parameterization using standardized instead of unstandardized loadings performed slightly better in terms of RMSE.

For the MCMC method, we implemented an adaptive Metropolis-Hastings algorithm in which the proposal distribution is adapted during the burn-in phase (see Equation 18; Draper, 2008). In this procedure, a desirable acceptance rate r along with a tolerance level $(r - \delta, r + \delta)$ is specified (in our case $r = 0.45$ and $\delta = 0.10$). Then, in the burn-in phase of the algorithm, the empirical acceptance rates r^* for each parameter are calculated in batches of 50 iterations. At the end of each batch, the proposal distribution standard deviation (e.g., for the latent correlation τ_ρ) is updated as follows:

$$\tau_\rho = \begin{cases} \tilde{\tau}_\rho \left(2 - \frac{(1-r^*)}{1-r}\right) \text{ if } r^* > r + \delta \\ \tilde{\tau}_\rho \left(1/\left(2 - \frac{r^*}{r}\right)\right) \text{ if } r^* < r - \delta \ , \\ \tilde{\tau}_\rho \text{ else} \end{cases} \tag{33}$$

Where r^* is the empirical acceptance rate from the most recent batch of iterations, and $\tilde{\tau}_\rho$ is the proposal distribution standard deviation that was used in the most recent batch. Thus, the proposal distribution standard deviations are modified if the acceptance rate is not within the tolerance level $(r - \delta, r + \delta)$. The modification of the proposal distributions was stopped after the burn-in phase (2,500 iterations). To evaluate the tuning phase for the proposal distribution standard deviations, we investigated the empirical acceptance rates for one condition of the main simulation. The average acceptance rate for the model parameters was close to the desired value of 0.45, which is considered optimal in the literature to achieve efficient MCMC chains (Hoff, 2009).

This algorithm was implemented using the function amh from the R package LAM (Robitzsch, 2020). Before running the main simulation study, we investigated the behavior of the MCMC sampler in preliminary simulations by inspecting two criteria: (a) the potential scale reduction factor (PSR; Gelman et al., 2014), and (b) the effective sample size (see Zitzmann and Hecht, 2019, for a discussion). Applying these two criteria suggested that an average chain length of 5,000 iterations with a burn-in period of 2,500 iterations was sufficient to provide a good approximation of the posterior distribution. The Bayesian point estimates were defined as the mean (EAP), mode (MAP), and median (Med) of the marginal posterior distribution. Furthermore, the Bayesian credible interval (BCI) was defined by the 2.5th and the 97.5th percentiles of the posterior distribution (Gelman et al., 2014).

For both the PML and the MCMC methods, we varied the prior distributions for the standardized loadings and the latent correlation. For each standardized loading, we specified a four-parameter beta distribution Beta4(μ_λ, ν_λ, ε, $1 - \varepsilon$) with a prior guess of $\mu_\lambda = 0.5$ and prior sample sizes of $\nu_\lambda = 1$ and $\nu_\lambda = 3$ (see **Figure 2**). In addition, we included a prior distribution with $\mu_\lambda = 0.5$ and $\nu_\lambda = 0$, which corresponds to a uniform distribution on $[\varepsilon, 1 - \varepsilon]$. For the latent correlation, we specified a four-parameter beta distribution Beta4(μ_ρ, ν_ρ, $-1 + \varepsilon$, $1 - \varepsilon$) with a prior guess that matched the true correlation of the data-generating model (i.e., $\mu_\rho = \rho$) and two levels of prior sample sizes ($\nu_\rho = 1$ and 3). We also specified a prior distribution with $\mu_\rho = 0$ and $\nu_\rho = 0$, which corresponds to a uniform distribution on $[-1 + \varepsilon, 1 - \varepsilon]$. This resulted in 3 (loadings) × 3 (correlations) = 9 different specifications of the prior distributions. Note that these prior distributions were correctly specified (i.e., prior guess matched the true population value or uniform prior distribution was specified) and only differed in the amount of information that was incorporated into the prior specification (i.e., prior sample size).

We also investigated the impact of misspecified prior distributions. To this end, we specified a wide range of four-parameter beta distributions for the standardized loadings and the correlation. For the standardized loadings, we included misspecified priors with a prior guess of $\mu_\lambda = 0.80$ and prior sample sizes of $\nu_\lambda = 1$ and $\nu_\lambda = 3$. For the latent correlation, we specified a prior distribution with a prior guess of $\mu_\rho = 0.50$ and prior sample sizes of $\nu_\rho = 1$, and $\nu_\rho = 3$. However, we also included misspecified priors that underestimated (with a prior guess of $\mu_\rho = 0.20$) or overestimated (with a prior guess of $\mu_\rho = 0.80$) the true correlation. Again, each misspecified prior was included with prior sample sizes of $\nu_\rho = 1$ and $\nu_\rho = 3$. In addition, we used an uniform distribution for the latent correlation (i.e., $\mu_\rho = 0$ and $\nu_\rho = 0$). These prior settings for correlations were fully crossed with the different prior settings for correctly and misspecified prior settings on standardized factor loadings. In total, we specified 5 (standardized loadings) × 7 (correlations) = 35 models with different prior specifications, and we estimated them with both the PML and the MCMC methods.

For the standard deviations of the indicator variables, we used improper prior distributions that are constant (Muthén and Asparouhov, 2012). The specification of the improper prior distribution for the standard deviation was held constant across the conditions of the simulation and the analysis models. The R code for the data-generating model and the different analysis models is provided in **Supplementary 2** at https://doi.org/fwr7.

We used three criteria to evaluate the different estimation approaches: bias, RMSE, and coverage rate. Bias was calculated by determining the difference between the mean parameter estimate and the true population parameter value from each design cell. We assessed the overall accuracy with the (empirical) RMSE, which combines the squared empirical bias and the variance of the parameter estimates into a measure of overall accuracy. Finally, we determined the coverage rate of the 95% confidence intervals. A coverage rate between 91% and 98% was considered acceptable (Muthén and Muthén, 2002).

Results

We first report the results for the two ML estimation approaches. Second, we compare the different Bayesian estimators in the case of correctly specified prior distributions. Third, we investigate the impact of misspecified prior distributions on the performance of the Bayesian approach.

ML Estimation

For unconstrained ML estimation, a solution was considered to show estimation problems when the algorithm did not converge using the defaults in the nlminb optimizer or when the algorithm converged to a solution that included an inadmissible estimate (i.e., correlation smaller than −1 or larger than 1). **Table 3** shows that the percentage of estimation problems for unconstrained ML estimation strongly depended on the sample size and the

TABLE 3 | Simulation study 1: percentage of solutions with estimation problems for unconstrained maximum likelihood (ML) estimation and constrained maximum likelihood (CML) estimation by magnitude of the true correlation (ρ) and sample size (N).

	ML Conv			ML Conv+Adm			CML Boundary			ML = CML		
	ρ			ρ			ρ			ρ		
N	**0.3**	**0.5**	**0.7**	**0.3**	**0.5**	**0.7**	**0.3**	**0.5**	**0.7**	**0.3**	**0.5**	**0.7**
30	54.5	64.4	67.9	51.3	57.3	53.4	6.3	11.1	21.6	28.0	33.0	34.8
50	73.1	82.7	88.0	70.9	77.5	73.9	3.4	6.3	17.0	53.1	61.7	64.3
100	94.0	97.6	99.6	93.9	96.4	92.8	0.4	1.4	7.2	87.1	92.8	91.6

ML Conv = percentage of converged solutions for unconstrained ML estimation; ML Conv+Adm = percentage of converged solutions with admissible values for unconstrained ML estimation, that is, estimated correlation was within the interval [−1, 1]; CML Boundary = percentage of solutions for constrained ML estimation with boundary estimate for correlation; ML = CML = percentage of solutions in which estimated correlation for ML and CML estimations were numerically identical (up to three decimal places). Note that ML Conv+Adm are not conditional percentages and, thus, values cannot exceed those of ML Conv.

magnitude of the true correlation. For example, with $N = 30$ and $\rho = 0.30$, only 54.5% of the replications converged, and 51.3% of the replications provided converged solutions with admissible estimates. In contrast, for constrained ML estimation, all replications converged. However, in small samples and with a large correlation, a substantial percentage of the solutions for constrained ML estimation showed values at the boundary of the parameter space (e.g., estimated correlation equals one). Furthermore, the results also show that with increasing sample size, the estimates of unconstrained ML and constrained ML converged to each other. For example, in the condition with $N = 100$ and a large correlation ($\rho = 0.70$), unconstrained and constrained estimation provided (numerically) identical estimates of the latent correlation in 91.6% of the replications.

Table 4 shows the bias and RMSE for unconstrained and constrained ML estimation as a function of the sample size and the true correlation. The results are presented for three different subsets of replications. First, we included only replications in which unconstrained ML estimation converged and estimated correlations had admissible values; that is, they fell within the range of −1 and 1 ("Conv+Adm" in **Table 4**). Second, we present results for all replications in which unconstrained ML estimation converged, and inadmissible values (i.e., correlations smaller than −1 or larger than 1) were truncated to −1 or 1 ("Conv"). Third, we show the results for all replications ("All"). Note that only constrained ML estimation converged for all replications. As can be seen, the two approaches performed very similarly across the different subsets of replications. The results also show that the estimates produced from replications without estimation problems ("Conv+Adm") are a highly selective set of estimates that strongly differ in terms of RMSE from the estimates that are provided by the full set of replications ("All"). In the following, we use constrained ML estimation, which is equivalent to PML estimation with uniform distributions on the admissible parameter space, and compare it with the Bayesian estimation approach.

Bayesian Estimation With Correctly Specified Priors

Table 5 shows the bias for PML and the EAP (obtained from the MCMC method) with uniform and different correctly specified prior distributions as a function of the sample size and the magnitude of the true correlation. In these specifications, the prior guesses for the correlation (i.e., μ_ρ = 0.30, 0.50, or 0.70) as well as the standardized loading (i.e., μ_λ = .50) were set to the true value (when the prior sample sizes of the corresponding priors were at least one). Note that when using a prior sample size of zero, μ_ρ was set to zero, and a uniform distribution was specified. As can be seen, both the PML and the EAP produced biased estimates of the correlation, particularly when the sample

TABLE 4 | Simulation study 1: bias and RMSE for unconstrained maximum likelihood (ML) estimation and constrained maximum likelihood (CML) estimation of the latent correlation as a function of the true correlation (ρ), the sample size (*N*), and different sets of replications.

		Bias					RMSE				
		Conv+Adm.		Conv.		All	Conv+Adm.		Conv.		All
ρ	***N***	**ML**	**CML**	**ML**	**CML**	**CML**	**ML**	**CML**	**ML**	**CML**	**CML**
0.3	30	0.022	**0.055**	0.043	**0.068**	−0.016	0.352	0.358	0.395	0.395	0.411
	50	0.025	0.035	0.040	0.049	−0.003	0.284	0.288	0.310	0.309	0.328
	100	0.003	0.006	0.004	0.006	−0.004	0.202	0.201	0.203	0.203	0.210
0.5	30	−0.037	−0.003	0.019	0.048	−0.025	0.299	0.295	0.332	0.322	0.360
	50	−0.042	−0.032	−0.015	−0.005	−0.049	0.266	0.262	0.299	0.285	0.316
	100	0.000	0.002	0.006	0.008	0.003	0.196	0.194	0.202	0.201	0.206
0.7	30	**−0.124**	**−0.088**	−0.034	−0.008	**−0.068**	0.299	0.269	0.299	0.274	0.344
	50	**−0.071**	**−0.063**	−0.012	−0.005	−0.035	0.242	0.234	0.252	0.245	0.275
	100	−0.026	−0.025	−0.004	−0.003	−0.004	0.172	0.171	0.184	0.183	0.185

RMSE = root mean square error; ρ = true latent correlation; N = sample size; ML = unconstrained ML estimation; CML = constrained ML estimation. Conv+Adm = all replications in which unconstrained ML converged with admissible values, that is, estimated correlation was within the interval [−1, 1]; Conv = all replications in which unconstrained ML estimation converged and estimated correlations outside [−1, 1] are truncated to −1 or 1; All = all replications (only constrained ML produced estimates for all replications). Biases smaller than −0.05 or larger than.05 are printed in bold.

TABLE 5 | Simulation study 1: bias and RMSE for the mode of the joint posterior (PML) and the mean of the marginal posterior (EAP) as Bayesian point estimates of the latent correlation with different correctly specified prior distributions as a function of the sample size and true correlation (ρ).

				Bias			**SD**			**RMSE Gain**		
				ρ			ρ			ρ		
N	***Meth***	ν_ρ	ν_λ	**0.3**	**0.5**	**0.7**	**0.3**	**0.5**	**0.7**	**0.3**	**0.5**	**0.7**
30	PML	0	0	−0.001	**−0.052**	**−0.069**	0.415	0.393	0.330	100	100	100
		0	1	0.034	−0.010	−0.017	0.434	0.399	0.313	105	101	93
		0	3	0.043	0.001	−0.003	0.424	0.389	0.302	103	98	90
		1	0	**0.086**	**0.153**	**0.158**	0.370	0.364	0.264	92	100	91
		1	1	**0.119**	**0.198**	**0.192**	0.374	0.347	0.230	95	101	89
		1	3	**0.121**	**0.213**	**0.209**	0.364	0.346	0.219	92	102	90
		3	0	0.049	**0.092**	**0.165**	0.283	0.257	0.233	69	69	85
		3	1	**0.086**	**0.117**	**0.190**	0.291	0.249	0.208	73	69	83
		3	3	**0.088**	**0.121**	**0.210**	0.280	0.245	0.196	71	69	85
	EAP	0	0	**−0.072**	**−0.146**	**−0.215**	0.284	0.263	0.236	71	76	95
		0	1	−0.021	**−0.077**	**−0.124**	0.331	0.300	0.245	80	78	81
		0	3	−0.008	**−0.057**	**−0.098**	0.342	0.310	0.248	83	79	79
		1	0	**0.058**	**0.129**	**0.104**	0.221	0.173	0.149	55	54	54
		1	1	**0.081**	**0.153**	**0.133**	0.266	0.209	0.137	67	65	57
		1	3	**0.089**	**0.164**	**0.136**	0.277	0.221	0.138	70	69	58
		3	0	−0.006	0.002	**0.087**	0.169	0.151	0.105	41	38	41
		3	1	0.018	0.035	**0.112**	0.191	0.178	0.116	46	46	48
		3	3	0.024	0.045	**0.122**	0.197	0.185	0.118	48	48	50
50	PML	0	0	−0.018	−0.010	−0.041	0.335	0.302	0.275	100	100	100
		0	1	0.004	0.021	−0.011	0.343	0.293	0.255	102	97	92
		0	3	0.006	0.027	0.001	0.331	0.286	0.247	99	95	89
		1	0	0.042	**0.139**	**0.148**	0.312	0.321	0.248	94	116	104
		1	1	**0.061**	**0.165**	**0.167**	0.310	0.309	0.228	94	116	102
		1	3	**0.060**	**0.170**	**0.188**	0.301	0.305	0.216	91	116	103
		3	0	0.021	**0.070**	**0.149**	0.250	0.234	0.228	75	81	98
		3	1	0.045	**0.094**	**0.161**	0.257	0.220	0.212	78	79	96
		3	3	0.045	**0.097**	**0.181**	0.249	0.217	0.202	75	79	98
	EAP	0	0	**−0.068**	**−0.091**	**−0.147**	0.251	0.230	0.213	77	82	93
		0	1	−0.033	−0.037	**−0.085**	0.278	0.245	0.213	83	82	82
		0	3	−0.026	−0.022	**−0.065**	0.283	0.250	0.215	85	83	81
		1	0	0.013	**0.103**	**0.095**	0.214	0.191	0.149	64	72	63
		1	1	0.027	**0.123**	**0.117**	0.243	0.217	0.141	73	82	66
		1	3	0.032	**0.131**	**0.127**	0.250	0.226	0.142	75	86	68
		3	0	−0.023	−0.004	**0.074**	0.170	0.158	0.131	51	52	54
		3	1	−0.004	0.027	**0.096**	0.191	0.175	0.138	57	59	60
		3	3	0.003	0.036	**0.102**	0.195	0.180	0.138	58	61	62
100	PML	0	0	−0.006	−0.004	0.005	0.222	0.211	0.183	100	100	100
		0	1	0.005	0.008	0.013	0.220	0.203	0.176	99	96	97
		0	3	0.009	0.012	0.017	0.216	0.199	0.173	97	95	95
		1	0	0.021	**0.099**	**0.165**	0.216	0.266	0.196	98	134	140
		1	1	0.032	**0.101**	**0.166**	0.211	0.252	0.191	96	129	139
		1	3	0.035	**0.100**	**0.168**	0.208	0.246	0.189	95	126	139
		3	0	0.016	0.040	**0.156**	0.193	0.185	0.187	87	89	134
		3	1	0.027	0.048	**0.161**	0.190	0.179	0.184	86	88	134
		3	3	0.030	**0.052**	**0.159**	0.188	0.177	0.181	85	87	132
	EAP	0	0	−0.039	**−0.052**	**−0.056**	0.191	0.188	0.166	88	92	96
		0	1	−0.019	−0.028	−0.033	0.198	0.190	0.161	90	91	90
		0	3	−0.014	−0.018	−0.022	0.200	0.189	0.160	90	90	88
		1	0	−0.006	0.045	0.105	0.183	0.199	0.136	82	96	94

(Continued)

TABLE 5 | Continued

				Bias			SD			RMSE Gain		
				ρ			ρ			ρ		
N	***Meth***	ν_ρ	ν_λ	**0.3**	**0.5**	**0.7**	**0.3**	**0.5**	**0.7**	**0.3**	**0.5**	**0.7**
		1	1	0.007	**0.059**	**0.114**	0.190	0.202	0.132	86	100	96
		1	3	0.013	**0.064**	**0.119**	0.192	0.203	0.131	87	101	97
		3	0	−0.020	−0.011	**0.074**	0.157	0.158	0.138	71	75	86
		3	1	−0.004	0.007	**0.085**	0.164	0.159	0.136	74	75	88
		3	3	0.002	0.015	**0.089**	0.167	0.159	0.135	75	76	88

SD = standard deviation of empirical sampling distribution; RMSE = root mean square error. PML = penalized maximum likelihood (mode of joint posterior); EAP = mean of marginal posterior; ν_ρ = prior sample size for latent correlation; ν_λ = prior sample size for standardized loading; biases smaller than -0.05 or larger than 0.05 are printed in bold. For the RMSE gain, the RMSE of PML estimation with uniform prior distributions is used as a reference method; values smaller than 100 indicate that the RMSE of the respective method is lower than the RMSE of the reference method.

sizes were small ($N \leq 50$). In addition, there was a tendency that increasing the prior sample size from $\nu_\rho = 1$ to $\nu_\rho = 3$, and thereby selecting a more informative prior distribution for the latent correlation, decreased the bias for both the PML ($\nu_\rho = 1$: $M = 0.129$, $SD = 0.053$, range = 0.021 to 0.213; $\nu_\rho =3$: $M = 0.099$, $SD = 0.044$, range = 0.016 to 0.210) and the EAP ($\nu_\rho =1$: $M = 0.086$, $SD = 0.038$, range = −0.006 to 0.164; $\nu_\rho =3$: $M = 0.036$, $SD = 0.023$, range = −0.023 to 0.122). Interestingly, the results were less clear for increasing the sample size from $\nu_\rho = 0$ to $\nu_\rho = 1$, particularly for the PML.

For the RMSE, which combines bias and the variability of an estimator, we used the PML method with uniform prior distributions on the standardized loadings and the latent correlation as a reference method. As this specification of the PML method is equivalent to constrained ML estimation, it allows a direct comparison of the Bayesian approaches with the best performing ML approach. The RMSE gain in **Table 5** reports the relative gain of an estimator compared to the reference method (i.e., values larger/smaller than 100 indicate that the RMSE for the respective method is larger/smaller than for the reference method). The results show that the EAP obtained from the MCMC method clearly outperformed PML estimation across all sample size conditions and true values of the latent correlation. As expected from the illustration, the differences between the mode of the joint posterior distribution (PML) and the mean (EAP) of the marginal posterior were most pronounced in conditions with a very small sample size ($N = 30$) and a small true correlation. For example, in the condition with $N = 30$, $\rho = 0.30$, and uniform prior distributions, the RMSE of the estimates produced by the EAP were only 71% as large as the estimates produced by the PML. This is an important finding because it clearly shows that, even with (diffuse) uniform distributions on the loadings and the correlation, using the EAP (obtained from MCMC) stabilizes the parameter estimates compared to the PML (or constrained ML) method.

To further understand the RMSE differences, we calculated the empirical standard deviation (*SD*) of the estimators across the 1000 replications within each cell. The results show that the estimates of the PML were consistently more variable (across the different prior specifications) than those of the EAP. For both estimators, PML and EAP, selecting a more informative prior distribution for the correlation (e.g., $\nu_\rho = 3$ instead of $\nu_\rho = 1$) had a large positive effect on the accuracy of the parameter estimates. By contrast, choosing a more informative prior distribution for the standardized loadings did not consistently influence the accuracy of the estimates of the latent correlation. Thus, adding information to the prior distribution for the parameter of interest was the only specification that helped to stabilize estimates of the latent correlation in small sample sizes.

The main findings for bias and RMSE are summarized in **Figure 4** for the case with uniform prior distributions on the standardized loadings and the correlation. We also show the results for the mode (MAP) and the median (Med) of the marginal posterior of ρ. As can be seen, the Med performed similar to the EAP but showed slightly larger RMSE values. By contrast, the MAP provided less accurate estimates of the correlation in terms of RMSE and was even outperformed by the PML in almost all conditions (except for $N = 100$ and $\rho = 0.3$).

Furthermore, we assessed the coverage rates for the PML and MCMC methods. As can be seen in **Table 6**, the PML method provided acceptable coverage rates with percentages close to the nominal 95% in conditions with $N = 100$. In addition, the coverage rates produced by the MCMC method were sometimes too low, even in conditions with $N = 100$. However, these low coverage rates can be explained by the fact that the MCMC method was also more biased in these conditions.

Finally, we also investigated the bias and RMSE of the different Bayesian estimates for a standardized loading. The main results are summarized in **Figure 5** for the case with uniform prior distributions (for the detailed results, see **Supplementary 3** at https://doi.org/fwr7). Overall, the findings are in line with the results for the correlation. The EAP produced the most accurate estimates of the loadings in terms of RMSE across the investigated conditions, even though the estimates were slightly negatively biased, particularly in conditions with $N = 30$. Interestingly, with smaller sample sizes, the univariate mode (MAP) was clearly outperformed by the multivariate mode (PML). Further simulation research should compare the different Bayesian point estimates for more extreme values of the loading (i.e., standardized loading of 0.3 or 0.9). It is possible that with smaller or larger loading values, the bias introduced by the EAP outweighs the gains in variability, resulting in different

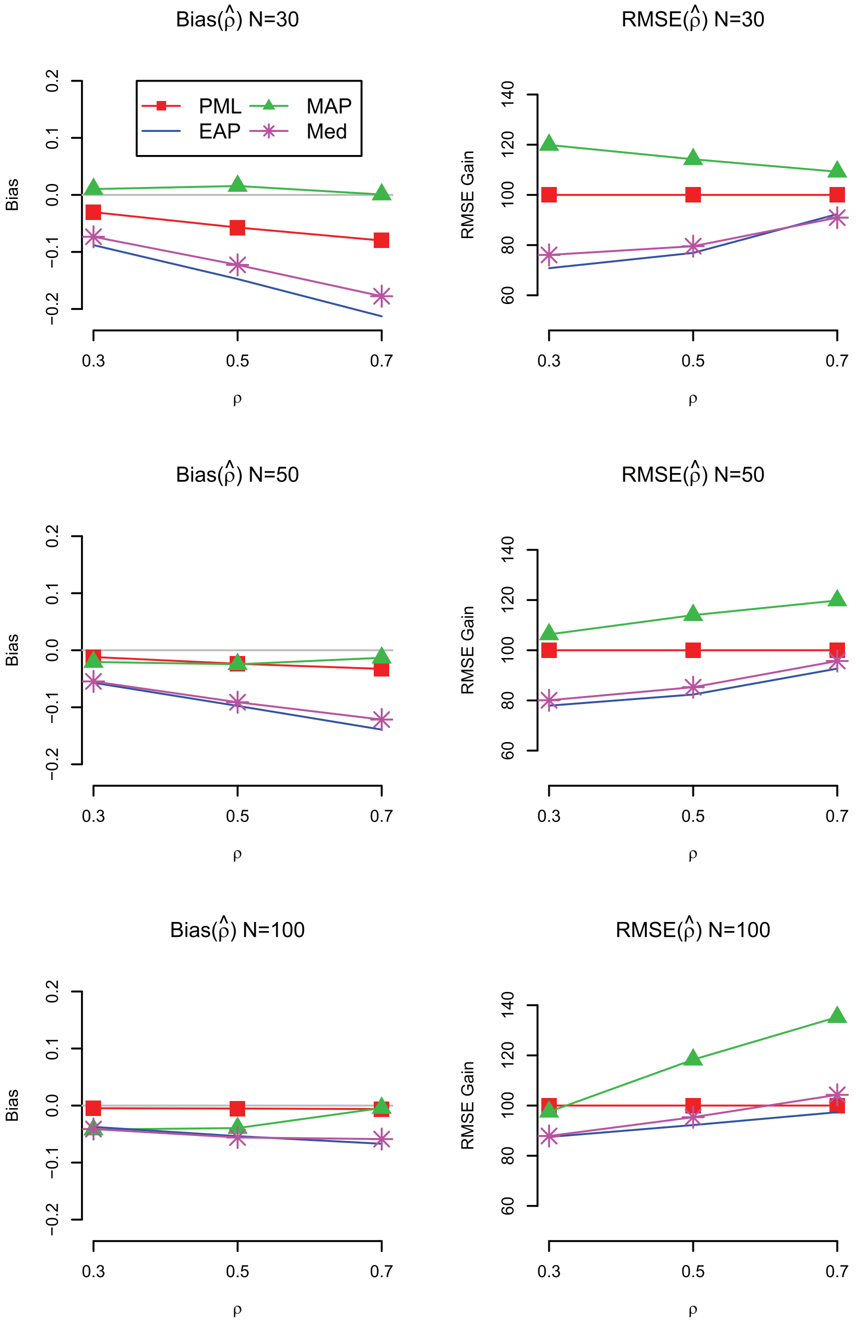


FIGURE 4 | Simulation Study 1: Bias **(left panels)** and RMSE gain **(right panels)** of the estimators of the correlation (ρ) as a function of the sample size and the magnitude of the true correlation. For the RMSE gain, PML is used as a reference method; values smaller than 100 indicate that the RMSE of the respective method is lower than the RMSE of the reference method; PML = mode of joint posterior; MAP = mode of marginal posterior; Med = median of marginal posterior; EAP = mean of marginal posterior. Results are shown for models with uniform prior distributions for the correlation and the standardized loadings.

conclusions about the overall accuracy of the different Bayesian point estimates.

Bayesian Estimation With Misspecified Priors

We also assessed the impact of misspecified prior distributions. **Table 7** shows the bias and RMSE for N = 30 and N = 100. The main findings can be summarized as follows. First, even in the case of misspecified prior distributions, the EAP outperformed the PML in terms of the RMSE and provided more accurate parameter estimates across most conditions and prior specifications. Second, a misspecified prior distribution for the loading had only a small and sometimes even positive effect on the RMSE. One possible explanation is that we only included misspecified priors that overestimated the true size of

TABLE 6 | Simulation study 1: coverage rates of the latent correlation for penalized maximum likelihood and Markov chain monte carlo methods with different correctly specified prior distributions as a function of the magnitude of the true correlation (ρ) and sample size.

			PML			**MCMC**		
			ρ			**ρ**		
N	ν_ρ	ν_λ	**0.3**	**0.5**	**0.7**	**0.3**	**0.5**	**0.7**
30	0	0	**85.8**	**87.0**	91.1	**98.9**	**98.6**	98.0
	0	1	**87.6**	91.2	96.8	97.5	97.9	97.9
	0	3	**88.5**	94.1	97.5	97.3	97.7	**98.1**
	1	0	**88.9**	**90.3**	94.7	**99.2**	98.0	**89.3**
	1	1	**90.0**	93.6	97.7	**98.2**	95.9	**84.3**
	1	3	**90.7**	94.2	**98.3**	**98.2**	95.5	**83.3**
	3	0	92.8	95.6	96.8	**99.6**	**99.5**	**99.4**
	3	1	95.2	97.0	**98.3**	**99.6**	**99.0**	**98.5**
	3	3	97.5	97.8	**98.8**	**99.2**	**99.1**	98.0
50	0	0	**86.6**	**90.1**	93.1	97.1	**98.2**	97.4
	0	1	**88.4**	92.7	95.7	95.5	96.1	96.7
	0	3	**90.1**	93.6	97.4	94.9	95.7	97.0
	1	0	**87.9**	**90.0**	96.5	97.5	96.9	**89.4**
	1	1	**89.9**	92.4	**98.2**	96.2	95.6	**87.3**
	1	3	92.0	91.8	**98.8**	95.6	93.8	**85.3**
	3	0	91.8	96.4	97.7	**98.8**	**98.8**	**98.3**
	3	1	95.1	97.2	**98.7**	98.0	**98.3**	96.5
	3	3	96.7	97.0	**99.3**	97.9	97.5	95.5
100	0	0	**89.9**	92.7	97.0	95.6	95.7	96.8
	0	1	92.1	94.8	97.0	95.3	96.1	97.2
	0	3	92.8	95.4	97.5	94.8	96.4	96.9
	1	0	91.0	92.3	94.4	96.8	96.7	**89.9**
	1	1	92.3	92.5	94.6	96.1	94.9	**88.5**
	1	3	93.0	91.0	95.1	95.5	94.5	**87.8**
	3	0	93.8	95.8	95.1	97.3	**98.5**	97.0
	3	1	95.5	96.3	95.1	97.1	97.8	96.4
	3	3	95.8	96.1	95.8	96.6	98.0	95.1

PML = mode of joint posterior (obtained from penalized maximum likelihood) with standard errors (calculated from observed information matrix); MCMC = Bayesian Credible Interval (BCI) based on MCMC; ρ = true latent correlation; ν_ρ = prior sample size for latent correlation; ν_λ = prior sample size for standardized loading. Coverage rates smaller than 91% or larger than 98% are printed in bold.

the loading (i.e., $\mu_\lambda = 0.80$). Overestimating the reliability of the indicators by assuming a large positive loading comes close to a manifest approach that ignores the unreliability of scale scores when calculating the correlation. However, with small sample sizes, it has been shown that a manifest approach can produce more accurate estimates of structural relationships than a latent approach that corrects for measurement error (e.g., Lüdtke et al., 2008; Ledgerwood and Shrout, 2011; Savalei, 2019). Third, for the prior distribution of the correlation, the results clearly show that overestimating the true size of the latent correlation (i.e., $\mu_\rho = 0.80$) had a more negative impact on the accuracy of the estimates in terms of RMSE than underestimating the size of the true correlation (i.e., $\mu_\rho = 0.20$). More importantly, choosing a small correlation of 0.20 as a prior guess for the prior distribution, even though misspecified, produced more accurate estimates of the correlation than the Bayesian approach with uniform priors on the loadings and the correlation, particularly when the sample size was $N = 30$. Thus, a conservative approach that uses smaller prior guesses for the latent correlation seems to be a promising strategy when the goal is to stabilize the estimates of the latent correlations with weakly informative prior distributions (i.e., prior sample sizes of 1 or 3).

SIMULATION STUDY 2

The previous simulation study assumed that the observed variables were multivariate normally distributed. However, the true distribution is rarely known for real data, and the CFA will likely be misspecified to some extent. In Simulation Study 2, we investigate how robust the Bayesian approach is against the misspecification of the distributional assumptions. More specifically, we consider the case of observed variables that are linearly related but have non-normal marginal distributions (Foldnes and Olsson, 2016). Again, we compared the different Bayesian point estimates obtained from the joint posterior (PML) or the marginal posterior distribution (MAP, EAP, and Med). As a benchmark, we also included ML approaches that are based on robust estimation approaches (Yuan et al., 2004; Yuan and Zhang, 2012). For further comparisons, we also considered an unweighted least squares (ULS) estimation method (Browne, 1974). Limited information methods such as ULS are expected to be more robust in modeling violations than ML estimators (MacCallum et al., 2007).

Simulation Model and Conditions

The data-generating model was again a two-factor CFA model with six observed variables. We generated a covariance structure (see Equation 2) that followed a CFA model with parallel and standardized loadings of 0.50 and a variance of one for the observed variables. The procedure of Foldnes and Olsson (2016) was used to generate six observed variables that preserved the covariance structure and had a prespecified level of skewness and kurtosis for the marginal distributions of observed variables. Six different combinations of skewness and kurtosis values were chosen to implement a range of non-normal distributions for the observed variables: 0/0 (skewness/excess kurtosis), 0/3, 0/7, 1/3, 1/7, and 2/7. Again, we manipulated the latent correlation between the two factors (ρ = 0.10, 0.30, 0.50, 0.70, and 0.90) and the sample size (N = 30, 50, 100, and 500). For each of the 5 × 5 × 4 = 100 conditions, we generated 1,000 simulated data sets.

Analysis Models

Each of the simulated data sets was analyzed with a two-factor CFA in which the loadings were freely estimated, and the variances of the two factors were set to one. We used PML estimation with a uniform prior distribution on the standardized loadings and the correlation. In addition, we included a robust version of PML (PMLR) in which the sufficient statistics $\bar{\mathbf{x}}$ and $\mathbf{S}$ were replaced by a robust sample mean vector $\bar{\mathbf{x}}_{\text{rob}}$ and a robust sample covariance matrix $\mathbf{S}_{\text{rob}}$ that were obtained with the R

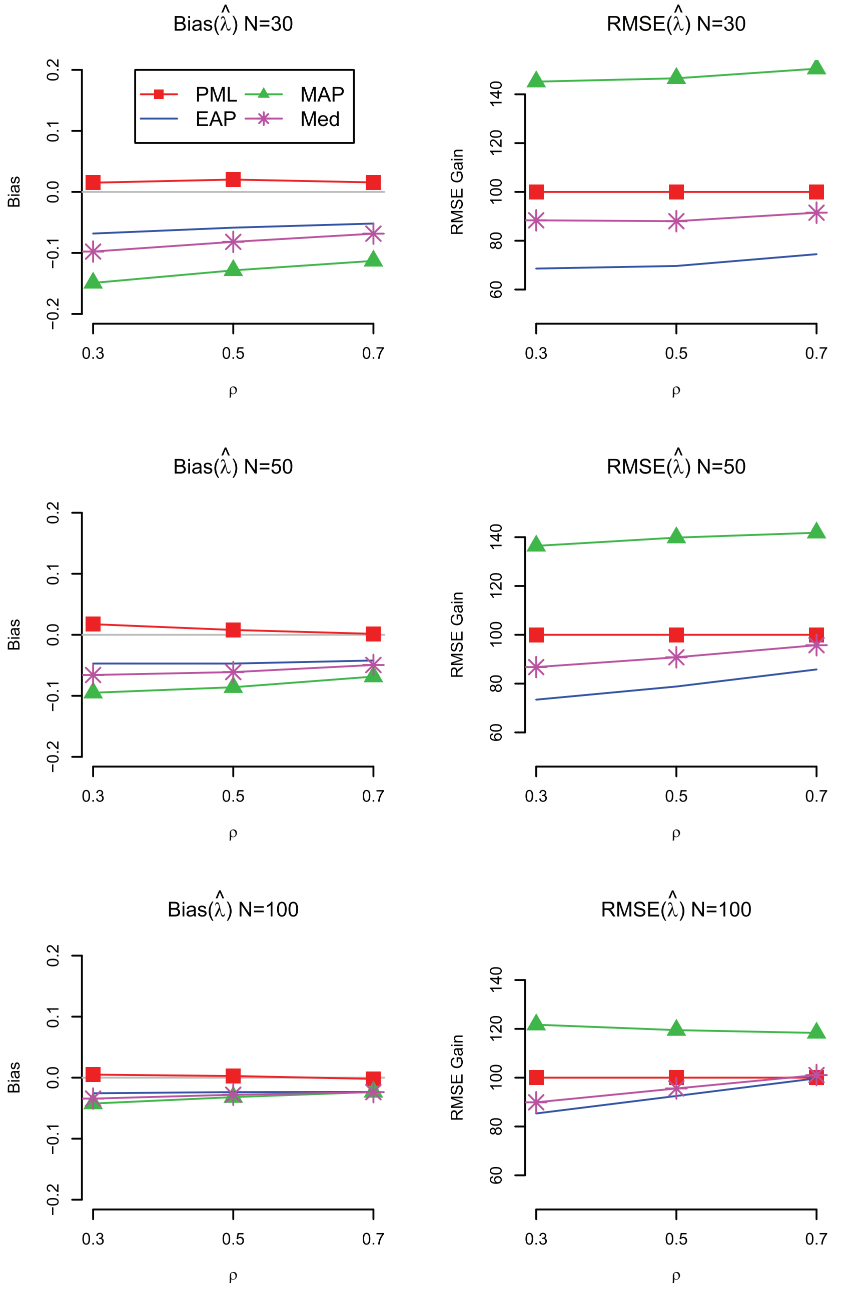


FIGURE 5 | Simulation Study 1: Bias **(left panels)** and RMSE gain **(right panels)** of the estimators of a loading (λ) as a function of the sample size and the magnitude of the true correlation. For the RMSE gain, PML is used as a reference method; values smaller than 100 indicate that the RMSE of the respective method is lower than the RMSE of the reference method; PML = mode of joint posterior; MAP = mode of marginal posterior; Med = median of marginal posterior; EAP = mean of marginal posterior. Results are shown for models with uniform prior distributions for the correlation and the standardized loadings.

package rsem (Yuan and Zhang, 2012). The robust estimation procedure provides Huber-Type M-estimates of means and covariances and has been shown to produce more efficient parameters, particularly for distributions with heavy tails (Yuan et al., 2004). For comparison purposes, we also included an ULS estimation method. The ULS estimate is defined as:

$$\widehat{\boldsymbol{\theta}}_{\mathrm{ULS}} = \arg\min_{\boldsymbol{\theta}} \operatorname{tr}\left\{(\mathbf{S} - \boldsymbol{\Sigma}(\boldsymbol{\theta}))^{\mathrm{T}}\,(\mathbf{S} - \boldsymbol{\Sigma}(\boldsymbol{\theta}))\right\} \qquad (34)$$

The ULS method was also specified with robustly estimated means and covariances (ULSR; Yuan et al., 2004). Finally, the MCMC method was applied to obtain the mode (MAP), mean (EAP), and median (Med) of the marginal posterior distributions. We specified uniform distributions for the standardized loadings and the correlation. For the standard deviations of the indicator variables, we used improper prior distributions that are constant for all conditions of the simulation and all (Bayesian) analysis

TABLE 7 | Simulation study 1: bias and RMSE for the latent correlation as a function of different misspecified prior distributions and the sample size.

