Bayesian multilevel structural equation modeling: An investigation into robust prior distributions.

Sara van Erp¹

William J. Browne²

¹Utrecht University

²University of Bristol

August 25, 2020

Abstract

Bayesian estimation of multilevel structural equation models (MLSEMs) offers advantages in terms of sample size requirements and computational feasibility, but does require careful specification of the prior distribution especially for the random effects variance parameters. The traditional “non-informative” conjugate choice of an inverse-Gamma prior with small hyperparameters has been shown time and again to be problematic. In this paper, we investigate alternative, more robust prior distributions. In contrast to multilevel models without latent variables, MLSEMs have multiple random effects variance parameters, both for the multilevel structure and for the latent variable structure. It is therefore even more important to construct reasonable priors for these parameters. We find that, although the robust priors outperform the traditional inverse-Gamma prior, their hyperparameters do require careful consideration.

Keywords: Bayesian, robust priors, multilevel models.

1 Introduction

Multilevel structural equation modeling (MLSEM) has become increasingly popular to test complex theories in samples that are hierarchically structured (Muthén, 1994). While SEM allows researchers to control for measurement error due to the fact that only a finite number of items are sampled, multilevel analysis takes into account sampling error due to the fact that only a finite number of individuals is sampled. Combined, MLSEM offers a powerful tool that is being used throughout educational, psychological, and sociological research.

Traditionally, maximum likelihood (ML) algorithms have been used to estimate MLSEMs. However, in the case of categorical indicators, such algorithms require multidimensional numerical integration which quickly becomes computationally infeasible (Asparouhov and Muthén, 2007). Weighted least squares (WLS) algorithms avoid the high dimensional integration, but are restricted to random intercept models (Asparouhov and Muthén, 2012). Additionally, both ML and WLS rely on the frequentist asymptotical framework and thus

require a substantial number of groups to obtain admissible and accurate parameter estimates (Hox and Maas, 2001). For example, Meuleman and Billiet (2009) concluded that for relatively simple MLSEMs, at least 60 groups are needed to detect large structural effects at the between level.

Due to these problems with frequentist estimation methods, Bayesian estimation of MLSEMs has become increasingly popular. In a Bayesian analysis, a prior distribution is specified for each of the parameters in the model. Combined with the likelihood of the data this results in a posterior distribution, which is generally obtained through Markov Chain Monte Carlo (MCMC) sampling. The Bayesian approach offers several advantages compared to the frequentist framework. First, MCMC sampling does not require multi-dimensional numerical integration enabling complex models with categorical indicators to be estimated (see e.g., Asparouhov and Muthén, 2012). Second, estimates of transformed parameters, such as indirect structural effects, can be easily obtained by transforming the MCMC samples and credible intervals are obtained automatically (Yuan and MacKinnon, 2009). Credible intervals are the Bayesian alternative of confidence intervals, but unlike confidence intervals do not rely on normality assumptions or asymptotical theory. Third, through the prior distribution, the problem of variances that are estimated to be negative (i.e., Heywood cases) can be solved; so long as the prior on the variance has zero mass at negative values, the estimate can never become negative. Moreover, prior information about parameters in the model can be included in the prior distribution. This information might be based on previous investigations or expert knowledge. Finally, it has been shown that Bayesian MLSEM can provide accurate estimates with a smaller number of groups compared to frequentist estimation (Hox et al., 2012; Zitzmann et al., 2016).

A main focus of the Bayesian literature on multilevel models is the specification of the prior for the random effects variance parameter. The random effects variance parameter decides how much pooling across observations takes place. A random effects variance parameter equal to zero implies complete pooling of observations, i.e., that there are no random

effects and thus that we fit the same regression line for all higher level units. As the value for the random effects variance increases, the amount of pooling decreases until there is no pooling at all which means that each higher level unit has its own regression line, a solution that will generally overfit the data. The prior for the random effects variance parameter determines where the model lies on this continuum from complete pooling to no pooling. Historically, inverse-Gamma priors were popular choices for variance parameters, including those at the between level, since they are conjugate and thus result in a posterior that is an inverse-Gamma distribution as well. Specifically, an inverse-Gamma(ϵ , ϵ) prior with small ϵ has often been used as a default uninformative prior distribution. However, it is now well-known in the literature that for random effects variances, this specification is actually very informative and highly sensitive to the exact choice of the hyperparameters (see e.g., Gelman, 2006; Klein and Kneib, 2016; Lunn et al., 2009). Moreover, as we will show later, the inverse-Gamma prior does not have support for random effects variances equal to zero and will therefore always favor more complex solutions with less pooling. A second often used class of default priors for variance parameters are uniform improper priors. This is in line with the “objective” Bayesian approach (Berger, 2006) since these priors will, in simple models, result in the same estimates as classical ML estimation. Various improper uniform prior options exist. The most straightforward approach is to specify a uniform prior directly on the variance, i.e., $p(\sigma^2) \propto 1$. A second option, advocated by Berger (2006), Berger and Strawderman (1996), and Gelman (2006), is to specify a uniform prior on the standard deviation, i.e., $p(\sigma^2) \propto \sigma^{-1}$. Finally, a uniform prior can be specified on the logarithm of the variance, i.e., $p(\sigma^2) \propto \sigma^{-2}$. The main issue with specifying improper priors for random effects variances is that the resulting posterior might be improper as well in cases where there is a limited amount of information in the data on the higher level (Berger, 2006; Gelman, 2006). This has been shown empirically in a SEM where improper priors resulted in lower convergence rates (van Erp et al., 2018). Note that the inverse-Gamma (ϵ , ϵ) prior approximates the improper prior $p(\sigma^2) \propto \sigma^{-2}$ as the ϵ goes to zero. As a result, the inverse-Gamma

prior can also lead to an unstable MCMC sampler.

Recently, several weakly-informative prior distributions have been proposed as alternative priors for random effects variances. Compared to the uniform improper priors, weakly-informative priors perform some shrinkage of the random effects variance parameter towards zero, i.e., the complete pooling situation, to avoid overfitting. For example, Gelman (2006) and Polson and Scott (2012) propose to use the half-Cauchy prior, which is proper and thus avoids the issue of uniform improper priors. Additionally, the half-Cauchy prior has heavy tails and is therefore considered robust to misspecification of the scale. Simpson et al. (2017) discuss a general approach to construct so-called “scale dependent” priors, which can be straightforwardly applied to random effects variances. Although some studies exist that compare some of these priors in general multilevel models (see e.g., Klein and Kneib, 2016), they have yet to be investigated in the context of MLSEM. Studies investigating prior choice in MLSEM generally focus on conjugate prior distributions that vary in the degree of informativeness (Depaoli and Clifton, 2015; Helm, 2018; Zitzmann et al., 2016). These studies show that informative priors, when correctly specified, perform best. However, in reality we do not know the true population value and therefore require prior distributions that are robust against possible prior misspecification. Moreover, MLSEMs have multiple random effects variance parameters, both for the multilevel structure and for the latent variable structure. It is therefore even more important to construct reasonable and robust priors for these parameters.

The goal of this study is to compare robust prior distributions for random effects variances in MLSEM. We will compare default specifications of the robust priors as well as informative (in)accurate specifications. The outline of this paper is as follows: in Section 2 we describe the categorical MLSEM used throughout this paper along with the motivating application from education. The robust prior distributions are presented and compared in Section 3 and Section 4 presents two simulation studies to investigate the performance of the priors, followed by a discussion in Section 5.

2 Bayesian doubly latent categorical multilevel model

2.1 Model

The MLSEM used in this investigation is based on the doubly latent multilevel model (DLMM) proposed in Marsh et al. (2009) to estimate contextual effects. A well-known application of the DLMM is to estimate the “Big-Fish-Little-Pond-Effect” (BFLPE) that predicts that at the individual (or within) level, student achievement has a positive effect on academic self-concept, whereas at the between-level, school-average achievement has a negative effect on academic self-concept. The term “doubly latent” arises from the fact that the model takes into account both measurement and sampling error. Measurement error, which is a consequence of the fact that only a finite number of items are sampled is controlled for by specifying a measurement model for each of the factors. Sampling error, on the other hand, is the result of sampling only a finite number of individuals and is taken into account by including random effects in the model. The model used is depicted in Figure 1. We use a similar representation as in Marsh et al. (2009).