			Bias					**RMSE**				
			μ_λ = 0.5			**μ_λ = 0.8**		**μ_λ = 0.5**			**μ_λ = 0.8**	
Meth	**μ_ρ**	**ν_ρ**	**ν_λ = 0**	**ν_λ = 1**	**ν_λ = 3**	**ν_λ = 1**	**ν_λ = 3**	**ν_λ = 0**	**ν_λ = 1**	**ν_λ = 3**	**ν_λ = 1**	**ν_λ = 3**
			N = 30									
PML	0	0	–0.038	0.009	0.019	–**0.139**	–0.022	0.391	**0.393**	0.379	0.361	0.367
	0.5	1	**0.161**	**0.196**	**0.201**	–0.008	**0.119**	0.388	**0.396**	**0.397**	0.325	0.344
	0.5	3	**0.094**	**0.120**	**0.123**	–0.019	**0.089**	0.281	0.281	0.278	0.261	0.263
	0.2	1	–0.027	0.021	0.026	–**0.114**	–0.002	0.320	0.322	0.317	0.323	0.318
	0.2	3	–**0.086**	–0.040	–0.039	–**0.136**	–0.046	0.269	0.259	0.255	0.290	0.272
	0.8	1	**0.297**	**0.328**	**0.347**	**0.112**	**0.237**	**0.432**	**0.444**	**0.447**	0.358	0.384
	0.8	3	**0.337**	**0.360**	**0.392**	**0.153**	**0.271**	**0.427**	**0.436**	**0.453**	0.329	0.373
EAP	0	0	–**0.141**	–**0.070**	–0.050	–**0.180**	–**0.103**	0.303	0.307	0.314	0.313	0.314
	0.5	1	**0.131**	**0.151**	**0.161**	–0.035	**0.056**	0.222	0.264	0.276	0.247	0.268
	0.5	3	0.004	0.038	0.048	–**0.066**	0.011	0.157	0.188	0.196	0.199	0.202
	0.2	1	–**0.121**	–**0.071**	–**0.058**	–**0.163**	–**0.088**	0.242	0.248	0.252	0.277	0.268
	0.2	3	–**0.178**	–**0.140**	–**0.131**	–**0.192**	–**0.131**	0.245	0.231	0.228	0.266	0.243
	0.8	1	**0.360**	**0.340**	**0.330**	**0.145**	**0.222**	0.387	0.377	0.375	0.282	0.319
	0.8	3	**0.309**	**0.320**	**0.327**	**0.151**	**0.231**	0.334	0.346	0.352	0.252	0.300
			N = 100									
PML	0	0	–0.011	0.002	0.006	–0.039	–0.024	0.203	0.196	0.194	0.194	0.181
	0.5	1	**0.089**	**0.090**	**0.090**	0.031	0.035	**0.273**	**0.261**	**0.255**	**0.231**	**0.212**
	0.5	3	0.032	0.042	0.046	0.002	0.014	0.183	0.180	0.179	0.173	0.162
	0.2	1	–0.008	0.004	0.008	–0.034	–0.020	0.190	0.184	0.183	0.186	0.173
	0.2	3	–0.036	–0.024	–0.020	–0.055	–0.040	0.172	0.165	0.164	0.176	0.164
	0.8	1	**0.188**	**0.185**	**0.163**	**0.091**	**0.082**	**0.332**	**0.324**	**0.302**	**0.262**	**0.234**
	0.8	3	**0.218**	**0.216**	**0.203**	**0.135**	**0.118**	**0.340**	**0.332**	**0.321**	**0.286**	**0.253**
EAP	0	0	–**0.059**	–0.034	–0.024	–**0.113**	–**0.069**	0.191	0.187	0.187	**0.207**	0.185
	0.5	1	0.039	0.054	**0.059**	–0.053	–0.016	0.199	**0.207**	**0.208**	0.195	0.180
	0.5	3	–0.016	0.003	0.011	–**0.069**	–0.030	0.155	0.156	0.157	0.171	0.154
	0.2	1	–**0.062**	–0.039	–0.029	–**0.109**	–**0.066**	0.176	0.171	0.171	0.197	0.174
	0.2	3	–**0.095**	–**0.072**	–**0.061**	–**0.128**	–**0.086**	0.174	0.163	0.160	0.195	0.170
	0.8	1	**0.191**	**0.182**	**0.178**	0.038	**0.059**	**0.278**	**0.270**	**0.266**	**0.213**	**0.204**
	0.8	3	**0.177**	**0.173**	**0.176**	0.041	**0.064**	**0.248**	**0.249**	**0.253**	0.195	0.189

RMSE = root mean square error. PML = penalized maximum likelihood (mode of joint posterior); EAP = mean of marginal posterior; μ_ρ = prior guess for latent correlation; ν_ρ = prior sample size for latent correlation; μ_λ = prior guess for loading ν_λ = prior sample size for loading; biases smaller than –0.05 or larger than 0.05 are printed in bold. For each sample size condition, RMSE values larger than the RMSE for the PML method with uniform priors on loadings (μ_λ = 0.5, ν_λ = 0) and the correlation (μ_ρ = 0, ν_ρ = 0) are printed in bold. The true correlation and loadings were set to ρ = 0.50 and λ = 0.50, respectively.

models. The R code for the data-generating model and the different analysis models is provided in **Supplementary 4** at https://doi.org/fwr7.

Results

Table 8 shows the bias and RMSE for the different estimators of the correlation for conditions with a true correlation of ρ = 0.50 (see **Supplementary 5** for detailed information about the other conditions). We again report RMSE gain with PML as the reference method (i.e., values larger/smaller than 100 indicate that the RMSE for the respective method is larger/smaller than for PML). Overall, the results confirm the previous findings that the EAP and Med produce (negatively) bias estimates of the correlation. However, with smaller sample sizes ($N \leq 50$), the estimates of the EAP and the Med were also more accurate in terms of the RMSE gain. When the variables strongly deviated from normality and the sample size was large, the robust estimation approaches (PMLR, ULSR, and ULSR) were slightly more efficient (i.e., smaller SD of the parameter estimates) than the different Bayesian point estimates. However, the results also reveal that, for moderate deviations from normality, the conclusions about the performance of the different Bayesian point estimates are relatively robust against distributional misspecifications. In addition, it should be mentioned that the multivariate mode (PML) consistently outperformed the univariate mode (MAP) across all conditions.

Furthermore, we obtained similar results for the estimates of the loadings (see **Supplementary 6**); that is, the estimates produced by the EAP and Med were slightly biased but overall more accurate in terms of RMSE than the other approaches. Again, the performance differences between the Bayesian point estimates were relatively robust against deviations

TABLE 8 | Simulation study 2: bias and RMSE for the latent correlation as a function of the distribution of the observed variables (skewness and kurtosis) and the sample size.

		Bias							RMSE Gain						
skew/kurt	***N***	**PML**	**PMLR**	**ULS**	**ULSR**	**MAP**	**Med**	**EAP**	**PML**	**PMLR**	**ULS**	**ULSR**	**MAP**	**Med**	**EAP**
0/0	30	–**0.065**	–**0.067**	–**0.010**	–**0.011**	–**0.003**	–**0.131**	–**0.154**	100	**101**	99	**101**	**116**	80	77
	50	–0.019	–0.020	0.016	0.017	–0.018	–**0.085**	–**0.092**	100	**101**	96	98	**114**	85	82
	100	–0.014	–0.016	0.008	0.008	–**0.051**	–**0.063**	–**0.061**	100	**101**	94	95	**116**	95	92
	500	–0.002	–0.002	0.002	0.001	–0.011	–0.012	–0.013	100	**101**	99	100	**102**	**101**	**101**
0/3	30	–0.038	–0.033	0.000	0.007	0.027	–**0.106**	–**0.130**	100	100	**101**	98	**119**	80	77
	50	–0.042	–0.036	0.002	0.001	–0.037	–**0.102**	–**0.109**	100	98	97	94	**108**	84	81
	100	–0.007	–0.004	0.015	0.016	–0.041	–**0.058**	–**0.056**	100	94	96	91	**117**	95	92
	500	–0.003	–0.005	0.001	–0.002	–0.011	–0.013	–0.014	100	97	99	95	**104**	**101**	**101**
0/7	30	–0.043	–0.024	0.007	0.016	0.024	–**0.108**	–**0.133**	100	95	**101**	98	**116**	82	79
	50	–0.021	–0.011	0.019	0.021	–0.021	–**0.085**	–**0.092**	100	96	97	93	**115**	87	84
	100	–0.004	–0.004	0.018	0.013	–0.037	–**0.054**	–**0.052**	100	92	98	89	**117**	94	91
	500	–0.003	–0.002	0.001	0.002	–0.012	–0.013	–0.014	100	92	99	91	**101**	100	**101**
1/3	30	–0.048	–0.046	0.003	0.004	0.026	–**0.110**	–**0.136**	100	99	**103**	99	**119**	81	79
	50	–0.023	–0.018	0.016	0.019	–0.019	–**0.086**	–**0.093**	100	98	95	94	**111**	82	79
	100	–0.006	–0.012	0.013	0.008	–0.042	–**0.055**	–**0.054**	100	99	94	93	**116**	96	92
	500	0.002	–0.002	0.005	0.001	–0.006	–0.008	–0.009	100	97	99	96	**103**	**101**	**101**
2/7	30	–0.015	–0.009	0.033	0.029	**0.064**	–**0.078**	–**0.105**	100	95	99	93	**114**	77	74
	50	–0.017	–0.017	0.017	0.001	–0.014	–**0.079**	–**0.086**	100	94	97	94	**113**	87	84
	100	–0.003	–0.015	0.015	0.001	–0.035	–**0.051**	–0.049	100	93	98	89	**117**	96	93
	500	0.003	–0.014	0.007	–0.011	–0.006	–0.007	–0.008	100	95	99	94	**103**	100	100

RMSE = root mean square error. PML = penalized maximum likelihood; PMLR = penalized maximum likelihood with robustly estimated covariance matrix; ULS = unweighted least squares; ULSR = unweighted least squares with robustly estimated covariance; MAP = mode of marginal posterior; Med = median of marginal posterior; EAP = mean of marginal posterior; μ_ρ = prior guess for latent correlation; ν_ρ = prior sample size for latent correlation; skew = skewness; kurt = kurtosis; Biases smaller than –0.05 or larger than 0.05 are printed in bold. For the RMSE gain, the RMSE of PML estimation is used as a reference method; values smaller than 100 indicate that the RMSE of the respective method is lower than the RMSE of the reference method. The true correlation and loadings were set to ρ = 0.50 and λ = 0.50, respectively.

from normality, and larger sample sizes were needed to show the gains in efficiency for the robust estimation approaches (with the exception that the ULS method performed less favorably with $N = 500$).

DISCUSSION

In this article, we showed that a Bayesian approach can stabilize the parameter estimates of a CFA model in small sample size conditions. We discussed different Bayesian point estimators—the mode (PML) of the joint posterior distribution and the mean (EAP), median (Med), or mode (MAP) of the marginal posterior distribution—and evaluated their performance in two simulation studies from a frequentist point of view. The results showed that the EAP outperformed the PML in terms of RMSE and produced more accurate estimates of latent correlations in many conditions. These performance gains can be explained by the fact that the EAP pulls large estimates toward zero (i.e., shrinkage effect), resulting in less variable estimates of the correlation. However, there is a turning point at which, with a larger true correlation, the EAP is less accurate than the PML because the bias introduced by the shrinkage effect outweighs the gains in efficiency (see Choi et al., 2011). As expected, with larger sample sizes, the differences between the Bayesian point estimates vanished, and the different Bayesian estimators performed similarly. We also suggested the four-parameter beta distribution as a prior distribution for loadings and correlations and argued that it could often be advantageous to choose a parameterization in which the main parameters of interest are bounded (Muthén and Asparouhov, 2012; Merkle and Rosseel, 2018). Another finding of our simulation study was that selecting weakly informative four-parameter beta distributions as priors helped stabilize parameter estimates (e.g., Depaoli and Clifton, 2015; van Erp et al., 2018). Importantly, this was also the case when the prior was mildly misspecified.

The main limitation of our simulation study is that we used a very simple CFA model with only two latent factors and a small number of items with no cross-loadings (i.e., simple structure). It would be straightforward to extend the discussed approaches to models with more latent factors. In constrained ML estimation and PML estimation, appropriate determinant constraints could be implemented to ensure the positive definiteness of the correlation matrix of latent variables (Wothke, 1993; Rousseeuw and Molenberghs, 1994). For the MCMC method, determinant constraints could be introduced in the Metropolis-Hastings step to check for the positive definiteness of the correlation matrix (see Browne, 2006).

In addition, we only assessed the quality of statistical inferences (i.e., coverage rates) with normally distributed variables. It would be an interesting topic for future research also to investigate robust estimation approaches for Bayesian

CFA models. First, one could use robustly estimated covariance matrices as input for Bayesian CFA models. In this case, robust standard errors must also be applied in Bayesian estimation because the model is misspecified (Müller, 2013; Walker, 2013; Bissiri et al., 2016). Second, models with more flexible distributions for the latent factors and residuals could be applied (Lin et al., 2018). For example, Zhang et al. (2014) proposed a Bayesian factor analysis model with scaled t-distributions (and freely estimated degrees of freedom) that are less sensitive to outlier values.

The results of the present study could be extended into several directions. First, it would be interesting to explore further the potential of PML estimation for CFA models in challenging data constellations (e.g., small samples, complex models; Rosseel, 2020). PML estimation seems to be particularly promising when researchers do not need full access to the posterior distribution and are only interested in obtaining stable point estimates for the parameters of interest. In contrast to simulation-based MCMC techniques, which can be slow and challenging to implement, PML estimation shares the advantage of traditional ML estimation that a deterministic optimization of the log-posterior is performed with clear convergence criteria and reasonable computational efficiency (Cousineau and Helie, 2013). In Simulation Study 1, for example, the average run time was about 2 min for MCMC but only two seconds for PML. The run time differences could be considerably larger with more complex models (Chung et al., 2013). Second, data sets in psychological research often have a multilevel structure (e.g., individuals are nested within clusters/groups) and, in many applications, it is of interest to analyze relationships among latent constructs at both levels of analysis (e.g., individual level and group level; Heck and Thomas, 2015). However, a notable finding in the multilevel literature is that a substantial number of groups is needed to obtain stable parameter estimates of group-level relationships (Lüdtke et al., 2011; Li and Beretvas, 2013; Kelava and Brandt, 2014; Can et al., 2015). Thus, an important topic for future research could be to extend the Bayesian approaches discussed here to multilevel CFA models (Kim et al., 2016). Finally, it would be interesting to compare the different Bayesian estimators to other approaches that have been suggested as solutions for estimation problems in small sample size conditions (Rosseel, 2020). For example, a two-step approach, such as factor score regression, has been suggested as a robust alternative to SEMs in challenging data constellations (Smid and Rosseel, 2020). Besides, alternative error correction approaches could be used that introduce lower bounds to circumvent small estimates of reliability in order to stabilize the estimation of latent correlations (Grilli and Rampichini, 2011). Using these lower bounds, l can be translated into a uniform distribution of standardized loadings on the interval $[l, 1]$. However, using lower bounds for the indicator-specific reliability larger than zero possibly introduces too much information. Besides, parameter estimates could be pretty sensitive to the subjective choice of lower bounds.

To conclude, this article showed that the Bayesian approach has great potential for estimating CFA models with small sample sizes. Using simulated data, we showed that the four-parameter beta distribution can be used as a prior distribution for standardized loadings and latent correlations to stabilize parameter estimates in challenging data constellations. However, in real applications, the specification of prior distributions should be accompanied by a sensitivity analysis that tests how sensitive the resulting parameter estimates are to different specifications of prior information (Depaoli and van de Schoot, 2017).

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://doi.org/fwr7

REFERENCES

Anderson, J. C., and Gerbing, D. W. (1984). The effect of sampling error on convergence, improper solutions, and goodness-of-fit indices for maximum likelihood confirmatory factor analysis. *Psychometrika* 49, 155–173. doi: 10.1007/bf02294170

Arbuckle, J. L. (2017). *Amos (Version 25.0) [Computer Program]*. Chicago: IBM SPSS.

Azevedo, C. L. N., and Andrade, D. F. (2013). CADEM: a conditional augmented data EM algorithm for fitting one parameter probit models. *Braz. J. Probab. Stat.* 27, 245–262. doi: 10.1214/11-BJPS172

Azevedo, C. L. N., Andrade, D. F., and Fox, J.-P. (2012). A Bayesian generalized multiple group IRT model with model-fit assessment tools. *Comput. Stat. Data Anal.* 56, 4399–4412. doi: 10.1016/j.csda.2012.03.017

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

Bissiri, P. G., Holmes, C. C., and Walker, S. G. (2016). A general framework for updating belief distributions. *J. R. Stat. Soc. Series B Stat. Methodol.* 78, 1103–1130. doi: 10.1111/rssb.12158

Bollen, K. A. (1989). *Structural Equations With Latent Variables*. New York, NY: Wiley.

Bolstad, W., and Curran, J. M. (2017). *Introduction to Bayesian Statistics*. Hoboken, NJ: Wiley.

Boomsma, A. (1985). Nonconvergence, improper solutions, and starting values in LISREL maximum likelihood estimation. *Psychometrika* 50, 229–242. doi: 10.1007/bf02294248

Browne, M. W. (1974). Generalized least squares estimators in the analysis of covariance structures. *S. Afr. Stat. J.* 8, 1–24.

Browne, W. J. (2006). MCMC algorithms for constrained variance matrices. *Comput. Stat. Data Anal.* 50, 1655–1677. doi: 10.1016/j.csda.2005.02.008

Browne, W. J., and Draper, D. (2006). A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Anal.* 1, 473–514. doi: 10.1214/06-BA117

Bürkner, P.-C. (2020). Analysing standard progressive matrices (SPM-LS) with Bayesian item response models. *J. Intell.* 8:5. doi: 10.3390/jintelligence8010005

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

Carlin, B. P., and Louis, T. A. (2009). *Bayesian Methods for Data Analysis*. Boca Raton, FL: Chapman and Hall–CRC.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., et al. (2017). Stan: a probabilistic programming language. *J. Stat. Softw.* 76, 1–32. doi: 10.18637/jss.v076.i01

Chen, F., Bollen, K. A., Paxton, P., Curran, P. J., and Kirby, J. B. (2001). Improper solutions in structural equation models: causes, consequences, and strategies. *Soc. Methods Res.* 29, 468–508. doi: 10.1177/0049124101029004003

Chen, J., Choi, J., Weiss, B. A., and Stapleton, L. (2014). An empirical evaluation of mediation effect analysis with manifest and latent variables using markov chain monte carlo and alternative estimation methods. *Struct. Equ. Modeling* 21, 253–262. doi: 10.1080/10705511.2014.882688

Choi, J., Kim, S., Chen, J., and Dannels, S. (2011). A comparison of maximum-likelihood and Bayesian estimation for polychoric correlation using monte carlo simulation. *J. Educ. Behav. Stat.* 36, 523–549. doi: 10.3102/1076998610381398

Choi, J., and Levy, R. (2017). Markov chain monte carlo estimation methods for structural equation modeling: a comparison of subject-level data and moment-level data approaches. *Biometr. Biostat. Int. J.* 6, 463–474. doi: 10.15406/bbij.2017.06.00182

Chung, Y., Rabe-Hesketh, S., Dorie, V., Gelman, A., and Liu, J. (2013). A nondegenerate penalized likelihood estimator for variance parameters in multilevel models. *Psychometrika* 78, 685–709. doi: 10.1007/s11336-013-9328-2

Cole, S. R., Chu, H., and Greenland, S. (2014). Maximum likelihood, profile likelihood, and penalized likelihood: a primer. *Am. J. Epidemiol.* 179, 252–260. doi: 10.1093/aje/kwt245

Cousineau, D., and Helie, S. (2013). Improving maximum likelihood estimation using prior probabilities: a tutorial on maximum a posteriori estimation and an examination of the Weibull distribution. *Tutor. Quant. Methods Psychol.* 9, 61–71. doi: 10.20982/tqmp.09.2.p061

Cowles, M. K., and Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: a comparative review. *J. Am. Stat. Assoc.* 91, 883–904. doi: 10.1080/01621459.1996.10476956

de Valpine, P., Turek, D., Paciorek, C. J., Anderson-Bergman, C., Temple Lang, D., and Bodik, R. (2017). Programming with models: writing statistical algorithms for general model structures with NIMBLE. *J. Comput. Graph. Stat.* 26, 403–413. doi: 10.1080/10618600.2016.1172487

DeCarlo, L. T., Kim, Y., and Johnson, M. S. (2011). A hierarchical rater model for constructed responses, with a signal detection rater model. *J. Educ. Meas.* 48, 333–356. doi: 10.1111/j.1745-3984.2011.00143.x

Depaoli, S., and Clifton, J. P. (2015). A Bayesian approach to multilevel structural equation modeling with continuous and dichotomous outcomes. *Struct. Equat. Model.* 22, 327–351. doi: 10.1080/10705511.2014.937849

Depaoli, S., and van de Schoot, R. (2017). Improving transparency and replication in Bayesian statistics: the WAMBS-checklist. *Psychol. Methods* 22, 240–261. doi: 10.1037/met0000065

Dolan, C. V., and Molenaar, P. C. M. (1991). A comparison of four methods of calculating standard errors of maximum-likelihood estimates in the analysis of covariance structure. *Br. J. Math. Stat. Psychol.* 44, 359–368. doi: 10.1111/j.2044-8317.1991.tb00967.x

Draper, D. (2008). "Bayesian multilevel analysis and MCMC," in *Handbook of Multilevel Analysis*, eds J. de Leuuw and E. Meijer (New York, NY: Springer), 77–139. doi: 10.1007/978-0-387-73186-5_2

Efron, B. (2015). Frequentist accuracy of Bayesian estimates. *J. R. Stat. Soc. Ser. B* 77, 617–646. doi: 10.1111/rssb.12080

Erosheva, E. A., and Curtis, S. M. (2017). Dealing with reflection invariance in Bayesian factor analysis. *Psychometrika* 82, 295–307. doi: 10.1007/s11336-017-9564-y

Fan, J., Li, R., Zhang, C.-H., and Zou, H. (2020). *Statistical Foundations of Data Science*. Boca Raton, FL: CRC Press.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika* 80, 27–38. doi: 10.1093/biomet/80.1.27

Foldnes, N., and Olsson, U. H. (2016). A simple simulation technique for nonnormal data with prespecified skewness, kurtosis, and covariance matrix. *Multivar. Behav. Res.* 51, 207–219. doi: 10.1080/00273171.2015.1133274

Fox, J.-P. (2010). *Bayesian Item Response Modeling: Theory and Applications*. New York, NY: Springer.

Fox, J.-P., Mulder, J., and Sinharay, S. (2017). Bayes factor covariance testing in item response models. *Psychometrika* 82, 979–1006. doi: 10.1007/s11336-017-9577-6

Gagné, P., and Hancock, G. R. (2006). Measurement model quality, sample size, and solution propriety in confirmatory factor models. *Multivar. Behav. Res.* 41, 65–83. doi: 10.1207/s15327906mbr4101_5

Galindo-Garre, F., and Vermunt, J. (2006). Avoiding boundary estimates in latent class analysis by Bayesian posterior mode estimation. *Behaviormetrika* 33, 43–59. doi: 10.2333/bhmk.33.43

Galindo-Garre, F., Vermunt, J. K., and Bergsma, W. P. (2004). Bayesian posterior estimation of logit parameters with small samples. *Sociol. Methods Res.* 33, 88–117. doi: 10.1177/0049124104265997

Gao, F., and Chen, L. (2005). Bayesian or non-bayesian: a comparison study of item parameter estimation in the three-parameter logistic model. *Appl. Meas. Educ.* 18, 351–380. doi: 10.1207/s15324818ame1804_2

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2014). *Bayesian Data Analysis*. London: Chapman & Hall.

Gerbing, D. W., and Anderson, J. C. (1987). Improper solutions in the analysis of covariance structures: Their interpretability and a comparison of alternate respecifications. *Psychometrika* 52, 99–111. doi: 10.1007/bf02293958

Gill, J. (2007). *Bayesian Methods for the Social and Behavioral Sciences*. Boca Raton, FL: CRC Press.

Gokhale, D. V., and Press, S. J. (1982). Assessment of a prior distribution for the correlation coefficient in a bivariate normal distribution. *J. R. Stat. Soc. Ser. A* 145, 237–249. doi: 10.2307/2981537

Gonzalez, R., and Griffin, D. (2001). Testing parameters in structural equation modeling: every "one" matters. *Psychol. Methods* 6, 258–269. doi: 10.1037/1082-989X.6.3.258

Grilli, L., and Rampichini, C. (2011). The role of sample cluster means in multilevel models. *Methodology* 7, 121–133. doi: 10.1027/1614-2241/a000030

Harwell, M. R., and Baker, F. B. (1991). The use of prior distributions in marginalized Bayesian item parameter estimation: a didactic. *Appl. Psychol. Meas.* 15, 375–389. doi: 10.1177/014662169101500409

Hayashi, K., and Arav, M. (2006). Bayesian factor analysis when only a sample covariance matrix is available. *Educ. Psychol. Meas.* 66, 272–284. doi: 10.1177/0013164405278583

Hecht, M., Gische, C., Vogel, D., and Zitzmann, S. (2020). Integrating out nuisance parameters for computationally more efficient Bayesian estimation–an illustration and tutorial. *Struct. Equat. Model.* 27, 483–493. doi: 10.1080/10705511.2019.1647432

Heck, R. H., and Thomas, S. L. (2015). *An Introduction to Multilevel Modeling Techniques: MLM and SEM Approaches Using Mplus*. New York, NY: Routledge.

Heinze, G., and Schemper, M. (2002). A solution to the problem of separation in logistic regression. *Stat. Med.* 21, 2409–2419. doi: 10.1002/sim.1047

Held, L., and Bové, S. (2014). *Applied Statistical Inference*. New York, NY: Springer.

Hoeschele, I., and Tier, B. (1995). Estimation of variance components of threshold characters by marginal posterior modes and means *via* Gibbs sampling. *Genet. Sel. Evol.* 27, 519–540. doi: 10.1186/1297-9686-27-6-519

Hoff, P. D. (2009). *A First Course in Bayesian Statistical Methods*. New York, NY: Springer.

Holtmann, J., Koch, T., Lochner, K., and Eid, M. (2016). A comparison of ML, WLSMV and Bayesian methods for multilevel structural equation models in small samples: a simulation study. *Multivariate Behav. Res.* 51, 661–680. doi: 10.1080/00273171.2016.1208074

Hoogland, J. J., and Boomsma, A. (1998). Robustness studies in covariance structure modeling: an overview and a meta analysis. *Sociol. Methods Res.* 26, 329–368. doi: 10.1177/0049124198026003003

Hox, J., van de Schoot, R., and Matthijsse, S. (2012). How few countries will do? Comparative survey analysis from a Bayesian perspective. *Surv. Res. Methods* 6, 87–93. doi: 10.18148/srm/2012.v6i2.5033

Hox, J. J., Moerbeek, M., Kluytmans, A., and 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. *Front. Psychol.* 5:78. doi: 10.3389/fpsyg.2014.00078

Hoyle, R. H. (2012). *Handbook of Structural Equation Modeling*. New York, NY: Guilford Press.

Hoyle, R. H., and Kenny, D. A. (1999). "Sample size, reliability, and tests of statistical mediation," in *Statistical Strategies for Small Sample Research*, ed. R. H. Hoyle (Thousand Oaks, CA: Sage), 195–222.

Huang, P.-H. (2018). A penalized likelihood method for multi-group structural equation modelling. *Br. J. Math. Stat. Psychol.* 71, 499–522. doi: 10.1111/bmsp.12130

Jackman, S. (2009). *Bayesian Analysis for the Social Sciences*. New York, NY: Wiley.

Jackson, D. L., Gillaspy, J. A. Jr., and Purc-Stephenson, R. (2009). Reporting practices in confirmatory factor analysis: an overview and some recommendations. *Psychol. Methods* 14, 6–23. doi: 10.1037/a0014694

Jacobucci, R., and Grimm, K. J. (2018). Comparison of frequentist and Bayesian regularization in structural equation modeling. *Struct. Equat. Model.* 25, 639–649. doi: 10.1080/10705511.2017.1410822

Jin, S., Moustaki, I., and Yang-Wallentin, F. (2018). Approximated penalized maximum likelihood for exploratory factor analysis: an orthogonal case. *Psychometrika* 83, 628–649. doi: 10.1007/s11336-018-9623-z

Johnson, M. J., and Sinharay, S. (2016). "Bayesian estimation," in *Handbook of Item Response Theory: Statistical Tools*, Vol. 2, ed. W. J. van der Linden (Boca Raton, FL: Chapman & Hall/CRC), 237–257.

Johnson, N. L., Kotz, S., and Balakrishnan, N. (1994). *Continuous Univariate Distributions*, Vol. 2. New York, NY: Wiley.

Johnson, T. R., and Kuhn, K. M. (2015). Simulation-based Bayesian inference for latent traits of item response models: introduction to the ltbayes package for R. *Behav. Res. Methods* 47, 1309–1327. doi: 10.3758/s13428-014-0540-5

Junker, B. W., Patz, R. J., and VanHoudnos, N. M. (2016). "Markov chain Monte Carlo for item response models," in *Handbook of Item Response Theory: Statistical Tools*, Vol. 2, ed. W. J. van der Linden (Boca Raton, FL: Chapman & Hall/CRC), 271–312. doi: 10.1201/b19166-15

Kaplan, D., and Depaoli, S. (2012). "Bayesian structural equation modeling," in *Handbook of Structural Equation Modeling*, ed. R. Hoyle (New York, NY: Guilford Press), 650–673.

Kelava, A., and Brandt, H. (2014). A general non-linear multilevel structural equation mixture model. *Front. Psychol.* 5:748. doi: 10.3389/fpsyg.2014.00748

Kenny, D. A., and Zautra, A. (1995). The trait-state-error model for multiwave data. *J. Consult. Clin. Psychol.* 63, 52–59. doi: 10.1037/0022-006x.63.1.52

Kieftenbeld, V., and Natesan, P. (2012). Recovery of graded reponse model parameters: a comparison of marginal maximum likelihood and markov chain monte carlo estimation. *Appl. Psychol. Meas.* 36, 399–419. doi: 10.1177/0146621612446170

Kim, E. S., Dedrick, R. F., Cao, C., and Ferron, J. M. (2016). Multilevel factor analysis: reporting guidelines and a review of reporting practices. *Multiv. Behav. Res.* 51, 881–898. doi: 10.1080/00273171.2016.1228042

Kline, R. B. (2016). *Principles and Practice of Structural Equation Modeling*. New York, NY: Guilford Press.

Ledgerwood, A., and Shrout, P. E. (2011). The trade-off between accuracy and precision in latent variable models of mediation processes. *J. Pers. Soc. Psychol.* 101, 1174–1188. doi: 10.1037/a0024776

Lee, S.-Y. (1981). A Bayesian approach to confirmatory factor analysis. *Psychometrika* 46, 153–160. doi: 10.1007/bf02293896

Lee, S.-Y. (1992). Bayesian analysis of stochastic constraints in structural equation models. *Br. J. Math. Stat. Psychol.* 45, 93–107. doi: 10.1111/j.2044-8317.1992.tb00979.x

Lee, S.-Y. (2007). *Structural Equation Modeling: A Bayesian Approach*. West Sussex: Wiley.

Lee, S.-Y., and Song, X.-Y. (2004). Evaluation of the Bayesian and maximum likelihood approaches in analyzing structural equation models with small sample sizes. *Multivar. Behav. Res.* 39, 653–686. doi: 10.1207/s15327906mbr3904_4

Li, X., and Beretvas, S. N. (2013). Sample size limits for estimating upper level mediation models using multilevel SEM. *Struct. Equat. Model.* 20, 241–264. doi: 10.1080/10705511.2013.769391

Lin, T. I., Wang, W. L., McLachlan, G. J., and Lee, S. X. (2018). Robust mixtures of factor analysis models using the restricted multivariate skew-t distribution. *Stat. Model.* 18, 50–72. doi: 10.1177/1471082x17718119

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

Lüdtke, O., Marsh, H. W., Robitzsch, A., and Trautwein, U. (2011). A 2 × 2 taxonomy of multilevel latent contextual models: accuracy–bias trade-offs in full and partial error correction models. *Psychol. Methods* 16, 444–467. doi: 10.1037/a0024376

Lüdtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., and Muthén, B. (2008). The multilevel latent covariate model: a new, more reliable approach to group-level effects in contextual studies. *Psychol. Methods* 13, 203–229. doi: 10.1037/a0012869

Lüdtke, O., Robitzsch, A., Kenny, D. A., and Trautwein, U. (2013). A general and flexible approach to estimating the social relations model using Bayesian methods. *Psychol. Methods* 18, 101–119. doi: 10.1037/a0029252

Lüdtke, O., Robitzsch, A., and Wagner, J. (2018). More stable estimation of the STARTS model: a Bayesian approach using Markov chain Monte Carlo techniques. *Psychol. Methods* 23, 570–593. doi: 10.1037/met0000155

MacCallum, R. C., Browne, M. W., and Cai, L. (2007). "Factor analysis models as approximations," in *Factor Analysis at 100*, eds R. Cudeck and R. C. MacCallum (Mahwah, NJ: Lawrence Erlbaum), 153–175.

Martin, J. K., and McDonald, R. P. (1975). Bayesian estimation in unrestricted factor analysis: a treatment for heywood cases. *Psychometrika* 40, 505–517. doi: 10.1007/bf02291552

Maydeu-Olivares, A. (2017). Maximum likelihood estimation of structural equation models for continuous data: standard errors and goodness of fit. *Struct. Equat. Model.* 24, 383–394. doi: 10.1080/10705511.2016.1269606

McNeish, D. (2016). On using Bayesian methods to address small sample problems. *Struct. Equat. Model.* 23, 750–773. doi: 10.1080/10705511.2016.1186549

Merkle, E. C., Fitzsimmons, E., Uanhoro, J., and Goodrich, B. (2020). Efficient Bayesian structural equation modeling in Stan. *arXiv* [Preprint] arXiv:2008.07733v1,

Merkle, E. C., Furr, D., and Rabe-Hesketh, S. (2019). Bayesian comparison of latent variable models: conditional versus marginal likelihoods. *Psychometrika* 84, 802–809. doi: 10.1007/s11336-019-09679-0

Merkle, E. C., and Rosseel, Y. (2018). blavaan: Bayesian structural equation models via parameter expansion. *J. Stat. Softw.* 85, 1–30. doi: 10.18637/jss.v085.i04

Miocevic, M., Levy, R., and MacKinnon, D. P. (2020). Different roles of prior distributions in the single mediator model with latent variables. *Multivar. Behav. Res.* 56, 20–40. doi: 10.1080/00273171.2019.1709405

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika* 51, 177–195. doi: 10.1007/bf02293979

Müller, U. K. (2013). Risk of Bayesian inference in misspecified models, and the sandwich covariance matrix. *Econometrica* 81, 1805–1849. doi: 10.3982/ECTA9097

Muthén, B. O. (2010). *Bayesian Analysis in Mplus: A Brief Introduction (Version 3)*. Available online at: https://www.statmodel.com/download/IntroBayesVersion%203.pdf (accessed October 8, 2020).

Muthén, B. O., and Asparouhov, T. (2012). Bayesian structural equation modeling: a more flexible representation of substantive theory. *Psychol. Methods* 17, 313–335. doi: 10.1037/a0026802

Muthén, L. K., and Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Struct. Equat. Model.* 9, 599–620. doi: 10.1207/s15328007sem0904_8

Muthén, L. K., and Muthén, B. O. (2012). *Mplus User's Guide*, 7th Edn. Los Angeles, CA: Muthén & Muthén.

Natesan, P. (2015). Comparing interval estimates for small sample ordinal CFA models. *Front. Psychol.* 6:1599. doi: 10.3389/fpsyg.2015.01599

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., et al. (2006). *Uncertain Judgements: Eliciting Expert Probabilities*. Chichester: Wiley.

Plummer, M. (2003). "JAGS: a program for analysis of Bayesian graphical models using Gibbs sampling," in *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*, Vol. 124, eds K. Hornik, F. Leisch, and A. Zeileis (Vienna: Technische Universität Wien), 1–10.

Poon, W.-Y. (1999). Bayesian analysis of square ordinal-ordinal tables. *Br. J. Math. Stat. Psychol.* 52, 111–124. doi: 10.1348/000711099158991

Press, S. J., and Shigemasu, K. (1989). "Bayesian inference in factor analysis," in *Contributions to Probability and Statistics*, eds L. J. Gleser, M. D. Perlman, S. J. Press, and A. R. Sampson (New York, NY: Springer), doi: 10.1007/978-1-4612-3678-8_18

Rindskopf, D. (1984). Structural equation models: empirical identification, Heywood cases, and related problems. *Sociol. Methods Res.* 13, 109–119. doi: 10.1177/0049124184013001004

Rindskopf, D. (2012). Next steps in Bayesian structural equation models: comments on, variations of, and extensions to Muthén and Asparouhov (2012). *Psychol. Methods* 17, 336–339. doi: 10.1037/a0027130

Robitzsch, A. (2020). *LAM: Some Latent Variable Models. R Package Version 0.5-15*. Available online at: https://CRAN.R-project.org/package=LAM (accessed May 9, 2020).

Rosseel, Y. (2012). lavaan: an R package for structural equation modeling. *J. Stat. Softw.* 48, 1–36. doi: 10.1002/9781119579038.ch1

Rosseel, Y. (2020). "Small sample solutions for structural equation modeling," in *Small Sample Size Solutions*, eds R. van de Schoot and M. Miocevic (London: Routledge), 226–238. doi: 10.4324/9780429273872-19

Rousseeuw, P. J., and Molenberghs, G. (1994). The shape of correlation matrices. *Am. Stat.* 48, 276–279. doi: 10.1080/00031305.1994.10476079

Savalei, V. (2014). Understanding robust corrections in structural equation modeling. *Struct. Equat. Model.* 21, 149–160. doi: 10.1080/10705511.2013.824793

Savalei, V. (2019). A comparison of several approaches for controlling measurement error in small samples. *Psychol. Methods* 24, 352–370. doi: 10.1037/met0000181

Savalei, V., and Kolenikov, S. (2008). Constrained vs. unconstrained estimation in structural equation modeling. *Psychol. Methods* 13, 150–170. doi: 10.1037/1082-989x.13.2.150

Schoenberg, R. (1997). Constrained maximum likelihood. *Comput. Econ.* 10, 251–266. doi: 10.1023/A:1008669208700

Silverman, B. W. (1998). *Density Estimation for Statistics and Data Analysis*. Boca Raton, FL: CRC.

Smid, S. C., McNeish, D., Miocevic, M., and van de Schoot, R. (2020). Bayesian versus frequentist estimation for structural equation models in small sample contexts: a systematic review. *Struct. Equat. Model.* 27, 131–161. doi: 10.1080/10705511.2019.1577140

Smid, S. C., and Rosseel, Y. (2020). "SEM with small samples: two-step modeling and factor score regression versus Bayesian estimation with informative priors," in *Small Sample Size Solutions*, eds R. van de Schoot and M. Miocevic (London: Routledge), 239–254. doi: 10.1080/10705511.2014.882686

Song, X.-Y., and Lee, S.-Y. (2009). *Basic and Advanced Bayesian Structural Equation Modeling: With Applications in the Medical and Behavioral Sciences*. Chichester: Wiley.

Spiegelhalter, D. J., Thomas, A., Best, N. G., Gilks, N. R., and Lunn, D. (2003). *BUGS: Bayesian Inference Using Gibbs Sampling*. Cambridge: MRC Biostatistics Unit.

Stark, P. B. (2015). Constraints versus priors. *SIAM/ASA J. Uncert. Quant.* 3, 586–598. doi: 10.1137/130920721

Taylor, J. M. (2019). Overview and illustration of Bayesian confirmatory factor analysis with ordinal indicators. *Pract. Assess. Res. Evaluat.* 24, 1–27.

Traub, R. E. (1994). *Reliability for the Social Sciences: Theory and Applications*. Thousand Oaks, CA: Sage Publications.

van de Schoot, R., Depaoli, S., King, R., Kramer, B., Märtens, K., Tadesse, M. G., et al. (2021). Bayesian statistics and modelling. *Nat. Rev. Methods Prim.* 1, 1–26. doi: 10.1038/s43586-020-00001-2

van Erp, S., Mulder, J., and Oberski, D. L. (2018). Prior sensitivity analysis in default Bayesian structural equation modeling. *Psychol. Methods* 23, 363–388. doi: 10.1037/met0000162

van Erp, S., Oberski, D. L., and Mulder, J. (2019). Shrinkage priors for Bayesian penalized regression. *J. Math. Psychol.* 89, 31–50. doi: 10.1016/j.jmp.2018.12.004

Walker, S. G. (2013). Bayesian inference with misspecified models. *J. Stat. Plan. Inference* 143, 1621–1633. doi: 10.1016/j.jspi.2013.05.013

Waller, N. G., and Feuerstahler, L. (2017). Bayesian modal estimation of the four-parameter item response model in real, realistic, and idealized data sets. *Multivar. Behav. Res.* 52, 350–370. doi: 10.1080/00273171.2017.1292893

Wolf, E. J., Harrington, K. M., Clark, S. L., and Miller, M. W. (2013). Sample size requirements for structural equation models: an evaluation of power, bias, and solution propriety. *Educ. Psychol. Meas.* 73, 913–934. doi: 10.1177/0013164413495237

Wothke, W. (1993). "Nonpositive definite matrices in structural modeling," in *Testing Structural Equation Models*, eds K. A. Bollen and J. S. Long (Newbury Park, CA: Sage), 256–293.

Yao, L. (2014). "Multidimensional item response theory for score reporting," in *Advances in Modern International Testing: Transition from Summative to Formative Assessment*, eds Y. Cheng and H.-H. Chang (Charlotte, NC: Information Age).

Yuan, K.-H., and Bentler, P. M. (2007). "Structural equation modeling," in *Handbook of Statistics 26: Psychometrics*, eds C. R. Rao and S. Sinharay (Amsterdam: North-Holland), 297–358.

Yuan, K.-H., Bentler, P. M., and Chan, W. (2004). Structural equation modeling with heavy tailed distributions. *Psychometrika* 69, 421–436. doi: 10.1007/BF02295644

Yuan, K.-H., and Chan, W. (2008). Structural equation modeling with near singular covariance matrices. *Comput. Stat. Data Anal.* 52, 4842–4858. doi: 10.1016/j.csda.2008.03.030

Yuan, K.-H., and Zhang, Z. (2012). Robust structural equation modeling with missing data and auxiliary variables. *Psychometrika* 77, 803–826. doi: 10.1007/s11336-012-9282-4

Zeng, L. (1997). Implementation of marginal Bayesian estimation with four-parameter beta prior distributions. *Appl. Psychol. Meas.* 21, 143–156. doi: 10.1177/01466216970212004

Zhang, J., Li, J., and Liu, C. (2014). Robust factor analysis using the multivariate t-distribution. *Stat. Sin.* 24, 291–312.

Zitzmann, S., and Hecht, M. (2019). Going beyond convergence in Bayesian estimation: why precision matters too and how to assess it. *Struct. Equat. Model.* 26, 646–661. doi: 10.1080/10705511.2018.1545232

Zitzmann, S., Lüdtke, O., Robitzsch, A., and Hecht, M. (2021). On the performance of Bayesian approaches in small samples: a comment on Smid, McNeish, Miocevic, and van de Schoot (2020). *Struct. Equat. Model.* 28, 40–50. doi: 10.1080/10705511.2020.1752216

Zitzmann, S., Lüdtke, O., Robitzsch, A., and Marsh, H. W. (2016). A Bayesian approach for estimating multilevel latent contextual models. *Struct. Equat. Model.* 23, 661–679. doi: 10.1080/10705511.2016.1207179

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

METHODS
published: 11 May 2021
doi: 10.3389/fpsyg.2021.614236

A Similarity-Weighted Informative Prior Distribution for Bayesian Multiple Regression Models

*Christoph König**

Department of Educational Psychology, Institute of Psychology, Goethe University Frankfurt, Frankfurt, Germany

Specifying accurate informative prior distributions is a question of carefully selecting studies that comprise the body of comparable background knowledge. Psychological research, however, consists of studies that are being conducted under different circumstances, with different samples and varying instruments. Thus, results of previous studies are heterogeneous, and not all available results can and should contribute equally to an informative prior distribution. This implies a necessary weighting of background information based on the similarity of the previous studies to the focal study at hand. Current approaches to account for heterogeneity by weighting informative prior distributions, such as the power prior and the meta-analytic predictive prior are either not easily accessible or incomplete. To complicate matters further, in the context of Bayesian multiple regression models there are no methods available for quantifying the similarity of a given body of background knowledge to the focal study at hand. Consequently, the purpose of this study is threefold. We first present a novel method to combine the aforementioned sources of heterogeneity in the similarity measure ω. This method is based on a combination of a propensity-score approach to assess the similarity of samples with random- and mixed-effects meta-analytic models to quantify the heterogeneity in outcomes and study characteristics. Second, we show how to use the similarity measure ω as a weight for informative prior distributions for the substantial parameters (regression coefficients) in Bayesian multiple regression models. Third, we investigate the performance and the behavior of the similarity-weighted informative prior distribution in a comprehensive simulation study, where it is compared to the normalized power prior and the meta-analytic predictive prior. The similarity measure ω and the similarity-weighted informative prior distribution as the primary results of this study provide applied researchers with means to specify accurate informative prior distributions.

Keywords: informative prior distributions, prior information, heterogeneity, similarity, Bayesian multiple regression, comparability

OPEN ACCESS

Edited by:
Kwanghee Jung,
Texas Tech University, United States

Reviewed by:
Navid Feroze,
University of Azad Jammu and Kashmir, Pakistan
Kyungtae Kim,
Tennessee Department of Education, United States

****Correspondence:***
Christoph König
koenig@psych.uni-frankfurt.de

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *05 October 2020*
Accepted: *13 April 2021*
Published: *11 May 2021*

Citation:
König C (2021) A Similarity-Weighted Informative Prior Distribution for Bayesian Multiple Regression Models. Front. Psychol. 12:614236. doi: 10.3389/fpsyg.2021.614236

INTRODUCTION

Informative prior distributions are a crucial element of Bayesian statistics, and play a pivotal role for scientific disciplines that aim at constructing a cumulative knowledge base. Informative prior distributions are background knowledge quantified and introduced in a Bayesian analysis. Their use allows studies to build upon each other, hence to update the knowledge base of a scientific discipline continuously. This is also a central tenet of the new statistics (Cumming, 2014). Despite the increase of Bayesian statistics in various scientific disciplines over the last years, the use of informative prior distributions is still relatively rare (for instance in Psychology, see van de Schoot et al., 2017; for Educational Science see König and van de Schoot, 2018). Thus, the potential of Bayesian statistics for cumulative science is not fully realized yet.

Goldstein (2006) states that the tentative use of informative prior distributions is due to their frequently criticized subjective nature. Vanpaemel (2011) adds the lack of methods to formalize background knowledge as another reason. From an applied viewpoint, this is more severe: if the background knowledge is inaccurate, which is the case if the prior mean does not equal the population mean, parameter estimates may be biased (McNeish, 2016; Finch and Miller, 2019). Specifying accurate informative prior distributions is a question of carefully selecting studies that comprise the body of comparable background knowledge. Psychological research, however, consists of studies that are being conducted under different circumstances, with different samples and varying instruments. Thus, results of previous studies include different sources of heterogeneity, and not all available results can and should contribute equally to an informative prior distribution (Zhang et al., 2017). This implies a necessary weighting of background information based on the similarity of the previous studies to the focal study at hand. Current approaches to account for heterogeneity by weighting informative prior distributions are either not easily accessible or incomplete. For example, the power prior weighs the likelihood of the data and requires complicated intermediate steps in order to use the quantified heterogeneity properly (Ibrahim et al., 2015; Carvalho and Ibrahim, 2020). The meta-analytic predictive prior (Neuenschwander et al., 2010) is more intuitive by weighting the informative prior distribution directly, but uses heterogeneity in outcomes only. To complicate matters further, to date there are no methods available for investigating and quantifying the similarity of a given body of background knowledge to the focal study at hand. Specifying accurate informative prior distributions, however, requires an approach that quantifies all sources of heterogeneity in a body of background knowledge into a measure of similarity, and using this measure to weight the associated informative prior distribution in a direct and intuitive way.

Consequently, the purpose of this study is threefold. We first present a novel method to combine the aforementioned sources of heterogeneity in the similarity measure ω. This method is based on a combination of a propensity-score approach to assess the similarity of samples with random- and mixed-effects meta-analytic models to quantify the heterogeneity in outcomes and study characteristics (e.g., Tipton, 2014; Cheung, 2015). Second, we show how to use the novel similarity measure ω as a weight for informative prior distributions for the substantial parameters (regression coefficients) in Bayesian multiple regression models. Third, we investigate the performance and the behavior of the similarity–weighted informative prior distribution in a comprehensive simulation study, where it is compared to the normalized power prior (Carvalho and Ibrahim, 2020) and the meta-analytic predictive prior (Weber et al., 2019). The similarity measure ω and the similarity-weighted informative prior distribution as the primary results of this study provide applied researchers with means to specify accurate informative prior distributions.

The structure of this paper is as follows. First, the conceptual background of similarity is illustrated. Next, it is shown how these sources of heterogeneity can be quantified and combined in the similarity measure ω. Based on this, the similarity-weighted informative prior distribution is described. The design and results of the simulation investigating the performance and behavior of this distribution is presented next, followed by a discussion of how the similarity measure ω and the similarity-weighted informative prior distribution contribute to building confidence in and to systemizing the use of informative prior distributions in Psychological research. Please note that, in order to keep the manuscript as accessible as possible, mathematical details are kept at a minimum.

CONCEPTUAL BACKGROUND

The Concept of Similarity

When specifying informative prior distributions, researchers are confronted with a body of background knowledge comprised of conceptual replications of studies (Schmidt, 2009). Conceptual replications focus on the general theoretical process, without copying the methods of previously conducted studies (Makel et al., 2012). Thus, the studies differ in samples, variables, and other characteristics. Without assessing their similarity to the focal study at hand, using studies for informative prior distributions might imply an unwarranted generalization; excluding studies might be too restrictive and imply that no background knowledge is available, when in truth there is. Hence, an adequate similarity measure should take into account all relevant sources of heterogeneity in research results. Consequently, the conceptual framework of the similarity measure ω follows Shadish et al. (2002), who build upon Cronbach (1982), and distinguishes between units and treatments (*UT*), outcomes (*O*), and settings (*S*) of the studies as sources for heterogeneity. More specifically, we conceptualize *UT* as samples and predictor variables, *O* as outcome variables or effect sizes, and *S* as study characteristics commonly investigated as moderators in mixed-effects meta-analytic models. Thus, we define similarity as the variability in research results due to the three sources of heterogeneity. This differentiation takes into account that heterogeneity in outcomes is not sufficient for an adequate assessment of similarity (Lin et al., 2017). The quantification of the three sources of heterogeneity is addressed next.

Quantifying Sources of Heterogeneity

For a similarity measure to work adequately, it is pivotal that the different sources of heterogeneity can be quantified accurately with state-of-the-art methods. More specifically, the similarity measure ω is based on three components:

(a) the modified generalizability index $\overline{B}$ that is based on Tipton (2014), (b) the between-study heterogeneity τ^2 resulting from (Bayesian) random-effects meta-analytic models, and (c) δ_{τ^2}, the difference between the residual variance τ^2_{res} of (Bayesian) mixed-effects meta-analytic models and τ^2 (for an overview see, for instance, Jak, 2015). Each individual measure quantifies important aspects of the comparability of research results.

Quantifying Similarity in Predictors and Samples With $\overline{B}$

The first component of the similarity measure ω is the modified generalizability index $\overline{B}$. In its original form, the generalizability index B is a propensity score-based measure of distributional similarity between a sample and a population (Tipton and Olsen, 2018). We modified it so that it describes the similarity between the samples of the focal study and a previously conducted study that is part of the body of available background knowledge. The generalizability index and its modified version takes values between zero and one, which indicate no and perfect similarity of the two samples, respectively. It is based on $s(\mathbf{X})$, a theoretical sampling propensity score defined as $s(\mathbf{X}) = \Pr(Z = 1|\mathbf{X})$, and describes the probability Z of an individual being in the sample of the focal study (vs. being in the sample of the previously conducted study) based on a set of covariates $\mathbf{X}$ (Tipton, 2014). The sampling propensity score can be estimated by a logistic regression model $\log[\mathrm{s(X)/1\text{-}s(X)}] = \alpha_0 + \alpha_m + \mathrm{X}_m$, where $m = 1, , m$ is the number of covariates. Adapting Tipton (2014), for a set of covariates $\mathbf{X}$ and sampling propensity score $s(\mathbf{X})$, the modified generalizability index is then defined as $\beta = \int \sqrt{f_f(s) f_p(s)} ds$, where $f_f(s)$ and $f_p(s)$ are the distributions of sampling propensity scores in the sample of the focal and previously conducted study, respectively. An estimator of β is provided by a discrete version of the generalizability index $B = \sum_h \sqrt{w_{fh} w_{ph}}$, where h is the number of bins and w_{fh} and w_{ph} are the proportions of the focal and previously conducted study samples, respectively (Tipton, 2014). In case of multiple previously conducted studies, the modified version of the generalizability index B is calculated for each comparison of the samples of the focal and previously conducted studies. It is the average of the individual indices $\overline{B} = \frac{1}{k} \sum_k B_k$, with k being the number of previously conducted studies. We implemented this procedure as a kernel density estimation with a Gaussian kernel and a non-parametric bandwidth selector (Moss and Tveten, 2019), so that the number of bins does not have to be chosen a priori.

Quantifying Heterogeneity in Outcomes With τ^2

The second component of the similarity measure ω is the between-study heterogeneity τ^2, which is a measure for the variance in effect sizes, such as standardized mean differences, log-odds ratios, and more recently, partial and semi-partial correlations as effect sizes for regression coefficients (Aloe and Thompson, 2013). It is the variance component of random-effects meta-analytic models, which assume that the population effect sizes are not equal across the studies. Several studies show that this assumption is usually correct: the typical between-study heterogeneity in outcomes ranges from 0.13 to 0.24 (van Erp et al., 2017; Stanley et al., 2018; Kenny and Judd, 2019). Random-effects meta-analytic models allow individual studies to have their own effect (e.g., Cheung, 2015). Let y_k be the effect found in study k. The study-specific model is then $y_k = \overline{\beta} + u_k + \varepsilon_k$ where $\overline{\beta}$ is the average effect size, u_k are deviations from the average effect size, ε_k is the study-specific error term and $Var(\varepsilon_k)$ is the known sampling variance. The variance of these deviations $Var(u_k)$ is the between-study heterogeneity τ^2 indicating the variability of the effect sizes across the studies included in the meta-analysis. The between-study heterogeneity is strictly positive $\tau^2 > 0$. When τ^2 increases, consensus in the average effect decreases. This lack of consensus in the average effect, the uncertainty quantified by τ^2, should be represented in a weight of an informative prior distribution. However, only the meta-analytic predictive prior distribution uses τ^2 as weight. Both the average effect and the between-study heterogeneity τ^2 can be estimated by Maximum Likelihood, Restricted Maximum Likelihood and Bayesian estimation methods (for overviews, see Veroniki et al., 2016; Williams et al., 2018). For situations with a small number of studies, and the known problems of ML and REML estimators regarding τ^2 in these cases, we implemented a hierarchical Bayesian random-effects meta-analytic model to estimate $\tau^{\,2}$ accurately.

Quantifying Heterogeneity in Study Characteristics with δ_{τ^2}

The third component of the similarity measure ω is δ_{τ^2}, the difference between the residual variance τ^2_{res} in the effect sizes, estimated by a (Bayesian) mixed-effects meta-analytic model, and their estimated between-study heterogeneity τ^2. Mixed-effects meta-analytic models extend random-effects meta-analytic models by introducing study characteristics as potential moderators of the effects. The study-specific model is then $y_k = \beta x_k + u_k + \varepsilon_k$, where $\mathbf{x}_k$ is a vector of predictors including a constant of one (Cheung, 2015). Under the mixed-effects meta-analytic model, the variance of the deviations $Var(u_k)$ is the residual variance τ^2_{res} in the effect sizes after controlling for study characteristics as moderators. If $\tau^2_{res} < \tau^2$, the study characteristics explain variance in the effect sizes. This implies that the effect sizes not only vary across studies, but also across specific study characteristics. For example, it is possible that effects found in the 1980s differ systematically from effects found in the 2010s. Thus, there is additional uncertainty in the average effect that is quantified by δ_{τ^2}. If $\tau^2_{res} \geq \tau^2$, the study characteristics do not explain any variance in the effect sizes, and δ_{τ^2} is truncated to zero. Hence, $\delta_{\tau^2} > 0$ if $\tau^2_{res} < \tau^2$, and 0 otherwise. Similar to the random-effects meta-analytic models, for situations with a small number of studies we implemented a hierarchical Bayesian mixed-effects meta-analytic model to estimate τ^2_{res} and, subsequently, calculate δ_{τ^2} accurately.

The Similarity Measure ω

The similarity measure ω integrates the three components into a single index. It is conceptually similar to the variance component of a Bayesian hierarchical model (comparable to the a_0-parameter of the power prior; Ibrahim et al., 2015; Neuenschwander et al., 2009). Thus, its use as weight for informative prior distributions places certain demands on the measure, both mathematically and conceptually. First, similar to the a_0-parameter of the power prior (Ibrahim et al., 2015), the similarity measure ω needs to take values between zero and one, $\omega \in [0, 1]$. This avoids any potential overweighting of the quantified background knowledge, compared to the information contained in the data of the focal study. Moreover, the similarity measure $\omega \rightarrow 1$ as the comparability of the previously conducted studies in the body of background knowledge and the focal study increases. On the one hand, when $\omega = 0$ the previously conducted studies and the focal study are not comparable, and no information contained in the informative prior distribution is used. On the other hand, when $\omega = 1$, the focal study is a direct replication of the previously conducted studies in the body of background knowledge, and the information contained in the prior distribution is used fully. Second, the similarity measure ω needs to adequately reflect the inverse relation between B, and τ^2 and δ_{τ^2}. While an increasing B indicates an increased comparability, increasing τ^2 and δ_{τ^2} indicate a decreasing comparability. Thus, the similarity measure needs to align the conceptual meaning of the three indices to reflect the comparability of the focal study with the study in the body of background knowledge adequately. Third, the similarity measure ω needs to be flexible in specification and discriminate strongly across the range of plausible values especially for τ^2 and δ_{τ^2}, which we know to typically range between 0.13 and 0.24 (van Erp et al., 2017; Stanley et al., 2018; Kenny and Judd, 2019). This aims at conservative estimates of ω, again to avoid the informative prior distribution overwhelming the likelihood of the data of the focal study. Considering all these requirements, the similarity measure ω can be expressed formally as,

$$\omega = \left(\frac{1}{1 + \exp\left[10 * \left(\sqrt{\tau^2 + \delta_{\tau^2}} - 0.24\right)\right]} \right) * \overline{B} \qquad (1)$$

Thus, the similarity measure ω essentially is a logistic function of τ^2 and δ_{τ^2} with maximum value $L = 1$, midpoint $\omega_0 = 0.24$ and slope $s = 10$, weighted by $\overline{B} = \frac{1}{K}\sum_k B_k$, where $k = 1 \ldots K$ is the number of previously conducted studies. The parameters of this weighted logistic function are chosen so that the resulting values of the similarity measure ω adequately reflects the characteristics of Psychological research: the midpoint is carefully chosen following van Erp et al. (2017), and the slope is chosen to discriminate adequately across the typical range of between-study heterogeneity (Stanley et al., 2018; Kenny and Judd, 2019). We assume an additive relationship between τ^2 and δ_{τ^2}. Taken together, the behavior of the similarity measure is as required: $\omega \rightarrow 1$ as τ^2 and δ_{τ^2} decrease and $\overline{B}$ increases. Applying equation (1) to a situation of a Bayesian multiple regression model with three predictors and ten previously conducted studies yields three parameter-specific similarity measures, which can be used to weigh an informative prior distribution.

Applying ω – The Similarity-Weighted Informative Prior Distribution

The similarity measure ω can now be used to weight an informative prior distribution and integrate it, without any necessary intermediary calculations, in a usual Bayesian analysis. Contrary to the power prior of Ibrahim et al. (2015), who weight the likelihood of the previously conducted studies, in this case it involves raising the informative prior distribution to the power ω, $p(\theta \mid D) \propto p(D \mid \theta)\ \pi(\theta)^{\omega}$ where $p(\theta \mid D)$ is the posterior distribution of a parameter θ, $p(D \mid \theta)$ is the likelihood of the data, and $\pi(\beta\theta)^{\omega}$ is the similarity-weighted informative prior distribution. Because this prior distribution utilizes data from previously conducted studies, it belongs to the class of evidence-based informative prior distributions (Kaplan, 2014). We illustrate the use of the similarity measure ω as weight for an informative prior distribution with an example of a simple Bayesian multiple regression with three predictors. Let $\mathbf{y}$ be a n × 1-vector of outcomes, and $\mathbf{X}$ a n × p predictor matrix, where n is the sample size of the focal study and $p = 3$ the number of predictors. Then,

$$y \sim N(\beta_0 + \mathbf{X\beta},\ \sigma^2) \qquad (2)$$

is the likelihood of the Bayesian multiple regression model, with β_0 being the intercept, $\boldsymbol{\beta}$ a p × 1-vector of regression coefficients, and σ^2 being the error variance. The prior specification is as follows:

$$\beta_0 \sim N(0,\ 10) \qquad (3)$$

$$\boldsymbol{\beta} \sim N(\mu_p, SE_p^2)^{\omega_p} \qquad (4)$$

$$\sigma^2 \sim half - Cauchy(0,\ 2.5) \qquad (5)$$

Both β_0 and σ^2 receive weakly informative prior distributions, and the hyperparameters of the informative prior distributions (means and standard deviations) for the regression coefficients β_p are the average effects μ_p and their standard errors SE_p^2 estimated by multiple univariate or a single multivariate random-effects meta-analysis (Cheung, 2015; Smid et al., 2020). They are weighted by the parameter-specific similarity measures ω_p. Generally speaking, as $\omega \rightarrow 0$ the peak around the mean of the informative prior distribution flattens, and the distribution becomes broader. A broader prior distribution carries less information about the parameter of interest; hence, the broader the distribution the lesser its informativeness.

SIMULATION

We conducted a comprehensive simulation to assess the behavior of the similarity measure ω and to investigate the performance of the similarity-weighted informative prior distribution. R-code, functions, and data of the simulation are available at https://doi.org/10.17605/OSF.IO/8AEF4.

Design

The design consisted of the following, systematically varied factors. First, the number of previously conducted studies that are part of the available body of background knowledge ($K = 3, 5, 10$). Second, the sample sizes of the previously conducted studies, indicated by the difference between the average sample sizes of these studies and the sample size of the focal study (smaller and larger $\triangle_N = -100, 100$). Third, the similarity of the predictors, indicated by the differences in means of the respective distributions (i.e., their overlap) between the previously conducted studies and the focal study (from large overlap to no overlap $\triangle_\mu = 0.25, 0.5, 1, 2, 3$). Fourth, the between-study heterogeneity in the effect sizes, thus the (lack of) consensus in the background knowledge (small to large $\tau^2 = 0.025, 0.05, 0.10, 0.15, 0.20, 0.35, 0.5$). Moreover, we simulated one moderator variable that explained 10% of the between-study heterogeneity in the effect sizes. Thus, the simulated amount of variance in outcomes and study characteristics is $\tau^2 + \delta_{\tau^2} = 0.0275, 0.055, 0.110, 0.165, 0.275, 0.385, 0.550$. In total, the design of the simulation consisted of 210 conditions.

Data Generation and Analysis

We applied the following procedure to generate the datasets in each condition. First, we simulated the dataset of the focal study, according to the multiple regression model in equation (2), with fixed sample size $N_F = 200$, true regression coefficients $\beta_F = (0.5, 0.25, -0.5)$ and a normally distributed error $\sigma^2_F \sim N(0, 1)$. Predictors in $\mathbf{X}_F$ were drawn from standard normal distributions. Next, we constructed the database of previously conducted studies, also according to the multiple regression model in equation (2) with normally distributed error $\sigma^2_D \;\; N(0, 1)$. As a first step, the sample size for the k-th ($k = 1 \ldots K$) study of the database was drawn from a normal distribution $N(N_{P_i},\ 25)$, where $N_{P_i} = N_F + \triangle_N$. In the second step, for the k-th study of the database a vector of regression coefficients β_k was drawn from a multivariate normal distribution with mean vector $\mu_{\beta_k} = (0.4,\ 0.0,\ 0.3)$, i.e., their meta-analytic means, and variance τ^2. Compared to β_F, the mean coefficients in μ_{β_k} represent certainty, disagreement, and contradiction in the size of the effect. Predictors in $\mathbf{X}_k$ were drawn from normal distributions $N(\mu_{N_P}, 1)$, where $\mu_{N_P} = \triangle_{\mu_P}$. This procedure was repeated one hundred times in each condition, resulting in 21,000 datasets (i.e., the simulated dataset of the focal study and the databases of the previously conducted studies).

Each dataset was analyzed with a Bayesian multiple regression model with (a) non-informative priors for the regression coefficients (pooled analysis), (b) the normalized power prior (NPP), (c) the meta-analytic predictive prior (MAP), and (d) the similarity-weighted informative prior distribution (SWIP). For the non-informative model, the datasets of the focal and previously conducted studies were pooled into a single dataset. The NPP was implemented as a standard normal-inverse gamma model as described in Carvalho and Ibrahim (2020). For both the MAP and SWIP a Bayesian random-effects meta-analysis was run with the generated database of previously conducted studies to calculate the meta-analytic mean effect, its standard error, and the between-study heterogeneity τ^2. The meta-analytic mean effect and its standard error were used as hyperparameters of the MAP and SWIP. The meta-analysis was based on Fisher's r-to-z transformed partial correlation coefficients using the metafor-package (Viechtbauer, 2010). This follows Aloe and Thompson (2013) who introduced partial or semi-partial correlations as adequate effect sizes for regression coefficients. The specification of the MAP model and its robustification procedure followed the standard implementation of the RBesT-package outlined in Weber et al. (2019). Prior to the SWIP analysis, the modified generalizability index $\overline{B}$ for the previously conducted studies and the similarity measure ω was calculated as in equation (1). The similarity measure ω was then introduced as parameter-specific weight for the informative prior distributions for the regression coefficients as in equation (4). All models were specified with Stan and its R interface *RStan* (Stan Development Team, 2020). Four chains each of length 2,000 with 1,000 burn-in cycles were set up. Different random starting values were supplied to each chain. Convergence was assessed using the Gelman-Rubin *R*-statistic (Gelman and Rubin, 1992), where $R < 1.02$ indicated convergence. All solutions converged.

Evaluation Criteria

To assess the behavior of the similarity measure ω we focused on its relation to $\tau^2 + \delta_{\tau^2}$ and $\triangle_\mu$, and its relation to the shrinkage in the parameter estimates. Therefore, we estimated linear models. Shrinkage was defined as the difference between the focal-study estimates (the true values β_F) and the estimates obtained by the similarity-weighted informative prior distribution. Moreover, comparing the performance of the different prior distributions involved, for each condition, averaging the parameter estimates and their standard errors over replications, $\overline{\beta} = \frac{1}{R}\sum_R \beta$ and $\overline{SE_\beta} = \sqrt{\frac{1}{R}\sum_R SE^2_\beta}$, respectively. The similarity measure behaves as expected if it decreases as $\tau^2 + \delta_{\tau^2}$ and $\triangle_\mu$ increase. Moreover, shrinkage should increase as the similarity increases. Good performance of the different informative prior distributions is indicated by increasing shrinkage of the parameter estimates toward their meta-analytic means, as well as decreasing standard errors of the parameter estimates, depending on the degree of similarity.

RESULTS

Behavior of the Similarity Measure ω

Figure 1 illustrates the behavior of the similarity measure ω conditional on $\tau^2 + \delta_{\tau^2}$ for different levels of $\triangle_\mu$ combined for all three regression coefficients (left panel), and the behavior of the shrinkage of the estimates of the three regression coefficients, conditional on the similarity measure ω (right panel), across all simulation conditions. The similarity measure ω behaves as expected; as both $\tau^2 + \delta_{\tau^2}$ and $\triangle_\mu$ increase, i.e., the similarity between the focal and the previously conducted studies decreases, the similarity measure ω decreases as well. Moreover, we have a non-compensatory relation between the

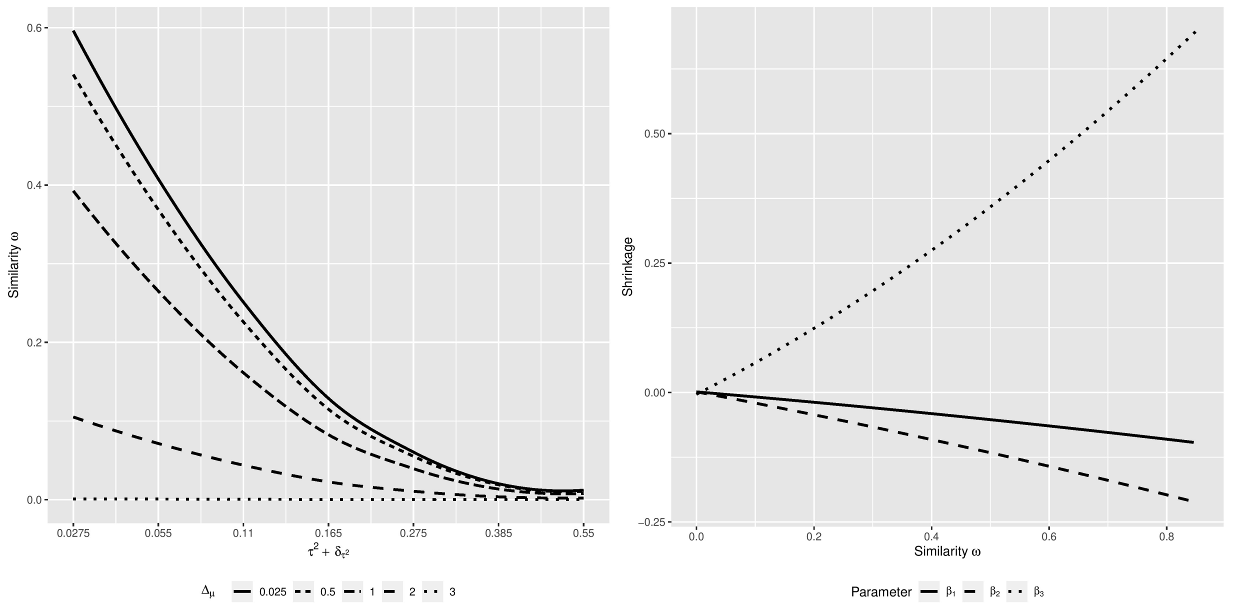


FIGURE 1 | Regression curves of the relation between the similarity measure ω and $\tau^2 + \delta_{\tau^2}$ for different levels of $\triangle_\mu$ (**left panel**), and the relation between the shrinkage and the similarity measure ω, for each β-parameter (**right panel**), based on estimates from 21,000 simulated datasets.

components of the similarity measure. High similarity in samples and predictors does not compensate for a lack of similarity regarding outcomes and study characteristics, and vice versa. The shrinkage of the parameter estimates behaves accordingly: as the focal and the previously conducted studies become more similar, indicated by an increasing similarity measure ω, the estimates of the regression coefficients shrink toward their meta-analytic means. If the focal and previously conducted studies are highly dissimilar, shrinkage is close to zero, and the estimates of the regression coefficients remain at estimates resulting from the focal study. Lastly, shrinkage is stronger when the meta-analytic means and the focal-study estimates of the regression coefficients are considerably apart (see β_3, compared to the other two parameters). This is, however, just an effect of the distance between the values of $\beta_3 = -0.5$ and its meta-analytic mean $\mu_{\beta_3} = 0.3$. With an increasing distance between a parameter estimate and it meta-analytic mean, the potential amount of shrinkage increases as well. Moreover, the different direction of the shrinkage in case of β_3 is due to the meta-analytic mean being larger than the focal-study estimate. In case of the other regression coefficients, their meta-analytic means are smaller than their focal-study estimates, thus the shrinkage is negative.

Performance of the Similarity-Weighted Informative Prior Distribution

Figures 2, **3** illustrate the behavior of the estimates of the three regression coefficients and their standard errors, respectively, obtained from the pooled Bayesian analysis, the NPP, the MAP, and the SWIP, conditional on the simulated factors. The estimated regression coefficients obtained with the SWIP lie consistently between their true values β_F and their true meta-analytic means μ_{β_k}. Shrinkage toward the true meta-analytic means is sensitive to changes in both $\tau^2 + \delta_{\tau^2}$ and $\triangle_\mu$. In contrast, the MAP consistently yields parameter estimates close to the true values β_F, except for β_3 when $\tau^2 + \delta_{\tau^2} < .10$. Thus, the MAP is largely insensitive to changes in both $\tau^2 + \delta_{\tau^2}$ and $\triangle_\mu$. Compared to the NPP, shrinkage of the parameter estimates of the SWIP is comparably sensitive to changes in both $\tau^2 + \delta_{\tau^2}$ and $\triangle_\mu$, but more conservative. For example, when $\triangle_\mu$ is large, the NPP sometimes yields overestimated parameters. Moreover, while the SWIP shrinks the parameters never beyond their estimates obtained with the pooled analysis, the NPP shrinks the parameter estimates in some cases beyond their meta-analytic means.