[FIGURE 1 HERE]

We assume that academic self-concept is measured by six binary indicators y for individual i in group j measured on item k ¹ We assume an underlying continuous latent response variable $\tilde{y}_{ikj}$ that is linked to the binary observed responses y_{ikj} as follows:

$$y_{ikj} = \begin{cases} 0 & \text{if } -\infty < \tilde{y}_{ikj} \leq 0 \\ 1 & \text{if } 1 < \tilde{y}_{ikj} \leq \infty \end{cases} \quad (1)$$

The measurement model can then be defined for the continuous latent responses $\tilde{y}_{ikj}$ at

¹Note that, generally, empirical data investigating the BFLPE, such as the PISA OECD (2007) data uses 4-point Likert scale indicators for academic self-concept. However, we assume binary indicators instead to keep the computation time feasible in our simulation studies.

the within level:

$$\tilde{y}_{ikj} = \mu_{kj} + \lambda_k^W \eta_{ij}^W + \epsilon_{ikj}^W, \quad (2)$$

where μ_{kj} denotes the intercept for item k in group j and is fixed to 0 for identification, λ_k^W is the loading for item k at the within level, η_{ij}^W is the factor score for individual i in group j , and ϵ_{ikj}^W is the residual at the within level for individual i in group j measured on item k . Throughout this paper, we assume a logistic distribution for the residuals at the within level, i.e., $\epsilon_{ikj}^W \sim \text{logistic}(0, \sigma_{W,k}^2)$ and we fix the variance to that of the standard logistic distribution, i.e., $\sigma_{W,k}^2 = \pi^2/3$ for identification².

We focus on the random intercept model, such that the measurement model at the between level is defined as:

$$\mu_{kj} = \mu_k + \lambda_k^B \eta_j^B + \epsilon_{kj}^B, \quad (3)$$

where μ_k reflects the overall intercept for item k , λ_k^B is the loading for item k at the between level, η_j^B is the factor score at the between level for group j , and ϵ_{kj}^B denotes the residual at the between level for item k in group j , which we assume to be normally distributed, i.e., $\epsilon_{kj}^B \sim N(0, \sigma_{B,k}^2)$.

The factor scores on self-concept are predicted by the academic achievement scores x_{ij} leading to the following structural model:

$$\eta_{ij}^W \sim N(\beta^W x_{ij}, \omega_W^2) \quad (4)$$

$$\eta_j^B \sim N(\alpha + \beta^B x_{b,j}, \omega_B^2) \quad (5)$$

Here, β^W represents the effect of achievement on self-concept at the within level, whereas β^B represents the effect of achievement on self-concept at the between level. To identify

²Alternatively, a probit link function can be used, in which case $\epsilon_{ikj}^W \sim N(0, 1)$.

the latent variables, we restrict the mean, i.e., $\alpha = 0$ and one loading at each level³, i.e., $\lambda_1^B = \lambda_1^W = 1$. As recommended by Asparouhov and Muthén (2019), we use latent mean centering which estimates the group mean achievement in each group j , i.e., $x_{b,j}$ to take into account measurement error. The parameter of interest is the contextual effect, which equals $\beta^B - \beta^W$.

An important quantity in multilevel models is the variance partition coefficient (VPC; Goldstein et al., 2002), which is a measure of the correlation between students in the same school. In multilevel SEM, the VPC ρ_k for item k is defined as (see e.g., Muthén, 1994):

$$\rho_k = \frac{\lambda_k^{2,B} \omega_B^2 + \sigma_{B,k}^2}{(\lambda_k^{2,B} \omega_B^2 + \sigma_{B,k}^2) + (\lambda_k^{2,W} \omega_W^2 + \sigma_{W,k}^2)} \quad (6)$$

2.2 Prior distributions

In a Bayesian analysis, prior distributions need to be specified for each parameter in the model, i.e., $p(\lambda_k, \sigma_{B,k}^2, \beta^W, \omega_W^2, \beta^B, \omega_B^2)$. The focus of this paper is on priors for the random effects variances $\sigma_{B,k}^2, \omega_W^2$, and ω_B^2 and the next section discusses the priors we consider for these parameters in detail. For the other parameters in the model, we use the following independent, uninformative or weakly informative prior distributions throughout this paper: $p(\lambda_k) \sim \text{Normal}(0, 5)$, for $k = 1, \dots, 6$, $p(\beta^W) \sim \text{Normal}(0, 1)$, $p(\beta^B) \sim \text{Normal}(0, 1)$. Note that other choices are possible and it is always important to investigate the sensitivity of the results to the specification of the prior distribution (van Erp et al., 2018).

3 Robust prior distributions for random effects variances

The problems with the traditional inverse-Gamma family and the improper uniform priors for random effects variances have inspired many researchers to investigate alternative,

³These restrictions are, in theory, sufficient to identify the model. However, it is generally recommended to additionally impose cross-level invariance, i.e., $\lambda^W = \lambda^B$ to improve interpretability of the factors at both levels and to avoid estimation issues (Jak, 2018).

more robust prior distributions. Here, we focus on three such alternative classes of priors: 1) Student’s t priors; 2) F priors; and 3) Scale dependent priors. Note that some priors are specified on the variances while others are specified on the standard deviations. Throughout this section, we will use θ to denote standard deviations and θ^2 to denote variance parameters.

3.1 Student’s t family

A widespread proposal for the prior on random effects is to specify a proper, weakly-informative half- t distribution on the standard deviation (see for example, Gelman, 2006; Polson and Scott, 2012). The probability density function of the half- t distribution is given by:

$$p(\theta; \nu, \sigma_0^2) = \frac{2\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\nu\pi}\sigma_0^2} \left(1 + \frac{\theta^2}{\nu\sigma_0^2}\right)^{-\frac{\nu+1}{2}}, \quad (7)$$

where σ_0^2 denotes the scale and ν the degrees of freedom. Different choices for the scale parameter σ_0^2 are discussed in Subsection 3.4. For the degrees of freedom ν , smaller values result in heavier tails. For $\nu = \infty$, we obtain the half-normal distribution, which has been proposed, among others, by Frühwirth-Schnatter and Wagner (2010). Roos et al. (2011) have shown that the half-normal prior indeed results in estimates of the random effects variances that are less sensitive to the chosen hyperparameters compared to the inverse-Gamma prior. Nevertheless, the half-normal distribution still has very light tails and will therefore be less robust to variance parameters that are larger than expected. Therefore, a more robust proposal in the literature is to specify the degrees of freedom to be smaller. Gelman (2006) notes the special case of the half-Cauchy prior with $\nu = 1$ as a reasonable default option. The half-Cauchy prior has been further investigated by Polson and Scott (2012), who have shown that it has excellent frequentist risk properties. There are some reports, however, that the heavy tails of the half-Cauchy prior can lead to numerical difficulties in MCMC sampling (Ghosh et al., 2018; Piironen and Vehtari, 2015). This issue generally arises in situations when there is not enough information in the data such that the parameter is weakly or non-

identified. The heavy tails of the (half-)Cauchy prior do not perform enough shrinkage in such cases to identify the parameter.

3.2 F family

A popular family of distributions for variance parameters in the Bayesian literature is the family of F priors, also known as scaled Beta2 or generalized beta prime priors (Fúquene et al., 2014; Mulder and Pericchi, 2018; Pérez et al., 2017; Polson and Scott, 2012). The probability density function of the F prior is given by (Mulder and Pericchi, 2018)⁴:

$$p(\theta^2; \nu_1, \nu_2, \sigma_0^2) = \frac{\Gamma(\frac{\nu_2 + \nu_1}{2})}{\Gamma(\frac{\nu_1}{2})\Gamma(\frac{\nu_2}{2})(\sigma_0^2)^{\frac{\nu_1}{2}}} (\theta^2)^{\frac{\nu_1}{2}-1} (1 + \theta^2/\sigma_0^2)^{-\frac{\nu_1 + \nu_2}{2}}, \quad (8)$$

where $\sigma_0^2 > 0$ denotes a scale parameter, the degrees of freedom $\nu_1 > 0$ determines the behaviour around zero, and the degrees of freedom $\nu_2 > 0$ influences the tail behaviour in a similar manner as the degrees of freedom for Student's t distribution. Specifically, $\nu_2 = 2$ results in similar tails as a Cauchy distribution. With regard to the behaviour at zero, $\frac{\nu_1}{2} > 1$ results in a density that is zero at the origin, for $\frac{\nu_1}{2} = 1$ the density is bounded at the origin, and for $\frac{\nu_1}{2} < 1$ the density goes to infinity at the origin.

Interestingly, if $p(\theta^2) \sim F(\nu_1 = 1, \nu_2, \sigma_0^2)$, then the prior on the standard deviation $p(\theta)$ corresponds to a half-Student's t distribution with degrees of freedom equal to ν_2 and scale $\sqrt{\sigma_0^2}$. Thus, the F family can be seen as a generalization of the Student's t family. Moreover, if $p(\theta^2) \sim F(\nu_1, \nu_2, \sigma_0^2)$, then the prior on the precision $h = \frac{1}{\theta^2}$ will also be an F distribution, with a scale of $\frac{1}{\sigma_0^2}$ (i.e., the reciprocity property).