This general pattern is similar in case of the standard error of the parameter estimates. In case of the SWIP, the standard errors decrease as the similarity of the focal and previously conducted studies increases. More specifically, they converge to the standard errors of the pooled Bayesian analysis. This implies a similarity-dependent borrowing of information from the previously conducted studies that increases the precision of the parameter estimates of the focal study. This is true for all simulation conditions, although it is most distinct when the number of available studies is large ($K = 10$). In contrast, the standard errors of the estimates of the MAP do not converge; they largely remain at around 0.7. Thus, the MAP does not borrow information from the previously conducted studies. The standard errors of the estimates of the NPP tend to be smaller than the standard errors of the SWIP, especially when the number of previously conducted studies is large ($K = 10$). Thus, the NPP borrows more information. When the focal-study estimates and their meta-analytic means

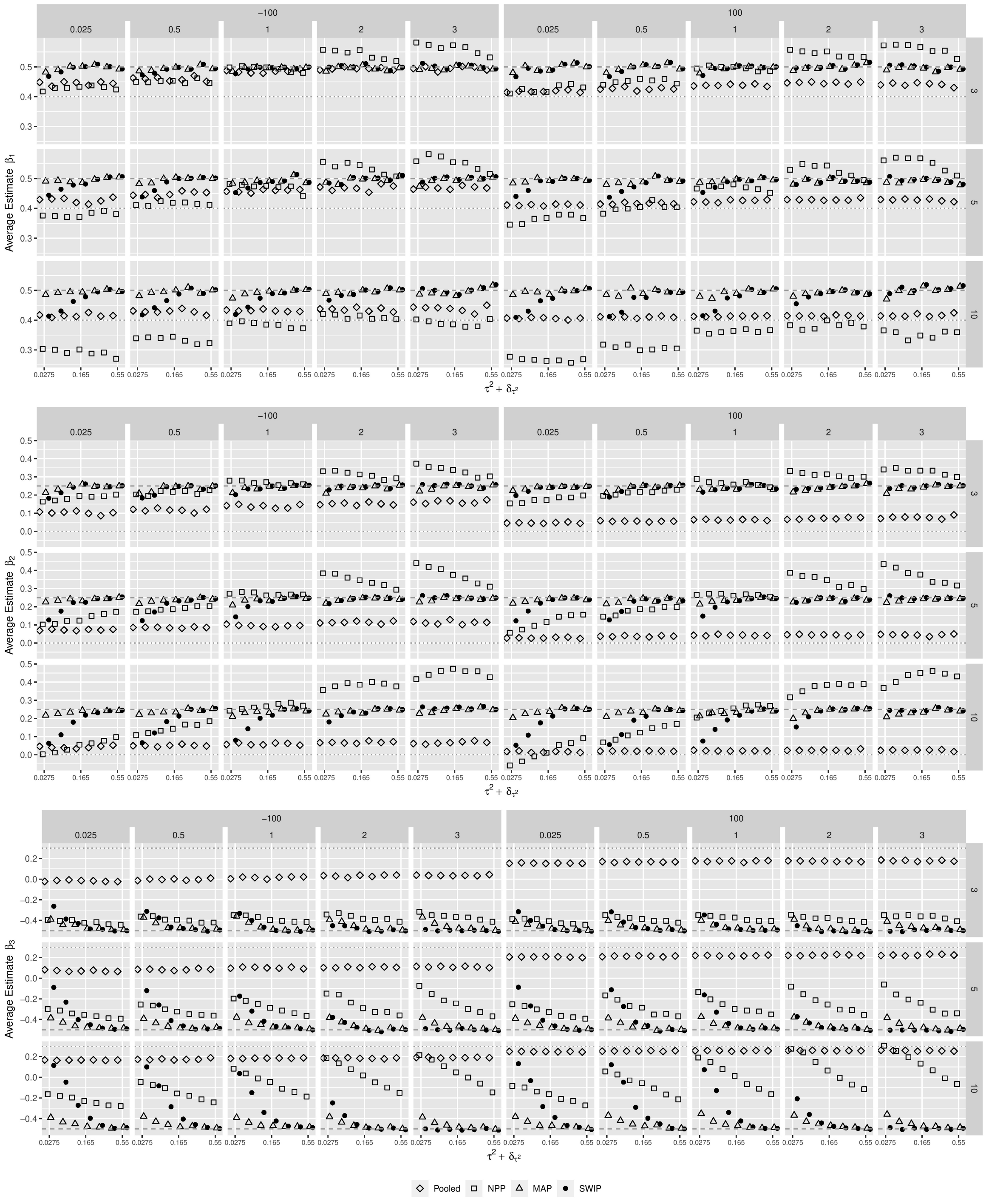


FIGURE 2 | The behavior of the parameter estimates across simulation conditions. The similarity of the focal and the previously conducted studies decreases from **left** to **right**. Pooled = pooled Bayesian analysis; NPP = normalized power prior; MAP = meta-analytic predictive prior; SWIP = similarity-weighted informative prior distribution. The dashed horizontal line represents the true value of the respective regression coefficient of the focal study. The dotted horizontal line represents the true (generating) meta-analytic mean of the respective regression coefficient.

contradict (in case of β_3), however, the standard errors of the estimates of the NPP tend to be larger, especially when the number of previously conducted studies is small and $\triangle_\mu$ is large.

Overall, the performance of the SWIP is more consistent and sensitive to changes in similarity between the focal and previously conducted studies, compared to both the NPP and MAP, while yielding conservative estimates. As the similarity

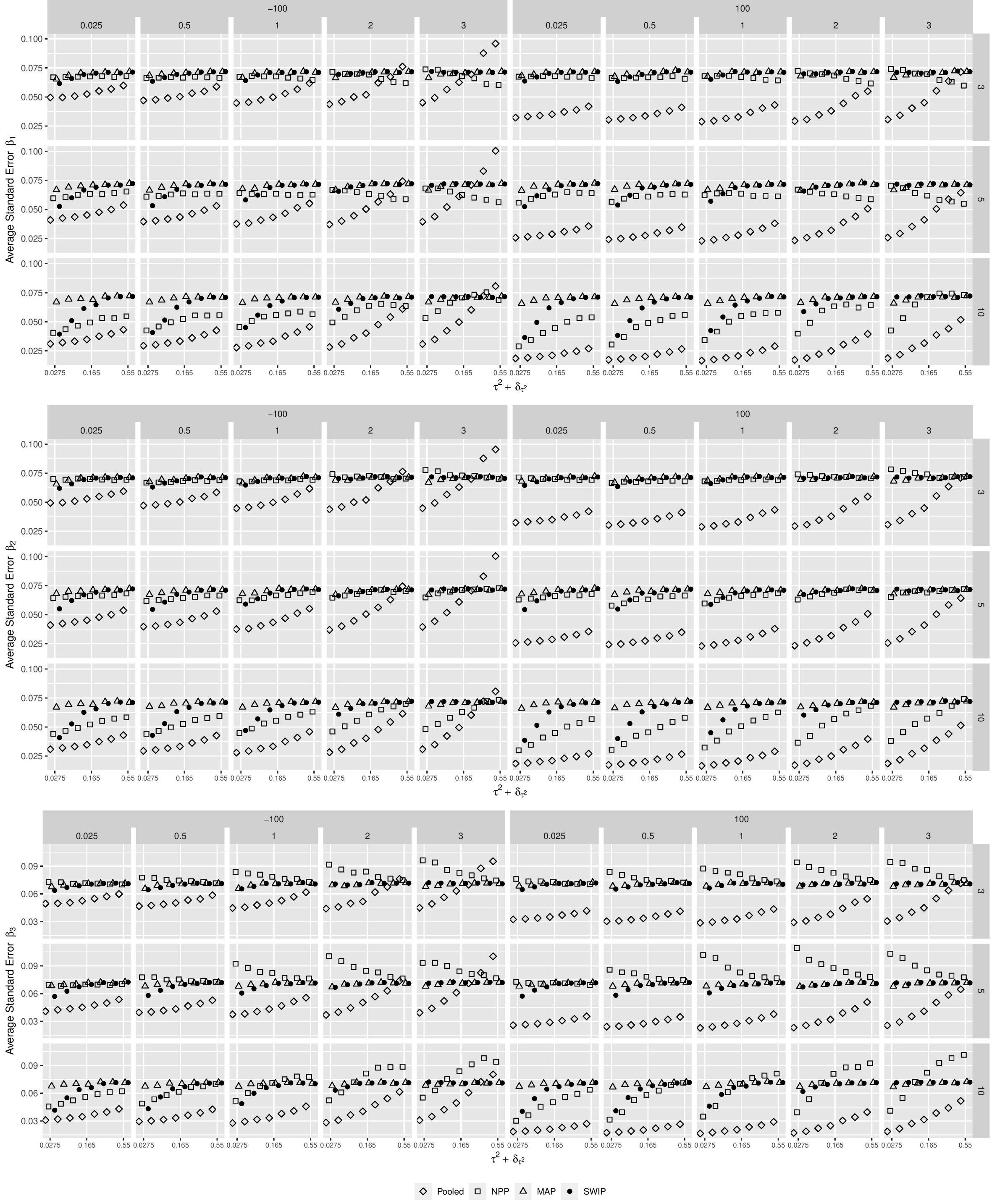


FIGURE 3 | The behavior of the standard errors of the parameter estimates across simulation conditions. The similarity of the focal and the previously conducted studies decreases from **left** to **right**. Pooled = pooled Bayesian analysis; NPP = normalized power prior; MAP = meta-analytic predictive prior; SWIP = similarity-weighted informative prior distribution.

increases, the parameter estimates of the SWIP shrink toward the estimates of the pooled Bayesian analysis, and more information is borrowed from the body of available background knowledge. Thus, the standard errors of the parameter estimates decrease, and the estimates are more precise. In this context, the number of previously conducted studies plays a vital role. When the number is small, i.e., when there is less information to borrow, both shrinkage and precision are less distinct.

DISCUSSION

The purpose of this study was to illustrate a novel method to assess the similarity of studies in the context of specifying informative prior distributions for Bayesian multiple regression models. We illustrated the quantification, based on a propensity-score approach and random- and mixed-effects meta-analytic models (e.g., Tipton, 2014; Cheung, 2015), and combination of heterogeneity in samples and predictors, outcomes, and study characteristics in the novel similarity measure ω. We showed how to use the similarity measure ω as a weight for informative prior distributions for the regression coefficients, and investigated the behavior of the similarity measure ω and the similarity–weighted informative prior distribution, comparing its performance to the normalized power prior and meta-analytic predictive prior.

The Performance of the Similarity-Weighted Informative Prior Distribution

The results of our simulation show that the parameter estimates of the similarity-weighted informative prior distribution behave similar to those of hierarchical Bayesian models: as the similarity of the focal and previously conducted studies increases, they shrink toward their pooled, meta-analytic means. Simultaneously, the precision of the parameter estimates increases because more information is borrowed from the previously conducted studies. From the perspective of cumulative knowledge creation, this behavior is desired. As evidence from comparable studies accumulates, our knowledge of the size of an effect becomes incrementally more certain until, over time, it represents the best knowledge we have (unless the evidence contradicts; Kruschke et al., 2012; König and van de Schoot, 2018). The meta-analytic predictive prior, on the one hand, does not provide this increasing certainty in the size of an effect. Compared to the similarity–weighted informative prior distribution, the similarity-dependent shrinkage is much less distinctive. Since the meta-analytic predictive prior only considers the heterogeneity in outcomes, it may be an indication that, echoing Lin et al. (2017), this is not sufficient for an adequate assessment of similarity of the focal and previously conducted studies. Parameter estimates of the normalized power prior, on the other hand, exhibit a stronger, but inconsistent shrinkage toward the pooled, meta-analytic means. From the perspective of cumulative knowledge creation, this is problematic, because the normalized power prior provides parameter estimates that are biased, and the precision of the estimates does not increase consistently as evidence accumulates.

Since the performance of the similarity-weighted informative prior distribution stands or falls with the accuracy of the components of the similarity measure ω, it is essential to estimate the random and mixed-effects meta-analytic models as unbiased as possible. This is usually based on either maximum likelihood (ML) or restricted maximum likelihood (REML) estimation (e.g., Cheung, 2015). These likelihood-based methods, however, exhibit poor performance especially when the number of previously conducted studies is small (Bender et al., 2018), additionally to the general underestimation of the between-study heterogeneity of ML-based random-effects meta-analytic models (Cheung, 2015). Several studies show a superior performance of Bayesian approaches, especially hierarchically specified random and mixed-effects meta-analytic models, in terms of the accuracy of the (residual) variance components (Williams et al., 2018; Seide et al., 2019). Thus, when using the similarity measure ω to specify the similarity-weighted informative prior distributions, we recommend using these Bayesian approaches to estimate both the mean effect size and its variance components, as illustrated in this study.

On the one hand, the similarity-weighted informative prior distribution simplifies the concept of the normalized power prior. The similarity measure is used to weight the informative prior distribution directly, which is more intuitive and less challenging than weighting the likelihood of the data from the previously conducted studies (Ibrahim et al., 2015). The complex calculation of multiple marginal likelihoods by means of bridge sampling approaches (see Carvalho and Ibrahim, 2020) is not necessary. Calculating marginal likelihoods can be complicated and time-consuming especially when the underlying models are complex (for instance, structural equation models), and their likelihood is analytically intractable (Ibrahim et al., 2015). On the other hand, the similarity-weighted informative prior distribution extends both the normalized power prior and meta-analytic predictive prior by taking into account multiple sources of heterogeneity in previously conducted studies, and quantifying these sources in the similarity measure ω. The benefits of this holistic approach are illustrated by the performance of the similarity-weighted informative prior distribution.

Future Directions

The similarity measure ω and the similarity-weighted informative prior distribution offer various opportunities for further research. First, the inconsistent behavior of the normalized power prior may be due to the limited number of available small-sample studies (Neuenschwander et al., 2009). Thus, a limitation of this study is that we only considered sample sizes of the focal and previously conducted studies that are of a comparable order of magnitude. Investigating the performance of the similarity-weighted informative prior distribution in situations where these sample sizes differ by orders of magnitude, and where the sample sizes of the previously conducted studies vary considerably, is an important topic for further research. If the sample sizes of the focal and previously conducted studies vary considerably in size (especially when $N_P \gg N_F$), it is possible to multiply the scale parameter of the informative prior distribution SE_p^2 by the ratio N_P/N_F. This can be understood as a mechanism to avoid that the prior information overwhelms the likelihood, because it flattens the distribution and makes it less informative. Second, the similarity measure can be used as the a_0-parameter of the normalized power prior. Investigating the behavior of the normalized power prior in the context of a fixed–a_0 approach, where the study-specific a_0-parameters are fixed to the values of the study-specific similarity measures may be an interesting topic for future research. Especially because the fixed–a_0 approach is considered superior to the random–a_0 approach, where the comparability of the focal and previously conducted studies is inferred from the data, and the prior distribution for the a_0-parameter has to be chosen carefully (Neuenschwander et al., 2009; Ibrahim et al., 2015). Third, comparing ML-based and Bayesian meta-analytic or other approaches in the context of assessing the similarity of studies, i.e., regarding their impact on the behavior of the similarity-weighted informative prior distribution, is another important topic for future studies. As mentioned above, the precision of the average effect sizes that are used as the hyperparameters of the informative prior distributions, are pivotal for the accuracy of these distributions. Identifying the correct approach, especially when the number of previously conducted studies is small (Bender et al., 2018), is crucial for the performance of the similarity-weighted informative prior distribution. Fourth, the calculation of the modified generalizability index $\overline{B}$ still requires the availability of the raw data of the previously conducted studies. This remains a limitation for the applicability of the similarity measure. Extending its applicability is a question of being able to calculate the modified generalizability index $\overline{B}$ in situations when only summary data are available. It is possible, however, to simulate a number of datasets based on correlation matrices, or means and standard deviations, and to calculate $\overline{B}$ for each of the simulated datasets. The pooled $\overline{B}$ can then be used to calculate the similarity measure. Such an approach, similar to multiple imputation or the estimation of plausible values, will be addressed and investigated in a future study. Fifth, both the similarity measure and the similarity-weighted informative prior distribution are currently only available for multiple regression models, i.e., univariate methods. It may be fruitful to extend and adapt both to multivariate methods, for example structural equation models.

Concluding Remarks

As mentioned in the introduction to this study, specifying accurate informative prior distributions is a question of carefully selecting studies that comprise the body of comparable background knowledge. Given the considerable heterogeneity of studies that are being conducted in Psychological research (different circumstances, with different samples and instruments), the results of these studies are heterogeneous, and not all available results can and should contribute equally to an informative prior distribution. The similarity measure ω and the similarity-weighted informative prior distribution developed in this study provide researchers with tools to (a) justify the selection of studies that contribute to the informative prior distribution, and (b) to accomplish the necessary similarity-based weighting of the available background knowledge. On the one hand, the quantification of the similarity of studies, and the similarity-based weighting of prior information, are important elements of a systematization of the specification and use of informative prior distribution. Being able to justify empirically the use of previously conducted studies for the specification of informative prior distributions, on the other hand, helps building confidence in the use of informative prior distributions. The theoretical rationale of the similarity measure ω and the evidence-based nature of the similarity-weighted informative prior distribution may help to supersede the subjective notion of informative prior distributions. We hope that the similarity measure ω and the similarity-weighted informative prior distribution stimulates further research, eventually helping researchers in Psychology to move beyond non-informative prior distributions, and to finally exploit the full potential of Bayesian statistics for cumulative knowledge creation.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: Open Science Framework—http://doi.org/10.17605/OSF.IO/8AEF4.

AUTHOR CONTRIBUTIONS

CK developed the conceptual background, designed, programmed, and ran the simulation, analyzed the data, and wrote the manuscript.

REFERENCES

Aloe, A., and Thompson, C. (2013). The synthesis of partial effect sizes. *J. Soc. Soc. Work Res.* 4, 390–405. doi: 10.5243/jsswr.2013.24

Bender, R., Friede, T., Koch, A., Kuss, O., Schlattmann, P., Schwarzer, G., et al. (2018). Methods for evidence synthesis in the case of very few studies. *Res. Synthesis Methods* 9, 382–392. doi: 10.1002/jrsm.1297

Carvalho, L. M., and Ibrahim, J. (2020). On the normalized power prior. *arxiv [Preprint]*

Cheung, M. W.-L. (2015). metaSEM: an R package for meta-analysis using structural equation modeling. *Front. Psychol.* 5:1521. doi: 10.3389/fpsyg.2014.01521

Cronbach, L. J. (1982). *Designing Evaluations of Educational and Social Programs*. San Francisco, CA: Jossey Bass.

Cumming, G. (2014). The new statistics: why and how. *Psychol. Sci.* 25, 7–29.

Finch, W. H., and Miller, J. E. (2019). The use of incorrect informative priors in the estimation of MIMIC Model parameters with small sample sizes. *Struct. Equation Modeling Multidisciplinary J.* 26, 497–508. doi: 10.1080/10705511.2018.1553111

Gelman, A., and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Stat. Sci.* 7, 457–511. doi: 10.1214/ss/1177011136

Goldstein, M. (2006). Subjective bayesian analysis: principles and practice. *Bayesian Anal.* 1, 403–420. doi: 10.1214/06-BA116

Ibrahim, J., Chen, M.-H., Gwon, Y., and Chen, F. (2015). The power prior: theory and applications. *Stat. Med.* 34, 3724–3749. doi: 10.1002/sim.6728

Jak, S. (2015). *Meta-Analytic Structural Equation Modeling*. Berlin: Springer.

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

Kenny, D. A., and Judd, C. M. (2019). The unappreciated heterogeneity of effect sizes: implications for power, precision, planning of research, and replication. *Psychol. Methods* 24, 578–589. doi: 10.1037/met0000209

König, C., and van de Schoot, R. (2018). Bayesian statistics in educational research: a look at the current state of affairs. *Educ. Rev.* 70, 486–509. doi: 10.1080/00131911.2017.1350636

Kruschke, J., Aguinis, H., and Joo, H. (2012). The time has come: bayesian methods for data analysis in the organizational sciences. *Organ. Res. Methods* 15, 722–752. doi: 10.1177/0956797613504966

Lin, L., Chu, H., and Hodges, J. (2017). Alternative measures of between-study heterogeneity in meta-analysis: reducing the impact of outlying studies. *Biometrics* 73, 156–166. doi: 10.1111/biom.12543

Makel, M. C., Plucker, J. A., and Hegarty, B. (2012). Replications in psychology research: how often do they really occur? *Perspect. Psychol. Sci.* 7, 537–542. doi: 10.1177/1745691612460688

McNeish, D. (2016). On using bayesian methods to address small sample problems. *Struct. Equation Modeling Multidisciplinary J.* 23, 750–773. doi: 10.1080/10705511.2016.1186549

Moss, J., and Tveten, M. (2019). kdensity: an R package for kernel density estimation with parametric starts and asymmetric kernels. *J. Open Sour. Softw.* 4:1566. doi: 10.21105/joss.01566

Neuenschwander, B., Branson, M., and Spiegelhalter, D. (2009). A note on the power prior. *Stat. Med.* 28, 3562–3566. doi: 10.1002/sim.3722

Neuenschwander, B., Capkun-Niggli, G., Branson, M., and Spiegelhalter, D. (2010). Summarizing historical information on controls in clinical trials. *Clin. Trials* 7, 5–18. doi: 10.1177/1740774509356002

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Rev. General Psychol.* 13, 90–100. doi: 10.1037/a0015108

Seide, S., Röver, C., and Friede, T. (2019). Likelihood-based random-effects meta-analysis with few studies: empirical and simulation studies. *BMC Med. Res. Methodol.* 19:16. doi: 10.1186/s12874-018-0618-3

Shadish, W. R., Cook, T. D., and Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston, MA: Houghton-Mifflin.

Smid, S. C., McNeish, D., Miocevic, M., and van de Schoot, R. (2020). Bayesian versus frequentist estimation for structural equation models in small sample contexts: a systematic review. *Struct. Equation Modeling* 27, 131–161. doi: 10.1080/10705511.2019.1577140

Stan Development Team, (2020). *Rstan: The R interface to Stan, Version 2.19.3*. Available online at: http://mc-stan.org/users/interfaces/rstan.html (accessed September 1, 2020).

Stanley, T. D., Carter, E. C., and Doucouliagos, H. (2018). What meta-analyses reveal about the replicability of psychological research. *Psychol. Bull.* 144, 1325–1346. doi: 10.1037/bul0000169

Tipton, E. (2014). How generalizable is your experiment? An index for comparing experimental samples and populations. *J. Educ. Behav. Stat.* 39, 478–501. doi: 10.3102/1076998614558486

Tipton, E., and Olsen, R. B. (2018). A review of statistical methods for generalizing from evaluations of educational interventions. *Educ. Res.* 47, 516–524. doi: 10.3102/0013189X1878152

van de Schoot, R., Winter, S. D., Ryan, O., Zondervan-Zwijnenburg, M., and Depaoli, S. (2017). A systematic review of Bayesian articles in psychology: the last 25 years. *Psychol. Methods* 22, 217–239. doi: 10.1037/met0000100

van Erp, S., Verhagen, A. J., Grasman, R. P. P. P., and Wagenmakers, E.-J. (2017). Estimates of be-tween-study heterogeneity for 705 meta-analyses reported in Psychological Bulletin from 1990–2013. *J. Open Psychol. Data* 5:4. doi: 10.5334/jopd.33

Vanpaemel, W. (2011). Constructing informative model priors using hierarchical methods. *J. Math. Psychol.* 55, 106–117. doi: 10.1016/j.jmp.2010.08.005

Veroniki, A. A., Jackson, D., Viechtbauer, W., Bender, R., Bowden, J., Knapp, G., et al. (2016). Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Res. Synthesis Methods* 7, 55–79. doi: 10.1002/jrsm.1164

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *J. Stat. Softw.* 36, 1–48.

Weber, S., Li, Y., Seaman Iii, J. W., Kakizume, T., and Schmidli, H. (2019). Applying meta-analytic predictive priors with the R Bayesian evidence synthesis tools. *arxiv [Preprint]*

Williams, D. R., Rast, P., and Bürkner, P.-C. (2018). Bayesian meta-analysis with weakly informative prior distributions. *PsyArXiv [Preprint]* doi: 10.31234/osf.io/7tbrm

Zhang, Z., Jiang, K., Liu, H., and Oh, I.-S. (2017). Bayesian meta-analysis of correlation coefficients through power prior. *Commun. Stat. Theory Methods* 46, 11988–12007. doi: 10.1080/03610926.2017.1288251

Conflict of Interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Psychology

ORIGINAL RESEARCH
published: 22 July 2021
doi: 10.3389/fpsyg.2021.624032

Approximate Measurement Invariance of Willingness to Sacrifice for the Environment Across 30 Countries: The Importance of Prior Distributions and Their Visualization

Ingrid Arts, Qixiang Fang, Rens van de Schoot and Katharina Meitinger*

Department of Methodology and Statistics, Faculty of Social Sciences, Utrecht University, Utrecht, Netherlands

OPEN ACCESS

Edited by:
Alexander Robitzsch,
IPN—Leibniz Institute for Science and Mathematics Education, Germany

Reviewed by:
Niki De Bondt,
University of Antwerp, Belgium
Artur Pokropek,
Institute of Philosophy and Sociology (PAN), Poland

***Correspondence:**
Ingrid Arts
i.j.m.arts@uu.nl

Specialty section:
This article was submitted to Quantitative Psychology and Measurement,
a section of the journal Frontiers in Psychology

Received: *30 October 2020*
Accepted: *21 June 2021*
Published: *22 July 2021*

Citation:
Arts I, Fang Q, van de Schoot R and Meitinger K (2021) Approximate Measurement Invariance of Willingness to Sacrifice for the Environment Across 30 Countries: The Importance of Prior Distributions and Their Visualization. Front. Psychol. 12:624032. doi: 10.3389/fpsyg.2021.624032

Nationwide opinions and international attitudes toward climate and environmental change are receiving increasing attention in both scientific and political communities. An often used way to measure these attitudes is by large-scale social surveys. However, the assumption for a valid country comparison, measurement invariance, is often not met, especially when a large number of countries are being compared. This makes a ranking of countries by the mean of a latent variable potentially unstable, and may lead to untrustworthy conclusions. Recently, more liberal approaches to assessing measurement invariance have been proposed, such as the alignment method in combination with Bayesian approximate measurement invariance. However, the effect of prior variances on the assessment procedure and substantive conclusions is often not well understood. In this article, we tested for measurement invariance of the latent variable "willingness to sacrifice for the environment" using Maximum Likelihood Multigroup Confirmatory Factor Analysis and Bayesian approximate measurement invariance, both with and without alignment optimization. For the Bayesian models, we used multiple priors to assess the impact on the rank order stability of countries. The results are visualized in such a way that the effect of different prior variances and models on group means and rankings becomes clear. We show that even when models appear to be a good fit to the data, there might still be an unwanted impact on the rank ordering of countries. From the results, we can conclude that people in Switzerland and South Korea are most motivated to sacrifice for the environment, while people in Latvia are less motivated to sacrifice for the environment.

Keywords: measurement invariance, visualization, Bayes, group ranking, MGCFA, prior sensitivity, Bayesian approximate measurement invariance (BAMI)

INTRODUCTION

One of the main issues the world population faces today is climate and environmental change. Some of the challenges that have to be faced include floods, droughts, food insecurity, and biodiversity loss. These challenges may give rise to socioeconomic problems such as refugee crises, relocating populations and cities, and famines (Zhang et al., 2020). As the challenges will differ across regions, but are not limited by national borders, international cooperation is required. At the same time, a "one size fits all" solution is unlikely to solve these issues (Andonova and Coetzee, 2020). Several studies have been conducted on how the inhabitants of different countries perceive the subject of climate and environmental change, and the different aspects of social behavior regarding this subject: e.g., knowledge of climate change, risk perception, and the willingness to act (van Valkengoed and Steg, 2019). Hadler and Kraemer (2016) showed that the inhabitants of different countries do not assess all these threats in the same way: in some countries air pollution is seen as a major threat, while in others water shortages are considered a hazard.

The term "environmental concern" has been used widely to explain environmental behavior (e.g., Dunlap and Jones, 2002; Bamberg, 2003; Schultz et al., 2005; Franzen and Meyer, 2010; Marquart-Pyatt, 2012a; Fairbrother, 2013; Pampel, 2014; Mayerl, 2016; Pisano and Lubell, 2017; Shao et al., 2018). However, a clear definition of this concept is lacking (e.g., Dunlap and Jones, 2002; Schultz et al., 2005). Bamberg (2003, p. 21) described environmental concern as "the whole range of environmentally related perceptions, emotions, knowledge, attitudes, values, and behaviors," while Dunlap and Jones (2002, p. 485) described environmental concern as "the degree to which people are aware of problems regarding the environment and support efforts to solve them and/or indicate the willingness to contribute personally to their solution." Following the latter definition, environmental concern consists of at least two parts: on the one hand, perceptions of environmental problems (e.g., risks and beliefs), and, on the other hand, the willingness to contribute to the solution (e.g., to pay more taxes or higher prices, or to fly less). This translates into two latent variables that operationalize environmental concern: "environmental attitude" (EA) and "willingness to sacrifice (or pay) for the environment" (WTS). These two latent variables have been used both individually and in combination to operationalize environmental concern (Mayerl and Best, 2019). The latent variable WTS is frequently used to measure the extent to which people are willing to sacrifice something in their daily life (money, goods, time, comfort) to save the environment, and has been examined by several authors (e.g., Ivanova and Tranter, 2008; Fairbrother, 2013; Franzen and Vogl, 2013; Pampel, 2014; Sara and Nurit, 2014; Shao et al., 2018). The relation with cultural, sociological, economic, or political factors has been studied quite extensively (e.g., Marquart-Pyatt, 2012b; Franzen and Vogl, 2013; Pampel, 2014; Bozonnet, 2016; McCright et al., 2016; Shao et al., 2018).

Large-scale surveys are often used for exploring knowledge, attitudes, and (intentional) behavior regarding climate and environmental change (e.g., Bamberg, 2003; Franzen and Meyer, 2010; Marquart-Pyatt, 2012a; Hadler and Kraemer, 2016; Knight, 2016; Pisano and Lubell, 2017; Libarkin et al., 2018). One precondition for the valid comparison of attitudes toward climate and environmental change across many countries is that measurement properties are equivalent across countries (Jöreskog, 1971; Vandenberg and Lance, 2000). This means that all participants in all countries should interpret both the survey questions and the underlying latent variables in the same way. This equivalence of measurement properties is also called Measurement Invariance (MI). Establishing whether MI holds is usually done by conducting a maximum-likelihood (ML) Multi-Group Confirmatory Factor Analysis (MGCFA). There are at least four types of MI: configural (also referred to as "weak"), metric, scalar ("strong"), and residual ("strict") invariance. Configural invariance allows for the comparison of latent variables among groups, metric invariance allows for a comparison of the items (questions) that make up the latent variable(s) among groups, and scalar invariance allows for the comparison of latent means across groups. Scalar invariance, however, is rarely established, especially when many groups are compared (e.g., Muthen and Asparouhov, 2013; Lommen et al., 2014; Kim et al., 2017; Marsh et al., 2018)[1].

Measurement invariance of the latent variable WTS has been investigated by Mayerl and Best (2019), and they established both configural and metric invariance, but not scalar invariance. Using ML MGCFA, Marquart-Pyatt (2012b) also found configural and metric invariance, but not scalar invariance. To our knowledge, scalar invariance for the latent variable WTS has not been found by other authors, rendering the substantive interpretation of results from country rankings potentially untrustworthy (Byrne and van de Vijver, 2017; Marsh et al., 2018).

Alternative approaches have been proposed, such as alignment optimization, which allows for few but larger parameters differences between some groups (Asparouhov and Muthén, 2014), Bayesian Approximate MI (Muthén and Asparouhov, 2012; van de Schoot et al., 2013), hereinafter referred to as BAMI[2], which allows multiple but small differences between all groups, or a combination of both, BAMI alignment (Asparouhov and Muthén, 2014). When BAMI alignment is used, small variances are allowed for each group, while a few groups are allowed to have large variances. This leads to fewer noninvariant parameters than when the ML alignment method is applied, facilitating the interpretation of the model (Asparouhov and Muthén, 2014). Although this might be a highly interesting approach when a comparison of many groups is desired, it seems that, at least up until now, this approach has not been applied often: we only found two studies in which BAMI and alignment are combined: De Bondt and Van Petegem (2015) and van de Vijver et al. (2019), and certainly not in the field of environmental change.

The key to using Bayesian methods is the use of priors: some "wiggle room" is defined between which the variances of different groups are allowed to vary. However, the selection of these priors (from simulation studies, literature, or experience)

[1] Residual invariance means that the sum of specific variance (variance of the item that is not shared with the factor and error variance) are also equal across groups (Putnick and Bornstein, 2016). Since this is not a requirement for comparing means across groups, we do not report it in this article.

[2] In previous research, the term Approximate MI (van de Schoot et al., 2013) has sometimes been used as a collective term for any method that can be used when the criteria for the exact scalar model are not fulfilled (Russell et al., 2016; Flake and McCoach, 2018), and sometimes to mention a specific method (e.g., Byrne and van de Vijver, 2017; Amérigo et al., 2020). To prevent any further confusion, we propose to use the term Bayesian Approximate Measurement Invariance (BAMI) when using a Bayesian model with strong informative priors on differences between factor loadings and/or intercepts, thus excluding non-Bayesian (ML or empirical Bayes) type of methods like random item effects (Fox and Verhagen, 2018).

is not an easy task. It seems that researchers applying Bayesian methods are not always fully aware of the potential impact of specifying priors (e.g., Spiegelhalter et al., 2000; Rupp et al., 2004; Ashby, 2006; Kruschke et al., 2012; Rietbergen et al., 2017; van de Schoot et al., 2017; König and van de Schoot, 2018; Smid et al., 2020). Nonetheless, for the verification and reproducibility of research (Munafò et al., 2017; van de Schoot et al., 2021), it is crucial to evaluate the influence of varying priors on the impact of substantive conclusions, which is referred to as sensitivity analysis. Some general guidelines regarding prior sensitivity can be found in the literature (e.g., Depaoli and van de Schoot, 2017; van Erp et al., 2018; van de Schoot et al., 2019; Pokropek et al., 2020). Although a sensitivity analysis of different prior settings helps to determine the impact of prior variances on substantive conclusions, it has, to our knowledge, never been applied for BAMI with empirical data.

The goals of our article are to apply the method of BAMI to the concept of "willingness to sacrifice (or pay) for the environment," compare the results of different prior settings to each other and to other methods of dealing with measurement invariance (i.e., ML MGCFA and the ML alignment method) through visualization, and to provide an example for a transparent workflow.

In what follows, we first provide a technical introduction to the four methods we used to assess MI. As it can be difficult to interpret multiple models and methods, and because we want to be as transparent as possible in our decision-making process, we summarize our design choices and possible alternatives in a decision tree. We test the models to evaluate whether and how different prior variances influence the ranking of the countries on the latent variable WTS. We visualize the results to facilitate a comparison of the latent means of different models and methods without the use of complex and elaborate tables. All appendices, the scripts to reproduce our results, the final output files and additional material can be found on website of the Open Science Framework (OSF) (Arts et al., 2021).

TECHNICAL BACKGROUND

In this section, we introduce the four methods we used to evaluate measurement invariance: (1) ML MGCFA, (2) ML MGCFA using the alignment optimization, (3) BAMI, and (4) BAMI in combination with the alignment method.

MGCFA

The MGCFA model is defined as:

$$y_{ipg} = \nu_{pg} + \lambda_{pg}\eta_{ig} + \epsilon_{ipg} \tag{1}$$

where $p = 1, ...P$ is the number of observed indicator variables, $g = 1, ...G$ is the number of groups, $i = 1, ...N$ is the number of individual observations, λ_{pg} is a vector of factor loadings, ν_{pg} is a vector of intercepts and η_{ig} is a vector of latent variables. Furthermore, ϵ_{ipg} is a vector of error terms that is assumed to be normally distributed with $N(0, \theta_{pg})$, and η_{ig} is assumed to have a distribution of $N(\alpha_g, \varphi_g)$. θ_{pg} is the variance of ϵ_{ipg}, α_g is the mean of normally distributed latent variable η_{ig}, and φ_g is the variance of η_{ig}. For WTS let P = 3 (3 items) and G = 30 (30 countries), which means that λ_{pg} is a 3 × 30 matrix. The same is true for ν_{pg}.

In the configural model, both λ and ν are allowed to vary across groups[3], but the factor structure is equal for all groups, that is, in all 30 countries the latent variable WTS is covered by the same three items.

When both the number of latent variables and the factor loading λ are held equal across groups but the intercept ν is allowed to vary, one is testing for metric invariance: $\lambda_{11} = \lambda_{12} = \lambda_{13}$, etc. This means that for every group, the latent variable η_g contributes equally to item y_{pg}.

If metric invariance holds, it is possible to test for scalar invariance. In this case both loadings λ and intercepts ν are held equal across groups: $\lambda_{11} = \lambda_{12} = \lambda_{13}$ etc. and $\nu_{11} = \nu_{12} = \nu_{13}$ etc., so that Equation (1) becomes:

$$y_p = \nu_p + \lambda_p\eta + \epsilon_{ipg} \tag{2}$$

When scalar invariance holds, the latent means of WTS can be compared between groups, and a ranking of the latent means can be made. However, scalar, or strong, invariance is very rare, especially when comparing many groups (Asparouhov and Muthén, 2014; Byrne and van de Vijver, 2017; Kim et al., 2017; Marsh et al., 2018). This is due to the fact that with increasing number of countries, the probability increases that countries substantially deviate in answering behavior. When many groups with small deviations are being compared, these small deviations add up to the non-invariance of the scale assessing WTS.

Alignment Optimization

To reduce the impact of a lack of measurement invariance for many groups, the alignment optimization method has been introduced (Muthen and Asparouhov, 2013; Asparouhov and Muthén, 2014). Alignment optimization consists of two steps (Asparouhov and Muthén, 2014). First, a null model M_0 is estimated with loadings and intercepts allowed to vary across groups. As loadings and intercepts are freed across groups, factor means and factor variances are set to 0 and 1 for every group: $\alpha_g = 0$ and $\varphi_g = 1$. Now, the latent variable for the null model η_{g0} can be calculated.

Second, the method divides groups G into pairs Q and tries to find, for every Q, the intercepts and loadings that yield the same likelihood as the M_0 model (Asparouhov and Muthén, 2014; Flake and McCoach, 2018). Now, λ_{pg} and ν_{pg} can be calculated, where α_g and φ_g have to be chosen in such a way that they minimize the amount of measurement non-invariance and $q1$, $q2$, etc. are the different pairs of groups in the data. For the full set of equations, see Asparouhov and Muthén (2014), Flake and McCoach (2018). This means that, for the latent variable WTS, $q = 1...435$ for every item (for every item there are 435 possible pairs).

[3] Technically speaking, this is not entirely correct: for identification of the model, Mplus by default fixes the loading/intercept of the first item of every group to 1. For more details about parameterization of CFA models we refer the interested reader to Little et al. (2006).

The total amount of measurement non-invariance is shown by the total loss/simplicity function F:

$$F = \sum_{p} \sum_{g_1 < g_2} w_{g_1,g_2} f(\lambda_{pg_1,q_1} - \lambda_{pg_2,q_1}) + \sum_{p} \sum_{g_1 < g_2} w_{g_1,g_2} f(\nu_{pg_1,q_1} - \nu_{pg_2,q_1}) \quad (3)$$

In Equation (3), for the intercepts and loadings of every Q, the differences between the parameters are summed and then scaled by the Component Loss Function (CLF) f. Group sizes are appointed by weight factors w_{g_1} and w_{g_2}, where w_{g_1} is the weight factor of group 1 and w_{g_2} is the weight factor of the, differently sized group 2. In this way, bigger pairs of groups contribute more to the total loss function than smaller pairs. The weight factor can be calculated as follows:

$$w_q = w_{g_1,g_2} = \sqrt{N_{g_1} N_{g_2}} \quad (4)$$

The CLF has been used in exploratory factor analysis (EFA) to estimate factor loadings with the simplest possible structure (Jennrich, 2006). For the alignment the CLF is:

$$f(x) = \sqrt{\sqrt{x^2 + \epsilon}} \quad (5)$$

with ϵ being a small number, for example, 0.01 (Asparouhov and Muthén, 2014). This positive number ensures that $f(x)$ has a continuous first derivative, making the optimization of the total loss function F easier. As ϵ is so small, $f(x) \approx \sqrt{|x|}$, which leads to no loss if $x = 0$, amplified loss if $x < 1$, and attenuated loss if $x > 1$ (Asparouhov and Muthén, 2014). Due to this CLF, F will be minimized when there are a few large non-invariant loadings and intercepts and a majority of approximately non-invariant loadings and intercepts (Kim et al., 2017). When there are many medium-sized non-invariant parameters, the total loss function does not optimize (Flake and McCoach, 2018). If F does optimize, the parameters α_g and φ_g will be identified for all groups except the first one. For the first group, the variance can be calculated using the following parameter constraints, making the number of estimated parameters $(2G - 1)$:

$$\varphi_1 \times ... \times \varphi_g = 1 \quad (6)$$

α_1 can be set to 0, although this is not always needed and might lead to untrustworthy estimates (Asparouhov and Muthén, 2014). When α_1 and φ_1 are both constrained, the alignment is called FIXED in Mplus, and when only φ_1 is constrained, that alignment is said to be FREE.

Although the alignment optimization allows for some large invariances between groups, all other groups are assumed to have the same loadings and intercepts. In other words, small variances between groups cannot be taken into account. To ensure that mean and variance can be fixed for one country (as it is in the MGCFA and BAMI), we have to opt for the FIXED alignment and specify one country to be fixed.

BAMI

A synopsis of Bayesian statistics, including the most important aspects of determining prior distributions, likelihood functions and posterior distributions, in addition to discussing different applications of the method across disciplines can be found in van de Schoot et al. (2021).

With BAMI, priors with a mean of zero and some small variance are put on the differences between factor loadings and the differences between intercepts across groups: the terms λ_{pg} and ν_{pg} from Equation (1) are now estimated being approximately equal across groups instead of exactly equal: $\lambda_{11} \approx \lambda_{12} \approx \lambda_{13}$ etc. instead of $\lambda_{11} = \lambda_{12} = \lambda_{13}$, etc. and $\nu_{11} \approx \nu_{12} \approx \nu_{13}$ etc. instead of $\nu_{11} = \nu_{12} = \nu_{13}$, etc.

The prior is not put directly on the differences between parameters, but on the covariances between parameters. This means that, for instance

$$V(\lambda_{11} - \lambda_{12}) = V(\lambda_{11}) + V(\lambda_{12}) - 2Cov(\lambda_{11}, \lambda_{12}) \quad (7)$$

where $V(\lambda_{11} - \lambda_{12})$ is the difference between the variances of the first loading of the first group and the first loading of the second group. If we assume that these prior variances are small, for instance 0.5, and the covariance is 0.495, that would lead to a value of 0.01 for $V(\lambda_{11} - \lambda_{12})$, or V^d.

BAMI uses strong informative priors on cross-group variances of loadings λ and intercepts ν. It is important to carefully select these priors since they have a strong impact on the posterior results. Large values of V^d will result in decreasing the chance of model convergence, as they do not impose enough information on the model (Muthén and Asparouhov, 2012). Smaller values of V^d, on the other hand, might bring the model too close to a scalar model, reducing flexibility of the model to deal with the existing non-invariance.

BAMI With Alignment

The alignment method and BAMI can be combined. In that case, small variances are allowed for each group, while a few groups are allowed to have large variances. The alignment method for BAMI is similar to that for the exact method:

In the first step, an M_0 model is estimated, from which the optimal set of measurement parameters from the configural model is calculated. Now the M_0 model is a model where the intercepts and loadings are approximately equal across groups and the factor means and variances are estimated as free parameters in all groups but the first one.

In the second step, this M_{B0} model, the posterior of the configural factor loadings and intercepts are computed using the following equations:

$$\lambda_{pg,0} = \lambda_{pg,B} \sqrt{\varphi_{Bg}} \quad (8)$$

$$\nu_{pg,0} = \nu_{pg,B} + \alpha_{Bg} \lambda_{pg,B} \quad (9)$$

where $\lambda_{pg,0}$ and $\nu_{pg,0}$ are the configural loadings and intercepts and $\alpha_{Bg}, \varphi_{Bg}, \lambda_{pg,B}$, and $\nu_{pg,B}$ are the BAMI parameters. Using the BAMI parameters and Equations (8) and (9), the configural loadings and intercepts are computed for every iteration. These are then used to form the posterior distribution for $\lambda_{pg,0}$ and $\nu_{pg,0}$.

In the third and final step, the aligned estimates are obtained for every iteration using the configural factor loading and intercept values to minimize the simplicity function of Equation (3). The aligned parameter values obtained from one iteration are used as starting values in the next iteration. Finally, the aligned parameter values from all iterations are then used to estimate the aligned posterior distribution as well as the point estimates and the standard errors for the aligned parameters (Asparouhov and Muthén, 2014). This leads to fewer non-invariant parameters than when the ML alignment method is applied, facilitating the interpretation of the model (Asparouhov and Muthén, 2014).

METHODS AND DATA

Data

We used the data from the 2010 Module on Environment of the ISSP (ISSP Research Group, 2019). For the full report on this module, see GESIS (2019). The latent variable WTS consists of three questions, see **Table 1** for the exact wording, with answers on a five-point response scale (1 being *very unwilling* and 5 being *very willing*) and a *cannot choose* option for participants who could not or would not answer the question. WTS has, in combination with EA, been tested for MI by Mayerl and Best (2019) to explain the concept "environmental concern" when applied to 30 countries: Austria, Belgium, Bulgaria, Canada, Chile, Croatia, Czech Republic, Denmark, Finland, France, Germany, Great Britain, Israel, Japan, Latvia, Lithuania, Mexico, New Zealand, Norway, Philippines, Russia, Slovakia, Slovenia, South Africa, South Korea, Spain, Sweden, Switzerland, Turkey, and the United States. They found that, although metric invariance was achieved, scalar invariance was not. When we repeated this analysis we came to the same conclusion, for the results of this analysis, see Appendix A in Arts et al. (2021). For simplicity reasons, we only focus on the latent variable WTS, just like Ivanova and Tranter (2008), Fairbrother (2013), Franzen and Vogl (2013), Pampel (2014), Sara and Nurit (2014), and Shao et al. (2018). To further analyze this scale, we first ensured that we used the exact same data from the ISSP 2010 environment module and we followed the identical procedure as in the original study to handle missingness (i.e., listwise deletion—correspondence with author, November 26 2019), resulting in the same sample (n = 24,583). For the exact procedure and all code, see Appendix A in Arts et al. (2021). The sample sizes per country ranges from 798 (Iceland) to 3,112 (South Africa) with an average group size of 1,401, see for more details **Table 2**.

TABLE 1 | Exact wording of the questions in WTS.

Number	Question
Q12a	How willing would you be to pay much higher prices in order to protect the environment?
Q12b	How willing would you be to pay much higher taxes in order to protect the environment?
Q12c	How willing would you be to accept cuts in your standard of living in order to protect the environment?

TABLE 2 | Participating countries in the ISSP environmental module.

Country	Sample size	Country	Sample size
Argentina	1,130	Lithuania	1,023
Australia	1,946	Mexico	1,637
Austria	1,019	Netherlands	1,472
Belgium (Flanders)	1,142	New Zealand	1,172
Bulgaria	1,003	Norway	1,382
Canada	985	Philippines	1,200
Chile	1,436	Portugal	1,022
Croatia	1,210	Russia	1,619
Czech Republic	1,428	Slovakia	1,159
Denmark	1,305	Slovenia	1,082
Finland	1,211	South Africa	3,112
France	2,253	South Korea	1,576
Germany	1,407	Spain	2,560
Great Britain	928	Sweden	1,181
Iceland	798	Switzerland	1,212
Israel	1,216	Taiwan	2,209
Japan	1,307	Turkey	1,665
Latvia	1,000	United States	1,430
Total			50,437

Analytical Strategy

We assessed the measurement invariance of the latent variable WTS by applying four methods for detecting MI: ML MGCFA, the ML alignment optimization, BAMI, and BAMI with alignment optimization. For all analyses, one reference country was selected for which the factor mean and factor variance are held to 0 and 1, respectively (Spain). By fixing the mean and variance for a specific reference country for every model, it is ensured that any differences in outcomes are due to a method and not due to a difference in default settings of the model (for some models by default the parameters are fixed for the first group, while for other models it is the last group). We selected Spain as the reference country since the results presented by Mayerl and Best (2019) indicate that the results for WTS from this country can be seen as "average" within the group of thirty countries.

For the BAMI method, both with and without alignment, we tested the effect of different priors on the models. One way of selecting priors for new data is by using the results of simulation studies. **Table 3** shows an overview of simulation studies that have investigated BAMI and the priors that were used. As can be seen from this table, the simulation results are not entirely conclusive: The authors of these articles report that they achieve the best results when using priors with a variance of 0.001, 0.005, 0.01, or 0.05. However, the number of groups, group sizes and invariance criteria in these studies vary, complicating a comparison of the best performing prior variance(s).

We also searched for empirical studies in which BAMI was applied to empirical data. In a total of 30 empirical studies,

TABLE 3 | Simulation studies using Bayesian approximate measurement invariance.

Article	Number of groups	Group size	Prior variance	Invariance criteria
Muthén and Asparouhov, 2012	40	500	0.10, 0.05, 0.01	PPP
van de Schoot et al., 2013	2	1,000	0.50, 0.05, 0.01, 0.005, 0.0005	PPP, 95% CI
Kim et al., 2017	25, 50	50, 100, 1,000	0.05, 0.001	DIC, PPP, 95% CI, BIC
Lek et al., 2018	2	50, 100, 200, 1000	0.10, 0.05, 0.01, 0.001	95% CI
Shi et al., 2017	2	500	0.10, 0.05, 0.01	PPP, 95%CI
Pokropek et al., 2019	24	1500	0.10, 0.05, 0.01, 0.005	cor, RMSEA, 95%CI
Pokropek et al. (2020)	4, 24, 50	400, 1500, 3,000	0.05, 0.025, 0.01, 0.005, 0.001, 0.000*	BIC, DIC, PPP

**A Bayesian model with a prior variance of 0 is the scalar model.*

PPP, posterior predictive p*-value; DIC, deviance information criterion; 95% CI, 95% credibility interval; cor, correlation; RMSEA, root mean square error of approximation; BIC, Bayesian information criterion.*

there were 13 in which only one prior was used, and in eight of these 13 studies, no specification was given as to why that specific prior was used. In the 17 studies where multiple priors were tested, three did not provide any information on why these priors were selected. The 14 other studies based the priors used on Muthén and Asparouhov (2012), van de Schoot et al. (2013), Asparouhov et al. (2015), or Seddig and Leitgöb (2018). For more information about these empirical studies and their variances see the additional material (Arts et al., 2021). The most frequently used prior variance in these studies is 0.01, followed by 0.05 as recommended by Muthén and Asparouhov (2012) and van de Schoot et al. (2013), respectively. However, other priors were also included in the different sensitivity analyses, ranging from 0.000000001 to 0.5.

We decided to estimate five different models, with priors with a variance of 0.05–0.01 (decreasing at 0.01 per prior) and three models with priors with variances of 0.001, 0.0005, and 0.0001. This includes the prior variances that are used most often in both simulation and empirical studies. Using such a large number of priors should create a clear overview of the influence of different prior variances on the rank order stability of the countries when ranked on their latent factor means. In addition to priors on the differences between loadings and intercepts, there are also priors on other parameters, such as the residuals. However, we will not discuss these priors in this article and we relied on the Mplus default values which can be found in Muthén and Muthén (2019). To ensure that the chains reached their target distributions, we checked whether all iterations after burn-in met the Gelman-Rubin criterion. Therefore, we set the convergence criterion to a rather strict 0.01 instead of the default 0.05 (Muthén and Muthén, 2019) and the maximum and minimum number of iterations to 100,000 and 40,000, respectively.

For the analysis in this article, we used the software Mplus version 8.4 (Muthén and Muthén, 2019). The results were analyzed using R version 6.3.2 (R Development Core Team, 2017) and MplusAutomation version 0-7.3 (Hallquist and Wiley, 2018) was used for the exchange between the two programs. More information about the analysis and the exact Mplus and R code can be found in Appendix B on Arts et al. (2021).

Model Fit

To assess model fit for ML MGCFA, the indices that are most widely used are the χ^2-value, root mean square error of approximation (RMSEA), comparative fit index (CFI), Tucker-Lewis index (TLI), and the standardized root mean square residual (SRMR) (Gallagher and Brown, 2013, p. 298). When testing for configural invariance cutoff values of CFI $\geq$ 0.95, TLI $\geq$ 0.95, RMSEA $\leq$ 0.06, and SRMR $\leq$ 0.08 have been proposed by Hu and Bentler (1999). When checking for metric and scalar invariance, relative fit indices are more useful than absolute fit indices (Chen, 2007). These relative fit indices are a comparison of configural with metric and metric with scalar fit indices. Depending on these fit indices, the model for metric invariance can be assumed to perform better or worse than the model for configural invariance (and the same is true for metric and scalar invariance). For sample sizes above 300, ΔRMSEA $\leq$0.015 and ΔCFI $\leq$0.01 or ΔSRMR $\leq$0.03 indicate invariance when moving from the configural to the metric model, and ΔRMSEA $\leq$0.015 and ΔCFI $\leq$0.01 or ΔSRMR $\leq$0.01 indicate noninvariance when moving from the metric to the scalar model (Chen, 2007).

For the alignment method, fit indices have not been specified. Muthen and Asparouhov (2014) propose that the results can be considered trustworthy when no more than 25% of the parameters are non-invariant. However, Kim et al. (2017) have argued that this way the degree and location of non-invariance cannot be taken into account.

When BAMI is used, the model fit may be indicated by the posterior predictive p-value (PPP-value). This value indicates the ratio between the iterations for which the replicated χ^2 value exceeds the observed χ^2 value (Pokropek et al., 2020). A PPP-value of 0.50 indicates perfect model fit; a value below 0.50 indicates an underfit of the model, and a value above 0.50 indicates an overfit. Furthermore, the 95% credibility interval (CI) should include 0, preferably with 0 in the middle of the interval (Muthén and Asparouhov, 2012; van de Schoot et al., 2013). As PPP-values decline, the model fits the data less well. However, a specific cutoff value at which the model no longer fits the data is hard to determine. Muthén and Asparouhov (2012) suggest that models with PPP-values lower than 0.10, 0.05, or 0.01. do not fit the data anymore. In the literature, PPP-values above 0.05 are often seen as an indication for good model fit. A drawback of the PPP-value is that it might not identify a model with good fit correctly when using different priors with large sample sizes (Asparouhov and Muthén, 2010, 2019; Hoijtink and van de Schoot, 2018; Hoofs et al., 2018)[4].

Recently, Bayesian versions of fit indices have been proposed: Bayesian RMSEA (BRMSEA), Bayesian CFI (BCFI), and Bayesian TLI (BTLI) can be computed based on differences between the observed and replicated discrepancy functions (Liang, 2020). These Bayesian fit statistics have been implemented in Mplus version 8.4, making it more convenient to identify good model fit (Asparouhov and Muthén, 2019). The calculation of these fit indices is very similar to that of the fit indices of an exact model, and therefore, the same cutoff values can be used (Asparouhov and Muthén, 2019; Garnier-Villarreal and Jorgensen, 2020). This means that BCFI $\geq$ 0.95, BTLI $\geq$ 0.95, and BRMSEA $\leq$ 0.06 indicate good model fit. However, just as with the ML models, a combination of cutoff values must be used to indicate good or bad model fit. Other criteria that are being used to determine model fit are the BIC (Schwarz, 1978) and the DIC (Spiegelhalter et al., 2002). These information theoretic indices are less self-explanatory than the other fit indices: when selecting the best performing model from a series of models (the model that fits the data best and is the least complex), the model with the lowest BIC or DIC is preferred. This does not mean that the model with the lowest BIC or DIC is a good fit to the data: it is simply preferable to models with a higher BIC or DIC. Asparouhov et al. (2015) stated that, when sample sizes are large, coupled with a large number of observed indicators, DIC is preferable to BIC and Pokropek et al. (2020) concluded that DIC is a good indicator to identify the preferred prior mean and variance. On the other hand, Hoijtink and van de Schoot (2018) stated that the DIC is not suitable for evaluating models with small priors. This makes the use of the DIC as fit index promising, but also shows that its value should be treated with care. At a minimum, DIC should always be combined with other fit indices.

BAMI with alignment has, similar to the ML alignment method, no guidelines to determine model fit. Both De Bondt and Van Petegem (2015) and van de Vijver et al. (2019) tested a model with multiple small prior variances. De Bondt and Van Petegem (2015) used a prior variance of 0.01 and conducted a sensitivity analyses with prior variances decreasing with a factor 10, and van de Vijver et al. (2019) used a prior variance of 0.05 and conducted a sensitivity analyses with prior variances of 0.001, 0.005, 0.01, 0.05, and 0.1. Both De Bondt and Van Petegem (2015) and van de Vijver et al. (2019) analyzed the alignment part of the model by comparing, for each item, the intercepts and loadings across paired groups. This can be a very laborious process when multiple items and multiple groups are concerned. One could also use the rule of thumb that, to obtain trustworthy results, no more than 25% of the parameters can be invariant, as proposed by Muthen and Asparouhov (2014).

As shown above, there are many different criteria and cut-off values that provide insight into whether a model fits the data. Since there are so many different indicators these cutoff values should be treated with care: fit statistics can be influenced by, e.g., sample size or model complexity (Chen, 2007). Additionally, having one indication of good model fit is not enough to conclude that the model is a good fit to the data, and multiple fit statistics may even contradict each other. This exact point was addressed by Lai and Green (2016), who showed that RMSEA and CFI can contradict each other. Even when there is sufficient evidence that a model is a good fit to the data, this does not necessarily mean that it is the best model.

RESULTS

MGCFA

The fit indices for the metric and scalar MGCFA are shown in **Table 4**. Since the configural model was saturated, the results are not shown here. Therefore, for the metric model we asses the absolute fit indices instead of the relative fit indices. The metric model shows good fit, with a CFI and TLI of 0.993 and 0.989, respectively. With 0.069 the RMSEA value is above 0.06 but still below 0.08, indicating at least a reasonable fit. The fit indices for the scalar model all point to rejection of the scalar model: ΔRMSEA, ΔSRMR and ΔCFI are well above the cutoff values of 0.015, 0.01, and 0.01 (0.085, 0.057, and 0.065, respectively). Based on these results we conclude that scalar invariance is absent, and that a comparison of the latent variable WTS across countries may not be trustworthy. However, this exact approach could be too strict in its assessment.

Alignment Optimization

Regarding the alignment optimization, the invariant, and non-invariant parameters are shown in **Table 5** with non-invariant parameters bolded and in brackets. Most non-invariant parameters can be found in the intercepts, with 48 non-invariant parameters, while for the loadings only seven parameters are non-invariant. However, a total of 55 parameters are non-invariant, which is 30.55% of all parameters. This is well above 25%, a rough cut-off value proposed by Muthen and Asparouhov (2014), implying that, for these data, a valid rank order comparison cannot be made if the ML alignment method is used.

BAMI

For BAMI, only the results for the models that converged are presented here (models with a prior variance of 0.02, 0.01, 0.001, 0.0005, and 0.0001). These models also converged when the

[4] Hoijtink and van de Schoot (2018) demonstrated that, with increasing sample sizes, the PPP-value does not decrease, but increases. Therefore, the prior-posterior predictive *p*-value (PPPP-value) was proposed by Hoijtink and van de Schoot (2018), and a generalized version was implemented in Mplus by Asparouhov and Muthén (2017). Whereas, the original PPP-value is a test of model fit which tests the fit of the model to the data and is based on comparing the model with the unrestricted covariance model, the PPPP-value is a test for the approximate parameters in the model. The PPPP-value is not a test for model fit and should not be interpreted as evidence that the model fits the data. The proper interpretation of the PPPP-value is given by Asparouhov and Muthén (2017): "If the test does not reject, the minor parameters (represented by θ_1) can be assumed to come from N(0, v) distribution, with v being a small variance. More broadly speaking, if the PPPP does not reject, that means that there is no evidence in the data for the minor parameters in model M(θ_1, θ_2) to be outside the N(0, v) distribution" (p. 10). Here, θ_2 represents the large parameters of model M. Unfortunately, the PPPP-value is not yet available for BAMI in Mplus.

TABLE 4 | Fit statistics of the MGCFA model.

	χ^2 (df)	$\Delta\chi^2$(Δ df)	*p*-value	RMSEA	Δ RMSEA	SRMR	Δ SRMR	CFI	Δ CFI	TLI	Δ TLI
Configural*											
Metric	287.324 (58)		0.00	0.069		0.051		0.993		0.989	
Scalar	2382.434 (116)	2095.110 (58)	0.00	0.154	0.085	0.108	0.057	0.928	0.065	0.944	0.045

**This model was saturated. df, degrees of freedom; RMSEA, root mean square error of approximation; SRMR, standardized root mean square residual; CFI, comparative fit index; TLI, Tucker-Lewis index. Numbers are absolute.*

TABLE 5 | (Non)invariant parameters for ML alignment optimization.

Intercepts/Thresholds	
Q12a	33 **(40)** **(56)** **(100)** **(124)** 152 191 203 **(208)** 246 **(250)** **(276)** **(376)** **(392)** **(410)** 428 440 484 **(554)** **(578)** **(608)** 643 703 705 710 752 **(756)** 792 **(826)** **(840)**
Q12b	33 40 56 **(100)** 124 **(152)** **(191)** **(203)** 208 246 250 276 376 392 410 **(428)** **(440)** **(484)** 554 578 **(608)** **(643)** **(703)** 705 **(710)** **(752)** 756 **(792)** **(826)** **(840)**
Q12c	33 **(40)** 56 **(100)** 124 **(152)** 191 **(203)** 208 **(246)** 250 276 376 **(392)** **(410)** **(428)** **(440)** 484 **(554)** 578 **(608)** 643 703 705 **(710)** **(752)** **(756)** **(792)** **(826)** **(840)**
Loadings	
Q12a	33 40 56 100 124 152 191 203 208 246 250 276 376 392 410 428 440 484 554 578 608 643 703 705 710 752 756 792 826 840
Q12b	33 40 56 100 124 152 191 **(203)** 208 246 250 **(276)** 376 392 410 428 440 484 554 **(578)** 608 643 703 705 **(710)** 752 756 792 826 840
Q12c	33 40 **(56)** 100 124 152 191 203 208 246 250 276 376 392 **(410)** 428 440 484 554 578 608 643 703 705 710 752 **(756)** 792 826 840

Noninvariant parameters are in bold and within parentheses.

number of iterations was doubled, which was not the case for the models with other prior settings.

To select the model(s) with a good fit, one could use model fit indices, but just as with regular SEM there is not one single statistic that should be used, and only a combination of fit indices should be used to indicate model fit. **Table 6** shows fit statistics for the models with prior variances 0.02, 0.01, 0.001, 0.0005, and 0.0001. **Table 6** shows that only for models with a prior variance of 0.02 and 0.01 the PPP $>$ 0 (0.36 and 0.12, respectively) and the 95% CI contains 0. BRMSEA is 0.014 for the model with a prior variance of 0.02, and it is 0.049 for the model with prior variance of 0.01. For the other models, BRMSEA $>$ 0.1. For the model with prior variance 0.02, both BCFI and BTLI are 1.00, and for the model with prior variance 0.01 BCFI is 0.999 and BTLI is 0.994, indicating good model fit. Using PPP-value, the model with variance 0.02 comes closest to 0.5 with a PPP-value of 0.36. However, the PPP-value might be untrustworthy because of the large sample size of our study (24,583 respondents) (Asparouhov and Muthén, 2019). Hoijtink and van de Schoot (2018) stated that the PPP-value is not suitable for evaluating small priors. Concerning both CI and BRMSEA, only the models with prior variances of 0.02 and 0.01 indicate a good fit. When looking at BCFI and BTLI, however, the models with prior variances of 0.02, 0.01, 0.001, and 0.005 all indicate good fit, although fit statistics approach to their cutoff values as prior variances decline. When combining the above results with the DIC for the different models, the values for the model with prior variance 0.02 is the lowest (19,7428.86), indicating that this is the best fitting model based on *post-hoc* fit indices.

Figure 1 shows the means of the latent variable for the BAMI models with variances of 0.02, 0.01, 0.001, 0.0005, and 0.0001. From this figure it can be seen that with declining prior variance, the outcome of the model approaches that of the scalar model (on the right). This is to be expected, as the scalar model is a model of priors with a mean and variance of 0.

Figure 2 is a graph of the means per country per model (scalar invariance, ML alignment, and all BAMI models). For illustrative purposes, we present the results for BAMI both with and without alignment in one figure. **Figure 1** shows that the overall mean differences between latent means of the different models are small but increase as the prior variance decreases: $\Delta_{0.01-0.02}$ is 0.007 and $\Delta_{0.0005-0.0001}$ is 0.069. For individual countries, this is not always the case: **Figure 2** shows that, for the 15 lowest ranking countries this same pattern is visible, but for the top 15 countries the means increase with prior variance. However, as the countries rank lower, the differences between models increase. For the lowest-ranking country (Latvia) the difference between the model with prior variance 0.02 and that with prior variance 0.0001 is 0.616, while for the highest-ranking country (Switzerland) the difference is 0.208. For the three highest-ranking countries (Switzerland, South Korea and Denmark) the model with the highest prior variance (0.02) shows larger differences from the model with a prior variance of 0.01 (0.173, 0.161, and 0.128, respectively) than do other models with consecutively lower prior variances.

BAMI With Alignment

When the BAMI model with alignment is applied, first, the BAMI model is estimated. The outcome of the BAMI models is given in the description above (**Table 6**). From this BAMI model, a configural model is estimated, which is then aligned. This means that fit indices cannot be used to indicate model fit of the final

TABLE 6 | Fit statistics of the BAMI models.

Prior variance	PPP	95% CI	BRMSEA	BCFI	BTLI	BIC	DIC
0.02	0.363	[−52.706 to 79.836]	0.014	1.000	1.000	200288.68	197428.66
0.01	0.117	[−25.495 to 109.968]	0.049	0.999	0.994	200324.54	197514.18
0.001	0.000	[669.517 to 852.222]	0.117	0.976	0.968	201095.93	198186.01
0.0005	0.000	[1100.629 to 1286.848]	0.130	0.962	0.960	201546.82	198601.39
0.0001	0.000	[1873.721 to 2022.907]	0.148	0.938	0.949	202327.90	199334.78

PPP, posterior predictive probability; CI, credibility intervals; BRMSEA, Bayesian root mean square error of approximation; BCFI, Bayesian comparative fit index; BTLI, Bayesian Tucker-Lewis index; BIC, Bayesian information criterion; DIC, deviance information criterion.

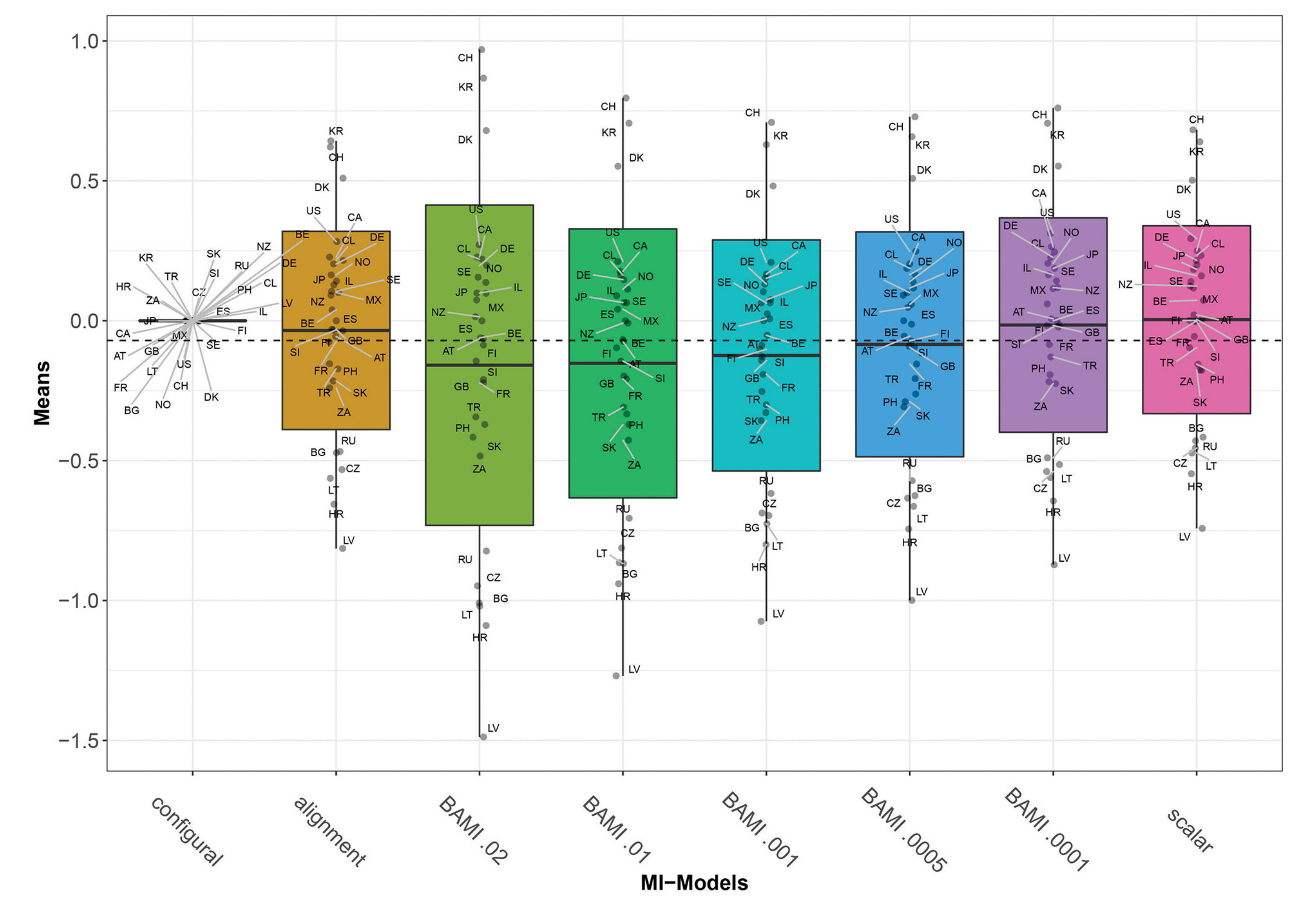


FIGURE 1 | Means for configural invariance, scalar invariance, ML alignment, and BAMI models. AT, Austria; BE, Belgium; BG, Bulgaria; CA, Canada; CL, Chile; HR, Croatia; CZ, Czech Republic; DK, Denmark; FI, Finland; FR, France; DE, Germany; GB, Great Britain; IL, Israel; JP, Japan; LV, Latvia; LT, Lithuania; MX, Mexico; NZ, New Zealand; NO, Norway; PH, Philippines; RU, Russia; SK, Slovakia; SI, Slovenia; ZA, South Africa; KR, South Korea; ES, Spain; SE, Sweden; CH, Switzerland; TR, Turkey; US, United States. The x-axis shows the different models—configural, scalar, ML alignment, and BAMI—with their specific variances. The dashed black line shows the overall mean.

model. Instead, just as with the ML alignment model, we use the percentage of non-invariant parameters to determine good model fit. **Table 7** shows the number of non-invariant parameters per model.

From this table, it can be seen that, as prior variances decrease, so does the number of non-invariant parameters. The three models with prior variances of 0.02, 0.01, and 0.001 all have a percentage of non-invariant parameters above 25% (although the model with prior variance 0.001 is only slightly above), making the results, and thus a group ranking from these models, unreliable. For the models with prior variances of 0.0005 and 0.0001 the percentages of non-invariant groups are 16.67 and 1.67, respectively, implying good model fit and a valid group ranking.

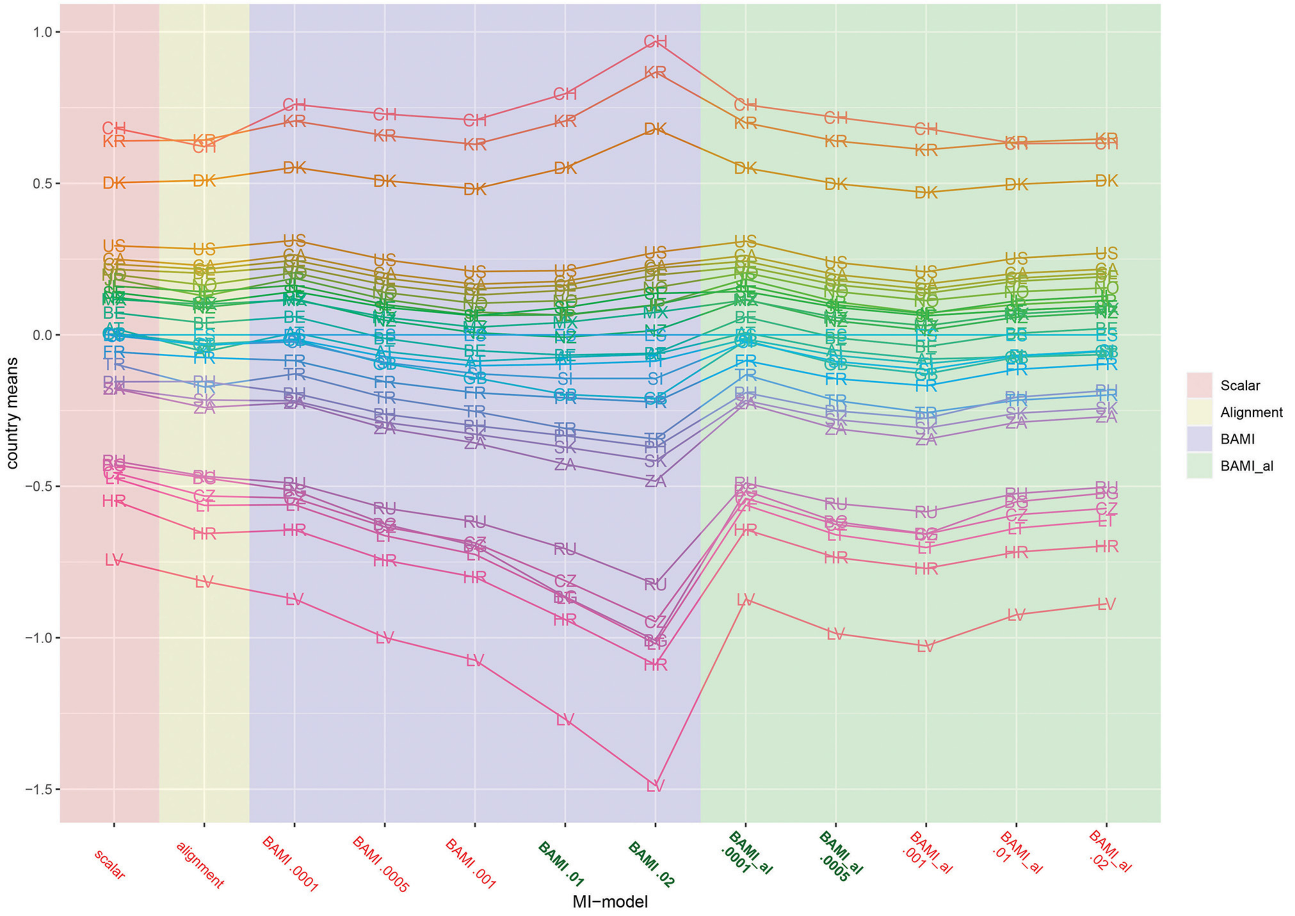


FIGURE 2 | Means per country for scalar invariance, ML alignment, and BAMI models with and without alignment. AT, Austria; BE, Belgium; BG, Bulgaria; CA, Canada; CL, Chile; HR, Croatia; CZ, Czech Republic; DK, Denmark; FI, Finland; FR, France; DE, Germany; GB, Great Britain; IL, Israel; JP, Japan; LV, Latvia; LT, Lithuania; MX, Mexico; NZ, New Zealand; NO, Norway; PH, Philippines; RU, Russia; SK, Slovakia; SI, Slovenia; ZA, South Africa; KR, South Korea; ES, Spain; SE, Sweden; CH, Switzerland; TR, Turkey; US, United States. The x-axis shows the different models—scalar, ML alignment, and BAMI with and without alignment—with their specific variances. Models that appear to ba a good fit to the data are indicated in bold green, models with bad fit in red.

TABLE 7 | The number of non-invariant parameters for the BAMI models with alignment.

Prior Variance	Number of non-invariant intercepts			Number of non-invariant loadings			Total number	
	Q12a	Q12B	Q12C	Q12a	Q12B	Q12C	Sum	%
0.02	15	16	19	2	5	5	62	34.44
0.01	15	15	19	0	5	5	59	32.78
0.001	9	15	16	0	2	4	46	25.56
0.0005	4	11	12	0	1	2	30	16.67
0.0001	0	3	0	0	0	0	3	1.67

As with the ML alignment model, most non-invariant parameters are the intercept parameters. For the model with the lowest number of non-invariant parameters (prior variance 0.0001), these parameters belong to the intercepts of question 12b (are you willing to pay higher taxes to save the environment), for the countries Lithuania, South Africa, and Turkey. When taking into account only the models with a percentage of non-invariant parameters below 25%, there are 12 countries for which all parameters are invariant for both models: Spain, Austria, Canada, Denmark, Finland, Germany, Israel, Mexico, Russia, Slovakia, Slovenia, and Sweden. **Figures 2**, **3** show that, as the priors decrease, so do the mean differences of the model outcomes (both the overall means and the means per country).