To implement the F prior, the following Gamma mixture of inverse-Gamma parametrization can be used:

$$p(\theta^2) \sim \text{IG}(\frac{\nu_2}{2}, \tau^2) \quad (9)$$

$$p(\tau^2) \sim \text{G}(\frac{\nu_1}{2}, \sigma_0^{-2}), \quad (10)$$

⁴This corresponds to the parametrization used in Pérez et al. (2017) when $p = \frac{\nu_1}{2}$, $q = \frac{\nu_2}{2}$, and $b = \sigma_0^2$

which results in the F prior when integrating out τ^2 .

3.3 Scale dependent family

Simpson et al. (2017) have proposed the use of penalized complexity priors as general default priors. The basic idea of these priors is that deviations from a simpler base model are penalized thereby avoiding a model that overfits. This characteristic is especially useful in the context of random effects variance parameters which are generally weakly identified due to the small amount of information available at the higher level. Simpson et al. (2017) derive the penalized complexity prior for the precision of a random effect as follows: first, define the base model. For example, for the random effects component $\epsilon_{kj}^B \sim N(0, \sigma_{B,k}^2)$ in (3), the base model, which is the simplest model in its class, would correspond to $\sigma_{B,k}^2 = 0$ or the absence of random effects. Next, the distance of the flexible extension of the base model, i.e., $\epsilon_{kj}^B \sim N(0, \sigma_{B,k}^2)$ to the base model is computed based on the Kullback-Leibler divergence (KLD). Simpson et al. (2017) define the distance between the two models with densities f and g as: $d(f||g) = \sqrt{2KLD(f||g)}$. Then a prior is specified for the distance d such that deviations from the base model are penalised. By assuming a constant rate of penalisation, Simpson et al. (2017) specify an exponential prior for the distance $p(d) = \lambda \exp(-\lambda d)$, although they note that it is possible to relax this assumption if needed. For example, when interest lies in variable selection, the exponential tails are too light and a heavier tailed prior can be more sensible, for example a half-Cauchy prior on the distance will recover the well-known horseshoe prior. Transforming the prior on the distance to the original space leads to a type-2 Gumbel prior for the precision or, equivalently, an exponential distribution with scale σ_0^2 for the standard deviation⁵:

$$p(\theta; \sigma_0^2) = \frac{1}{\sigma_0^2} \exp\left(-\frac{1}{\sigma_0^2} \theta\right) \quad (11)$$

Lower values for σ_0^2 will result in an increased penalisation for deviating from the base

⁵We parametrise the exponential distribution in terms of the scale σ_0^2 for consistency, but generally it is defined using the rate parameter λ , which is the inverse of the scale, i.e., $\lambda = \frac{1}{\sigma_0^2}$.

model. The choice of σ_0^2 will be discussed in Subsection 3.4.

3.4 Specification of the hyperparameters

Prior	Parameter	Scale	Tail behaviour	Behaviour origin
half-Normal	θ	σ_0^2	$\nu = \infty$	-
half-Student's t	θ	σ_0^2	ν	-
half-Cauchy	θ	σ_0^2	$\nu = 1$	-
F	θ^2	σ_0^2	ν_2	ν_1
Exponential	θ	σ_0^2	-	-

Note. θ denotes a standard deviation parameter and θ^2 denotes a variance parameter.

Table 1: Overview of robust prior distributions for random effects variances

An overview of the different priors considered is provided in Table 1. It is clear from Table 1 that the priors vary in flexibility. Each prior has a parameter that determines the scale, or how spread out the prior is. In addition, the half-Student's t and F priors have a parameter that influence the tail behaviour of these priors. Specifically, this degrees of freedom hyperparameter determines how heavy the tails are and thus how much prior mass is put on extreme values. In general, a prior with heavier tails is more robust since extreme values will not be shrunk towards zero as much compared to a prior with thinner tails. For the half-Student's t prior, we consider the special cases of the half-normal prior ($\nu = \infty$) and the half-Cauchy prior ($\nu = 1$), but other values for ν can be used, with higher degrees of freedom leading to thinner tails. For the F prior, we will consider $\nu_2 = 0.5$ throughout this paper to obtain tails that are heavier than the Cauchy. The larger ν_2 , the thinner the tails will be. Additionally, the F prior has a parameter that determines the behaviour at the origin. This degrees of freedom hyperparameter, ν_1 , is set to 1 so that the density goes to infinity at zero. Although other values for ν_1 are possible, it is important to ensure that $\frac{\nu_1}{2} \leq 1$ so that the prior has support at zero.

Throughout this paper, we consider and compare various settings for the scales in each prior (σ_0^2). A general informal recommendation for a default prior is a half-Student's t prior

with scale equal to 1 (see, for example, Stan Development Team, 2015, the section on priors for scale parameters in hierarchical models). Therefore, we use $\sigma_0^2 = 1$ throughout this paper as default prior specification. Additionally, if prior information is available, this can be included in the priors as follows. Specify a value for the standard deviation C and a threshold α such that the probability $P(\theta > C) = \alpha$. This equation can then be solved for θ using the cumulative distribution function. Section 4.2 describes which values for C and α we will consider in the simulation study and provides some examples of the resulting prior scales θ .

3.5 Comparison of the priors

In this section, we compare the probability density functions of the various priors to better understand their behaviour. Figure 2 shows the densities for the standard deviation. We use a scale of 1 for the robust priors and hyperparameters equal to 0.1 for the inverse-Gamma prior. The main difference between the inverse-Gamma prior and the more robust alternatives is that the inverse-Gamma prior has zero density at zero. As a result, the prior favors standard deviations that are away from zero which is problematic especially for standard deviations of random effects which are often very close to zero. The inverse-Gamma prior thus automatically favors more complex models with less pooling. The other priors do have prior probability around standard deviations equal to zero and are therefore more in line with what we might expect in reality and are less likely to overfit.

[FIGURE 2 HERE]

There are multiple standard deviations in the model for which a prior needs to be specified, specifically: $\sigma_{B,k}$, ω_W , and ω_B . Together with the factor loadings at the between and within level, λ_k^B and λ_k^W , these standard deviations play a role in computing the VPC ρ_k for each item (see equation (6)). As a result, specifying a prior on the standard deviations and factor loadings implies a certain prior on the VPC. This prior is considered in Figure 3. Recall from Section 2.1 that we fixed the loading of one item to 1 for identification purposes. The left figure shows the VPC of the restricted item while the figure on the right shows

the VPC of the unrestricted items. It is clear that this restriction has consequences for the implied prior of the VPC. This is due to the fact that the VPC depends on the loadings (see equation (6)). Specifically, we see that some priors put less probability on a VPC close to 1 for the restricted item compared to the free items. This is most pronounced for the priors with the thinnest tails, i.e., the half-normal and exponential priors. Furthermore, the implied priors on the VPC are not symmetrical. This is due to the fact that we assume a standard logistic distribution for the residuals at the within level. As a result, $\sigma_{W,k}^2$, which arises in the denominator of equation (6) is fixed as well.

[FIGURE 3 HERE]

4 Simulation studies

In order to investigate the performance of the robust priors for the random effects variances, we conduct two simulation studies. In the first study, we vary the population values for various parameters while keeping the number of groups fixed. We focus on comparing the traditional priors for variance parameters to a default setting of the more robust alternatives. In the second study, we use the same generated data sets as in the first study, but now we focus on informative prior settings with either correct or incorrect information to determine how robust the priors are to misspecifications. In both studies we generate data according to the model in Section 2 using the following population values:

1. ***Population values for the variance parameters.*** The population value for the latent variable variance parameters ω^2 is equal to 0.01 or 1. The sensitivity of the inverse Gamma priors to the hyperparameter choice arises especially in situations in which the random effects variance is close to zero (Gelman, 2006). Depending on which value is specified for each variance parameter, the VPC will vary. Hedges and Hedberg (2007) provide an overview of VPC values for educational achievement based on longitudinal surveys conducted in the United States. They found VPCs ranging from .03 to .3. These values are in line with VPCs used in other simulation studies

of multilevel SEMs (Depaoli and Clifton, 2015; Helm, 2018; Zitzmann et al., 2016). Therefore, we consider several combinations of the population values for the variance parameters that result in VPCs within this range. Note that the VPC depends on the following parameters: $VPC = f(\lambda_k^{2,B}, \omega_B^2, \sigma_{B,k}^2, \lambda_k^{2,W}, \omega_W^2, \sigma_{W,k}^2)$ (see (6)). We fix the population values for $\lambda_k^{2,B}$ and $\lambda_k^{2,W}$ to 1, and $\sigma_{W,k}^2$ to $\frac{\pi^2}{3}$ for all items. We consider different combinations of population values for the variances of the latent variables, ω_B^2 and ω_W^2 and subsequently compute $\sigma_{B,k}^2$ given these values and a VPC of .03 or .3. An overview of the different population values for the variance parameters and the resulting VPC is presented in Table 2.