Figure 3 shows that for the first three models with decreasing prior variance, the overall means also decrease. However, as prior variances decrease further (0.0005 and 0.0001), they rise

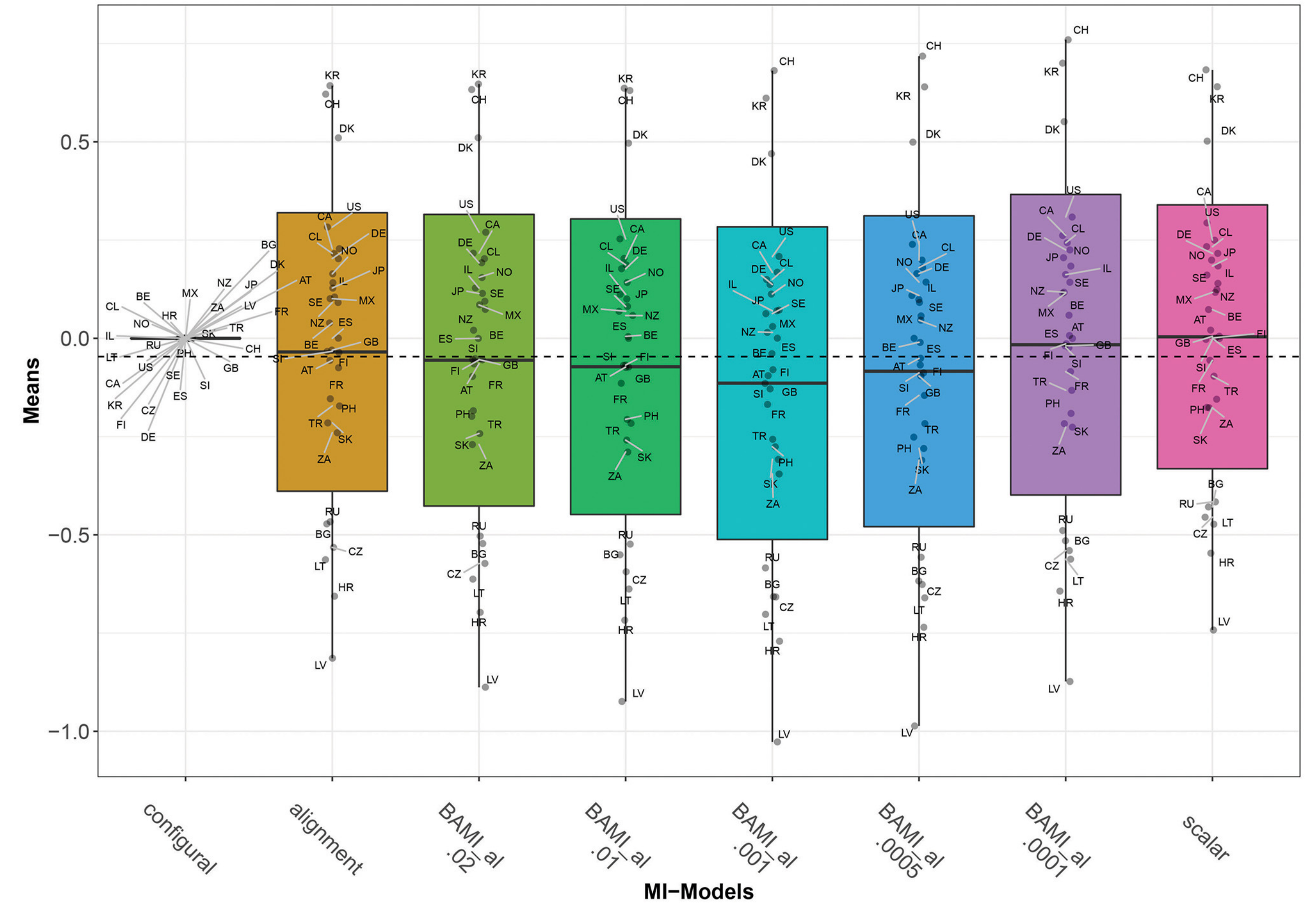


FIGURE 3 | Means for configural invariance, scalar invariance, ML alignment, and BAMI models with alignment. AT, Austria; BE, Belgium; BG, Bulgaria; CA, Canada; CL, Chile; HR, Croatia; CZ, Czech Republic; DK, Denmark; FI, Finland; FR, France; DE, Germany; GB, Great Britain; IL, Israel; JP, Japan; LV, Latvia; LT, Lithuania; MX, Mexico; NZ, New Zealand; NO, Norway; PH, Philippines; RU, Russia; SK, Slovakia; SI, Slovenia; ZA, South Africa; KR, South Korea; ES, Spain; SE, Sweden; CH, Switzerland; TR, Turkey; US, United States. The x-axis shows the different models—configural, scalar, ML alignment, and BAMI with alignment—with their specific variances. The dashed black line shows the overall mean.

slowly toward the means of the scalar model. The differences between means of models with consecutive priors are less clear than for the BAMI models (**Figure 1**). Now, $\Delta_{0.01-0.02}$ 0.016, $\Delta_{0.001-0.01}$ 0.042, $\Delta_{0.0005-0.0001}$ 0.0675, and $\Delta_{0.001-0.0005}$ 0.031. Although this pattern is visible in the means per model (**Figure 3**), it is less distinctive when looking at the means of individual countries (**Figure 2**). In that case, this pattern is most pronounced for Latvia and, to a lesser extent, Bulgaria, Lithuania, Hungary, Russia, South Africa, Slovakia, Turkey, the Philippines, Denmark, Croatia, and Switzerland. **Figures 2**, **3** show that, as prior variances decrease, so do the mean differences of the model outcomes (both the overall means and the means per country). Again, Latvia is the country with the most pronounced differences when comparing different priors.

Ranking

Figures 1–3 show that the latent means vary depending on the choice of prior variance. Models with smaller prior variances seem to have outcomes that approach the outcome of the scalar model. However, there is some variation at the country level. From **Figure 2**, we observe that there appear to be four different groups of countries with similar means: Switzerland, South Korea and Denmark at the top, then a large group with the United States, Canada, Chile, Germany, Norway, Japan, Israel, Sweden, New Zealand, Mexico, Austria, Great Britain, Finland, Spain, Slovenia, France, Turkey, Philippines, Slovakia, and South Africa. The third group comprises Russia, Czech Republic, Bulgaria, Lithuania, and Croatia, while the bottom group consists of only one country: Latvia. In particular for the second group, means are very close together, and it can be difficult to distinguish individual country means. **Figure 4** shows the ranking of the 30 countries for the analyzed models that converged. This figure shows that for nearly all the models, ranking changes somewhat when a different prior variance is used. Upon closer inspection, 13 of the 30 countries occupy the same place in the ranking for all the models, and most changes appear to be in the middle of the ranking. When combined with **Figure 2**, it becomes clear that country mean differences are small, especially for the BAMI model with alignment. For the BAMI models, country means differ slightly more, especially at the top and the bottom of the ranking. Mean differences for Latvia decrease with decreasing prior variance, and these differences are larger than the overall mean difference per model.

Comparing the individual country means of the BAMI and BAMI with alignment model shows that for all countries the differences between the models decrease with prior variance: differences between models are lowest when models with the lowest prior variances are compared. For the models with a prior variance of 0.0001 the country rankings are the same.

Focusing on only the models that appear to have a good fit to the data, according to their fit statistics, only the BAMI models with a prior variance of 0.02 and 0.01 and BAMI models with alignment with a prior variance of 0.0005 and 0.0001 are of importance. When comparing the rankings of these models, the ranking for the BAMI models is almost identical: only Spain and New Zealand switch places when changing models. The ranking of the BAMI models with alignment shows more variation: 23 countries rank the same for both models, while Mexico, New Zealand, Belgium, Austria, Great Britain, and Slovenia all shift one place up or down and Spain moves two places in the ranking.

These figures show that, regardless of prior variance or even model fit, people in Switzerland and South Korea are most motivated to sacrifice for the environment, while people in Bulgaria and Latvia are less motivated to sacrifice for the environment.

Decision Tree

As it can be difficult to draw conclusions from the means and rankings as shown in **Figures 1–4**, we devised a decision tree (**Figure 5**). This tree provides some insight into the decisions that we had to make regarding group means, group rankings and the influence of priors. Based on this decision tree, other readers might come to different conclusions. The tree comprises the entire process needed to evaluate the information contained in **Figures 1–4**, starting with the MGCFA test for scalar invariance:

1. We started with an ML MGCFA test for scalar invariance. To test for scalar invariance it is necessary that configural and metric invariance are met.
   a. Yes: It is now possible to compare ranks.
   b. No: Try another method to make means comparison valid. Go to step 2.
   We did not find scalar invariance, so we followed the "no" arrow to step 2.

2. Do you expect a large difference in parameters for some groups and equality for the rest of the groups[5]?
   a. Yes: Equality for almost all groups. Go to step 3.
   b. No: Only small differences or small and large differences. Go to step 4.
   We assumed there would be some differences in the parameters, although we did not know how large these differences would be or how many groups would differ from each other, so we first followed the "yes" arrow to step 3 and tested for ML alignment.

3. Does the alignment optimization yield $<$ 25% non-invariant parameters?
   a. Yes: The rank order can be trusted
   b. No: Try another method to make means comparison valid. Go top step 4.
   This yielded $>$ 25% non-invariant parameters, so we followed the "no" arrow to step 4.

4. Do you expect only small differences in parameters for different groups[6]?
   a. Yes: Only small differences. Use BAMI, Go to step 4A.
   b. No: Both small and large differences. Use BAMI in combination with alignment optimization. Go to step 4B.
   Now, we had to decide whether we expected small or large differences in parameters for different groups. Since we did not know how large the differences per group were, we used both, as they lead to step 5 in this decision tree; which option we chose would not make a difference.

5. Then, we needed to decide which priors to use and run the different models. We based our priors on previous literature on the use of BAMI (both simulations and empirical examples). We then moved to step 6.
6. We visualized the outcomes of the different models [scalar, ML alignment, and BAMI models (with and without alignment) that converged] in **Figures 1–4** (means and group rankings). The code to create these Figures can be found in Appendix C on Arts et al. (2021). We moved to step 7.
7. Is the rank order **as a whole** stable? (**Figure 4**)
   a. Yes: No or only minor changes in rank. Then the rank order is not at all or only slightly influenced by the choice of priors.
   b. No: Many changes across groups and models. Go to step 8.
   Since there were numerous changes in the rank order, we did not consider the rank order stable and we followed the "no" arrow to step 8.

8. Does the pattern of the rank order across different models make sense?
   a. Yes: Many changes, but all changes in the same section of the ranking (e.g., top) or the same groups change rank. Go to step 9.
   b. No: The pattern seems erratic. Go to step 10.
   From **Figure 5** we concluded that the upper and the lower parts of different rankings hardly change, and most rank changes take place in the middle part of the ranking. We considered this a logical pattern and followed the "yes" arrow.

[5] When in doubt whether large difference in parameters for some groups are to be expected, it is advisable to consult a substantive expert of your field. The decision whether large parameters can be expected should be based on previous research and/or expertise.

[6] Similarly, when in doubt whether small differences for many groups are to be expected, it is advisable to consult a substantive expert of your field. The decision whether small parameter differences can be expected should be based on previous research and/or expertise.

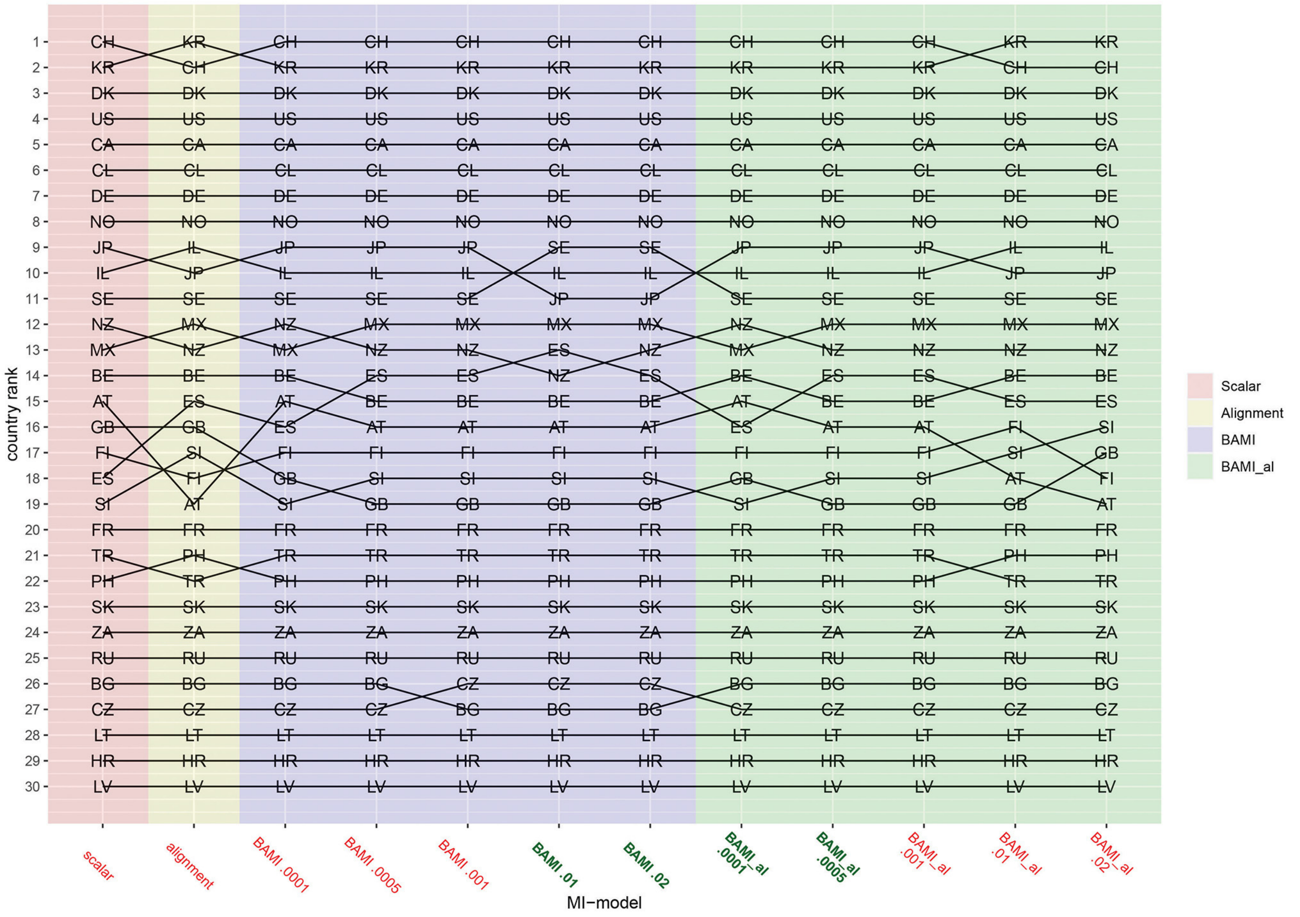


FIGURE 4 | Rankings per model. AT, Austria; BE, Belgium; BG, Bulgaria; CA, Canada; CL, Chile; HR, Croatia; CZ, Czech Republic; DK, Denmark; FI, Finland; FR, France; DE, Germany; GB, Great Britain; IL, Israel; JP, Japan; LV, Latvia; LT, Lithuania; MX, Mexico; NZ, New Zealand; NO, Norway; PH, Philippines; RU, Russia; SK, Slovakia; SI, Slovenia; ZA, South Africa; KR, South Korea; ES, Spain; SE, Sweden; CH, Switzerland; TR, Turkey; US, United States. The x-axis shows the different models—scalar, ML alignment, and BAMI with and without alignment—with their specific variances. Models that appear to ba a good fit to the data are indicated in bold green, models with bad fit in red.

9. Are individual groups stable across models?
   a. Yes: Individual groups never move more than one place up or down in the ranking across different models. Then the rank order is not or only slightly influenced by the choice of priors
   b. No: Individual groups continue moving up or down the ranking across different models.

   Changes in rank nearly always applied to the same countries, making the pattern rather stable. However, as some countries moved up or down two or three positions in the ranking across models, we found that stability of the groups could not be guaranteed. We followed the "no" arrow.

10. Are the mean differences per group per model small?
    a. Yes. There is almost no difference between groups, and the influence of the priors is small.
    b. No: Do not use rank order.

    Figure 3 shows that the differences per group are quite small, especially in the middle part of the ranking where most changes in rank take place. We therefore conclude that there is almost no difference between groups.

CONCLUSION AND DISCUSSION

The latent variable "willingness to sacrifice for the environment" (WTS) is an important aspect of environmental concern. It can provide insights into the intentional behavior regarding environmental concern, which, in turn, provides more insight into the willingness of the respondents to take action to protect the environment. Given that country rankings of latent means of WTS are frequently used in comparative studies, it is important to assess whether substantive findings are indeed trustworthy or are methodological artifacts due to lack of metric or scalar

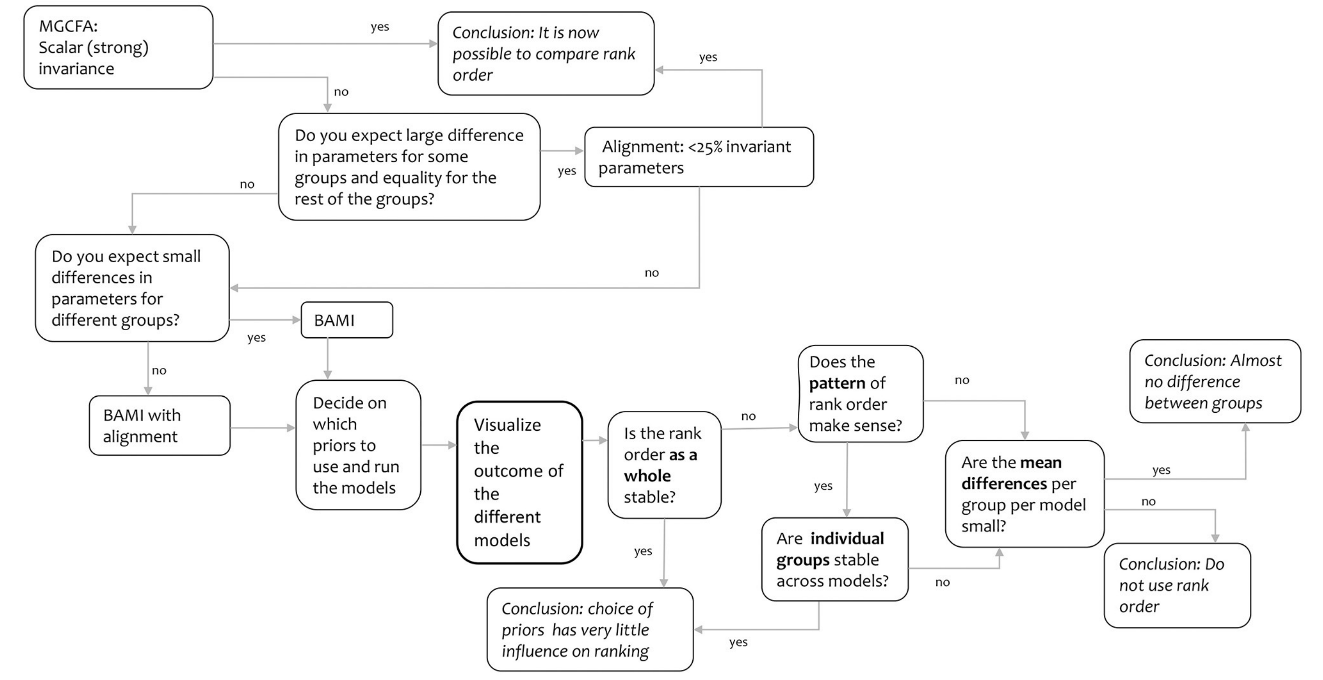

MGCFA:
Scalar (strong)
invariance
yes
no
Conclusion: It is now possible to compare rank order
yes
Do you expect large difference in parameters for some groups and equality for the rest of the groups?
yes
Alignment: <25% invariant parameters
no
Do you expect small differences in parameters for different groups?
no
yes
BAMI
no
BAMI with alignment
Decide on which priors to use and run the models
Visualize the outcome of the different models
Is the rank order as a whole stable?
no
Does the pattern of rank order make sense?
no
yes
Conclusion: Almost no difference between groups
Are the mean differences per group per model small?
yes
no
Are individual groups stable across models?
no
Conclusion: Do not use rank order
yes
Conclusion: choice of priors has very little influence on ranking
yes

FIGURE 5 | Decision tree.

invariance. The latent variable WTS was, in combination with the latent variable environmental attitude (EA), previously tested for MI by Mayerl and Best (2019). Using MGCFA, they did not find scalar invariance, questioning comparisons of the latent means across countries. However, recent discussions in MI point out that the approach of ML MGCFA may be too strict, and approaches such as alignment or BAMI, or a combination of both, may be a viable solution when exact scalar invariance tests fail (e.g., van de Schoot et al., 2013; Asparouhov and Muthén, 2014). In this article, we examined WTS in 30 different countries, using the 2010 ISSP data. We did not establish scalar invariance when using MGCFA, which is in line with the findings of Mayerl and Best (2019). In addition to MGCFA, we also assessed MI using ML alignment, BAMI and BAMI with alignment method.

Based on our results, we can determine which countries consistently rank high on the latent variable WTS (Switzerland and South Korea) and which countries consistently rank low (Latvia). However, we cannot say that, e.g., respondents in Sweden are more or less willing to sacrifice for the environment than respondents in Mexico. Thus, a more general conclusion about these country rankings can be drawn (high, low), but when exact ranking (e.g., fourth or fifth), or even exact means, are important, these country rankings should not be used. In conclusion, only with BAMI plus alignment optimization we were able to obtain stable results. From these, we can conclude that people in Switzerland and South Korea are most motivated to sacrifice for the environment, while people in Latvia are less motivated to sacrifice for the environment.

Regarding the use of different prior variances when using the BAMI method, models with a prior variance of 0.02 and 0.01 showed good model fit for most fit statistics (PPP, 95% CI, BRMSEA, BCFI, BTLI). For the model with variances of 0.001 and 0.0005 BCFI and BTLI were within limits. When taking into account that PPP might incorrectly identify model fit for models with large sample sizes (van de Schoot et al., 2012; Mulder, 2014), the results of BAMI models with a variance of 0.001 and 0.0005 might still fit the data. BIC and DIC are lowest for the BAMI model with a prior variance of 0.02, but the use of DIC for models with small prior variances has been disputed by Hoijtink and van de Schoot (2018). For the BAMI models with alignment the models with the smallest prior variances (0.0005 and 0.0001) give trustworthy results with a percentage of non-invariant parameters of 16.67 and 1.67%, respectively. This indicates that the BAMI models with a prior variance of 0.02 and 0.01 are a good fit to the data, while for the BAMI with alignment models, the models with a prior variance of 0.0001 and 0.0005 give trustworthy results.

Concerning comparing means of the BAMI models with different prior variances, both with and without alignment, the means are very similar for the models with prior variances of 0.0001 (difference of the overall mean per model is 0.001). These country rankings are, with the exception of Great Britain and Spain (rank 16 and 18, respectively) also the same for the scalar model. However, the scalar model and the BAMI model with a prior variance of 0.0001 cannot be assumed to be a good fit to the data (see **Tables 4**, **6**), while the BAMI model with alignment with the same prior can. When comparing two models that indicate reliable outcomes—the BAMI with prior variance of 0.02 and the BAMI with alignment model with a prior variance of 0.0001—differences per country are much larger (ranging from 0.616 to

0.029, with the exception of 0 for Spain), thus indicating that prior variances can have a large influence on model outcomes, and that the model results that appear to be reliable, can be very close or even equal to the results of a model that should be rejected. This also shows the difficulty of comparing BAMI models to BAMI models with alignment: because the ground on which models should be rejected are very different, it is difficult to say which model should be preferred, if any.

It is our opinion that visualizing the results facilitates determining of the effect of different prior variances. A visual presentation of the results could be a valuable addition to the presentation of results in elaborated tables that can be challenging to interpret, especially when many groups are compared. The visualization approach that we used in this article is, however, not the only possibility to visualize (MI) results. Depending on e.g., research question, group size, and personal preferences, the researcher can choose other ways to visualize the results. For example, van de Schoot et al. (2015) chose to display the effect of different prior variances on the differences between groups in several line charts, while Pokropek et al. (2020) decided to use color-coded tables to identify the most suitable fit statistic to identify the optimal prior, Zercher et al. (2015) used scatter plots to represent latent means per country, and van de Vijver et al. (2019) used a 3-dimensional plot showing the Euclidean distances between different groups. However, most researchers still use (mainly) tables to present their results (e.g., Chiorri et al., 2014; De Bondt and Van Petegem, 2015; Gucciardi et al., 2016; Seddig and Leitgöb, 2018; Solstad et al., 2020; Vilar, 2020). We propose that a visual presentation of the results can improve the comprehension of test results, and can serve as a useful addition to previous presentations of results. This can be particularly useful when a researcher is faced with contradictory priors: a visualization of model outcomes with these different priors immediately shows the effect that the priors might or might not have. In particular researchers who are less familiar with the subject of Bayesian modeling might benefit from such a visual presentation.

Limitations

This study has some limitations, that need to be mentioned. First, the latent variable WTS is linked to intentional behavior, but intentional behavior alone cannot explain environmental concern. This one-scale, three-item model is an oversimplification of real-world data: multiple latent variables are required to provide insight into environmental concern. If MI holds for WTS in combination with other latent variables (e.g., EA) a country ranking would be more meaningful when determining nationwide environmental concern. Also, WTS is mainly financially driven (two out of three question refer to paying to protect the environment: see **Table 1**). This would mean that respondents who cannot or do not want to contribute financially but are willing to contribute in some other way affect this latent variable differently than those who are willing to sacrifice financially. Second, in the analysis of the different methods, most settings were Mplus default settings. Using different settings might lead to different outcomes: e.g., a different simplicity function when using alignment could affect model outcome when the alignment optimization is used. It is also possible to choose Bayes as an estimator instead of ML. In that case, analysis starts with the same M_0 model as for ML alignment, and then loadings and intercepts are estimated using noninformative priors using Equation (3) (Asparouhov and Muthén, 2014). For other Mplus settings see Muthén and Muthén (2019). However, testing the many different settings in Mplus is beyond the scope of this article. Third, we used a decision tree (**Figure 5**) to interpret the results based on a visual inspection of country means and rankings. In this decision tree, several choices have to be made based on the ranking order and pattern (**Figure 4**). This pattern might not be interpreted by everyone in the same way: at Step 7 (is the rank order as a whole stable), Step 8 (does the pattern of the rank order make sense), and Step 9 (are individual groups stable across models) the reader has to decide whether a pattern is stable, the rank order makes sense, and individual groups are stable across models. It is also up to the reader to decide if the differences between group means are small or large (step 10). So, depending on the reader, conclusions might be different. However, we argue that using a decision tree always involves arbitrary decisions, like non-testable identification constraints, see for a discussion Little et al. (2006). We also believe that, in this case, the benefits of a decision tree (transparency to the workflow) outweigh the disadvantages.

Future Research

In this article we show that, for WTS, MI is present, making a ranking of countries possible. We also show that country means are not independent of specified priors and that, although the differences are small, an exact country ranking cannot be assumed. A combination of multiple latent variables (EA, knowledge of environmental concern, risk perception) might provide more insight into environmental concern. This would complicate model specification and analysis somewhat, since it would, e.g., make the use of priors on cross-loadings an important part of the model (Xiao et al., 2019; Liang et al., 2020). Another potential promising area of future study is the method of Robitzsch (2020), which improves the alignment optimization. A comparison with BAMI has not been made, yet. Future research may provide more insight into WTS and the topic of environmental concern. Looking at **Figure 2**, we see that four groups of countries have very similar means. It would be interesting to further investigate why this division into four groups appears. Is this also the case when other latent variables are investigated (separately or combined with WTS)? Is it purely data driven or are there underlying reasons that can explain these four groups (psychological, sociological, political, economic, etc.)? A multilevel model that includes such factors, might shed more light on why these four groups exist. The prior variances that we used in this article were based on previous literature on prior selection (both simulation and empirical studies). Another approach to selecting priors could be to consult literature on environmental change to determine the factors that drive environmental concern. Priors could then be based on e.g., socioeconomic status of a country, geography (and thus the environmental threat a particular country faces) or political system and stability. Multilevel tests aiming to unravel these factors have been conducted by e.g., Marquart-Pyatt (2012a), Fairbrother (2013), Pampel (2014), and Pisano and Lubell (2017), but, to our knowledge, such factors have not yet been included in a Bayesian model. However, the use of such priors would emphasize the strength of Bayesian modeling.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found at: doi: 10.4232/1.13271. All appendices, the scripts to reproduce our results, the final output files and additional material can be found on website of the Open Science Framework (OSF): https://osf.io/mvyws/ (Arts et al., 2021).

AUTHOR CONTRIBUTIONS

IA, RvdS, and KM designed the study and wrote the article. IA conducted the analyses. QF provided the necessary code for visualization. All authors contributed to the article and approved the submitted version.

FUNDING

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

REFERENCES

Amérigo, M., García, J. A., Pérez-López, R., Cassullo, G., Ramos, A., Kalyan-Venumbaka, S., et al. (2020). Analysis of the structure and factorial invariance of the Multidimensional Environmental Concern Scale (MECS). *Psicothema* 32, 275–283. doi: 10.7334/psicothema2019.281

Andonova, L., and Coetzee, K. (2020). "Chapter 12: Does successful emissions reduction lie in the had of non-state rather than stat actors?," in *Contemporary Climate Change, 1st Edn*, ed M. Hulme (New York, NY: Routledge), 177–190. doi: 10.4324/9780429446252-13

Arts, I., Fang, Q., Meitinger, M., van de Schoot, R. (2021). *Approximate Measurement Invariance of Willingness to Sacrifice for the Environment Across 30 Countries: the Importance of Prior Distributions and Their Visualization*. OSF. doi: 10.17605/OSF.IO/MVYWS

Ashby, D. (2006). Bayesian statistics in medicine: a 25 year review. *Stat. Med.* 25, 3589–3631. doi: 10.1002/sim.2672

Asparouhov, T., and Muthén, B. (2010). *Bayesian Analysis of Latent Variables Models Using Mplus*. Mplus Web Notes, 1–60.

Asparouhov, T., and Muthén, B. (2014). Multiple-group factor analysis alignment. *Struct. Equat. Model.* 21, 495–508. doi: 10.1080/10705511.2014.919210

Asparouhov, T., and Muthén, B. (2017). *Prior-Posterior Predictive P-Values*. Mplus Web Notes, 1–16.

Asparouhov, T., and Muthén, B. (2019). *Advances in Bayesian Model Fit Evaluation for Structural Equation Models*. Mplus Web Notes, 1–36. doi: 10.1080/10705511.2020.1764360

Asparouhov, T., Muthén, B., and Morin, A. (2015). Bayesian structural equation modeling with cross-loadings and residual covariances: comments on Stromeyer et al. *J. Manage.* 41, 1561–1577. doi: 10.1177/0149206315591075

Bamberg, S. (2003). How does environmental concern influence specific environmentally related behaviors? A new answer to an old question. *J. Environ. Psychol.* 23, 21–32. doi: 10.1016/S0272-4944(02)00078-6

Bozonnet, J. P. (2016). "Explaining environmental activism by national cultures: the hypothesis of hysteresis," in *Green European: Environmental Behaviour and Attitudes in Europe in a Historical and Cross-Cultural Comparative Perspective* (New York, NY), 91–110.

Byrne, B., and van de Vijver, F. (2017). The maximum likelihood alignment approach to testing for approximate measurement invariance. *Psicothema* 29, 539–551. doi: 10.7334/psicothema2017.178

Chen, F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Struct. Equat. Model.* 14, 464–504. doi: 10.1080/10705510701301834

Chiorri, C., Day, T., and Malmberg, L. (2014). An approximate measurement invariance approach to within-couple relationship quality. *Front. Psychol.* 5:983. doi: 10.3389/fpsyg.2014.00983

De Bondt, N., and Van Petegem, P. (2015). Psychometric evaluation of the overexcitability questionnaire-two applying Bayesian Structural Equation Modeling (BSEM) and multiple-group BSEM-based alignment with approximate measurement invariance. *Front. Psychol.* 6:1963. doi: 10.3389/fpsyg.2015.01963

Depaoli, S., and van de Schoot, R. (2017). Improving transparency and replication in Bayesian statistics: the WAMBS-checklist. *Psychol. Methods* 22, 240–261. doi: 10.1037/met0000065

Dunlap, R. E., and Jones, R. E. (2002). "Chapter 15: Environmental concern: conceptual and measurement issues," in *Handbook of Environmental Sociology, 1st Edn*, eds R. E. Dunlap and W. Michelson (Westport, WA: Abc-clio), 482–524.

Fairbrother, M. (2013). Rich people, poor people, and environmental concern: evidence across nations and time. *Eur. Sociol. Rev.* 29, 910–922. doi: 10.1093/esr/jcs068

Flake, J., and McCoach, D. (2018). An Investigation of the alignment method with polytomous indicators under conditions of partial measurement invariance. *Struct. Equat. Model.* 25, 56–70. doi: 10.1080/10705511.2017.1374187

Fox, J. P., and Verhagen, J. (2018). "Random item effects modeling for cross-national survey data," in *Cross-Cultural Analysis: Methods and Applications, 2nd Edn*, eds E. Davidov, P. Schmidt, J. Billiet, and B. Meuleman (New York, NY: Routledge), 529–550. doi: 10.4324/9781315537078-19

Franzen, A., and Meyer, R. (2010). Environmental attitudes in cross-national perspective: a multilevel analysis of the ISSP 1993 and 2000. *Eur. Sociol. Rev.* 26, 219–234. doi: 10.1093/esr/jcp018

Franzen, A., and Vogl, D. (2013). Acquiescence and the willingness to pay for environmental protection: a comparison of the ISSP, WVS, and EVS. *Soc. Sci. Q.* 94, 637–659. doi: 10.1111/j.1540-6237.2012.00903.x

Gallagher, M. W., and Brown, T. A. (2013). "Introduction to confirmatory factor analysis and structural equation modeling," in *Handbook of Quantitative Methods for Educational Research*, ed T. Teo (Rotterdam: Sense Publishers), 289–314. doi: 10.1007/978-94-6209-404-8

Garnier-Villarreal, M., and Jorgensen, T. (2020). Adapting fit indices for bayesian structural equation modeling: comparison to maximum likelihood. *Psychol. Methods* 25, 46–70. doi: 10.1037/met0000224

GESIS (2019). *ISSP 2010 Environment III, Variable Report: Documentation release 2019/06/13, related to the international dataset Archive-Study-No. ZA5500 Version 3.0.0, Variable Reports 2019|05*. Technical report, GESIS, Cologne.

Gucciardi, D., Zhang, C.-Q., Ponnusamy, V., Si, G., and Stenling, A. (2016). Cross-cultural invariance of the mental toughness inventory among Australian, Chinese, and malaysian athletes: a bayesian estimation approach. *J. Sport Rehabil.* 38, 187–202. doi: 10.1123/jsep.2015-0320

Hadler, M., and Kraemer, K. (2016). "Chapter 1: The perception of environmental threats in a global and European perspective," in *Green European: Environmental Behaviour and Attitudes in Europe in a Historical and Cross-Cultural Comparative Perspective*, eds A. Telesiene and M. Gross (New York, NY: Taylor & Francis), 13–30.

Hallquist, M., and Wiley, J. (2018). MplusAutomation: an R package for facilitating large-scale latent variable analyses in Mplus. *Struct. Equat. Model.* 25, 621–638. doi: 10.1080/10705511.2017.1402334

Hoijtink, H., and van de Schoot, R. (2018). Testing small variance priors using prior-posterior predictive P-values. *Psychol. Methods* 23, 561–569. doi: 10.1037/met0000131

Hoofs, H., van de Schoot, R., Jansen, N., and Kant, I. (2018). Evaluating model fit in Bayesian confirmatory factor analysis with large samples: simulation study introducing the BRMSEA. *Educ. Psychol. Measure*. 78, 537–568. doi: 10.1177/0013164417709314

Hu, L., and Bentler, P. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Struct. Equat. Model*. 6, 1–55. doi: 10.1080/10705519909540118

ISSP Research Group (2019). *International Social Survey Programme: Environment III - ISSP 2010*. ISSP Research Group.

Ivanova, G., and Tranter, B. (2008). Paying for environmental protection in a cross- national perspective. *Austr. J. Polit. Sci*. 43, 169–188. doi: 10.1080/10361140802035705

Jennrich, R. I. (2006). Rotation to simple loadings using component loss functions: the oblique case. *Psychometrika* 71, 173–191. doi: 10.1007/s11336-003-1136-B

Jöreskog, K. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika* 36, 109–133. doi: 10.1007/BF02291393

Kim, E. S., Cao, C., Wang, Y., and Nguyen, D. T. (2017). Measurement invariance testing with many groups: a comparison of five approaches. *Struct. Equat. Model*. 24, 524–544. doi: 10.1080/10705511.2017.1304822

Knight, K. W. (2016). Public awareness and perception of climate change: a quantitative cross-national study. *Environ. Sociol*. 2, 101–113. doi: 10.1080/23251042.2015.1128055

König, C., and van de Schoot, R. (2018). Bayesian statistics in educational research: a look at the current state of affairs. *Educ. Rev*. 70, 486–509. doi: 10.1080/00131911.2017.1350636

Kruschke, J. K., Aguinis, H., and Joo, H. (2012). The time has come: bayesian methods for data analysis in the organizational sciences. *Organ. Res. Methods* 15, 722–752. doi: 10.1177/1094428112457829

Lai, K., and Green, S. B. (2016). The problem with having two watches: assessment of fit when RMSEA and CFI disagree. *Multivar. Behav. Res*. 51, 220–239. doi: 10.1080/00273171.2015.1134306

Lek, K., Oberski, D., Davidov, E., Cieciuch, J., Seddig, D., and Schmidt, P. (2018). "Approximate measurement invariance," in *Advances in Comparative Survey Methods: Multinational, Multiregional, and Multicultural Contexts (3MC)*, eds T. P. Johnson, B.-E. Pennell, I. A. L. Stoop, and B. Dorer, 911–929. doi: 10.1002/9781118884997.ch41

Liang, X. (2020). Prior sensitivity in Bayesian structural equation modeling for sparse factor loading structures. *Educ. Psychol. Measure*. 80, 1025–1058. doi: 10.1177/0013164420906449

Liang, X., Yang, Y., and Cao, C. (2020). The performance of ESEM and BSEM in structural equation models with ordinal indicators. *Struct. Equat. Model*. 27, 874–877. doi: 10.1080/10705511.2020.1716770

Libarkin, J. C., Gold, A. U., Harris, S. E., McNeal, K. S., and Bowles, R. P. (2018). A new, valid measure of climate change understanding: associations with risk perception. *Clim. Change* 150, 403–416. doi: 10.1007/s10584-018-2279-y

Little, T. D., Siegers, D. W., and Card, N. A. (2006). A non-arbitrary method of identifying and scaling latent variables in SEM and MACS models. *Struct. Equat. Model*. 13, 59–72. doi: 10.1207/s15328007sem1301_3

Lommen, M. J., van de Schoot, R., and Engelhard, I. M. (2014). The experience of traumatic events disrupts the measurement invariance of a posttraumatic stress scale. *Front. Psychol*. 5:1304. doi: 10.3389/fpsyg.2014.01304

Marquart-Pyatt, S. T. (2012a). Contextual influences on environmental concerns cross-nationally: a multilevel investigation. *Soc. Sci. Res*. 41, 1085–1099. doi: 10.1016/j.ssresearch.2012.04.003

Marquart-Pyatt, S. T. (2012b). Explaining environmental activism across countries. *Soc. Nat. Resour*. 25, 683–699. doi: 10.1080/08941920.2011.625073

Marsh, H. W., Guo, J., Parker, P. D., Nagengast, B., Asparouhov, T., Muthén, B., et al. (2018). What to do when scalar invariance fails: the extended alignment method for multi-group factor analysis comparison of latent means across many groups. *Psychol. Methods* 23, 524–545. doi: 10.1037/met0000113

Mayerl, J. (2016). "Chapter 8: Environmental concern in cross-national comparison: methodological threats and measurement equivalence," in *Green European: Environmental Behaviour and Attitudes in Europe in a Historical and Cross-Cultural Comparative Perspective*, eds A. Telesiene and M. Gross (New York, NY: Taylor & Francis), 182–204.

Mayerl, J., and Best, H. (2019). Attitudes and behavioral intentions to protect the environment: how consistent is the structure of environmental concern in cross-national comparison? *Int. J. Sociol*. 49, 27–52. doi: 10.1080/00207659.2018.1560980

McCright, A. M., Dunlap, R. E., and Marquart-Pyatt, S. T. (2016). Political ideology and views about climate change in the European Union. *Environ. Polit*. 25, 338–358. doi: 10.1080/09644016.2015.1090371

Mulder, J. (2014). Prior adjusted default Bayes factors for testing (in)equality constrained hypotheses. *Comput. Stat. Data Anal*. 71, 448–463. doi: 10.1016/j.csda.2013.07.017

Munafó, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., Percie Du Sert, N., et al. (2017). A manifesto for reproducible science. *Nat. Hum. Behav*. 1, 1–9. doi: 10.1038/s41562-016-0021

Muthén, B., and Asparouhov, T. (2012). Bayesian structural equation modeling: a more flexible representation of substantive theory. *Psychol. Methods* 17, 313–335. doi: 10.1037/a0026802

Muthen, B., and Asparouhov, T. (2013). *New Methods for the Study of Measurement Invariance With Many Groups*. Mplus Web Notes, 1–60.

Muthen, B., and Asparouhov, T. (2014). IRT studies of many groups: the alignment method. *Front. Psychol*. 5:978. doi: 10.3389/fpsyg.2014.00978

Muthén, L., and Muthén, B. (2019). *MPlus User' Guide*.

Pampel, F. C. (2014). The varied influence of SES on environmental concern. *Soc. Sci. Q*. 95, 57–75. doi: 10.1111/ssqu.12045

Pisano, I., and Lubell, M. (2017). Environmental behavior in cross-national perspective: a multilevel analysis of 30 countries. *Environ. Behav*. 49, 31–58. doi: 10.1177/0013916515600494

Pokropek, A., Davidov, E., and Schmidt, P. (2019). A Monte Carlo simulation study to assess the appropriateness of traditional and newer approaches to test for measurement invariance. *Struct. Equat. Model*. 26, 724–744. doi: 10.1080/10705511.2018.1561293

Pokropek, A., Schmidt, P., and Davidov, E. (2020). Choosing priors in Bayesian measurement invariance modeling: a Monte Carlo simulation study. *Struct. Equat. Model*. 27, 750–764. doi: 10.1080/10705511.2019.1703708

Putnick, D. L., and Bornstein, M. H. (2016). Measurement invariance conventions and reporting: the state of the art and future directions for psychological research. *Dev. Rev*. 41, 71–90. doi: 10.1016/j.dr.2016.06.004

R Development Core Team (2017). *R Version 6.3.2*. R Development Core Team.

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

Robitzsch, A. (2020). LP loss functions in invariance alignment and haberman linking with few or many groups alexander. *Stats* 3, 246–283. doi: 10.3390/stats3030019

Rupp, A. A., Dey, D. K., and Zumbo, B. D. (2004). To Bayes or not to Bayes, from whether to when: applications of Bayesian methodology to modeling. *Struct. Equat. Model*. 11, 424–451. doi: 10.1207/s15328007sem1103_7

Russell, J. D., Graham, R. A., Neill, E. L., and Weems, C. F. (2016). Agreement in youth-parent perceptions of parenting behaviors: a case for testing measurement invariance in reporter discrepancy research. *J. Youth Adolesc*. 45, 2094–2107. doi: 10.1007/s10964-016-0495-1

Sara, A., and Nurit, C. (2014). Pro-environmental behavior and its antecedents as a case of social and temporal dilemmas. *Br. J. Educ. Soc. Behav. Sci*. 4, 508–526. doi: 10.9734/BJESBS/2014/6573

Schultz, P. W., Gouveia, V. V., Cameron, L. D., Tankha, G., Schmuck, P., and Franěk, M. (2005). Values and their relationship to environmental concern and conservation behavior. *J. Cross Cult. Psychol*. 36, 457–475. doi: 10.1177/0022022105275962

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Stat*. 6, 461–464. doi: 10.1214/aos/1176344136

Seddig, D., and Leitgöb, H. (2018). Approximate measurement invariance and longitudinal confirmatory factor analysis: concept and application with panel data. *Survey Res. Methods* 12, 29–41. doi: 10.18148/srm/2018.v12i1.7210

Shao, S., Tian, Z., and Fan, M. (2018). Do the rich have stronger willingness to pay for environmental protection? New evidence from a survey in China. *World Dev*. 105, 83–94. doi: 10.1016/j.worlddev.2017.12.033

Shi, D., Song, H., Liao, X., Terry, R., and Snyder, L. A. (2017). Bayesian SEM for specification search problems in testing factorial invariance. *Multivariate Behav. Res*. 52, 430–444. doi: 10.1080/00273171.2017.1306432

Smid, S. C., McNeish, D., Miočević, M., and van de Schoot, R. (2020). Bayesian versus frequentist estimation for structural equation models in small sample contexts: a systematic review. *Struct. Equat. Model.* 27, 131–161. doi: 10.1080/10705511.2019.1577140

Solstad, B. E., Stenling, A., Ommundsen, Y., Wold, B., Heuzé, J. P., Sarrazin, P., et al. (2020). Initial psychometric testing of the coach-adapted version of the empowering and disempowering motivational climate questionnaire: a Bayesian approach. *J. Sports Sci.* 38, 626–643. doi: 10.1080/02640414.2020.1722575

Spiegelhalter, D., Myles, J., Jones, D., and Abrams, K. (2000). *Bayesian Methods in Health Technology Assessment: A Review*. Technical Report, National Coordinating Centre for Health Technology Assessment (NCCHTA). doi: 10.3310/hta4380

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit. *J. R. Stat. Soc. Ser. B* 64, 583–639. doi: 10.1111/1467-9868.00353

van de Schoot, R., Broere, J. J., Perryck, K. H., Zondervan-Zwijnenburg, M., and Van Loey, N. E. (2015). Analyzing small data sets using Bayesian estimation: the case of posttraumatic stress symptoms following mechanical ventilation in burn survivors. *Eur. J. Psychotraumatol.* 6, 1–13. doi: 10.3402/ejpt.v6.25216

van de Schoot, R., Depaoli, S., King, R., Kramer, B., Märtens, K., Tadesse, M. G., et al. (2021). Bayesian statistics and modelling. *Nat. Rev. Methods Primers.* 1:16. doi: 10.1038/s43586-021-00017-2

van de Schoot, R., Hoijtink, H., Romeijn, J. W., and Brugman, D. (2012). A prior predictive loss function for the evaluation of inequality constrained hypotheses. *J. Math. Psychol.* 56, 13–23. doi: 10.1016/j.jmp.2011.10.001

van de Schoot, R., Kluytmans, A., Tummers, L., Lugtig, P., Hox, J., and Muthén, B. (2013). Facing offwith Scylla and Charybdis: a comparison of scalar, partial, and the novel possibility of approximate measurement invariance. *Front. Psychol.* 4:770. doi: 10.3389/fpsyg.2013.00770

van de Schoot, R., Veen, D., Smeets, L., Winter, S. D., and Depaoli, S. (2019). "Chapter 3: A tutorial on using the WAMBS," in *Small Sample Size Solutions: A Guide for Applied Researchers and Practitioners, 1st edn*, eds R. van de Schoot and M. Miocevic (New York, NY: Routledge), 30–49. doi: 10.4324/9780429273872-4

van de Schoot, R., Winter, S. D., Ryan, O., Zondervan-Zwijnenburg, M., and Depaoli, S. (2017). A systematic review of Bayesian articles in psychology: the last 25 years. *Psychol. Methods* 22, 217–239. doi: 10.1037/met0000100

van de Vijver, F. J. R., Avvisati, F., Davidov, E., Eid, M., Fox, J.-P., Donné, N. L., et al. (2019). "Invariance analyses in large-scale studies," in *OECD Education Working Papers* (Paris: Organisation for Economic Co-operation and Development (OECD)). doi: 10.1787/254738dd-en

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

van Valkengoed, A. M., and Steg, L. (2019). Meta-analyses of factors motivating climate change adaptation behaviour. *Nat. Clim. Change* 9, 158–163. doi: 10.1038/s41558-018-0371-y

Vandenberg, R. J., and Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: suggestions, practices, and recommendations for organizational research. *Organ. Res. Methods* 3, 4–69. doi: 10.1177/109442810031002

Vilar, R. (2020). Basic Values Survey (BVS)-A 20-nation study. doi: 10.31234/osf.io/wvtj3

Xiao, Y., Liu, H., and Hau, K. T. (2019). A comparison of CFA, ESEM, and BSEM in test structure analysis. *Struct. Equat. Model.* 26, 665–677. doi: 10.1080/10705511.2018.1562928

Zercher, F., Schmidt, P., Cieciuch, J., and Davidov, E. (2015). The comparability of the universalism value over time and across countries in the European Social Survey: exact vs. approximate measurement invariance. *Front. Psychol.* 6:733. doi: 10.3389/fpsyg.2015.00733

Zhang, D., Pei, Q., Frölich, C., and Ide, T. (2020). "Chapter 4: Does climate change drive violence, conflict and human migration?," in *Contemporary Climate Change, 1st Edn*, ed M. Hulme (New York, NY: Routledge), 51–61. doi: 10.4324/9780429446252-5

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 29 November 2021
doi: 10.3389/fpsyg.2021.621547

The Use of Questionable Research Practices to Survive in Academia Examined With Expert Elicitation, Prior-Data Conflicts, Bayes Factors for Replication Effects, and the Bayes Truth Serum

Rens van de Schoot[1,2], Sonja D. Winter[3], Elian Griffioen[1], Stephan Grimmelikhuijsen[4], Ingrid Arts[1], Duco Veen[2,4], Elizabeth M. Grandfield[1] and Lars G. Tummers[5]*

[1] Department of Methods and Statistics, Utrecht University, Utrecht, Netherlands, [2] Optentia Research Program, Faculty of Humanities, North-West University, Vanderbijlpark, South Africa, [3] Missouri Prevention Science Institute, University of Missouri, Columbia, MO, United States, [4] Department of Global Health, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, Netherlands, [5] School of Governance, Utrecht University, Utrecht, Netherlands

OPEN ACCESS

Edited by:
Alexander Robitzsch,
IPN – Leibniz Institute for Science and Mathematics Education, Germany

Reviewed by:
Peter Adriaan Edelsbrunner,
ETH Zürich, Switzerland
Hilde E. M. Augusteijn,
Tilburg University, Netherlands

***Correspondence:**
Rens van de Schoot
a.g.j.vandeschoot@uu.nl

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: 26 October 2020
Accepted: 20 October 2021
Published: 29 November 2021

Citation:
van de Schoot R, Winter SD, Griffioen E, Grimmelikhuijsen S, Arts I, Veen D, Grandfield EM and Tummers LG (2021) The Use of Questionable Research Practices to Survive in Academia Examined With Expert Elicitation, Prior-Data Conflicts, Bayes Factors for Replication Effects, and the Bayes Truth Serum.
Front. Psychol. 12:621547.
doi: 10.3389/fpsyg.2021.621547

The popularity and use of Bayesian methods have increased across many research domains. The current article demonstrates how some less familiar Bayesian methods can be used. Specifically, we applied expert elicitation, testing for prior-data conflicts, the Bayesian Truth Serum, and testing for replication effects via Bayes Factors in a series of four studies investigating the use of questionable research practices (QRPs). Scientifically fraudulent or unethical research practices have caused quite a stir in academia and beyond. Improving science starts with educating Ph.D. candidates: the scholars of tomorrow. In four studies concerning 765 Ph.D. candidates, we investigate whether Ph.D. candidates can differentiate between ethical and unethical or even fraudulent research practices. We probed the Ph.D.s' willingness to publish research from such practices and tested whether this is influenced by (un)ethical behavior pressure from supervisors or peers. Furthermore, 36 academic leaders (deans, vice-deans, and heads of research) were interviewed and asked to predict what Ph.D.s would answer for different vignettes. Our study shows, and replicates, that some Ph.D. candidates are willing to publish results deriving from even blatant fraudulent behavior–data fabrication. Additionally, some academic leaders underestimated this behavior, which is alarming. Academic leaders have to keep in mind that Ph.D. candidates can be under more pressure than they realize and might be susceptible to using QRPs. As an inspiring example and to encourage others to make their Bayesian work reproducible, we published data, annotated scripts, and detailed output on the Open Science Framework (OSF).

Keywords: informative prior, Bayes truth serum, expert elicitation, replication study, questionable research practices, Ph.D. students, Bayes Factor (BF)

INTRODUCTION

Several systematic reviews have shown that applied researchers have become more familiar with the typical tools of the Bayesian toolbelt (Johnson et al., 2010a; König and van de Schoot, 2017; van de Schoot et al., 2017, 2021a; Fragoso et al., 2018; Smid et al., 2020; Hon et al., 2021). However, there remain many tools in the Bayesian toolbelt that are less familiar in the applied literature. In the current article, we illustrate how some less familiar tools can be applied to empirical data: A Bayesian expert-elicitation method (O'Hagan et al., 2006; Anca et al., 2021) – also described in van de Schoot et al. (2021b), a test for prior-data conflict using the prior predictive *p*-value (Box, 1980) and the Data Agreement Criterion (DAC) (Veen et al., 2018), a Bayes truth serum to correct for socially desirable responses (Prelec, 2004), and testing for replication effects via the Bayes Factor (Bayarri and Mayoral, 2002; Verhagen and Wagenmakers, 2014). These methods are applied to the case of how Ph.D. students respond to academic publication pressure in terms of conducting questionable research practices (QRPs).

In what follows, we first elaborate on QRPs, how Ph.D. candidates respond to scenarios of QRPs, and senior academic leaders (deans, heads of departments, and research directors, etc.) believe Ph.D. candidates will deal with this pressure. In four separate sections, we present the results of the different studies and illustrate how the Bayesian methods mentioned above can be applied to answer the substantive research questions, thereby providing an example of how to use Bayesian methods for empirical data. Also, **Supplementary Material**, including annotated code, part of the anonymized data, and more detailed output files, can be found on the Open Science Framework (OSF)[1]. The Ethics Committee of the Faculty of Social and Behavioral Sciences at Utrecht University approved the series of studies (FETC15-108), and the questionnaires were co-developed and pilot-tested by a university-wide organization of Ph.D. candidates at Utrecht University (Prout) and the Dutch National Organization of Ph.D. candidates (PNN). **Supplementary Appendix A–C** contains additional details referred to throughout the text.

THE CASE OF QUESTIONABLE RESEARCH PRACTICES TO SURVIVE IN ACADEMIA

Science has always been a dynamic process with continuously developing and often implicit rules and attitudes. While a focus on innovation and knowledge production are essential to academic progress, it is equally important to convey and stimulate the use of the most appropriate research practices within the academic community (Martinson et al., 2005; Fanelli, 2009; Tijdink et al., 2014). There is an intense pressure to publish since scientific publications are integral in obtaining grants or obtaining a tenured position in academia (Gopalakrishna et al., 2021; Haven et al., 2021). Ph.D. candidates have noted that the most critical factors related to obtaining an academic position were the number of papers presented, submitted, and accepted in peer-reviewed journals (Sonneveld et al., 2010; Yerkes et al., 2010). In an observational study by Tijdink et al. (2014), 72% of respondents reported pressure to publish was "too high" and was associated with higher scores on a scientific misconduct questionnaire measuring self-reported fraud and QRPs. With increasing publication pressure, a growing number of scholars, and ever more interdisciplinary and international studies being conducted, academic norms have become diverse and complicated. Publication pressure combined with the ambiguity of academic standards has contributed to QRPs such as data fabrication, falsification, or other modifications of research results (Fanelli, 2010). Early-career scientists may struggle to identify QRPs and, as Sijtsma (2016) noted, may even commit QRPs unintentionally. Anecdotally, statements such as "this is how we always do it," "get used to it," or "this is what it takes to survive in academia" may also be familiar to some researchers and students, which do not help develop a sense of ethical standards for research practices.

In response to these observations, the contemporary debate about appropriate scientific practices is fierce and lively and has extended to non-academic domains. Therefore, how we conduct research and, equally important, how we inform, mentor, and educate young scientists is essential to sound scientific progress and how science is perceived and valued (Anderson et al., 2007; Kalichman and Plemmons, 2015). An observational study by Heitman et al. (2007), for example, found that scholars who reported receiving education about QRPs scored 10 points higher on a questionnaire about these issues (reporting that they are less likely to participate in QRPs) compared to scholars without prior QRP training. A Ph.D. trajectory is essentially about educating someone to become an independent scientist, ethical research practices should be part of all graduate curricula. Still, early-career scientists mostly learn from observing the scientific norms and practices of academic leaders (Hofmann et al., 2013), most of whom are their direct supervisors. Ph.D. candidates are in a highly dependent relation with these senior faculty members. Senior faculty, therefore, is in the position to influence the Ph.D. candidate, which also holds for ethical issues concerning scientific behavior. At the same time, Ph.D. candidates compete with their peers for a limited number of faculty positions, a situation that may also be a factor in yielding to questionable scientific behavior.

The various potential sources of pressure from senior academic leaders and peer competition occur in an early stage of their academic career when Ph.D. candidates are susceptible to learning about ethical research practices. Senior researchers with a role-model function may not completely understand the pressure experienced by the current cohort of Ph.D. candidates. It has, so far, never been investigated how such pressure interacts with the occurrence of questionable research behavior among Ph.D. candidates, nor how academic leaders predict the behavior of Ph.D. candidates in such situations.

Therefore, in the current article, we present a series of four studies investigating these issues.

For the first study, we asked Ph.D. candidates from a wide range of Social Sciences faculties across Netherlands what they would do when faced with the three scenarios, how they would respond, to whom to talk about it, and whether they

[1] https://osf.io/raqsd/

had experienced a similar situation in their career. We also added experimental conditions: in the description of the senior scholar for the vignettes, we manipulated the level of ethical leadership (high/low) and research transparency (high/low). Ethical leadership and research transparency were used as a manipulation check to see if participants interpreted the vignettes correctly. These two factors were included because in the organizational sciences, ethical leadership is thought to be a way to improve employees' ethical conduct (Brown and Treviño, 2006), and increased research transparency is offered as a solution to prevent fraud and misconduct in many fields of science (Parker et al., 2016).

For the second study, we interviewed academic leaders about what they expected. Ph.D. candidates would do in the scenarios from Study 1. The social sciences within Netherlands had a real wake-up call with the Stapel case (Callaway, 2011; Levelt et al., 2012; Markowitz and Hancock, 2014). Hopefully, this case would have created awareness, at least in academic leaders. The question is whether the academic leaders would think the Ph.D. candidates, who mostly started their projects after the news about Stapel had faded away, also changed their attitude toward scientific fraud and QRPs. Therefore, after obtaining the results from the academic leaders, we tested for expert-data (dis)agreement (Bousquet, 2008; Veen et al., 2018) between the academic leaders and the Ph.D. candidates to see if the academic leaders over-or underestimated the replies given by the Ph.D. candidates.

The third study concerned a conceptual replication of the first vignette in Study 1 (data fabrication). Replication is not only an essential aspect of scientific research but has also been recommended as a method to help combat QRPs (Sijtsma, 2016; Sijtsma et al., 2016; Waldman and Lilienfeld, 2016). Study 3 participants were from a major university in Netherlands not included in Study 1 and represented Psychology and medical sciences. We also added two new scenarios (gift authorship and omitting relevant information) and a second experimental condition in which we manipulated peer and senior pressure by including cues in the vignette about the (imaginary) prevalence of QRPs of fellow Ph.D. candidates and professors at a different, fictional, university. It was based on the assumption that obedience to authority–from superiors or peers–influences questionable behavior, as evidenced by the large body of literature on the theory of planned behavior (Ajzen, 1991) and more general work on subjective norms and peer pressure (Terry and Hogg, 1996).

Finally, in Study 4, we replicated the experiment of Study 1 in a new sample outside the Netherlands, namely, in three Social Sciences faculties in Belgium. Replication studies are not only an essential aspect of science; as mentioned above, they may also aid in uncovering and potentially reducing QRPs.

STUDY 1–VIGNETTE STUDY A

There were two goals for Study 1: First, to investigate how Ph.D. candidates would respond to the vignettes about data fabrication, deleting outliers to get significant results, and salami slicing; see **Supplementary Appendix A** for the text used in the vignettes. Second, we used a randomized experiment to investigate whether characteristics in the description of the senior, in terms of ethical leadership and transparency, would influence their responses.

Methods

Participants, Procedure, and Design

The Ph.D. candidates for Study 1 were recruited from 10 Social Sciences or Psychology faculties at eight universities in Netherlands out of 10 universities with Social Sciences or Psychology faculties. Two more universities were invited, but one declined to participate, and at the other, the data collection never got started due to practical issues. We always asked a third party (usually a Ph.D. organization within the university) to send invitations to their Ph.D. candidates to participate in our study. This procedure ensured that we were never in possession of the email addresses of potential participants. We used the online survey application, LimeSurvey, to create a separate, individualized survey for each university involved. To further ensure our participants' privacy, we configured the surveys to save anonymized responses without information about IP address, the date and time they completed the survey, or the location of their computer (city and country). Furthermore, we ensured that all demographics questions were not mandatory for participants to complete to decide how much information they wished to share with us. Finally, participants were offered the possibility to leave an email address if they wanted to receive notice of the outcomes of our research. However, we never created a data file that contained both the email addresses and the survey data. Participants were randomly assigned to one of the four conditions within the survey.

In total, 440 Ph.D. candidates completed the questions for at least one scenario. Descriptive statistics about the sample can be found in **Table 1**. The survey focused on the three scenarios concerning QRPs/fraud: (1) data fabrication, (2) deleting outliers to get significant results, and (3) salami-slicing; see **Supplementary Appendix A** for the exact text we used. After presenting a scenario to the participant, we first asked an open-ended question: "What would you do in this situation?" Then we asked: "Would you (try to) publish the results coming from this research?" (Yes/No) followed by an open-ended question "If you want, you can elaborate on this below."

We compared responses of the Ph.D. candidates across four conditions, which were combinations of two two-level factors, Leadership and Data. To convey these conditions to the participant, we used different combinations of the introductory texts. LimeSurvey allowed us to automatically and randomly assign participants to one of the four conditions for the first experiment and then again in one of the four conditions of the experiment.

To check whether participants perceived the manipulations (high versus low ethical leadership and high versus low research transparency), we included scales for both ethical leadership (Yukl et al., 2013) (Cronbach's alpha 0.919) and research transparency (developed for this study, see **Supplementary Appendix A** for the questions used, Cronbach's alpha 0.888).

TABLE 1 | Descriptive statistics for Study 1 (*N* = 440), Study 3 (*N* = 198), and Study 4 (*N* = 127).

Variable		Study 1	Study 3	Study 4
Gender	Male	128 (29.09%)	48 (24.24%)	80 (62.99%)
	Female	247 (56.14%)	121 (61.11%)	35 (27.56%)
	Prefer not to disclose	12 (2.73%)	8 (4.04%)	2 (1.57%)
	Missing	53 (12.05%)	21 (10.61%)	10 (7.87%)
Age		31.65 (7.84; 24/70)	31.40 (6.15; 24/64)	29.06 (4.79; 23/48)
		n = 365	*n* = 166	*n* = 116
Employment type	Standard Ph.D. candidate	296 (67.27%)	150 (75.76%)	45 (35.43%)
	No Ph.D. candidate but Ph.D. scholarship	17 (3.86%)	10 (5.05%)	56 (44.09%)
	External Ph.D. candidate	15 (3.45%)	10 (5.10%)	7 (5.51%)
	Other	53 (12.01%)	20 (10.10%)	10 (7.87%)
	Missing	59 (13.41%)	8 (4.04%)	9 (7.09%)
Data: Collecting and/or analyzing	I collect and analyze	370 (84.09%)	139 (70.20%)	103 (81.10%)
	I collect, someone else analyses	20 (4.55%)	14 (7.07%)	4 (3.15%)
	I analyze existing data	37 (8.41%)	32 (16.16%)	14 (11.02%)
	My research is mainly theoretical	7 (1.59%)	9 (4.55%)	4 (3.15%)
	Missing	6 (1.36%)	4 (2.02%)	2 (1.57%)
Certainty career in academics	Scale 1–10	6.76 (2.27; 1/10)	6.82 (2.32; 1/10)	5.39 (2.56, 1/10)
		n = 440	*n* = 198	*n* = 127
Ambition career in academics	Scale 1–10	6.80 (2.20; 1/10)	6.91 (2.14; 1/10)	5.50 (2.49; 1/10)
		n = 440	*n* = 198	*n* = 127
Perceived publication pressure	Scale 1–6	4.64 (0.91; 1/6)		
Is publication pressure present in the research field?	Scale 1–10		7.11 (1.87; 1/10)	7.41 (1.77;1/10)

Data are mean (SD; min/max) or frequency (%).

In **Supplementary Appendix B**, we describe the results of the manipulation checks for Ethical Leadership and Data Transparency. We concluded that the manipulation resulted in a different score on both variables across conditions, indicating that our manipulation was effective.

Analytic Strategy

We first provide descriptive statistics about the responses of the Ph.D. candidates to each of the vignettes.

Second, we present the replies to the open-ended questions. We grouped the responses in several categories. Grouping of the open answer was made based on group discussions and consensus among the authors using an *ad hoc* bottom-up process. Multiple categories could be given to each answer. We discussed ambiguous responses and only classified participants' answers in one of the categories if all authors reached a consensus. We also examined whether, based on information in the open-ended questions, the Ph.D. candidates provided an honest reply to the yes/no question about publishing and recoded the item into a new variable next to the existing variable. For the first scenario, in 22 cases, the information in the open-ended answer did not correspond with the yes/no question. An equal number of responses was recoded from "yes" to "no" and from "no" to "yes." For the second scenario, we recoded 154 answers. In most of these cases (97%), the Ph.D. candidate indicated in the open-ended answers that they would publish the results only if the outliers were described in the article. Since the scenario was about publishing the data without providing more information, we recoded these answers to "no." As a result, the percentage of participants indicating that they would attempt to publish dropped from 48.8 to 12.5% (a 36.3% decline). In the third scenario, in 16 cases, the information in the open-ended answer did not correspond with the yes/no answer. It resulted in a decline of 1.5% in the participants' indication that they would attempt to publish. Again, the decisions were discussed and only changed if consensus was reached among all authors.

Third, we used Bayes Factors for contingency tables in JASP (JASP-Team, 2018) to examine whether the experimental conditions affected the participants' attitude toward publishing data or analyses that might have fallen victim to QRPs. When a hypothesis is tested against an alternative hypothesis, and the results indicate that BF $\approx$ 1 implies that both hypotheses are equally supported by the data. However, for example, when BF = 10, the support for one hypothesis is 10 times larger than the support for the alternative hypothesis. For interpretation of Bayes Factors, we refer interested readers to the classical paper of Kass and Raftery (1995).

Results

Most Ph.D. candidates in this study (96.6%) answered "yes" to the question of whether they consider the vignette scenario to be fraudulent (see **Table 2**). As for the first scenario, almost all Ph.D. candidates believe data fabrication is fraudulent; interestingly, 5.9% (25 students) would still publish the results, and some participants reported having experienced such a situation.

Most participants provided extensive answers to the open-ended questions. We grouped their responses into six categories. The first category comprised 34.6% of the Ph.D. candidates who indicated they would never publish such results because they feel *morally obliged* not to do so, as is implied by statements like "it wouldn't feel good to do so" or "I can't accept that for myself," or put more strongly:

> *"Never, this goes against all I stand for and this is not what research is about, I feel very annoyed that this question is even being asked."*

The second category of Ph.D. candidates (22.6%) reported that they would first *talk to someone else before taking any action*. Of these Ph.D. candidates, 23.9% would first talk to another Ph.D.

candidate, 23.9% to their daily supervisor, 20.7% to their doctoral advisor, 20.7% to the project leader, 7.6% to the confidential counselor, and 3.3% to someone else. The third category of Ph.D. candidates (15.5%) indicated they would first take a more *pragmatic approach* before doing anything else. They would only want to decide when, for example, more information is provided, new data is collected, or more analyses are conducted. The fourth category of Ph.D. candidates (10.5%) is afraid the situation *might backfire on them in a later stage of their career* which is their main argument for not proceeding with the paper, as is exemplified by this statement:

> *"I'd rather finish my thesis later than put my career at risk."*

The fifth category of Ph.D. candidates (8.7%) provides as the main argument that *they believe in* good scientific practice and a world where science serves to advance humanity:

> *"Producing science and knowledge is part of academia so that humans can get closer to the 'truth', producing fake stuff is not part of academia and I don't want to be part of that."*

> *"In the long-term, being honest provides the best answers to societal issues."*

Lastly, we identified a group of Ph.D. candidates (8%) as "*at-risk*." They either reported that if the pressure were high enough, they would proceed with the publication, as indicated by the following quote:

> *"It's not a solid yes, but a tentative one. I can image, just to be realistic, in terms of publishing pressures and not wanting to be out of contract, that this would be the best bet after all."*

Or, they would follow their supervisor:

> *"If the supervisors tell me it's okay, I would try to publish the data."*

Or, they simply have no qualms about it:

> *"Since it will get me closer to obtaining my Ph.D."*

The result of testing for manipulation effect was that for all scenarios, the null model, assuming no effect for condition, was preferred over the alternative model (all BF_{01}'s < 1); see **Supplementary Appendix C** for detailed results.

Intermediate Conclusion

The first study shows that at least some Ph.D. candidates are willing to publish results even if they know the data has been made up, the deletion of outliers is not adequately described, or if they are asked to split their papers into several sub-papers (i.e., salami-slicing). The percentage of Ph.D. candidates who actually experienced such a situation is low but not zero (see **Table 2**). Contrary to our expectations and although the manipulation checks were successful (see **Supplementary Appendix B**)–neither ethical leadership of the senior/supervisor nor transparency in the description resulted in differences in the Ph.D. candidates' intended publishing behavior.

TABLE 2 | Results in percentages of the vignette studies Study 1 (*N* = 440), Study 3 (*N* = 198), and Study 4 (*N* = 127).

	Study 1			Study 3			Study 4		
	"Is this fraud?" (% Yes)	"Yes, I would try to publish"	"Have you experienced a similar situation?" (% Yes)	"Is this fraud?" (% Yes)	"Yes, I would try to publish"	"Have you experienced a similar situation?" (% Yes)	"Is this fraud?" (% Yes)	"Yes, I would try to publish"	"Have you experienced a similar situation?" (% Yes)
Scenario 1: Data fabrication[a]	96.6%	5.9%	3.2%	92.4%	9.6%	5.5%	92.9%	13.4%	5.5%
	(*n* = 440)	(*n* = 440)	(*n* = 440)	(*n* = 198)	(*n* = 198)	(*n* = 198)	(*n* = 127)	(*n* = 127)	(*n* = 127)
Scenario 2: Deleting outliers to get significant results	56.4%	12.3%	12.9%						
	(*n* = 407)	(*n* = 407)	(*n* = 407)						
Scenario 3: Salami slicing[a]	65.2%	32.0%	9.3%	16.6%	38.9%	17.2%	23.6%	32.8%	17.3%
	(*n* = 397)	(*n* = 397)	(*n* = 397)	(*n* = 185)	(*n* = 185)	(*n* = 185)	(*n* = 119)	(*n* = 119)	(*n* = 119)
Scenario 4: Gift authorship				42.4%	59.2%	30.3%	40.6%	58.8%	16.7%
				(*n* = 184)	(*n* = 184)	(*n* = 184)	(*n* = 118)	(*n* = 118)	(*n* = 116)
Scenario 5: Excluding information				71.7%	12.1%	13.6%	72.4%	16.1%	15.8%
				(*n* = 182)	(*n* = 182)	(*n* = 182)	(*n* = 118)	(*n* = 118)	(*n* = 118)

[a] *For Studies 3 and 4, we modified the description based on feedback from the participants, see* ***Supplementary Appendix A****.*

STUDY 2–EXPERT ELICITATION AND PRIOR-DATA CONFLICTS

The goal of Study 2 was to investigate how academic leaders believed Ph.D. candidates would respond to the three scenarios and to test whether the beliefs of the seniors about Ph.D. candidates' behavior regarding QRPs conflicted with the observed data from Study 1.

Methods

Participants and Design

We invited 36 academic leaders working at 10 different faculties of Social and Behavioral Sciences or Psychology in Netherlands–deans, vice-deans, heads of departments, research directors, and confidential counselors–to participate in the study and share what they believed Ph.D. candidates would answer when facing the three scenarios. The design of the study and how confidentiality would be ensured (i.e., personal characteristics would not be disclosed, answers would not be connected to specific data or results or used as predictors for explaining possible disagreements with the collected data in Study (2) was described in a face-to-face interview with the first author (RS). All academic leaders answered at least one scenario and very few skipped questions (response per scenario was 34, 35, and 33 from the 36 different leaders).

Analytic Strategy

The method used to obtain the necessary information from the experts is referred to as prior elicitation (O'Hagan et al., 2006), which is the process of extracting and creating a representation of an expert's beliefs. During a face-to-face interview, we used the Trial-Roulette elicitation method to capture the beliefs of the seniors in a statistical distribution. This elicitation method was introduced by O'Hagan et al. (2006) and was validated by Johnson et al. (2010b); Veen et al. (2017), Zondervan-Zwijnenburg et al. (2017), and Lek and Van De Schoot (2018).

To obtain a proper representation of the experts' beliefs about the percentage of Ph.D. candidates answering "yes" to the questions whether to publish the paper in the three scenarios, participants had to place twenty stickers, each representing five percent of a distribution, on an axis representing the percentage of Ph.D. candidates answering "yes" from 0% (left) to 100% (right). The placement of the first sticker at a specific position on the axis should indicate perceived likeliness by the expert for that value. In contrast, the other stickers represented uncertainty around this estimate, thereby creating a stickered distribution. The elicitation procedure resulted in one stickered distribution per expert per scenario, for a total of 102 valid distributions (six distributions could not be transformed into a parametric beta distribution). See **Figure 1** for an example of such a stickered distribution and see **Figure 2** for all the statistical distributions per scenario. The method we used to obtain statistical distributions based on the stickered distributions is published in van de Schoot et al. (2021b).

To examine whether the beliefs expressed by the senior academic leaders conflict with the observed data of the Ph.D. candidates (Study 1), we tested for an expert-data conflict. Box (1980) proposed using prior predictive distributions to test if the collected data was unlikely for this predictive distribution. Evans and Moshonov (2006) presented a variation, the prior-predictive check (PPC) computed per expert, and results in a value reflecting the existence of prior-data conflict. With the PPC, the prior distribution itself is used to predict various proportions that could have been observed. These predicted proportions can be used to assess the probability that the actual data proportion can be found using the prior distribution resulting in a probability value. When the value is less than 0.05, it reflects a prior-data conflict; see **Figure 3**. To cross-validate the results, we also computed the DAC developed by Bousquet (2008) and extended by Veen et al. (2018), where values >1 indicate a conflict. Since the results of both measures are highly comparable, see **Figure 4**; the results section below presents only the detailed PPC results. For a comparison between the two methods, see Lek and Van De Schoot (2019). The complete results, including annotated syntax, can be found on OSF (see text footnote 1).

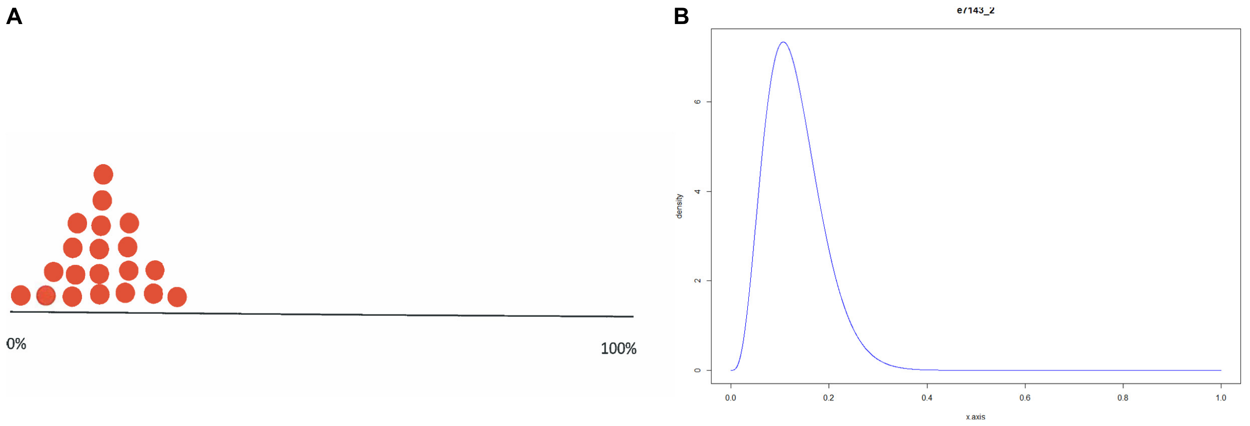


FIGURE 1 | Example of a stickered distribution using **(A)** the trial roulette method and **(B)** the probability distribution obtained with the SHELF software (Oakley and O'Hagan, 2010).