2. ***Population values for the contextual effect.*** When using group-mean centering, the contextual effect is defined as $\beta^B - \beta^W$. Marsh et al. (2009) provide three different standardized effect size measures that can be interpreted similarly to Cohen’s d . They recommend to use either one of the two more conservative effect sizes. Here, we rely on the second effect size in Marsh et al. (2009) to define the population values for the regression coefficients at the between and within level. This effect size standardizes the contextual effect with respect to the total variance of self-concept at the within level, i.e., η_W , and is defined as:

$$ES = \frac{2\beta^C \sigma_x^B}{\sigma_x^W (\beta^W)^2 + \omega^W}, \quad (12)$$

where β^C denotes the contextual effects which is equal to $\beta^B - \beta^W$ in the case of group-mean centering, and σ_x^B and σ_x^W denote the standard deviations of the predictor at the between and within level, respectively. We consider a small negative standardized contextual effect and no contextual effect, i.e., $ES = c(-0.2, 0)$. We choose these values such that we are able to investigate the power to detect a small contextual effect as well as the type 1 error rate. We fix β^W to 0.2 and σ_x^B and σ_x^W to 1 and compute the value for β^B to obtain the required standardized effect size. The resulting

values are shown in Table 2.

Setting	ω_B^2	ω_W^2	$\sigma_{B,k}^2$	VPC	ES	β^B
1	0.01	0.01	0.11	0.03	0.00	0.20
2	0.01	0.01	1.42	0.30	0.00	0.20
3	1.00	0.01	1.10	0.03	0.00	0.20
4	1.00	0.01	2.41	0.30	0.00	0.20
5	0.01	1.00	0.14	0.03	0.00	0.20
6	0.01	1.00	1.85	0.30	0.00	0.20
7	1.00	1.00	1.13	0.03	0.00	0.20
8	1.00	1.00	2.84	0.30	0.00	0.20
9	0.01	0.01	0.11	0.03	-0.20	0.19
10	0.01	0.01	1.42	0.30	-0.20	0.19
11	1.00	0.01	1.10	0.03	-0.20	0.19
12	1.00	0.01	2.41	0.30	-0.20	0.19
13	0.01	1.00	0.14	0.03	-0.20	0.10
14	0.01	1.00	1.85	0.30	-0.20	0.10
15	1.00	1.00	1.13	0.03	-0.20	0.10
16	1.00	1.00	2.84	0.30	-0.20	0.10

Table 2: Overview population values for the variance parameters and variance partition coefficient.

Combinations of the various population values result in 16 different conditions (2 values $\omega_B^2 \times 2$ values $\omega_W^2 \times 2$ values VPC $\times 2$ values ES). We consider these 16 conditions for a balanced design with 20 groups⁶ and a sample size of 20 within each group. We use 20 groups based on Hox et al. (2012) who concluded that 20 groups are sufficient for multilevel SEM. Moreover, 20 groups is still feasible in terms of data collection. The sample size of 20 within each group is based on educational contexts in which this can be seen as a realistic size for a school class.

4.1 Study 1: Influence of population values

In the first study, we analyze the data sets using the traditional uniform prior and the inverse-Gamma prior with $\epsilon = .1, .01, .001$ (denoted by: IG.1, IG.01, and IG.001, re-

⁶A third simulation study investigated the influence of the number groups and analyzed a subset of conditions for 50 and 100 groups. However, as the number of groups increased, we ran into more convergence issues. Results for converged conditions are available online at <https://osf.io/pq8gm/>.

spectively) as well as the robust half-Normal (HNdef), half-Cauchy (HCdef), F (Fdef), and Exponential (EXPdef) priors. For the robust priors we consider a default setting in which the scale equals $\sigma_0^2 = 1$, following a general recommendation often made for a robust default specification (see also Subsection 3.4).

4.2 Study 2: Influence prior misspecifications

In the second study, we focus on informative prior distributions. Specifically, we choose the prior scale σ_0 in such a way that $P(\sigma_0 > \sigma_{true}) = \alpha$, where σ_{true} equals the population value under which that parameter was simulated. We consider two settings: in the “correct” informative prior (i.e., HNinf, HCinf, Finf, and EXPinf), $\alpha = 0.5$ such that the resulting prior has sufficient probability around the population value. In the “incorrect” informative prior (i.e., HNinc, HCinc, Finc, and EXPinc), on the other hand, $\alpha = 0.05$ such that the prior has only very little mass on the population value. Since we never know the true value in practice, it is important to consider this situation to assess the robustness of the various priors to possible misspecifications. In total, we consider 8 different prior specifications (4 types of priors $\times$ 2 settings) for 16 population conditions in Study 2. Figures 4 and 5 show the various informative prior densities used in the simulation study when the population value for the standard deviation equals 0.1 or 1, respectively. Note that the incorrect F prior specification is missing when the population standard deviation equals 0.1, since this setting resulted in a scale σ_0^2 of 0, whereas the scale should be positive. However, in order to be consistent across priors in specifying the incorrect setting, we did not consider an alternative specification that did result in a positive scale. All priors are more peaked around zero in the incorrect specification, and therefore have less prior mass around the true population value compared to the correct specification. It might appear that both the half-Cauchy and F priors have almost no prior mass around the true population value, but this is just a visual consequence of these priors being very spread out due to their heavy tails.

[FIGURE 4 HERE] [FIGURE 5 HERE]

4.3 Outcomes

We consider the (relative) bias, mean squared error and 95% coverage rates⁷, which are all computed based on the median estimate. We use 500 replications per condition and have computed the Monte Carlo SE for every outcome measure to quantify the uncertainty in the simulation results (Morris et al., 2019). All Monte Carlo SEs were sufficiently small to conclude that 500 replications was enough (the maximum MCSE was 0.03 for the coverage percentages).

All analyzes are run using the R interface to Stan, **RStan** (Stan Development Team, 2018). For each analysis, we ran two MCMC chains with 3000 iterations each⁸, half of which was used as burnin⁹. We used a maximum treedepth of 10 and set `adapt_delta` to 0.90. We compared various convergence criteria, the results of which are available at <https://osf.io/pq8gm/>. The results presented below are based on “strict” convergence criteria in which a replication is considered converged if the split $\hat{R}$ (a version of the potential scale reduction factor, PSRF; Gelman and Rubin (1992)) is smaller than or equal to 1.05 and there are no divergent transitions. We removed all replications that did not meet these criteria and present the results for those conditions with at least 50% convergence. We only report the results for $ES = -0.2$ and $VPC = 0.03$ in cases where these variables did not have a substantial influence and refer to the online materials for the full results and tables on which the figures are based as well as all the code (<https://osf.io/pq8gm/>).

⁷We also considered the power and type 1 error rates. However, power to find a small contextual effect was low across the board (ranging from 0.01 to 0.09 even for 50 groups) and type 1 error rates were around the nominal 5% for all conditions, so these results are only available online at <https://osf.io/pq8gm/>. Initially, we also compared the precision of the various priors through the empirical SE, which is computed as $\sqrt{\frac{1}{N-1} \sum_{i=1}^N (\hat{\theta}_i - \bar{\theta})^2}$, with $\bar{\theta}$ denoting the mean of $\hat{\theta}_i$. However, the empirical SE was sufficiently small in all conditions and for all priors (in the order of 1e-14) that we do not include this outcome in the results.

⁸We used a larger number of iterations than the **RStan** default of 2000 iterations to increase the chances of convergence.

⁹To assess the influence of the number of iterations on convergence, we reran one condition with 6000 iterations. However, this did not increase the convergence percentage.

4.4 Results Study 1: Influence population values

4.4.1 Convergence

[FIGURE 6 HERE]

Figure 6 shows the percentages of converged replications for each prior, with dashed lines indicating 50% convergence. For $ES = 0$, the inverse-Gamma prior sometimes showed substantially different convergence percentages. In general, the inverse-Gamma prior is most problematic with multiple conditions resulting in less than 50% convergence. With regard to the VPC, the smaller VPC of 0.03 generally leads to higher convergence compared to the larger VPC of 0.3.

4.4.2 Bias

Table 3 shows the relative bias, which can be multiplied by 100 to obtain a percentage, with the absolute bias in brackets for the variance parameters and the parameter of interest, $\beta_B - \beta_W$. For the standard deviation of the latent variable at the between level, ω_B , all priors result in some bias, although the values are comparable across most priors with the exception of the inverse-Gamma priors. Most priors underestimate ω_B when the population value equals 1 and when the population value equals 0.1 combined with a VPC of 0.03. The IG(0.1, 0.1) prior, however, largely overestimates ω_B when the population value equals 0.1, regardless of the value of the VPC. For the standard deviation of the latent variable at the within level, ω_W , the robust priors show only small biases when the population values equal 0.1, with more substantial underestimation when the population values equal 1, whereas the IG(0.1, 0.1) and IG(0.01, 0.01) priors show a substantial overestimation when the population values equal 0.01. For the standard deviation of the items at the between level, σ_Y^B , we see a similar picture across all items. Specifically, the bias is generally small except for the IG(0.1, 0.1) prior when $\omega_B = 0.1$ and $VPC = 0.03$. All priors underestimate the parameter of interest, the contextual effect $\beta_B - \beta_W$, regardless of the population condition. The relative bias is close to negative 100% when ω_W equals 1 for all priors, but differs more across the priors when ω_W equals 0.1.

VPC	ω_B	ω_W	UN	IG.001	IG.01	IG.1	HNdef	HCdef	Fdef	EXPdef
Standard deviation latent variable between ω_B										
0.03	0.1	0.1	-0.427 (-0.043)	-0.393 (-0.039)	0.189 (0.019)	1.624 (0.162)	-0.429 (-0.043)	-0.435 (-0.043)	-0.428 (-0.043)	-0.46 (-0.046)
0.03	0.1	1	-0.342 (-0.034)	NA	NA	1.577 (0.158)	-0.351 (-0.035)	-0.356 (-0.036)	-0.358 (-0.036)	-0.39 (-0.039)
0.03	1	0.1	-0.834 (-0.834)	-0.876 (-0.876)	-0.801 (-0.801)	-0.609 (-0.609)	-0.827 (-0.827)	-0.832 (-0.832)	-0.829 (-0.829)	-0.842 (-0.842)
0.03	1	1	-0.84 (-0.84)	-0.882 (-0.882)	-0.814 (-0.814)	-0.637 (-0.637)	-0.835 (-0.835)	-0.84 (-0.84)	-0.838 (-0.838)	-0.85 (-0.85)
0.3	0.1	0.1	0.662 (0.066)	0.244 (0.024)	1.029 (0.103)	2.962 (0.296)	0.746 (0.075)	0.693 (0.069)	0.703 (0.07)	0.561 (0.056)
0.3	0.1	1	0.948 (0.095)	0.32 (0.032)	1.094 (0.109)	2.985 (0.298)	1.046 (0.105)	0.906 (0.091)	1 (0.1)	0.83 (0.083)
0.3	1	0.1	-0.772 (-0.772)	-0.843 (-0.843)	-0.758 (-0.758)	-0.546 (-0.546)	-0.755 (-0.755)	-0.773 (-0.773)	-0.771 (-0.771)	-0.784 (-0.784)
0.3	1	1	-0.76 (-0.76)	-0.837 (-0.837)	-0.758 (-0.758)	-0.555 (-0.555)	-0.734 (-0.734)	-0.749 (-0.749)	-0.749 (-0.749)	-0.768 (-0.768)
VPC	ω_B	ω_W	UN	IG.001	IG.01	IG.1	HNdef	HCdef	Fdef	EXPdef
Standard deviation latent variable within ω_W										
0.03	0.1	0.1	0.118 (0.012)	-0.179 (-0.018)	0.669 (0.067)	2.671 (0.267)	0.096 (0.01)	0.053 (0.005)	0.097 (0.01)	-0.017 (-0.002)
0.03	0.1	1	-0.818 (-0.818)	NA	NA	-0.584 (-0.584)	-0.826 (-0.826)	-0.833 (-0.833)	-0.827 (-0.827)	-0.842 (-0.842)
0.03	1	0.1	0.258 (0.026)	-0.197 (-0.02)	0.507 (0.051)	2.145 (0.215)	0.129 (0.013)	0.089 (0.009)	0.133 (0.013)	-0.001 (0)
0.03	1	1	-0.814 (-0.814)	-0.89 (-0.89)	-0.821 (-0.821)	-0.651 (-0.651)	-0.835 (-0.835)	-0.842 (-0.842)	-0.837 (-0.837)	-0.853 (-0.853)
0.3	0.1	0.1	0.227 (0.023)	-0.2 (-0.02)	0.481 (0.048)	2.085 (0.208)	0.072 (0.007)	0.05 (0.005)	0.094 (0.009)	-0.023 (-0.002)
0.3	0.1	1	-0.819 (-0.819)	-0.892 (-0.892)	-0.825 (-0.825)	-0.667 (-0.667)	-0.849 (-0.849)	-0.845 (-0.845)	-0.837 (-0.837)	-0.859 (-0.859)
0.3	1	0.1	0.295 (0.03)	-0.158 (-0.016)	0.481 (0.048)	2.032 (0.203)	0.084 (0.008)	0.103 (0.01)	0.174 (0.017)	0.003 (0)
0.3	1	1	-0.814 (-0.814)	-0.89 (-0.89)	-0.822 (-0.822)	-0.669 (-0.669)	-0.848 (-0.848)	-0.843 (-0.843)	-0.832 (-0.832)	-0.859 (-0.859)
VPC	ω_B	ω_W	UN	IG.001	IG.01	IG.1	HNdef	HCdef	Fdef	EXPdef
Standard deviation item 1 between σ_{y1}^B										
0.03	0.1	0.1	-0.007 (-0.002)	-0.248 (-0.083)	-0.092 (-0.031)	0.187 (0.062)	-0.036 (-0.012)	-0.051 (-0.017)	-0.032 (-0.011)	-0.068 (-0.023)
0.03	0.1	1	-0.038 (-0.014)	NA	NA	0.112 (0.042)	-0.077 (-0.029)	-0.091 (-0.035)	-0.082 (-0.031)	-0.122 (-0.046)
0.03	1	0.1	0.04 (0.042)	0 (0)	0.005 (0.005)	-0.016 (-0.017)	-0.023 (-0.024)	-0.004 (-0.005)	-0.001 (-0.001)	-0.013 (-0.014)
0.03	1	1	0.02 (0.022)	-0.037 (-0.039)	-0.035 (-0.037)	-0.031 (-0.033)	-0.046 (-0.048)	-0.039 (-0.042)	-0.035 (-0.037)	-0.034 (-0.036)
0.3	0.1	0.1	0.033 (0.04)	-0.014 (-0.017)	-0.01 (-0.012)	-0.014 (-0.017)	-0.034 (-0.04)	-0.025 (-0.029)	-0.003 (-0.003)	-0.018 (-0.022)
0.3	0.1	1	0.046 (0.063)	-0.005 (-0.007)	-0.003 (-0.004)	-0.002 (-0.003)	-0.047 (-0.065)	-0.019 (-0.026)	-0.008 (-0.01)	-0.027 (-0.037)
0.3	1	0.1	0.054 (0.084)	0.009 (0.014)	0.019 (0.03)	0.022 (0.035)	-0.049 (-0.077)	-0.012 (-0.019)	0.019 (0.03)	-0.015 (-0.023)
0.3	1	1	0.023 (0.039)	-0.009 (-0.014)	-0.022 (-0.037)	-0.016 (-0.028)	-0.093 (-0.157)	-0.037 (-0.062)	-0.017 (-0.029)	-0.049 (-0.083)
VPC	ω_B	ω_W	UN	IG.001	IG.01	IG.1	HNdef	HCdef	Fdef	EXPdef
Contextual effect $\beta_B - \beta_W$										
0.03	0.1	0.1	-1.12 (0.016)	-1.141 (0.016)	-1.266 (0.018)	-1.435 (0.02)	-1.103 (0.015)	-1.115 (0.016)	-1.165 (0.016)	-1.122 (0.016)
0.03	0.1	1	-0.982 (0.102)	NA	NA	-0.957 (0.1)	-0.969 (0.101)	-0.99 (0.103)	-0.985 (0.102)	-0.978 (0.102)
0.03	1	0.1	-1.165 (0.016)	-1.118 (0.016)	-1.393 (0.019)	-1.078 (0.015)	-1.202 (0.017)	-1.179 (0.017)	-1.252 (0.018)	-1.238 (0.017)
0.03	1	1	-0.979 (0.102)	-0.987 (0.103)	-0.949 (0.099)	-0.917 (0.095)	-0.99 (0.103)	-0.974 (0.101)	-0.933 (0.097)	-0.981 (0.102)
0.3	0.1	0.1	-0.693 (0.01)	-0.743 (0.01)	-0.718 (0.01)	-0.479 (0.007)	-0.831 (0.012)	-0.601 (0.008)	-0.878 (0.012)	-0.742 (0.01)
0.3	0.1	1	-0.976 (0.102)	-1.013 (0.105)	-0.997 (0.104)	-1.083 (0.113)	-1.03 (0.107)	-1.019 (0.106)	-1.022 (0.106)	-1.017 (0.106)
0.3	1	0.1	-1.011 (0.014)	-1.123 (0.016)	-1.248 (0.017)	-1.63 (0.023)	-1.219 (0.017)	-1.462 (0.02)	-1.172 (0.016)	-1.382 (0.019)
0.3	1	1	-1.053 (0.11)	-1.037 (0.108)	-0.973 (0.101)	-1.066 (0.111)	-0.96 (0.1)	-1.007 (0.105)	-0.985 (0.102)	-1.03 (0.107)

Note. NA indicates that results are not available since the convergence percentage $< 50\%$ in this condition.

Table 3: Relative bias (multiply by 100 to obtain percentages) with absolute bias in brackets for selected parameters based on strict convergence criteria and all converged replications in Study 1 for conditions in which $ES = -0.2$

4.4.3 Mean squared error (MSE)

[FIGURE 7 HERE]

Figure 7 shows the MSE for selected parameters in Study 1. For both ω_B and ω_W , lower population values for the corresponding parameter result in a smaller MSE. Differences between the priors are negligible, except for the $IG(.1, .1)$ prior, which shows a smaller MSE compared to the other priors when the population value equals 1. For σ_{y1}^B , MSE is slightly higher when $\omega_B = 1$ and when $VPC = 0.3$ (available online at: <https://osf.io/pq8gm/>) for all priors, while for the contextual effect, the MSE is close to zero across conditions and priors.

4.4.4 Coverage

[FIGURE 8 HERE]

Figure 8 shows the 95% coverage rates for selected parameters. For ω_B and ω_W , coverage is much too low when the corresponding population values equal 1 and it is around 100% when the population values equal 0.1 for all priors except the $IG(.1, .1)$ prior. Note that, for ω_B , the coverage is higher when $\omega_B = 1$ and $VPC = 0.3$ (available online at: <https://osf.io/pq8gm/>), although it is still not close to the nominal 95% in this situation. For σ_{y1}^B , coverage is above the nominal 95% for all priors.

[FIGURE 9 HERE]

Figure 9 shows the 95% coverage rates for the contextual effect $\beta_B - \beta_W$. For this parameter, coverage rates do depend on effect size and VPC. Specifically, coverage rates are too low when $\omega_W = 1$ (triangles), but only when $ES = -0.2$ and not for the $IG(.1, .1)$ prior.

4.5 Results Study 2: Influence prior misspecifications

We now turn to the results of the second simulation study in which we investigate the same population conditions as in Study 1, but with informative priors. The informative priors are specified such that they are either in line with the population values, or not.

4.5.1 Convergence

[FIGURE 10 HERE]

Figure 10 shows the percentages of converged replications for each prior in the second study. Again, only the results for $ES = -0.2$ are shown, since the percentages for $ES = 0$ are generally very similar, except for the correctly specified exponential prior (EXPinf). It is clear that, compared to the default priors in Study 1, the informative priors have more convergence problems. This is especially the case for the incorrectly specified F and half-Cauchy priors. For the incorrectly specified F prior, the convergence of 0% when at least one of the population standard deviations equals 0.1 is due to the fact that in these situations the prior scale equals zero whereas it should be positive (see also Subsection 4.2). Only the informative half-Normal priors and the incorrectly specified Exponential prior obtained more than 50% convergence in all conditions.

4.5.2 Bias

Table 4 shows the relative bias with the absolute bias in brackets for the variance parameters and the parameter of interest, $\beta_B - \beta_W$. As many conditions are missing due to non-convergence, it is difficult to compare the correctly and incorrectly specified informative priors. However, for the half-normal prior, the bias for ω_B is lower for the correctly specified prior when $\omega_B = 0.1$. Especially when the $VPC = 0.3$, the difference between the correct and incorrect specifications is substantial. When $\omega_B = 1$, the two specifications perform similarly. A similar picture arises for ω_W , with the bias being lower for the correctly specified priors when the population value is small, and similar performance when the population value equals 1. For σ_{y1}^B , the biases are lower for the correctly specified half-normal prior across all population conditions, while the contextual effect $\beta_B - \beta_W$ is underestimated across the board, regardless of the prior specification. Compared to the default half-normal prior in Study 1, the informative specification results in lower bias for ω_B mainly when $\omega_B = 0.1$ and $VPC = 0.3$, while in the other conditions the bias is comparable or slightly higher. For ω_W , on the other hand, the default half-normal leads to lower bias than the correct informative

specification when $\omega_W = 0.1$ and comparable results otherwise.

VPC	ω_B	ω_W	HNinf	HNinc	HCinf	HCinc	Finf	EXPinf	EXPinc
Standard deviation latent variable between ω_B									
0.03	0.1	0.1	-0.558 (-0.056)	-0.652 (-0.065)	-0.87 (-0.087)	NA	NA	-0.61 (-0.061)	-0.716 (-0.072)
0.03	0.1	1	-0.467 (-0.047)	-0.606 (-0.061)	-0.871 (-0.087)	NA	-0.91 (-0.091)	NA	-0.679 (-0.068)
0.03	1	0.1	-0.857 (-0.857)	-0.849 (-0.849)	-0.871 (-0.871)	NA	-0.876 (-0.876)	NA	-0.868 (-0.868)
0.03	1	1	-0.84 (-0.84)	-0.826 (-0.826)	-0.842 (-0.842)	NA	-0.868 (-0.868)	NA	-0.855 (-0.855)
0.3	0.1	0.1	-0.11 (-0.011)	-0.536 (-0.054)	NA	-0.999 (-0.1)	NA	NA	-0.613 (-0.061)
0.3	0.1	1	-0.007 (-0.001)	-0.553 (-0.055)	NA	NA	NA	NA	-0.653 (-0.065)
0.3	1	0.1	-0.801 (-0.801)	-0.804 (-0.804)	-0.828 (-0.828)	NA	NA	NA	-0.829 (-0.829)
0.3	1	1	-0.756 (-0.756)	-0.759 (-0.759)	-0.772 (-0.772)	NA	-0.822 (-0.822)	NA	-0.8 (-0.8)
VPC	ω_B	ω_W	HNinf	HNinc	HCinf	HCinc	Finf	EXPinf	EXPinc
Standard deviation latent variable within ω_W									
0.03	0.1	0.1	-0.31 (-0.031)	-0.677 (-0.068)	-0.902 (-0.09)	NA	NA	-0.408 (-0.041)	-0.78 (-0.078)
0.03	0.1	1	-0.844 (-0.844)	-0.888 (-0.888)	-0.877 (-0.877)	NA	-0.899 (-0.899)	NA	-0.904 (-0.904)
0.03	1	0.1	-0.249 (-0.025)	-0.668 (-0.067)	-0.899 (-0.09)	NA	-0.956 (-0.096)	NA	-0.772 (-0.077)
0.03	1	1	-0.824 (-0.824)	-0.868 (-0.868)	-0.838 (-0.838)	NA	-0.887 (-0.887)	NA	-0.896 (-0.896)
0.3	0.1	0.1	-0.299 (-0.03)	-0.69 (-0.069)	NA	-0.999 (-0.1)	NA	NA	-0.786 (-0.079)
0.3	0.1	1	-0.86 (-0.86)	-0.902 (-0.902)	NA	NA	NA	NA	-0.911 (-0.911)
0.3	1	0.1	-0.195 (-0.02)	-0.664 (-0.066)	-0.901 (-0.09)	NA	NA	NA	-0.766 (-0.077)
0.3	1	1	-0.828 (-0.828)	-0.872 (-0.872)	-0.835 (-0.835)	NA	-0.884 (-0.884)	NA	-0.899 (-0.899)
VPC	ω_B	ω_W	HNinf	HNinc	HCinf	HCinc	Finf	EXPinf	EXPinc
Standard deviation item 1 between σ_{y1}^B									
0.03	0.1	0.1	-0.113 (-0.038)	-0.477 (-0.16)	-0.427 (-0.143)	NA	NA	-0.128 (-0.043)	-0.558 (-0.187)
0.03	0.1	1	-0.121 (-0.046)	-0.451 (-0.17)	-0.376 (-0.142)	NA	-0.534 (-0.202)	NA	-0.513 (-0.194)
0.03	1	0.1	0.012 (0.013)	-0.119 (-0.125)	-0.003 (-0.003)	NA	-0.031 (-0.032)	NA	-0.089 (-0.093)
0.03	1	1	-0.018 (-0.019)	-0.144 (-0.153)	-0.028 (-0.03)	NA	-0.059 (-0.062)	NA	-0.126 (-0.134)
0.3	0.1	0.1	0.021 (0.025)	-0.109 (-0.13)	NA	-0.049 (-0.059)	NA	NA	-0.09 (-0.107)
0.3	0.1	1	0.023 (0.031)	-0.104 (-0.141)	NA	NA	NA	NA	-0.077 (-0.105)
0.3	1	0.1	0.035 (0.055)	-0.096 (-0.149)	0.015 (0.023)	NA	NA	NA	-0.063 (-0.098)
0.3	1	1	0.002 (0.004)	-0.123 (-0.208)	0 (0)	NA	-0.023 (-0.039)	NA	-0.097 (-0.163)
VPC	ω_B	ω_W	HNinf	HNinc	HCinf	HCinc	Finf	EXPinf	EXPinc
Contextual effect $\beta_B - \beta_W$									
0.03	0.1	0.1	-1.074 (0.015)	-1.125 (0.016)	-1.149 (0.016)	NA	NA	-1.053 (0.015)	-1.078 (0.015)
0.03	0.1	1	-0.986 (0.103)	-0.988 (0.103)	-0.981 (0.102)	NA	-0.99 (0.103)	NA	-0.996 (0.104)
0.03	1	0.1	-1.307 (0.018)	-1.235 (0.017)	-1.111 (0.016)	NA	-1.149 (0.016)	NA	-1.108 (0.016)
0.03	1	1	-1.016 (0.106)	-0.985 (0.102)	-0.983 (0.102)	NA	-0.99 (0.103)	NA	-0.988 (0.103)
0.3	0.1	0.1	-0.654 (0.009)	-0.819 (0.011)	NA	-0.895 (0.013)	NA	NA	-0.929 (0.013)
0.3	0.1	1	-1.001 (0.104)	-1.01 (0.105)	NA	NA	NA	NA	-1.025 (0.107)
0.3	1	0.1	-1.171 (0.016)	-1.174 (0.016)	-1.193 (0.017)	NA	NA	NA	-1.067 (0.015)
0.3	1	1	-1.039 (0.108)	-1 (0.104)	-1.052 (0.109)	NA	-1.052 (0.109)	NA	-1.05 (0.109)

Note. NA indicates that results are not available since the convergence percentage $< 50\%$ in this condition.

Table 4: Relative bias (multiply by 100 to obtain percentages) with absolute bias in brackets for selected parameters based on strict convergence criteria and all converged replications in Study 2 for conditions in which $ES = -0.2$

4.5.3 Mean squared error (MSE)

[FIGURE 11 HERE]

Figure 11 shows the MSE for selected parameters in Study 2. The results are very similar to those in Study 1 with negligible differences between the priors and mainly differences due to population values for the latent variable variance parameters, with higher population values (i.e., $\omega_B = \omega_W = 1$) resulting in larger MSEs for the corresponding parameter.

4.5.4 Coverage

[FIGURE 12 HERE]

Figure 12 shows the 95% coverage rates for selected parameters in Study 2. Again, the coverage for ω_B and ω_W is much too low when the corresponding population values equal 1 but, contrary to the coverage for the default priors (Study 1), the coverage remains too low for most priors when the population values equal 0.1, with the exception of the correctly specified exponential and half-normal priors. For σ_{y1}^B , coverage is close to the nominal 95% for all informative priors when $\omega_B = 1$, but generally lower when $\omega_B = 0.1$.

[FIGURE 13 HERE]

Figure 13 shows the 95% coverage rates for the contextual effect $\beta_B - \beta_W$. As in Study 1, coverage rates are close to the nominal 95% when $ES = 0$, but they are too low when $ES = -0.2$ and $\omega_W = 1$, especially when $VPC = 0.03$.

5 Discussion

The popularity of Bayesian MLSEM has inspired various investigations into the required number of groups (see e.g., Depaoli and Clifton, 2015; Helm, 2018; Hox et al., 2012). However, these studies mainly rely on traditional default prior specifications, that have been proven to be unreliable, or correctly specified informative priors that are not feasible in practice. The goal of this study was to investigate more robust prior distributions for the random effects variances in the context of MLSEM. In order to do so, we have investigated default and (in)accurate informative prior specifications of popular robust prior distributions in

MLSEM.

The differences between the robust prior distributions and the uniform prior were generally small. The main differences in results in the simulation studies appear to arise due to the population conditions rather than the prior distribution. Only the inverse-Gamma specifications did show substantially different, and often worse, performance. The default specification for the robust priors did show bias for the random effects variance parameters as well as the parameter of interest, the contextual effect. The MSE of the random effects variance parameters was high in certain conditions, but close to zero for the contextual effect across all conditions. The coverage for the random effects variance parameters was much too low, but only low for the contextual effect in certain population conditions. These results indicate that, although the robust prior distributions are better options for the random effects variance parameter than the inverse-Gamma prior, their hyperparameters still require careful consideration in order to obtain the best results. This was further investigated in Study 2, where we considered (in)accurate informative specifications for the robust priors. Unfortunately, the informative priors resulted in low convergence percentages which complicated a thorough investigation and comparison of their performance. Again, this illustrates the difficulty in specifying the hyperparameters for these robust prior distributions. Future research should further investigate these and other informative specifications to fully assess their robustness against misspecification.

One limitation of the simulation study lies in the fact that we removed the non-converged replications from the results. If the non-convergence in these removed replications is due to particular sample values this approach might bias the simulations. We tried to partially remedy this problem by also considering weaker convergence criteria, which resulted in less non-converged replications but generally did not show substantial differences in terms of bias, MSE, and coverage. Still, non-convergence and how to deal with it remains an issue in any simulation study investigating extreme conditions in terms of sample size or population values. Compared to other MCMC algorithms, the Hamiltonian Monte Carlo algorithm

used in Stan offers more convergence criteria. An advantage is that convergence can be assessed more accurately, however, it will also flag a replication as non-converged more easily. Especially for large models such as the model considered here, a few divergent transitions can occur rapidly resulting in a replication that is flagged as non-converged. It can therefore be useful to compare the results obtained using various convergence criteria.

Throughout this paper, we have focused on the prior for the random effects variance or standard deviation. In regular multilevel models, several authors have proposed to specify priors on the VPC, which leads to an implied prior on the random effects variance (see e.g., Daniels, 1999; Gustafson et al., 2006; Mulder and Fox, 2018; Natarajan and Kass, 2000). Although we have considered specifying a prior directly on the VPC, this approach is not straightforward in MLSEM. Specifically, since the VPC depends on multiple other model parameters (see Equation (6)), a choice needs to be made for which of those model parameters a prior is directly specified and for which of the model parameters the prior is implied. Future research is needed to investigate these choices and their consequences.

Bayesian MLSEM offers a powerful modeling framework to incorporate both measurement and sampling error within one model. However, the results presented in this paper indicate that some caution is warranted in applying these models. When the sample size is small, Bayesian estimation does not necessarily perform well (Smid et al., 2019). In order for Bayesian estimation to outperform classical estimation methods, the prior distribution needs to be well thought out. This is especially the case for the complex MLSEM considered in this paper: with only 20 groups, it is crucial for the prior distribution to add reasonable prior information to the analysis. More research is needed to investigate how we can achieve this.

6 Acknowledgements

This research was supported by a Research Talent Grant from the Netherlands Organisation for Scientific Research. We would like to thank Irene Klugkist, Duco Veen, and Mariëlle Zondervan-Zwijnenburg for helpful comments on an earlier version of this manuscript.

References

- Asparouhov, T. and Muthén, B. (2007). Computationally efficient estimation of multilevel high-dimensional latent variable models. In *proceedings of the 2007 JSM meeting in Salt Lake City, Utah, Section on Statistics in Epidemiology*, pages 2531–2535.
- Asparouhov, T. and Muthén, B. (2012). General random effect latent variable modeling: Random subjects, items, contexts, and parameters.
- Asparouhov, T. and Muthén, B. (2019). Latent variable centering of predictors and mediators in multilevel and time-series models. *Structural Equation Modeling: A Multidisciplinary Journal*, 26(1):119–142.
- Berger, J. (2006). The case for objective bayesian analysis. *Bayesian Analysis*, 1(3):385–402.
- Berger, J. O. and Strawderman, W. E. (1996). Choice of hierarchical priors: admissibility in estimation of normal means. *The Annals of Statistics*, 24(3):931–951.
- Daniels, M. J. (1999). A prior for the variance in hierarchical models. *Canadian Journal of Statistics*, 27(3):567–578.
- Depaoli, S. and Clifton, J. P. (2015). A bayesian approach to multilevel structural equation modeling with continuous and dichotomous outcomes. *Structural Equation Modeling: A Multidisciplinary Journal*, 22(3):327–351.
- Frühwirth-Schnatter, S. and Wagner, H. (2010). Stochastic model specification search for gaussian and partial non-gaussian state space models. *Journal of Econometrics*, 154(1):85–100.
- Fúquene, J., Pérez, M.-E., and Pericchi, L. R. (2014). An alternative to the inverted gamma

- for the variances to modelling outliers and structural breaks in dynamic models. *Brazilian Journal of Probability and Statistics*, 28(2):288–299.
- Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian Analysis*, 1(3):515–534.
- Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472.
- Ghosh, J., Li, Y., and Mitra, R. (2018). On the use of cauchy prior distributions for bayesian logistic regression. *Bayesian Analysis*, 13(2):359–383.
- Goldstein, H., Browne, W., and Rasbash, J. (2002). Partitioning variation in multilevel models. *Understanding Statistics*, 1(4):223–231.
- Gustafson, P., Hossain, S., and Macnab, Y. C. (2006). Conservative prior distributions for variance parameters in hierarchical models. *Canadian Journal of Statistics*, 34(3):377–390.
- Hedges, L. V. and Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, 29(1):60–87.
- Helm, C. (2018). How many classes are needed to assess effects of instructional quality? a monte carlo simulation of the performance of frequentist and bayesian multilevel latent contextual models. *Psychological Test and Assessment Modeling*, 60(2):265–285.
- Hox, J. J. and Maas, C. J. M. (2001). The accuracy of multilevel structural equation modeling with pseudobalanced groups and small samples. *Structural Equation Modeling: A Multidisciplinary Journal*, 8(2):157–174.
- Hox, J. J., van de Schoot, R., and Matthijsse, S. (2012). How few countries will do? comparative survey analysis from a bayesian perspective. *Survey Research Methods*, 6(2):87–93.
- Jak, S. (2018). Cross-level invariance in multilevel factor models. *Structural Equation Modeling: A Multidisciplinary Journal*, pages 1–16.
- Klein, N. and Kneib, T. (2016). Scale-dependent priors for variance parameters in structured additive distributional regression. *Bayesian Analysis*, 11(4):1071–1106.

- Lunn, D., Spiegelhalter, D., Thomas, A., and Best, N. (2009). The BUGS project: Evolution, critique and future directions. *Statistics in Medicine*, 28(25):3049–3067.
- Marsh, H. W., Lüdtke, O., Robitzsch, A., Trautwein, U., Asparouhov, T., Muthén, B., and Nagengast, B. (2009). Doubly-latent models of school contextual effects: Integrating multilevel and structural equation approaches to control measurement and sampling error. *Multivariate Behavioral Research*, 44(6):764–802.
- Meuleman, B. and Billiet, J. (2009). A Monte Carlo sample size study: How many countries are needed for accurate multilevel SEM? *Survey Research Methods*, 3(1):45–58.
- Morris, T. P., White, I. R., and Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38(11):2074–2102.
- Mulder, J. and Fox, J.-P. (2018). Bayes factor testing of multiple intraclass correlations. *Bayesian Analysis*.
- Mulder, J. and Pericchi, L. R. (2018). The matrix-f prior for estimating and testing covariance matrices. *Bayesian Analysis*, 13(4):1189–1210.
- Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research*, 22(3):376–398.
- Natarajan, R. and Kass, R. E. (2000). Reference bayesian methods for generalized linear mixed models. *Journal of the American Statistical Association*, 95(449):227–237.
- OECD (2007). *PISA 2006*.
- Pérez, M.-E., Pericchi, L. R., and Ramírez, I. C. (2017). The scaled beta2 distribution as a robust prior for scales. *Bayesian Analysis*, 12(3):615–637.
- Piironen, J. and Vehtari, A. (2015). Projection predictive variable selection using stan+ r. *arXiv preprint arXiv:1508.02502*.
- Polson, N. G. and Scott, J. G. (2012). On the half-cauchy prior for a global scale parameter. *Bayesian Analysis*, 7(4):887–902.
- Roos, M., Held, L., et al. (2011). Sensitivity analysis in bayesian generalized linear mixed models for binary data. *Bayesian Analysis*, 6(2):259–278.

- Simpson, D., Rue, H., Riebler, A., Martins, T. G., and Sørbye, S. H. (2017). Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical Science*, 32(1):1–28.
- 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. *Structural Equation Modeling: A Multidisciplinary Journal*, pages 1–31.
- Stan Development Team (2015). Prior choice recommendations. <https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations>. Accessed: 2019-08-05.
- Stan Development Team (2018). RStan: the R interface to Stan. R package version 2.18.2.
- van Erp, S., Mulder, J., and Oberski, D. L. (2018). Prior sensitivity analysis in default bayesian structural equation modeling. *Psychological Methods*, 23(2):363–388.
- Yuan, Y. and MacKinnon, D. P. (2009). Bayesian mediation analysis. *Psychological Methods*, 14(4):301–322.
- Zitzmann, S., Lüdtke, O., Robitzsch, A., and Marsh, H. W. (2016). A bayesian approach for estimating multilevel latent contextual models. *Structural Equation Modeling: A Multidisciplinary Journal*, 23(5):661–679.

List of Figures

1	Graphic representation of the doubly latent multilevel model (DLMM) used throughout this paper. SA stands for student achievement and SC stands for academic self-concept in the context of the BFLPE. Manifest variables are denoted by squares, while latent variables are denoted by circles. Black dots indicate a random intercept at the between level for that variable.	33
2	Densities of the priors on the standard deviations.	34
3	Densities of the implied priors on the variance partition coefficient (VPC) of the restricted and free items.	35
4	Prior densities for the informative specifications when the population value for the standard deviation equals 0.1.	36
5	Prior densities for the informative specifications when the population value for the standard deviation equals 1.	37
6	Percentages converged replications per condition according to the strict convergence criteria in Study 1 for $ES = -0.2$	38
7	Mean squared error (MSE) per condition according to the strict convergence criteria in Study 1 for $ES = -0.2$ and $VPC = 0.03$	39
8	95% coverage per condition according to the strict convergence criteria in Study 1 for $ES = -0.2$ and $VPC = 0.03$ and selected parameters. Dashed lines indicate the nominal 95%.	40
9	95% coverage per condition according to the strict convergence criteria in Study 1 for the contextual effect $\beta_B - \beta_W$. Dashed lines indicate the nominal 95%.	41
10	Percentages converged replications per condition according to the strict convergence criteria in Study 2 for $ES = -0.2$	42
11	Mean squared error (MSE) per condition according to the strict convergence criteria in Study 2 for $ES = -0.2$ and $VPC = 0.03$	43
12	95% coverage per condition according to the strict convergence criteria in Study 2 for $ES = -0.2$ and $VPC = 0.03$ and selected parameters. Dashed lines indicate the nominal 95%.	44
13	95% coverage per condition according to the strict convergence criteria in Study 2 for the contextual effect $\beta_B - \beta_W$. Dashed lines indicate the nominal 95%.	45

Figure 1: Graphic representation of the doubly latent multilevel model (DLMM) used throughout this paper. SA stands for student achievement and SC stands for academic self-concept in the context of the BFLPE. Manifest variables are denoted by squares, while latent variables are denoted by circles. Black dots indicate a random intercept at the between level for that variable.

Figure 2: Densities of the priors on the standard deviations.

Figure 3: Densities of the implied priors on the variance partition coefficient (VPC) of the restricted and free items.

Figure 4: Prior densities for the informative specifications when the population value for the standard deviation equals 0.1.

Figure 5: Prior densities for the informative specifications when the population value for the standard deviation equals 1.

Figure 6: Percentages converged replications per condition according to the strict convergence criteria in Study 1 for $ES = -0.2$.

Figure 7: Mean squared error (MSE) per condition according to the strict convergence criteria in Study 1 for $ES = -0.2$ and $VPC = 0.03$.

Figure 8: 95% coverage per condition according to the strict convergence criteria in Study 1 for $ES = -0.2$ and $VPC = 0.03$ and selected parameters. Dashed lines indicate the nominal 95%.

Figure 9: 95% coverage per condition according to the strict convergence criteria in Study 1 for the contextual effect $\beta_B - \beta_W$. Dashed lines indicate the nominal 95%.

Figure 10: Percentages converged replications per condition according to the strict convergence criteria in Study 2 for $ES = -0.2$.

Figure 11: Mean squared error (MSE) per condition according to the strict convergence criteria in Study 2 for $ES = -0.2$ and $VPC = 0.03$.

Figure 12: 95% coverage per condition according to the strict convergence criteria in Study 2 for $ES = -0.2$ and $VPC = 0.03$ and selected parameters. Dashed lines indicate the nominal 95%.

Figure 13: 95% coverage per condition according to the strict convergence criteria in Study 2 for the contextual effect $\beta_B - \beta_W$. Dashed lines indicate the nominal 95%.