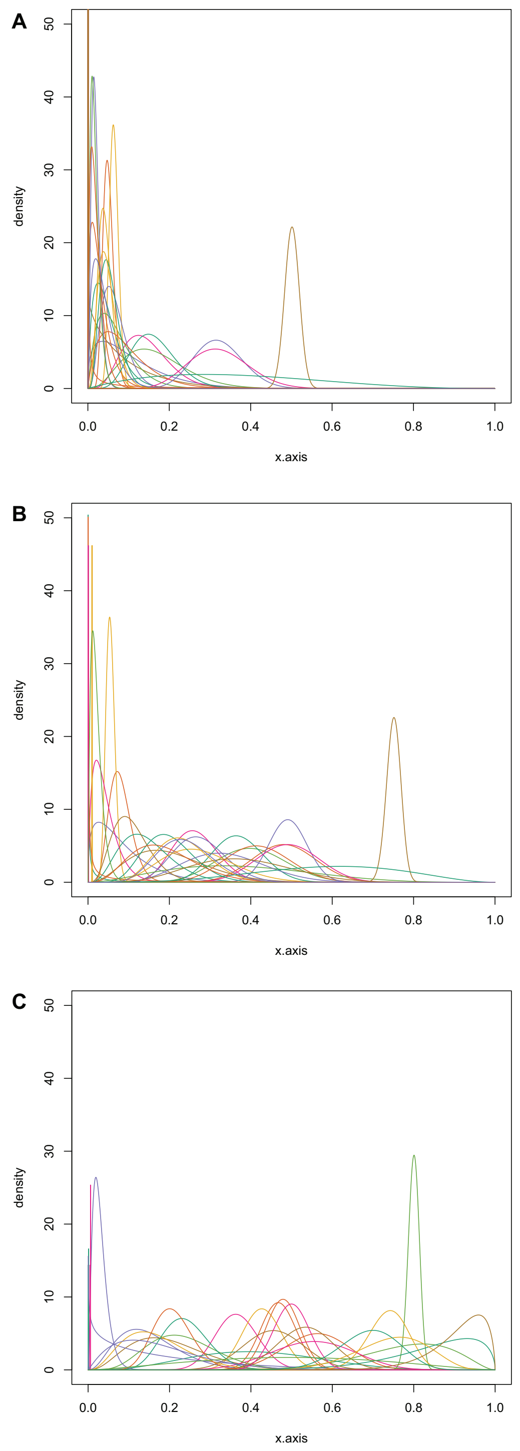


FIGURE 2 | The parametric beta distributions based on the experts' stickered distributions for Scenario 1 (**A**; *n* = 34), 2 (**B**; *n* = 35) and 3 (**C**; *n* = 33).

Results

As shown in van de Schoot et al. (2021b), 82% (40 and 18% for scenarios 2 and 3, respectively) of the academic leaders believed the percentage of Ph.D. candidates willing to publish a paper, even if they did not trust the data because of potential data fabrication, to be precisely zero (*n* = 8) or close to zero (*n* = 20).

When testing for prior-data conflicts for Scenario 1 (data fabrication), it appeared 20 experts (58.8%) showed no significant conflict with the data based on the PPC. Nine experts (26.5%) significantly underestimated the percentage of Ph.D. candidates willing to publish with fabricated data, while the remaining five (14.7%) overestimated this percentage. For Scenario 2 (Deleting Outliers), fewer experts (15; 42.9%) showed no significant conflict with the data. Only six experts (17.1%) significantly underestimated the percentage of Ph.D. candidates willing to publish with data that suppressed outliers, while 14 experts (40.0%) overestimated this percentage. For Scenario 3 (Salami Slicing), the lowest number of experts (11; 33.3%) showed no significant conflict with the data. Five experts (15.2%) significantly underestimated the percentage of Ph.D. candidates who would be willing to publish with data resulting from salami slicing, while most experts (17; 51.5%) overestimated this percentage.

Intermediate Conclusion

Some academic leaders overestimated the percentage, and some were in tune with the outcomes of Study 1. However, academic leaders (too) often underestimate the willingness of Ph.D. candidates to "survive academia" utilizing fraudulent or QRPs. Underestimation is far more problematic because one student or researcher conducting QRPs can have profound implications. It is not easy to predict such behavior but expecting it to be non-existent, as several academic leaders believed, is overly optimistic. These findings indicate an awareness gap with senior academic leaders, a worrisome conclusion, given their position in the academic hierarchy and their role in policy development.

STUDY 3–VIGNETTE STUDY B

There were three goals for Study 3: First, to conceptually reproduce and extend the vignette study (we modified the description of the scenarios based on feedback to study 1, and we added three new scenarios). Our second goal was to investigate the influence of peer and elite pressure. The third goal was to examine honesty about having committed a QRP through the Bayes truth serum (Prelec, 2004).

Methods

Participants, Procedure, and Design

For Study 3, we received a list of email addresses from one university of all Ph.D. candidates in two faculties (Psychology and Medicine), allowing us to send out our invitation email. We used the same online survey tool and set-up as study 1.

In total, 198 Ph.D. candidates completed the questions for at least one of the scenarios. The Ph.D. candidates were from two different faculties of one major university in Netherlands. Descriptive statistics on the sample can be found in **Table 1**.

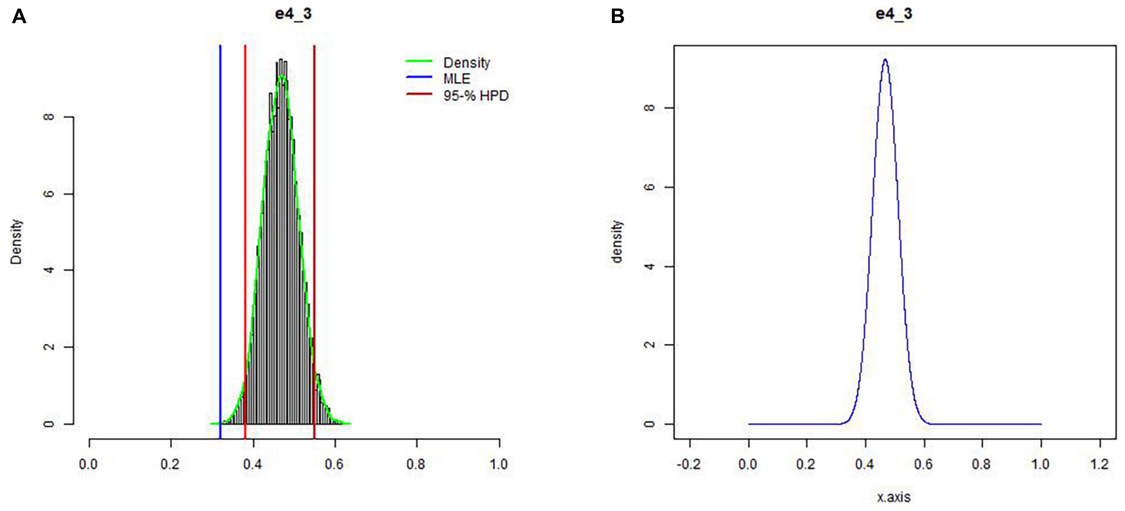


FIGURE 3 | (A) A histogram of predicted data is shown based on the prior derived from the expert (shown in **B**). The red lines indicate the credibility interval of the prior predictive distribution, and the blue line the observed percentage. The probability value appeared to be <0.001, showing there is a prior-data conflict. A table with results per expert can be found on the OSF.

Measures/Analytic Strategy

The first part of our survey was an adjusted version of the experiment applied in Study 1. Instead of three scenarios, we used only one scenario, an updated version of the Data Fabrication scenario adapted based on the Ph.D. candidates' feedback in Study 1; see **Supplementary Appendix A** for the new text. The conditions for this experiment remained the same as in Study 1.

The second part of our experiment concerned the effect of varying levels of Peer and Elite pressure on participants' publishing behavior when confronted with three QRPs: (1) Salami slicing (an adjusted version of the one used in Study 1), (2) gift authorship, i.e., adding an additional co-author who did not contribute to the article, and (3) leaving out relevant results. The effect of pressure was studied by adding vignettes that varied the pressure source (peer or elite) and the extent of pressure (low or high fictive percentages of the source of pressure partaking in QRPs). Again, we used Bayes Factors in JASP to test for the effects of the different conditions.

We also wanted to get a more accurate estimate of the prevalence of three QRPs (Salami Slicing, Gift Authorship, and Excluding Results) using the Bayesian truth serum (Prelec, 2004; John et al., 2012): a scoring algorithm that can be used to provide incentives for truthful responses. Participants were presented with an introductory text aimed at motivating participants to answer truthfully and asking them to answer three questions about the prevalence of each QRP in the department:

1. What percentage of your colleagues within your department has engaged in (QRP) on at least one occasion (on a scale from 0 to 100%)? (prevalence estimate).
2. Among those colleagues who have engaged in (QRP) on at least one occasion, what percentage would indicate that they have engaged in this research practice (on a scale from 0 to 100%)? (admission estimate).
3. Have you engaged in this research practice? (self-admission rate).

Based on responses to the questions above, it is possible to compute a more realistic Actual Prevalence. John et al. (2012) suggested calculating the geometric mean of the self-admission rate, the average admission rate, and the prevalence estimate derived from the admission rate to come to a conservative Actual Prevalence Rate. The geometric mean is based on the product of the individual numbers (as opposed to the arithmetic mean, which is based on their sum); see **Figure 5** and the OSF for annotated syntax (see text footnote 1).

Results

Ethical Leadership and Transparency Experiment

Similar to Study 1, most Ph.D. candidates in the sample (92.4%) considered the data fabrication scenario fraudulent, but almost 10% would try to publish the results, and 5.5% reported experiencing such a situation. Again, the manipulation check was successful (see **Supplementary Appendix B**); the null model was always preferred over the alternative model ($BF_{01} < 1$). Also, again, the results indicate that the experimental conditions did not differ in publishing behavior; see **Supplementary Appendix C** for details.

Peer and Elite Pressure Experiment

Compared to Study 1, a much lower percentage of Ph.D. candidates considered the vignette of salami-slicing to be fraud (65.2 versus 16.6%). In contrast, the percentage of candidates who had been in such a situation doubled to 17%. The overall rates of participants who answered "yes, I would try to publish" were comparable to Study 1. The new scenarios of gift authorship and excluding information are considered fraud by more Ph.D. candidates. A majority of the Ph.D. candidates would publish the results in the scenario of gift

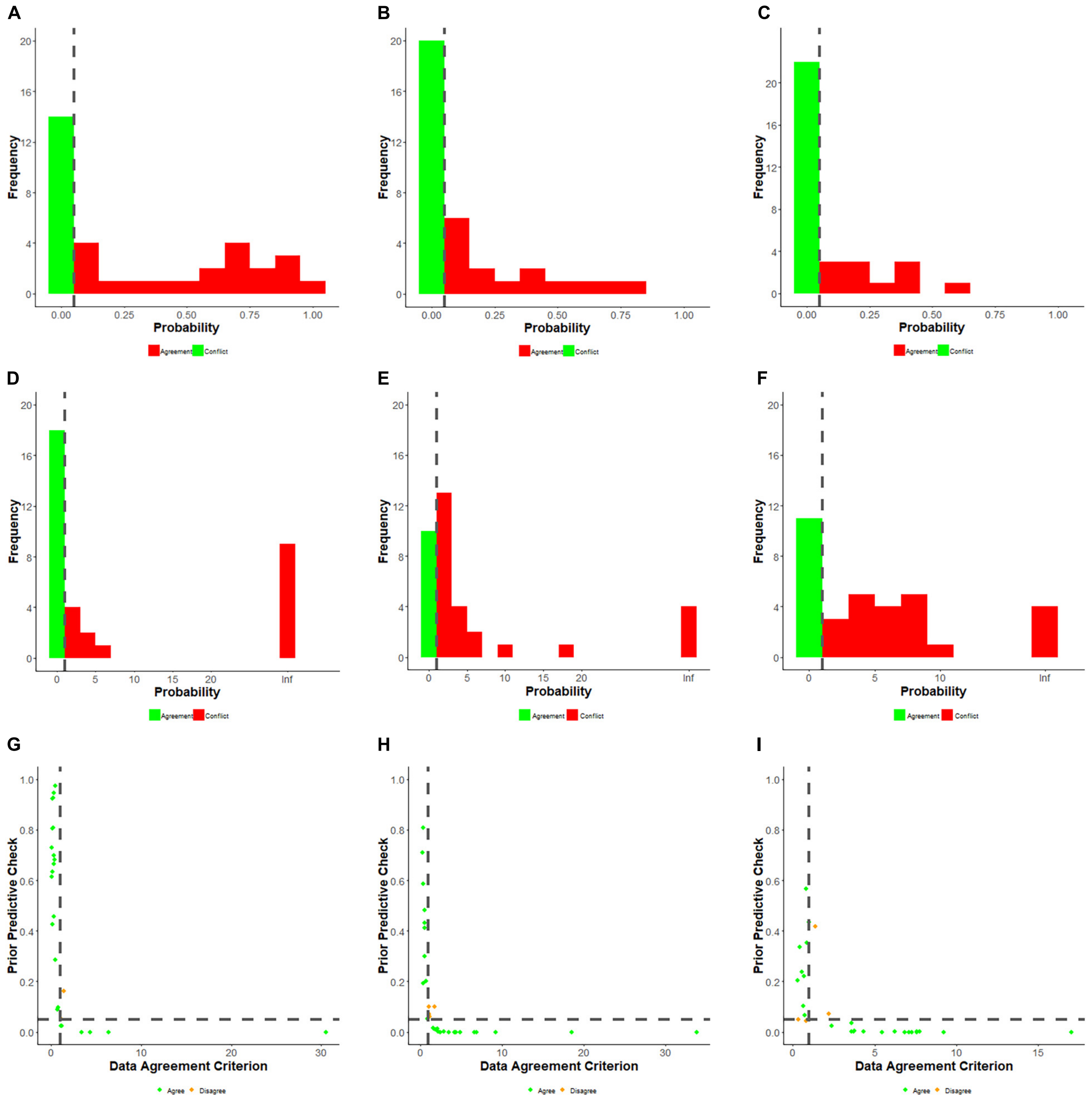


FIGURE 4 | Results for the prior predictive check **(A–C)**, the DAC **(D–F)**, and for the combination of the two **(G–I)** for each scenario separately. The dotted line represents the cut-off values used. The green dots in **(G–I)**, indicate identical conclusions for both measures, and the orange dots indicate numerical differences. It should be noted all of these are boundary cases, for example, a PPC of 0.049 (conflict) and a DAC score of 0.98 (no conflict).

authorship, but fewer had actually been in this situation, see **Table 2**. Concerning the *Peer and Elite Pressure* experiment, we did not find an effect for the experimental conditions (BFs < 1); see **Supplementary Appendix C** for detailed results. One exception was the model for salami slicing (Scenario 3r), which had a BF of 575, reflecting evidence in favor of a dependency in the contingency table. This result indicates higher pressure resulted in a higher percentage of Ph.D. candidates willing to publish the paper, especially when it concerned peer pressure.

Bayesian Truth Serum

Figure 5A shows our findings using the Bayesian truth serum. For example, 31% of the participants admitted to using the practice of gift authorship, much higher than for the other two scenarios. They expected that 40% of their colleagues did the same but that only 42% would admit doing so, leading to a Derived Prevalence Estimate of 73%. The conservative (geometric) prevalence rate would then be 46%, 14% more than the self-admission rate, comparable with the other two scenarios, 12, and 11%, respectively.

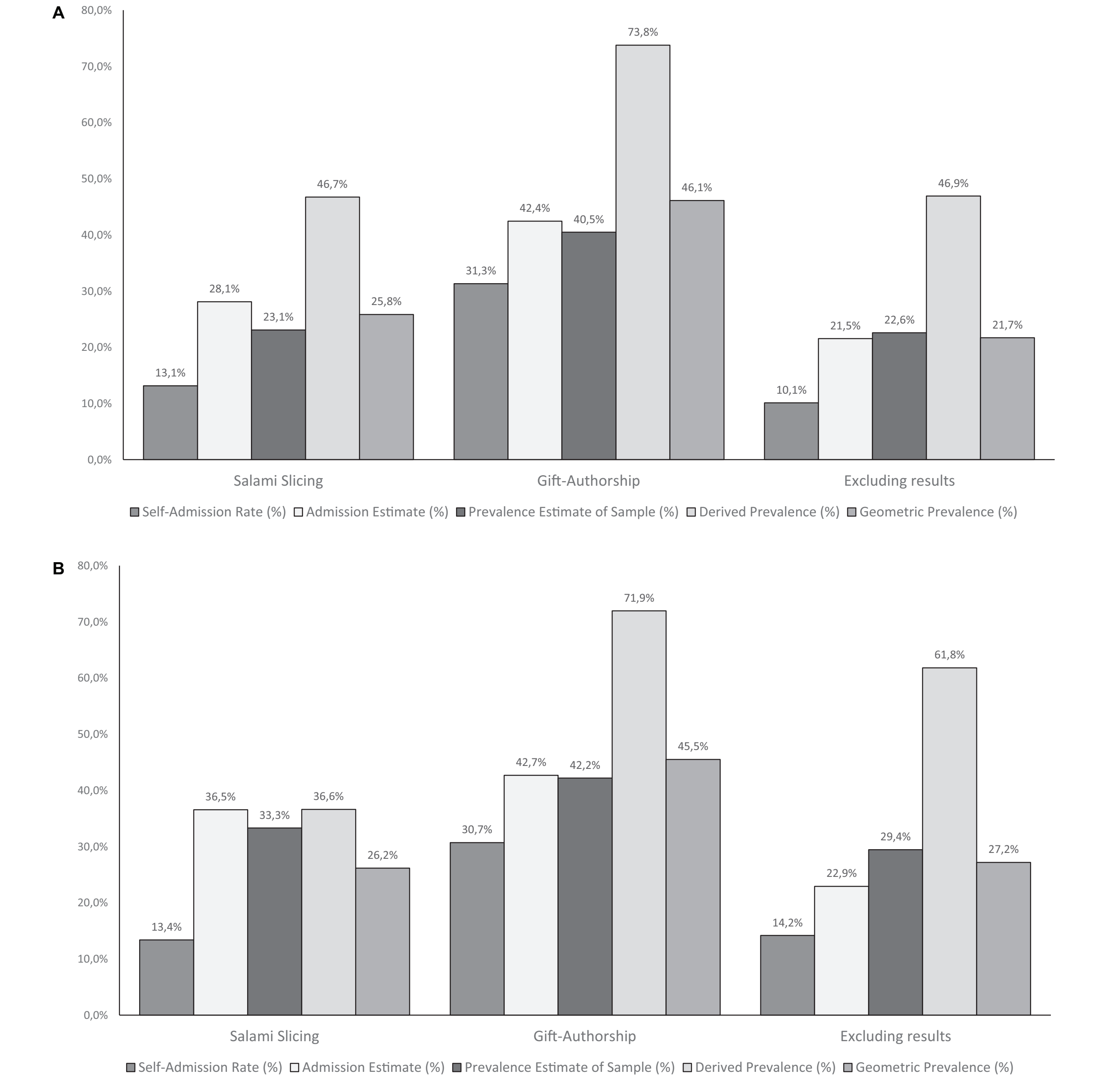


FIGURE 5 | Bayesian truth serum Results of Study 3 **(A)** and Study 4 **(B)**.

STUDY 4–INTERNATIONAL REPLICATION STUDY

The goal of the fourth study was to replicate the experiments of Study 3 and compute Bayes Factors for testing the replication effect of the Bayesian truth serum questions.

Methods

Participants, Procedure, and Design

The Ph.D. candidates were from 3 Social Sciences faculties in Belgium. We applied an identical procedure to Study 3. In total, 127 Ph.D. candidates completed the questions for at least one scenario. Descriptive statistics on the sample can be found in **Table 1**.

Analytic Strategy

First, we computed a Bayes Factor similar to the Bayes Factor we used in the previous sections to test the manipulation check and experimental conditions (H_0 = no effect). Second, we used the Equality of Effect Size Bayes Factor (Bayarri and Mayoral, 2002), which provides direct support, or lack thereof (i.e., H_0), to whether the effect size found in the original study (Study 3) equals

TABLE 3 | Results of the Bayesian test of replication where Original refers to Study 3 and Replication refers to Study 4.

Question	Scenario	Study	Mean	SD	*t*	BF1	BF2	BF3
Admission estimate	Salami	Original	28.10	32.45	12.18	4.51E + 22		
		Replication	36.54	31.19	13.20	2.69E + 22	5.30E + 22	0.67
	Gift authorship	Original	42.45	35.67	16.75	2.74E + 36		
		Replication	42.69	32.07	15.00	4.59E + 26	6.71E + 27	6.67
	Excluding results	Original	21.54	29.81	10.16	4.99E + 16		
		Replication	22.93	26.81	9.64	6.94E + 13	6.79E + 14	7.10
Prevalence estimate	Salami	Original	23.07	27.69	11.72	1.91E + 21		
		Replication	33.31	30.39	12.35	2.45E + 20	7.87E + 20	1.38
	Gift authorship	Original	40.48	34.78	16.38	2.11E + 35		
		Replication	42.19	29.96	15.87	4.63E + 28	3.12E + 29	2.16
	Excluding results	Original	22.58	27.47	11.57	6.48E + 20		
		Replication	29.44	29.52	11.24	5.05E + 17	3.96E + 18	4.65

BF1 refers to the Bayes Factor testing whether the estimate is zero or not. BF2 refers to the Bayes Factor Test for Replication Success. BF3 refers to the Equality of Effect Size Bayes Factor.

the effect size found in the replication attempt (Study 4). Third, we used the Bayes Factor Test for Replication Success (Verhagen and Wagenmakers, 2014) which is a test of the null hypothesis (H_0 = no replication) versus the alternative replication hypothesis (successful replication, H_{rep}). Annotated R-code to reproduce our results can be found on the OSF (see text footnote 1).

Results

The overall percentage of participants who answered "yes, I would try to publish" is shown in **Table 2**.

For the *Supervisor and Data Transparency* experiment, as shown in **Supplementary Appendix B**, the manipulation check worked, but, as before, we did not find an effect for the experimental conditions; see **Supplementary Appendix C** for detailed results. These results mean that the experimental conditions did not result in differences in publishing behavior.

The Bayes truth serum results can be found in **Figure 5B**, and the percentages are very similar to those of Study 3. **Table 3** displays the results of testing for a replication effect. For both studies and all three questions and scenarios, the Bayes Factors show extreme support of the percentages not being zero (see the results in the column titled BF1). The Bayes Factor for replication success (BF2) also shows great support for replicating the effects found in Study 3. The Equality of effect sizes Bayes Factor (BF3) provides support for some combinations, for example, the self-admission rate of the Salami slicing scenario with a BF of 13.74 and observed percentages of 13.13–13.39 [note that this Bayes Factor is typically much smaller (Verhagen and Wagenmakers, 2014)]. For some other conditions, there is less or even no support. In all, the percentages are pretty similar with similar effect sizes.

GENERAL DISCUSSION

The scientific community is where early career researchers such as Ph.D. candidates are socialized and develop their future norms of scientific integrity. Although there are positive indications in the public debate that QRPs are no longer acceptable, our results show that an alarming percentage of Ph.D. candidates still reported intentions to conduct fraud when under pressure, even when asked about it in hypothetical scenarios where social desirability is probably quite prevalent. QRPs can be a sensitive topic that may lead to social desirability response bias or untruthful responses (consciously or unconsciously), possibly due to obedience to authority. We consider even one Ph.D. candidate reporting intentions to commit fraud an alarming number. The Bayesian truth serum results gave far higher scores than the survey vignettes and are meant to be more trustworthy. So, the qualitative data indicates that publication pressure (surviving in academia) and supervisors' norms seem to drive the intention to conduct fraud.

Contrary to our expectations, and although the manipulation checks were all successful, neither ethical leadership of the senior/supervisor nor data transparency affected these vignettes on the Ph.D. candidates' intended publishing behavior. More worrying, academic leaders–such as deans and heads of departments—might have a blind spot for the pressure Ph.D. candidates may experience to conduct QRPs or even fraud. Academic leaders do not always have an accurate, up-to-date perception of Ph.D. candidates' willingness to engage in QRPs; eight leaders put all their density mass on exactly 0%, see Figure 5A in van de Schoot et al. (2021b). Some academic leaders in this study underestimated the inclination of Ph.D. candidates to conduct fraud or QRPs, although it must be said that some experts overestimated the percentage. It appeared not easy to predict such behavior but expecting it to be non-existent is overly optimistic.

All in all, the pressure to conduct QRPs or even commit fraud remains a significant problem for early-career scientists. We should keep an open eye for the possibility that early career researchers at least consider committing fraud when under pressure clears the way for discussing such practices. In this respect, it is imperative to inform senior academic leaders that their estimates of QRPs occurrence may be off. And although the awareness gap can go both ways in terms of over and underestimating the probability Ph.D. candidates would commit QRPs, it should be clear that underestimation could lead to more severe consequences in terms of scientific accuracy and rigor. Supervisors should take the initiative in having open discussions with the Ph.D. candidates in their department about good scientific practice versus unethical behavior. Leaders in general such as deans, vice-deans, heads of department, research directors, and confidential counselors should develop policies to address and prevent fraud and QRPs. It may seem an obvious statement to many academics. Still, as the responses in the current studies show, there are supervisors and academic leaders who do not think QRPs are a problem when they clearly still are.

The applied studies' strengths lay not only in the use of innovative Bayesian methods, but external validity is also supported using surveys and open answer formats, interviews, an experiment, and conceptual replication. The analyses focused on the quantitative aspects of the data to demonstrate the Bayesian methods outlined in the aim of the manuscript. We have added a report on the OSF (see text footnote 1) for interested readers with the descriptive qualitative responses and frequencies.

Although we expect these results to be generalizable (as supported by the replication study), the sample from Belgium may share similarities to Netherlands samples. Generalization to other countries and cultures will, of course, benefit from additional research and further future replication. Another limitation is the lack of a baseline condition without fraudulent research practices. Future studies could include conditions or scenarios without QRPs for comparison purposes. We also did not evaluate potential differences in "trying to publish" between Ph.D. candidates who reported encountering such QRP scenarios and those who have not. Future research may benefit by designing a study to examine whether experiencing these situations results in fraud beliefs or publishing decisions versus hypothetical scenarios.

In sum, supervisors, deans, and other faculty must keep in mind that Ph.D. candidates can be under more pressure than they realize and might be susceptible to using QRP.

CONCLUSION

More and more scientists have started to use Bayesian methods, and we encourage researchers to use the full potential of Bayesian methods. In this article, we demonstrated the application of some less commonly applied Bayesian methods by showcasing the use of expert elicitation, prior-data conflict tests, the Bayes truth serum, and testing for replication effects. As in all studies, many methodological and analytical decisions were made. While this could be seen as a limitation, we believe this is part of the transition toward Open Science. Therefore, to enable reproducibility, we shared all the underlying data and code following the FAIR principles: findability, accessibility, interoperability, and reusability. We hope our endeavor inspires other scientists to FAIR-ify their own work and provide the opportunity for other researchers to evaluate other alternative choices.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://osf.io/raqsd/.

ETHICS STATEMENT

The entire study was approved by the Ethics Committee of the Faculty of Social and Behavioral Sciences at Utrecht University (FETC15-108). The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

RS, SG, and LT designed the study, developed the questionnaires, and coded the open answers for Study 1. SW was in charge of data collection and conducted most of the analyses of Study 1, 3, and 4 together with IA. RS collected the expert data for Study 2. EG was in charge of the elicitation procedure for Study 2 under supervision of RS and SW and conducted the analyses for Study 2 together with DV. EMG dealt with all changes and updates required for the revision. All authors contributed significantly to the writing process and preparation of the **Supplementary Material**.

FUNDING

RS was supported by a grant from the Netherlands Organization for Scientific Research: NWO-VIDI-452-14-006. SG was supported by a grant from the Netherlands Organization for Scientific Research: NWO-VENI-451-15-024. LT was supported by grants from The Dutch Research Council (NWO): NWO-016.VIDI.185.017.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.621547/full#supplementary-material

REFERENCES

Ajzen, I. (1991). The theory of planned behavior. *Organ. Behav. Hum. Decis. Process.* 50, 179–211. doi: 10.1016/0749-5978(91)90020-T

Anca, M., Hanea, N., Gabriela, F., Bedford, T., and French, S. (2021). *Expert Judgement in Risk and Decision Analysis*. Berlin: Springer. doi: 10.1007/978-3-030-46474-5

Anderson, M. S., Horn, A. S., Risbey, K. R., Ronning, E. A., De Vries, R., and Martinson, B. C. (2007). What do mentoring and training in the responsible conduct of research have to do with scientists' misbehavior? Findings from a national survey of NIH-funded scientists. *Acad. Med.* 82, 853–860. doi: 10.1097/ACM.0b013e31812f764c

Bayarri, M., and Mayoral, A. (2002). Bayesian design of "successful" replications. *Am. Stat.* 56, 207–214. doi: 10.1198/000313002155

Bousquet, N. (2008). Diagnostics of prior-data agreement in applied Bayesian analysis. *J. Appl. Stat.* 35, 1011–1029. doi: 10.1080/02664760802192981

Box, G. E. (1980). Sampling and Bayes' inference in scientific modelling and robustness. *J. R. Stat. Soc. Ser. A* 143, 383–430. doi: 10.2307/2982063

Brown, M. E., and Treviño, L. K. (2006). Ethical leadership: a review and future directions. *Leadersh. Q.* 17, 595–616. doi: 10.1016/j.leaqua.2006.10.004

Callaway, E. (2011). *Report Finds Massive Fraud at Dutch Universities*. Berlin: Nature Publishing Group. doi: 10.1038/479015a

Evans, M., and Moshonov, H. (2006). Checking for prior-data conflict. *Bayesian Anal.* 1, 893–914. doi: 10.1214/06-BA129

Fanelli, D. (2009). How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PLoS One* 4:e5738. doi: 10.1371/journal.pone.0005738

Fanelli, D. (2010). Do pressures to publish increase scientists' bias? An empirical support from US States Data. *PLoS One* 5:e10271. doi: 10.1371/journal.pone.0010271

Fragoso, T. M., Bertoli, W., and Louzada, F. (2018). Bayesian model averaging: a systematic review and conceptual classification. *Int. Stat. Rev.* 86, 1–28. doi: 10.1111/insr.12243

Gopalakrishna, G., ter Riet, G., Vink, G., Stoop, I., Wicherts, J., and Bouter, L. M. (2021). Prevalence of questionable research practices, research misconduct and their potential explanatory factors: a survey among academic researchers in the Netherlands. *MetaArXiv* [Preprint]. doi: 10.31222/osf.io/vk9yt

Haven, T., Tijdink, J., Martinson, B., Bouter, L., and Oort, F. (2021). Explaining variance in perceived research misbehavior: results from a survey among academic researchers in Amsterdam. *Res. Integ. Peer Rev.* 6, 1–8. doi: 10.1186/s41073-021-00110-w

Heitman, E., Olsen, C. H., Anestidou, L., and Bulger, R. E. (2007). New graduate students' baseline knowledge of the responsible conduct of research. *Acad. Med.* 82, 838–845. doi: 10.1097/ACM.0b013e31812f7956

Hofmann, B., Myhr, A. I., and Holm, S. (2013). Scientific dishonesty—a nationwide survey of doctoral students in Norway. *BMC Med. Ethics* 14:3. doi: 10.1177/1556264615599686

Hon, C. K., Sun, C., Xia, B., Jimmieson, N. L., Way, K. A., and Wu, P. P.-Y. (2021). Applications of Bayesian approaches in construction management research: a systematic review. *Engineering, Construction and Architectural Management.* doi: 10.1108/ECAM-10-2020-0817

JASP-Team (2018). *JASP (Version 0.9. 0.1)[Computer software]*.

John, L. K., Loewenstein, G., and Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychol. Sci.* 23, 524–532. doi: 10.1177/0956797611430953

Johnson, S. R., Tomlinson, G. A., Hawker, G. A., Granton, J. T., and Feldman, B. M. (2010a). Methods to elicit beliefs for Bayesian priors: a systematic review. *J. Clin. Epidemiol.* 63, 355–369. doi: 10.1016/j.jclinepi.2009.06.003

Johnson, S. R., Tomlinson, G. A., Hawker, G. A., Granton, J. T., Grosbein, H. A., and Feldman, B. M. (2010b). A valid and reliable belief elicitation method for Bayesian priors. *J. Clin. Epidemiol.* 63, 370–383. doi: 10.1016/j.jclinepi.2009.08.005

Kalichman, M. W., and Plemmons, D. K. (2015). Research agenda: the effects of responsible-conduct-of-research training on attitudes. *J. Empir. Res. Hum. Res. Ethics* 10, 457–459. doi: 10.1177/1556264615575514

Kass, R. E., and Raftery, A. E. (1995). Bayes factors. *J. Am. Stat. Assoc.* 90, 773–795.

König, C., and van de Schoot, R. (2017). Bayesian statistics in educational research: a look at the current state of affairs. *Educ. Rev.* 70, 1–24. doi: 10.1080/00131911.2017.1350636

Lek, K., and Van De Schoot, R. (2018). Development and evaluation of a digital expert elicitation method aimed at fostering elementary school teachers' diagnostic competence. *Front. Educ.* 3:82. doi: 10.3389/feduc.2018.00082

Lek, K., and Van De Schoot, R. (2019). How the choice of distance measure influences the detection of prior-data conflict. *Entropy* 21:446. doi: 10.3390/e21050446

Levelt, W. J. M., Drenth, P., and Noort, E. (eds) (2012). *Flawed Science: The Fraudulent Research Practices of Social Psychologist Diederik Stapel*. Tilburg: Commissioned by the Tilburg University, University of Amsterdam, and the University of Groningen.

Markowitz, D. M., and Hancock, J. T. (2014). Linguistic traces of a scientific fraud: the case of Diederik Stapel. *PLoS One* 9:e105937. doi: 10.1371/journal.pone.0105937

Martinson, B. C., Anderson, M. S., and De Vries, R. (2005). Scientists behaving badly. *Nature* 435:737. doi: 10.1038/435737a

Oakley, J., and O'Hagan, A. (2010). *SHELF: The Sheffield Elicitation Framework (version 2.0)*. Sheffield: University of Sheffield.

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., et al. (2006). *Uncertain Judgements: Eliciting Experts' Probabilities*. New York, NY: John Wiley & Sons. doi: 10.1002/0470033312

Parker, T. H., Forstmeier, W., Koricheva, J., Fidler, F., Hadfield, J. D., Chee, Y. E., et al. (2016). Transparency in ecology and evolution: real problems, real solutions. *Trends Ecol. Evol.* 31, 711–719. doi: 10.1016/j.tree.2016.07.002

Prelec, D. (2004). A Bayesian truth serum for subjective data. *Science* 306, 462–466. doi: 10.1126/science.1102081

Sijtsma, K. (2016). Playing with data—or how to discourage questionable research practices and stimulate researchers to do things right. *Psychometrika* 81, 1–15. doi: 10.1007/s11336-015-9446-0

Sijtsma, K., Veldkamp, C. L., and Wicherts, J. M. (2016). Improving the conduct and reporting of statistical analysis in psychology. *Psychometrika* 81, 33–38. doi: 10.1007/s11336-015-9444-2

Smid, S. C., McNeish, D., Miočević, M., and van de Schoot, R. (2020). Bayesian versus frequentist estimation for structural equation models in small sample contexts: a systematic review. *Struct. Equ. Model. Multidiscipl. J.* 27, 131–161. doi: 10.1080/10705511.2019.1577140

Sonneveld, H., Yerkes, M. A., and Van de Schoot, R. (2010). *PhD Trajectories and Labour Market Mobility: A Survey of Recent Doctoral Recipients at Four Universities in the Netherlands*. Utrecht: Nederlands Centrum voor de Promotieopleiding/IVLOS.

Terry, D. J., and Hogg, M. A. (1996). Group norms and the attitude-behavior relationship: a role for group identification. *Pers. Soc. Psychol. Bull.* 22, 776–793. doi: 10.1177/0146167296228002

Tijdink, J. K., Verbeke, R., and Smulders, Y. M. (2014). Publication pressure and scientific misconduct in medical scientists. *J. Empir. Res. Hum. Res. Ethics* 9, 64–71. doi: 10.1177/1556264614552421

van de Schoot, R., Depaoli, S., King, R., Kramer, B., Märtens, K., Tadesse, M. G., et al. (2021a). Bayesian statistics and modelling. *Nat. Rev. Methods Prim.* 1, 1–26. doi: 10.1038/s43586-021-00017-2

van de Schoot, R., Griffioen, E., and Winter, S. D. (2021b). "Dealing with imperfect elicitation results," in *Expert Judgement in Risk and Decision Analysis*, eds A. M. Hanea, G. F. Nane, T. Bedford, and S. French (Cham: Springer), 401–417. doi: 10.1007/978-3-030-46474-5_18

van de Schoot, R., Winter, S. D., Ryan, O., Zondervan-Zwijnenburg, M., and Depaoli, S. (2017). A systematic review of Bayesian articles in psychology: the last 25 years. *Psychol. Methods* 22:217. doi: 10.1037/met0000100

Veen, D., Stoel, D., Schalken, N., Mulder, K., and van de Schoot, R. (2018). Using the data agreement criterion to rank experts'. *Beliefs Entr.* 20:592. doi: 10.3390/e20080592

Veen, D., Stoel, D., Zondervan-Zwijnenburg, M., and van de Schoot, R. (2017). Proposal for a five-step method to elicit expert judgement. *Front. Psychol.* 8:2110. doi: 10.3389/fpsyg.2017.02110

Verhagen, J., and Wagenmakers, E.-J. (2014). Bayesian tests to quantify the result of a replication attempt. *J. Exp. Psychol. Gen.* 143:1457. doi: 10.1037/a0036731

Waldman, I. D., and Lilienfeld, S. O. (2016). Thinking about data, research methods, and statistical analyses: commentary on Sijtsma's (2014)"Playing with Data". *Psychometrika* 81, 16–26. doi: 10.1007/s11336-015-9447-z

Yerkes, M. A., Van de Schoot, R., and Sonneveld, H. (2010). Who are the job seekers? Explaining unemployment among doctoral recipients. *Int. J. Doctor. Stud.* 7, 153–166. doi: 10.28945/1573

Yukl, G., Mahsud, R., Hassan, S., and Prussia, G. E. (2013). An improved measure of ethical leadership. *J. Leadersh. Organ. Stud.* 20, 38–48. doi: 10.1177/1548051811429352

Zondervan-Zwijnenburg, M., van de Schoot-Hubeek, W., Lek, K., Hoijtink, H., and van de Schoot, R. (2017). Application and evaluation of an expert judgment elicitation procedure for correlations. *Front. Psychol.* 8:90. doi: 10.3389/fpsyg.2017.00090

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership