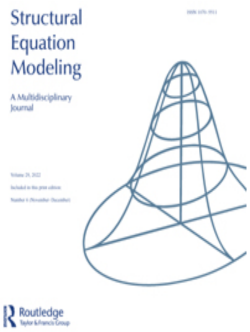

Structural Equation Modeling: A Multidisciplinary Journal

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/hsem20

The Influence of Using Inaccurate Priors on Bayesian Multilevel Estimation

Shufang Zheng, Lijin Zhang, Zhehan Jiang & Junhao Pan

To cite this article: Shufang Zheng, Lijin Zhang, Zhehan Jiang & Junhao Pan (2022): The Influence of Using Inaccurate Priors on Bayesian Multilevel Estimation, Structural Equation Modeling: A Multidisciplinary Journal, DOI: 10.1080/10705511.2022.2136185

To link to this article: https://doi.org/10.1080/10705511.2022.2136185

STRUCTURAL EQUATION MODELING: A MULTIDISCIPLINARY JOURNAL
https://doi.org/10.1080/10705511.2022.2136185

Routledge
Taylor & Francis Group

RESEARCH ARTICLE

Check for updates

The Influence of Using Inaccurate Priors on Bayesian Multilevel Estimation

Shufang Zheng[a], Lijin Zhang[a,b], Zhehan Jiang[c] and Junhao Pan[a]

[a]Sun Yat-sen University; [b]Stanford University; [c]Peking University

ABSTRACT
Researchers in psychology, education, and organizational behavior often encounter multilevel data with hierarchical structures. Bayesian approach is usually more advantageous than traditional frequentist-based approach in small sample sizes, but it is also more susceptible to the subjective specification of priors. To investigate the potentially detrimental effects of inaccurate prior information on Bayesian approach and compare its performance with that of traditional method, a series of simulations was conducted under a multilevel model framework with different settings. The results reveal the devastating impacts of inaccurate prior information on Bayesian estimation, especially in the cases of larger intraclass correlation coefficient, smaller level 2 sample size, and smaller prior variance. When the dependent variable is non-normal or binary, these negative effects are more noticeable. The present study investigated the impacts of inaccurate prior information and provides advice on the specification of priors.

KEYWORDS
Bayesian modeling; inaccurate prior information; multilevel modeling; prior specification; small samples

1. Introduction

Data with hierarchical structures are common in research on psychology, education, and organizational behavior (Aryee et al., 2012; Holtmann et al., 2016; Hox et al., 2017). In these data, participants tend to respond similarly to others within the same cluster. However, the dependence of these observations may violate the basic assumption of the model and are likely to result in biased parameter estimations (Rutterford et al., 2015; van Breukelen & Candel, 2012).

Considering the statistical problems caused by observations that are dependent within the same cluster, researchers are often suggested to conduct multilevel modeling (Ryu, 2015). Traditionally, multilevel models are estimated under the frequentist framework, relying on large samples and asymptotic normality (McNeish, 2016a). When the sample size is relatively small, frequentist estimators may encounter model convergence and parameter estimation problems (Depaoli & Clifton, 2015; McNeish, 2016a; McNeish & Stapleton, 2016; Schoeneberger, 2016).

As an alternative to the traditional frequentist approach, Bayesian estimation is advantageous for dealing with small sample sizes and can incorporate prior information to better estimate the parameters (McNeish, 2016a). However, when the sample size is small, the shapes of the posterior distributions are dominated by priors (Lynch, 2007). Considering the small-sample problems in multilevel modeling (Campbell & Walters, 2014), the issues of incorrect prior information of Bayesian estimation may be more prominent. Therefore, more relevant studies are needed to investigate the impacts of incorrect informative priors on Bayesian analysis (Finch & Miller, 2019; Holtmann et al., 2016), especially under the framework of multilevel modeling.

2. Multilevel Modeling

2.1. Multilevel Modeling and cRCT Trials

Multilevel modeling is frequently used to deal with dependent observations in the analysis of clustered data (Ryu, 2015). In multilevel models, the sources of variation are decomposed into different levels and can be further explained by predictors at different levels (Heck & Thomas, 2020). Multilevel modeling often requires large sample sizes at both levels to obtain unbiased estimates, but it can be challenging in collecting multilevel data because of high financial cost and geographic sparsity (Campbell & Walters, 2014).

As a typical multilevel research design, the cluster randomized controlled trial (cRCT) is also prone to the problems arising from small sample sizes (McNeish, 2016a). Generally, randomized controlled trials (RCT) are the gold standard of experimental study designs (Campbell & Walters, 2014; Rutterford et al., 2015). However, researchers are often encouraged to conduct experiments on groups (i.e., cRCT) rather than on each subject separately (i.e., RCT). This preference primarily occurs if there is a high risk of treatment contamination[1] (Campbell, 2019; Campbell & Walters, 2014; Donner & Zou, 2004; Rutterford et al., 2015). In addition, many researchers prefer cRCT because it can improve treatment efficiency, convenience, and compliance (Donner & Zou, 2004; Rutterford et al., 2015).

CONTACT Junhao Pan panjunh@mail.sysu.edu.cn Department of Psychology, Sun Yat-sen University, Guangzhou, P. R. China.
Supplemental data for this article is available online at https://doi.org/10.1080/10705511.2022.2136185.
[1]Individuals of control groups may be exposed to the intervention, thus impacting the evaluation of the effects of the treatment.

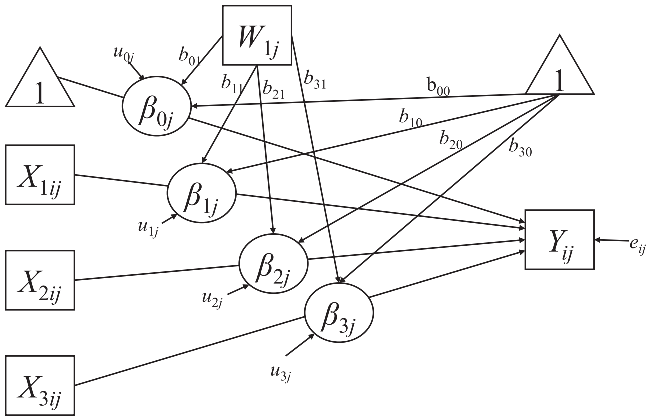

Figure 1. The path diagram of the multilevel model. *Note*: W_{1j} : the predictor at level 2; $X_{1ij}-X_{3ij}$: the predictors at level 1; Y_{ij}: the outcome variable at level 1. e_{ij}: the residual terms at level 1; u_{0j} : the residual term of the random intercept (β_{0j}); $u_{1j}-u_{3j}$: residual terms of the random slopes ($\beta_{1j}-\beta_{3j}$).

Therefore, although cRCTs are prone to the problems arising from small sample sizes, they are widely used for research in psychology (especially in clinical psychology), sociology, and education (Campbell, 2019; Ribeiro et al., 2018).

To increase the generalizability of the research, we focused on a multilevel model with a random intercept and random slopes. This model is also common in standard cRCT research. As shown in Figure 1, for the *ith* individual in the *jth* group, the outcome variable is defined as follows (McNeish, 2016a):

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{1ij} + \beta_{2j}X_{2ij} + \beta_{3j}X_{3ij} + e_{ij} \tag{1}$$

where

$$\beta_{0j} = b_{00} + b_{01}W_{1j} + u_{0j} \tag{2}$$

$$\beta_{1j} = b_{10} + b_{11}W_{1j} + u_{1j} \tag{3}$$

$$\beta_{2j} = b_{20} + b_{21}W_{1j} + u_{2j} \tag{4}$$

$$\beta_{3j} = b_{30} + b_{31}W_{1j} + u_{3j} \tag{5}$$

where e_{ij} denotes the residual terms at level 1 ($e_{ij}\sim N(0, \sigma_e^2)$), W_{1j} denotes the predictor at level 2 (e.g., treated/controlled groups in the cRCT), and $X_{1ij}-X_{3ij}$ denote the level 1 predictors (e.g., pre-test scores, gender, and other demographic variables, etc.). u_{0j} is the residual term of the random intercept (β_{0j}), and $u_{1j}-u_{3j}$ are the residual terms of the random slopes ($\beta_{1j}-\beta_{3j}$), respectively. These residual terms are jointly multivariate normal distributed ($u_j\sim MVN(0, \Theta)$, where Θ refers to the corresponding covariance matrix).

2.2. Traditional Frequentist Estimation

Traditionally, multilevel models are estimated with frequentist approaches (e.g., maximum likelihood (ML); Depaoli & Clifton, 2015). However, ML relies on large samples to hold asymptotic properties and obtain accurate parameter estimates (Asparouhov & Muthén, 2010; McNeish, 2016a). When the sample sizes at both the within- and between-levels are small, ML estimation is prone to convergence problems (Depaoli & Clifton, 2015; Hox et al., 2010; Hox & Maas, 2001; McNeish, 2016a; Schoeneberger, 2016), and may produce unreliable and unstable parameter estimates (Asparouhov & Muthén, 2010; Hox & Maas, 2001; McNeish, 2016a; McNeish & Stapleton, 2016).

In addition, multilevel models with categorical outcomes, which are common in psychological research, may bring additional challenges for traditional frequentist methods (Holtmann et al., 2016). ML requires a large number of dimensions in the process of multidimensional numerical integration (Muthén, 2010), which makes the estimation of model computationally demanding. Therefore, in this situation, the weighted least squares estimator (WLS) is often recommended instead (Muthén et al., 2015). WLS utilizes pairwise information (based on a polychoric correlation matrix) and requires fewer integration points for each dimension than ML (Asparouhov & Muthén, 2012a; Muthén et al., 2015), thus is efficient for categorical outcomes. Similar to WLS, weighted least squares estimator with mean and variance adjusted (WLSMV) produces accurate parameter estimate and fewer convergence problems in small sample sizes, and it is often recommended in multilevel models with categorical variables (Depaoli & Clifton, 2015; Hox et al., 2010). However, WLSMV may lead to poor model fitting and biased standard error estimation when the sample size is small (Muthén et al., 2015), and is limited to random intercept model (Asparouhov & Muthén, 2012b; van Erp & Browne, 2021).

2.3. Bayesian Estimation

Considering the limitations of the traditional frequentist approach in multilevel models with small sample sizes, researchers often recommend Bayesian approach as alternative (Depaoli & Schoot, 2017; Kadane, 2015; Van de Schoot et al., 2014). Bayesian estimation does not rely on asymptotic properties, making it attractive under the condition of small samples. In Bayesian approach, the unknown parameters are considered as random (rather than fixed) variables drawn from the corresponding posterior distributions (van de Schoot et al., 2021), which are jointly determined by the prior and likelihood distributions (Lynch, 2007). Based on the sampling algorithms (for example, Gibbs sampler and Markov chain Monte Carlo (MCMC) algorithms), Bayesian estimation can perform well in multilevel models, even with small sample sizes (McNeish, 2016a). It can obtain accurate parameter estimates with fewer clusters than the ML estimator (Hox et al., 2012).

Bayesian estimation is also advantageous in complex models (e.g., multilevel models with categorical variables). It does not conduct complex calculations for standard errors, which essentially eliminates convergence problems (Depaoli & Clifton, 2015; Levy & Choi, 2013). Thus, Bayesian estimation often outperforms traditional methods in model convergence and parameter estimation in multilevel models with small sample sizes (Asparouhov & Muthén, 2010; Hox et al., 2012). In addition, Bayesian approaches (e.g., MCMC; Gelfand & Smith, 1990; Gilks et al., 1996) do not require multidimensional numerical integration, which vastly improves their efficiency of model estimation. For traditional methods, some of the complex models are challenging to estimate, such as multilevel models with categorical variables (Depaoli & Clifton, 2015), the two-level nonlinear structural equation model (Song & Lee, 2004), and the multilevel structural equation model (MSEM) with categorical variables and random slopes (Asparouhov & Muthén, 2013).

2.3.1. Prior Distributions in Bayesian Estimation

Another advantage of the Bayesian method is that it can incorporate prior knowledge and evidence into model analysis. With appropriate informative priors, Bayesian methods generally outperform the traditional ML estimator (Depaoli & Clifton, 2015; McNeish, 2016a).

However, the procedure of the prior specification is somewhat subjective and may essentially change the estimation of the parameters. Generally speaking, priors with smaller prior variance have more significant impacts on the parameter estimation, especially when the sample sizes at both the within- and between-levels are small, or the dependent variables are categorical (Depaoli & Clifton, 2015; McNeish, 2016a). The specification of priors may significantly influence the posterior distribution of parameters (especially in small sample sizes; Lynch, 2007). Therefore, researchers should always be cautious about the specification of priors, especially in small samples (Holtmann et al., 2016; McNeish, 2016a).

2.3.2. Inaccurate Informative Priors in Bayesian Estimation

Inaccurate informative priors refer to priors in which the prior means deviate from the corresponding true values (Depaoli, 2014; Finch & Miller, 2019). It is usually challenging to correctly set the prior means in empirical studies because of the lack of previous knowledge and theoretical evidence. Thus, empirical researchers are prone to specifying inaccurate informative priors in practice. However, like McNeish (2016a), many researchers assumed that the prior means were all correctly specified when conducting simulations and ignored the possibility of inaccurate prior information.

Recently, more and more researchers have started to explore the potential influence of inaccurate prior information in different models, such as the latent growth model (Depaoli, 2014; Shi & Tong, 2017), the multiple indicators and multiple causes model (MIMIC; Finch & Miller, 2019), and the MSEM (Holtmann et al., 2016). Strongly informative priors with inaccurate prior information would lead to devastating parameter estimation, but weakly informative priors with inaccurate prior information may outperform the diffused priors in some situations (Depaoli, 2014; Holtmann et al., 2016). However, there are some limitations in the simulations of Holtmann et al. (2016), as they: (a) only investigated the measurement models and did not consider the structural model; (b) only investigated the impacts of inaccurate informative priors for categorical situations; (c) assumed the sample size to be large (50–200) at the between-level. However, in many empirical studies with clustered data structures (such as cRCT), the sample size at the between-level is generally less than 50 (e.g., Ha et al., 2017; Newton et al., 2018; Shen et al., 2019).

2.4. The Impacts of the Intraclass Correlation Coefficient

In multilevel modeling research, the impact of the intraclass correlation coefficient (ICC) is often of great interest (Depaoli & Clifton, 2015). The ICC refers to the proportion of total variation explained at the between-level ($ICC = \sigma_B^2/(\sigma_B^2 + \sigma_W^2)$), where σ_B^2 and σ_W^2 refer to the variance of the outcome at the between- and within-levels, respectively). A larger ICC indicates a more significant between-level variation compared to the within-level variation. Many previous studies have shown that when the level of ICC is small, traditional frequentist approach may lead to convergence problems and biased parameter estimation (Hox & Maas, 2001; Koch et al., 2015; Preacher & Kristopher, 2011). In these situations, Bayesian approach may overcome the problems of small ICC and accurately estimate the parameters (Hox et al., 2012; Koch et al., 2015).

However, in Bayesian approach, the impact of the ICC on parameter estimation with incorrect informative priors is still unclear. Many researchers have suggested that a smaller ICC may lead to worse performance (Can et al., 2015; Hox & Maas, 2001; Preacher & Kristopher, 2011), while others have provided contradictory results. For example, when ICC is large, the non-informative Bayesian estimator may provide seriously biased parameter estimation at the between-level in the MSEM (Depaoli & Clifton, 2015), suggesting that Bayesian estimation may be more sensitive to the specification of priors when the ICC is large. In addition, under the framework of multilevel MIMIC, bias toward the assessment of between-level parameters increased as the level of the ICC increased, and the power of the parameters decreased at the same time (Cao et al., 2019; Finch & French, 2011). Therefore, more research is needed to further explore the influence of ICC on Bayesian estimation in multilevel models (Fang et al., 2019), especially with different prior specifications.

2.5. Goals of the Present Research

To explore the impacts of incorrect informative priors on Bayesian estimation, the present study conducted a series of simulations based on multilevel models and cRCT. For Bayesian estimation with incorrect informative priors, we focused on the impacts of prior variance and prior mean deviation (from the population values). In addition, traditional methods (ML[2] for continuous data and WLSMV for binary data) were also considered for comparison. To cover as many situations in empirical studies as possible, different

[2]The MLR (robust maximum likelihood estimator) estimator was utilized as default.

data types, sample sizes, and different levels of ICCs were considered in the simulation studies. The goals of the present research were to investigate: (a) the impact of incorrect informative priors on Bayesian estimation in multilevel modeling; (b) whether and how these impacts vary with the characteristics of the datasets (i.e., the data types, sample sizes, and the levels of the ICC); (c) the situations in which Bayesian estimation with incorrect informative priors performs even better than traditional methods (if any). We hope that the present research might strengthen the theoretical and practical understanding of prior specifications in the multilevel modeling framework, especially in cRCT research, and provide guidelines for the specification of priors in empirical research.

3. Simulation Studies

Monte Carlo simulations were conducted for multilevel models to evaluate the performance of different estimators under different settings. The present study used R (R Core Team, 2017) to generate data and Mplus 8.4 (Muthén & Muthén, 1998–2017) to analyze the data.

3.1. Simulation Design

To maximize the generality of the research, we generated data from the multilevel model introduced in the previous section (Figure 1). Following McNeish (2016a), we generated X_{1ij} from a standard normal distribution ($X_{1ij} \sim N(0,1)$) and generated W_{1j}, X_{2ij}, and X_{3ij} as binary variables (with the proportion of 0 equaling 0.5, 0.5, and 0.75, respectively). The first diagonal element of Θ, σ^2_{u0}, changed with the ICC level. The other diagonals of Θ (i.e., $\sigma^2_{u1}-\sigma^2_{u3}$, the residual variances of random slopes) were set to 0.5, and the off-diagonals were set to 0. The population values of the intercepts of the random effects ($b_{00}-b_{30}$) were 0, 2.3, -0.25, and -0.75, and the effects of the treatment ($b_{01}-b_{31}$) were 1.5, 0.12, -0.05, and -0.75, respectively. In cRCT research, the treatment effects are usually of interest, so the present research focused on the estimation of $b_{01}-b_{31}$.

To investigate the influence of different priors on Bayesian estimators in various settings, data were generated with different types of dependent variables, different sample sizes at both levels, and different levels of ICCs. In addition, we analyzed the models with different estimators to compare the results of Bayesian approach with those of frequentist approach.

3.1.1. Types of Dependent Variables

Researchers often assume the data to be normally distributed in theoretical research, but this assumption is often too ideal to be satisfied in empirical studies (Fleishman, 1978). In general, compared to non-normal factor means, non-normal residuals may lead to more seriously biased parameter estimations (Song et al., 2010). In the current research, to investigate the impacts of priors on Bayesian estimators when normality is violated, we adopted the method proposed by Fleishman (1978) to generate non-normally distributed residuals for dependent variables. The distributions may be positively or negatively skewed.

Categorical (primarily binary) outcomes are common in cRCT research but may bring challenges to model estimation (Holtmann et al., 2016; Legare et al., 2015). Thus, in addition to continuous data (normally distributed for Study 1 and non-normally distributed for Study 2), the present research also considered conditions with binary outcomes (Study 3).

3.1.2. Estimators

To compare the performance of different estimators, we estimated the models with both frequentist methods (MLR for continuous data and WLSMV for binary data) and Bayesian methods. We also evaluated the impacts of different prior settings (including non-informative and informative priors) on the Bayesian analysis (see the example Mplus codes for Bayesian multilevel modeling in Appendix A). Since the treatment effects are usually of research interest, the present research only specified different informative priors for the regression parameters ($b_{10}-b_{30}$ and $b_{01}-b_{31}$) by adjusting the prior variances and prior means. Specifically, there are three levels of informative strengths: weakly informative priors, moderately informative priors, and strongly informative priors (with prior variances equaling 50%, 20%, and 10% of the corresponding true values of the parameters, respectively). The prior means may equal to (i.e., accurate informative priors), or deviate from the corresponding true values at different levels ($\pm 1SD$ or $\pm 3SD$) based on the prior variance terms (i.e., inaccurate informative priors; Depaoli, 2014). Figure 2 shows the informative priors for b_{10} as an example. The priors for the other parameters (e.g., residual variance) were set as the default. Thus, the number of different Bayesian estimators was $1 + 5 \times 3 = 16$, and the number of overall different estimators (including MLR or WLSMV) was 17 in total.

As mentioned before, WLSMV is relatively efficient in analyzing categorical data. However, to date, multilevel models with random slopes cannot be estimated with WLSMV in Mplus 8.4 (Barendse & Rosseel, 2020). Therefore, we conducted Study 3.1 and Study 3.2 separately for binary situation. In Study 3.1, we considered random slopes in the multilvel model like Study 1 and Study 2, and the MLR instead of WLSMV estimator was considered. To evaluate the performance of WLSMV in binary conditions, we additionally conducted Study 3.2, in which the simulated model only considered random intercepts (see Equation 6).

$$Y_{ij} = \beta_{0j} + b_{10}X_{1ij} + b_{20}X_{2ij} + b_{30}X_{3ij} + e_{ij} \tag{6}$$

3.1.3. Sample Sizes

Before deciding the sample sizes in the simulations, we explored typical sample sizes in cRCT research. We searched Web of Science using the keywords “cluster randomized controlled trials” and “multilevel” for relevant literature

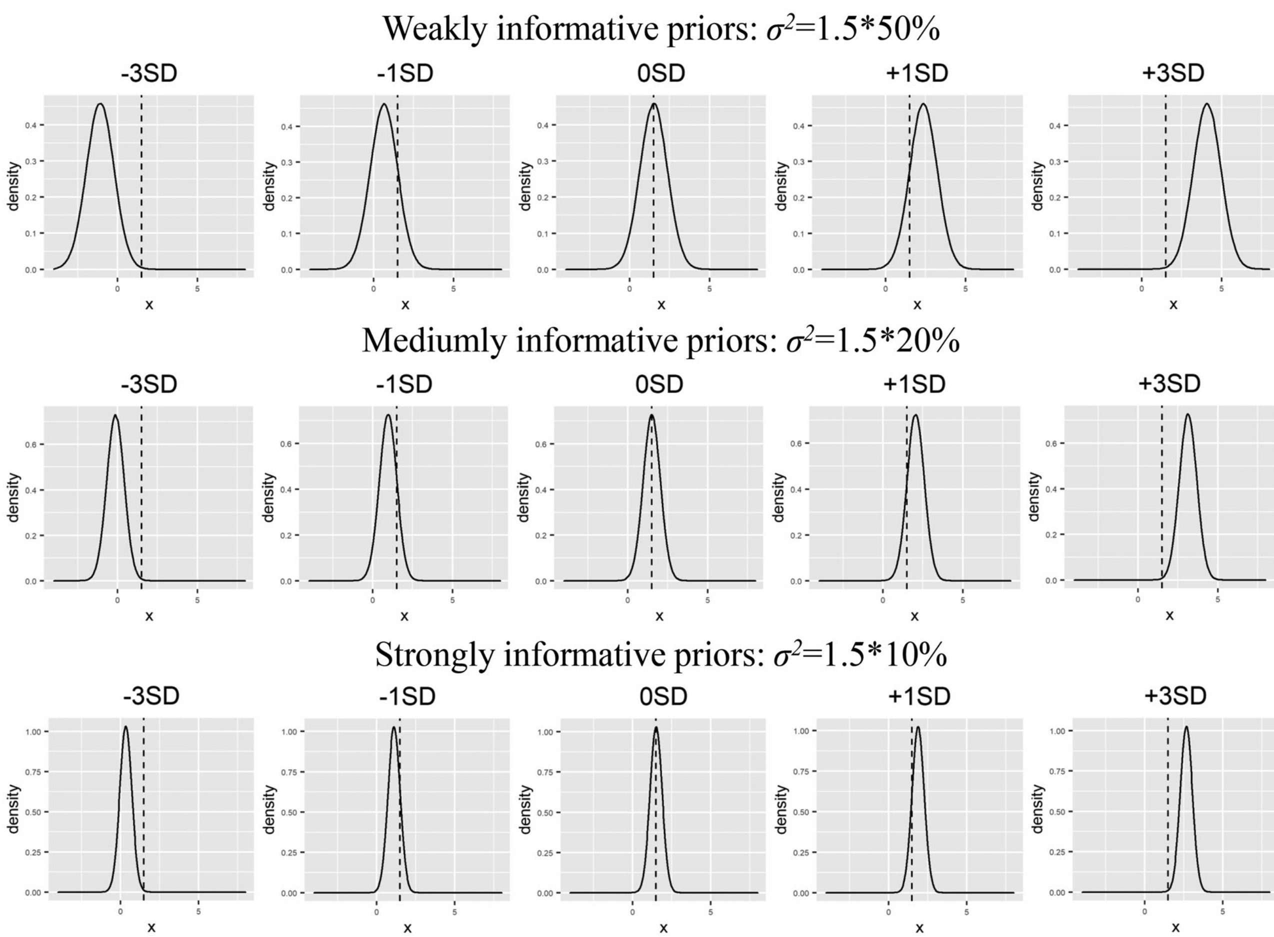

Figure 2. The example prior specifications for b_{01}. *Note*: The population value of b_{01} is 1.5 in the current simulations.

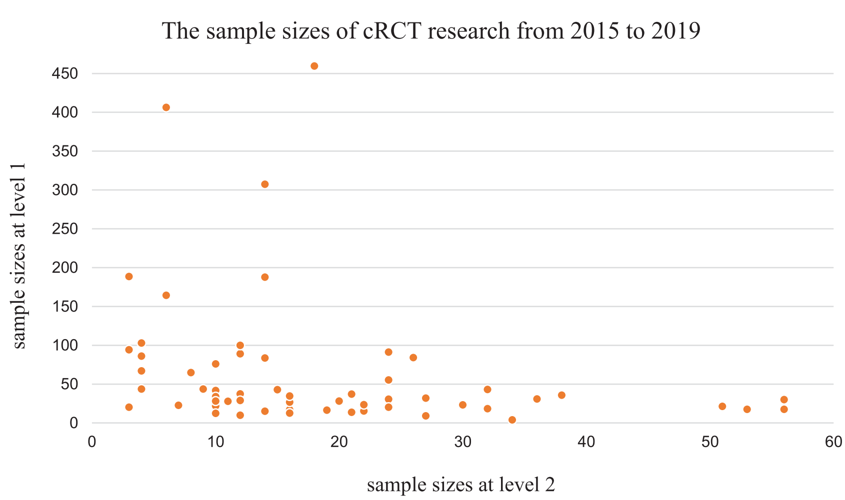

Figure 3. The sample sizes of cRCT research from 2015 to 2019.

published between 2015 and 2019. After screening the titles and abstracts, we obtained 95 cRCT papers and further coded their sample sizes at level 1 and level 2, if reported (Figure 3). As the number of clusters should be larger than the number of between parameters (Muthén, 2006), we set the sample sizes at level 2 as 20, 30, 40, and 50, and the sample sizes at level 1 were 30, 60, and 150, respectively.

3.1.4. ICC

As previously mentioned, ICC represents the proportion of the total variation explained at level 2. When ICC $\geq$0.059, it is necessary to conduct multilevel modeling to deal with hierarchical structures (Cohen, 1988). Following previous simulations (e.g., Depaoli & Clifton, 2015; Hox et al., 2010; Preacher & Kristopher, 2011), we set three levels of the ICC (0.05, 0.1, and 0.2) for the continuous data. For binary data

Table 1. The summary of simulation designs.

Study ID	Random-slope model	Dependent variables	ICC	NL2	NL1	Estimators
1	Yes	Normal	0.05, 0.1, 0.2	20, 30, 40, 50	30, 60, 150	1(MLR)+1(BD)+5*3 = 17
2	Yes	Non-normal	0.05, 0.1, 0.2	20, 30, 40, 50	30, 60, 150	1(MLR)+1(BD)+5*3 = 17
3.1	Yes	Binary	0.1, 0.2	20, 30, 40, 50	30, 60, 150	1(MLR)+1(BD)+5*3 = 17
3.2	No	Binary	0.1, 0.2	20, 30, 40, 50	30, 60, 150	1(WLSMV)+1(BD)+5*3 = 17

Note: MLR: robust maximum likelihood estimator; BD: Bayesian estimation with diffuse prior; WLSMV: weighted least squares estimator with mean and variance adjusted.

(Study 3.1 and Study 3.2), the residual variances at within level were fixed to 1 corresponding to the probit link function (Bauer & Sterba, 2011; Holtmann et al., 2016). To ensure that the residual variance is non-negative, we only considered two levels of ICCs (i.e., 0.1 and 0.2) in the binary conditions.

In summary, as shown in Table 1, the present research included four separate simulations (Study 1, Study 2, Study 3.1, and Study 3.2). Study 1 generated continuous and normally distributed outcomes, while Study 2 generated continuous and non-normal (including positively and negatively skewed) outcomes. Study 3.1 and Study 3.2 both generated binary outcomes, but Study 3.2 did not consider random slopes across the clusters like Study 1-Study 3.1. In addition, there are different levels of ICCs, sample sizes, and estimators in the simulations. Therefore, there were $3 \times 4 \times 3 \times 17 = 612$ cases in Study 1, $2 \times 3 \times 4 \times 3 \times 17 = 1224$ cases in Study 2, and $2 \times 4 \times 3 \times 17 = 408$ cases in Studies 3.1 and 3.2, respectively. Following previous research (Gates et al., 2017; Hsu et al., 2017; Jung et al., 2020; Li & Jacobucci, 2021), for each condition, 100 replications were performed in the current research.

3.2. Evaluation Criteria

In all the simulations, we assessed the convergence using the potential scale reduction (PSR), with the default criterion in Mplus (PSR <1.05). To evaluate the impacts of different factors on the estimators (e.g., ML, non-informative Bayesian estimation, and informative Bayesian estimation), we conducted ANOVAs for each study after excluding the non-converged replications. Then, the performance of parameter estimation was analyzed more specifically with different evaluation criteria, including the absolute relative bias (ARB) of parameter estimation, the ratio of standard error (SE) and SD (SE/SD), mean square error (MSE), 95% confidence interval coverage (95% CI coverage), and power.

3.2.1. ARB

ARB represents the absolute deviation of the parameter estimation from the population values, which is computed as follows:

$$ARB = \left|\frac{\hat{\theta} - \theta}{\theta}\right| \times 100\% \tag{7}$$

where $\hat{\theta}$ denotes the average parameter estimation of the converged replications, and θ denotes the corresponding true value of the parameter. ARB of less than 10% was considered acceptable (Flora & Curran, 2004).

3.2.2. SE/SD

SE/SD was calculated as the ratio of SE to SD for the converged replications. Asymptotically, the SE estimation (in other words, posterior SD) should be close to the empirical SDs, and the ratio between SE and SD should be approximately 1.0. In line with Cham et al. (2012), SE/SD between 0.9 and 1.1 was considered acceptable.

3.2.3. MSE

The MSE was computed as follows:

$$MSE = \frac{\sum \left(\hat{\theta}_i - \theta\right)^2}{N} \tag{8}$$

where N denotes the number of converged replications, $\hat{\theta}_i$ denotes the parameter estimation of the i^{th} converged sample. The smaller the MSE an estimator obtains, the more accurate the parameter estimation.

3.2.4. 95% CI Coverage

95% CI coverage refers to the percentage of converged replications in which the 95% confidence or credible interval (CI) contains the population value of the parameter. In general, 95% CI coverage should be close to 0.95.

3.2.5. Power

Power refers to the ability to successfully rejected the null hypothesis (H0). A powerful estimator is sensitive to detect truly existing effects. A higher power is typically more desirable.

3.3. Results of the Simulations

Before analyzing the simulation results, the Monte Carlo SE was calculated to quantify the uncertainty of the simulations (Morris et al., 2019, see in the Supplemental online material). The summarized Monte Carlo SE results for all the parameters across the conditions indicated that 100 replications per condition generally lead to a low Monte Carlo SE.

In Study 1, Study 2, and Study 3.2, all estimators converge successfully in all replications. However, in Study 3.1 (where the dependent variable is binary and there exist random slopes), the convergence rates of MLR decline to 0.62-1, while Bayesian estimation successfully converge in all the replications.

We only specify different priors for the regression coefficients (which are also of main interest), so only the performance of these parameters would be analyzed and compared in different situations. We checked the ANOVA

Table 2. Effect sizes (partial η^2) of study 1.

	ARB			SE/SD			MSE			95 Coverage			Power		
Effect	MLR	BD	INF	MLR	BD	INF	MLR	BD	INF	MLR	BD	INF	MLR	BD	INF
ICC	0.004	0.005	0.006	0.016	0.022	0.030	0.014	0.013	**0.481**	0.002	0.001	**0.067**	0.000	0.007	0.009
NL2	**0.092**	**0.093**	**0.620**	0.028	**0.152**	**0.948**	0.160	**0.159**	**0.515**	0.004	0.002	**0.256**	**0.124**	**0.284**	**0.253**
NL1	0.036	0.037	**0.182**	0.007	0.010	**0.761**	**0.064**	**0.064**	**0.073**	0.001	0.002	0.031	**0.065**	**0.085**	**0.078**
ICC * NL2	0.001	0.001	0.001	0.015	0.019	**0.143**	0.002	0.002	**0.087**	0.002	0.001	0.002	0.006	0.010	0.010
ICC * NL1	0.001	0.001	0.003	0.009	0.011	**0.178**	0.001	0.001	0.003	0.002	0.001	0.000	0.001	0.002	0.002
NL2 * NL1	0.001	0.001	0.004	0.027	0.042	**0.547**	0.007	0.007	0.009	0.001	0.002	0.001	0.003	0.004	0.006
ICC * NL2 * NL1	0.002	0.002	0.003	0.031	0.030	**0.133**	0.002	0.002	0.003	0.002	0.004	0.004	0.004	0.004	0.004
Mag	–	–	**0.925**	–	–	**0.941**	–	–	**0.664**	–	–	**0.871**	–	–	**0.568**
Mag * ICC	–	–	0.001	–	–	0.160	–	–	**0.255**	–	–	0.033	–	–	**0.078**
Mag * NL2	–	–	**0.231**	–	–	**0.849**	–	–	**0.277**	–	–	**0.179**	–	–	**0.069**
Mag * NL1	–	–	0.036	–	–	**0.105**	–	–	0.016	–	–	0.018	–	–	0.029
Mag * ICC * NL2	–	–	0.001	–	–	**0.413**	–	–	0.047	–	–	0.003	–	–	0.012
Mag * ICC * NL1	–	–	0.002	–	–	**0.158**	–	–	0.002	–	–	0.002	–	–	0.002
Mag * NL2 * NL1	–	–	0.004	–	–	**0.182**	–	–	0.009	–	–	0.003	–	–	0.004
Mag * ICC * NL2 * NL1	–	–	0.002	–	–	**0.389**	–	–	0.005	–	–	0.005	–	–	0.004
Strength	–	–	**0.957**	–	–	**0.999**	–	–	**0.531**	–	–	**0.678**	–	–	**0.646**
Strength * ICC	–	–	0.004	–	–	0.014	–	–	**0.153**	–	–	0.020	–	–	0.004
Strength * NL2	–	–	**0.383**	–	–	**0.972**	–	–	**0.183**	–	–	0.034	–	–	**0.064**
Strength * NL1	–	–	**0.061**	–	–	**0.897**	–	–	0.014	–	–	0.006	–	–	0.030
Strength * ICC * NL2	–	–	0.003	–	–	**0.297**	–	–	**0.145**	–	–	0.004	–	–	0.008
Strength * ICC * NL1	–	–	0.001	–	–	**0.296**	–	–	0.001	–	–	0.001	–	–	0.001
Strength * NL2 * NL1	–	–	0.003	–	–	**0.486**	–	–	0.001	–	–	0.002	–	–	0.003
Strength * ICC * NL2 * NL1	–	–	0.003	–	–	**0.248**	–	–	0.002	–	–	0.003	–	–	0.006
Mag * strength	–	–	**0.346**	–	–	**0.915**	–	–	0.014	–	–	**0.497**	–	–	**0.123**
Mag * strength * ICC	–	–	0.001	–	–	**0.159**	–	–	**0.068**	–	–	0.009	–	–	0.007
Mag * strength * NL2	–	–	0.090	–	–	**0.776**	–	–	0.056	–	–	0.010	–	–	0.043
Mag * strength * NL1	–	–	0.007	–	–	**0.116**	–	–	0.002	–	–	0.003	–	–	0.004
Mag * strength * ICC * NL2	–	–	0.001	–	–	**0.333**	–	–	**0.122**	–	–	0.004	–	–	0.011
Mag * strength * ICC * NL1	–	–	0.001	–	–	**0.288**	–	–	0.001	–	–	0.001	–	–	0.001
Mag * strength * NL2 * NL1	–	–	0.003	–	–	**0.135**	–	–	0.003	–	–	0.002	–	–	0.004
Mag * strength * ICC * NL2 * NL1	–	–	0.002	–	–	0.380	–	–	0.004	–	–	0.003	–	–	0.003

Note: MLR: robust maximum likelihood estimator; BD: Bayesian estimator with default priors; INF: Bayesian estimator with informative priors. ICC: intraclass coefficient; NL1: sample size at level 1; NL2: sample size at level 2; mag: the magnitude of prior mean deviation; strength: the strength of prior information. **Bold** effect sizes are larger than 0.06.

results for each study before analyzing the evaluation criteria (i.e., ARB, SE/SD, MSE, 95% CI coverage, and power). The ANOVA (e.g., Table 2[3]) indicated that the levels of ICC, the sample sizes at both level 1 and level 2, the prior variance, and the magnitude of deviation of prior means may impact parameter estimation differently. Additionally, the effects of these factors may interact with each other at the same time.

In the cRCT, the treatment effects on the outcome variable (b_{01}) are usually of interest. Therefore, the results for b_{01} are presented in more detail.

3.3.1. Study 1: Normally Distributed Outcomes

3.3.1.1. ARB. To show whether the estimators underestimated or overestimated the parameters, we present relative bias (RB) rather than ARB for b_{01} in Figure 4. MLR obtains a similar ARB as the Bayesian estimation with non-informative priors. For Bayesian estimators with incorrect informative priors, the ARB of parameter estimation decreases with both the prior variance and the magnitude of the prior mean deviation decrease. A large sample size at level 2 buffers the negative impacts of incorrect prior information. For the treatment effect on the outcome variable (b_{01}), the negative effects of incorrect prior information tended to increase as the ICC increased. However, this trend was not evident in the results of the other path coefficients.

3.3.1.2. SE/SD. Figure 5 presents the SE/SD results for b_{01} as an example. In general, Bayesian estimation with informative priors overestimates the SE/SD ratio. With an increase in prior variance, the ratio of SE/SD decreased toward 1. An increase in the sample size at level 2 can reduce the SE/SD ratio in the informative Bayesian estimation. Moreover, the ratio of SE/SD may be higher if the prior mean deviates from the population value more severely.

3.3.1.3. MSE. The MSE results are presented for b_{01} in Figure 6. MLR and non-informative Bayesian estimators tend to obtain large MSEs in the case of small samples. The Bayesian estimation with the correct informative priors obtains the lowest MSE. However, Bayesian methods with incorrect prior information led to a much larger MSE. The increase in the sample size at both levels decreased the MSE of the parameter estimation. Still, the impacts of the level 2 sample sizes are more evident than those of the level 1 sample sizes. In addition, a large prior variance decreases the negative impacts of incorrect prior information in Bayesian methods.

3.3.1.4. Coverage. Figure 7 presents the 95% CI coverage of parameter estimation for b_{01}. The coverage of parameter estimation increases significantly as the sample size at level 2

[3]Due to limited space, only the results of study 1 are presented. The ANOVA results of study 2 – study 3.2 are available in the Supplemental online material.

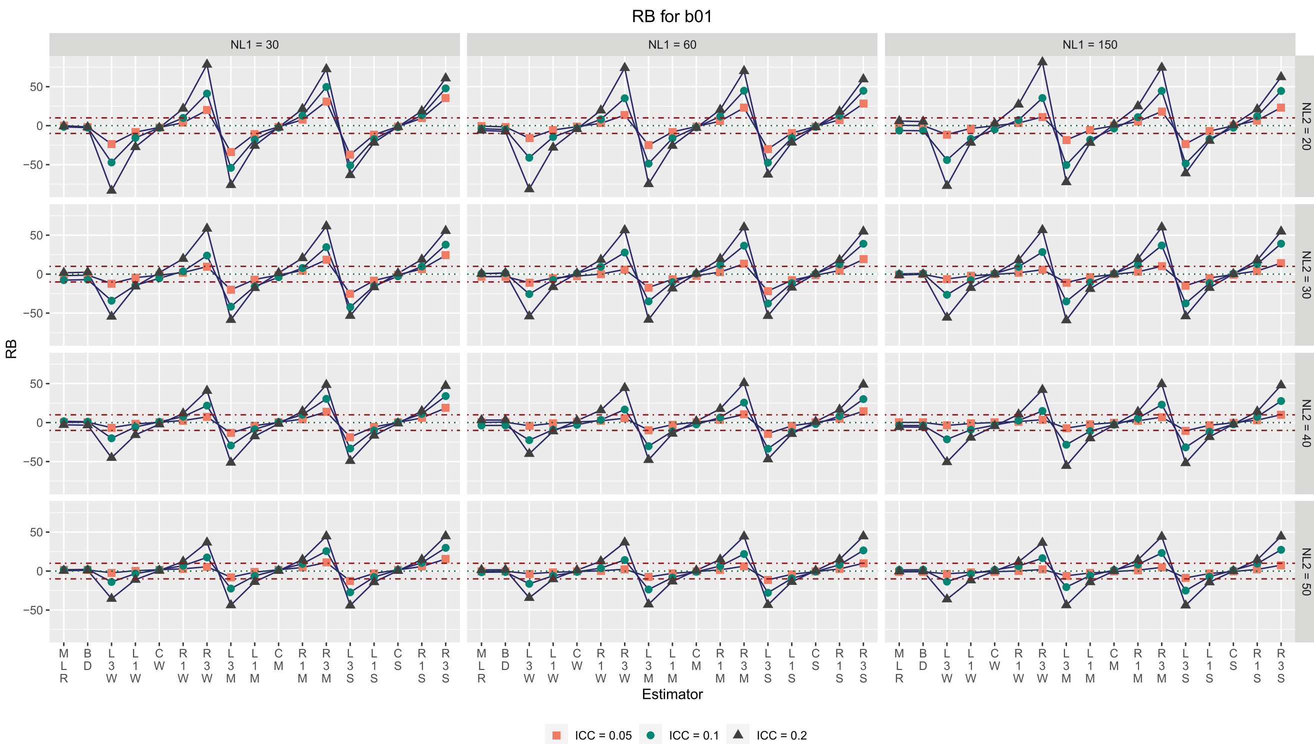

Figure 4. Relative bias (RB) for b_{01} in study 1. *Note*: MLR: robust maximum likelihood estimator; BD: Bayesian estimator with default priors. L3: Bayesian estimator with the prior means deviating 3SD from the true value to the left; L1: Bayesian estimator with the prior means deviating 1SD from the true value to the left; C: Bayesian estimator with the prior means equaling the true value; R1: Bayesian estimator with the prior means deviating 1SD from the true value to the right; R3: Bayesian estimator with the prior means deviating 3SD from the true value to the right. W: weakly informative priors (prior variance = population value*50%); M: mediumly informative priors (prior variance = population value*20%); S: strongly informative priors (prior variance = population value*10%). Dotted lines: $RB = 0$; dot dashed lines: $RB = \pm 0.1$.

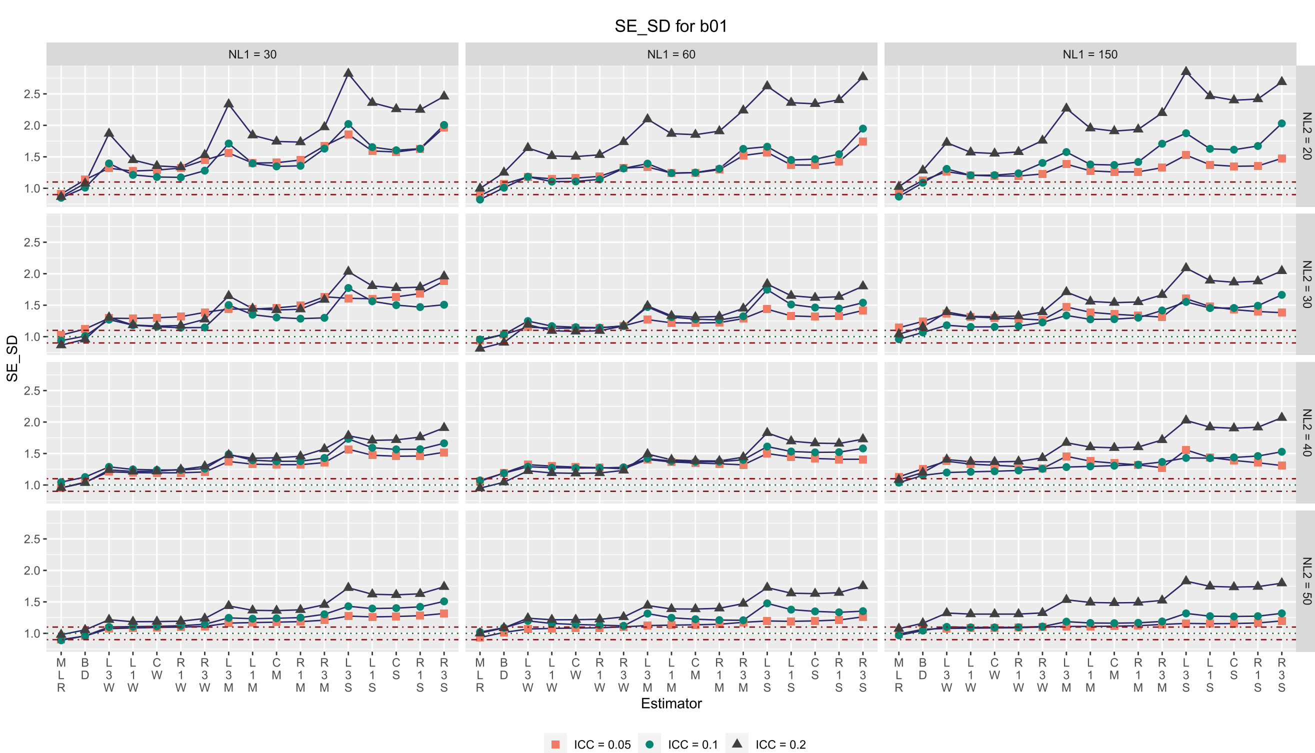

Figure 5. The ratio of SE and SD for b_{01} in study 1. *Note*: MLR: robust maximum likelihood estimator; BD: Bayesian estimator with default priors. L3: Bayesian estimator with the prior means deviating 3SD from the true value to the left; L1: Bayesian estimator with the prior means deviating 1SD from the true value to the left; C: Bayesian estimator with the prior means equaling the true value; R1: Bayesian estimator with the prior means deviating 1SD from the true value to the right; R3: Bayesian estimator with the prior means deviating 3SD from the true value to the right. W: weakly informative priors (prior variance = population value*50%); M: mediumly informative priors (prior variance = population value*20%); S: strongly informative priors (prior variance = population value*10%). Dot dashed lines: SE/SD = 0.9 or SE/SD = 1.1.

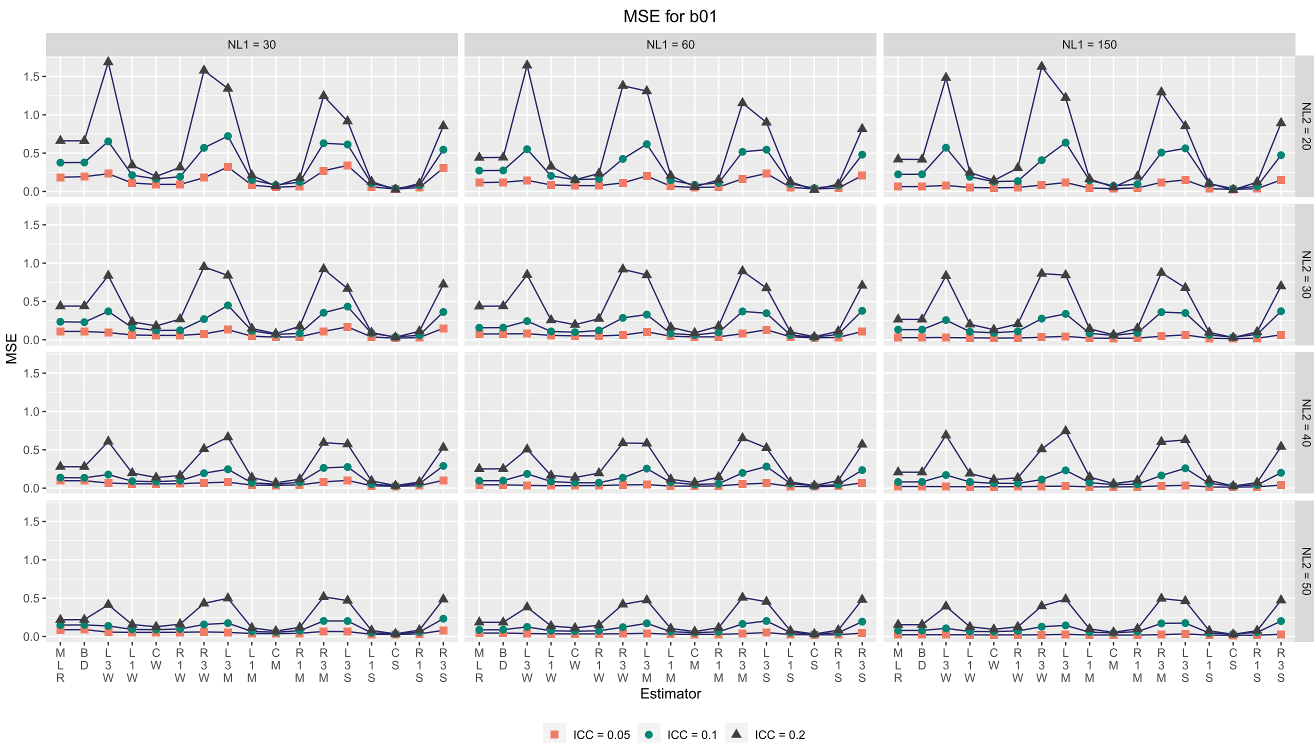

Figure 6. The MSE for b_{01} in study 1. *Note*: MLR: robust maximum likelihood estimator; BD: Bayesian estimator with default priors. L3: Bayesian estimator with the prior means deviating 3SD from the true value to the left; L1: Bayesian estimator with the prior means deviating 1SD from the true value to the left; C: Bayesian estimator with the prior means equaling the true value; R1: Bayesian estimator with the prior means deviating 1SD from the true value to the right; R3: Bayesian estimator with the prior means deviating 3SD from the true value to the right. W: weakly informative priors (prior variance = population value*50%); M: mediumly informative priors (prior variance = population value*20%); S: strongly informative priors (prior variance = population value*10%).

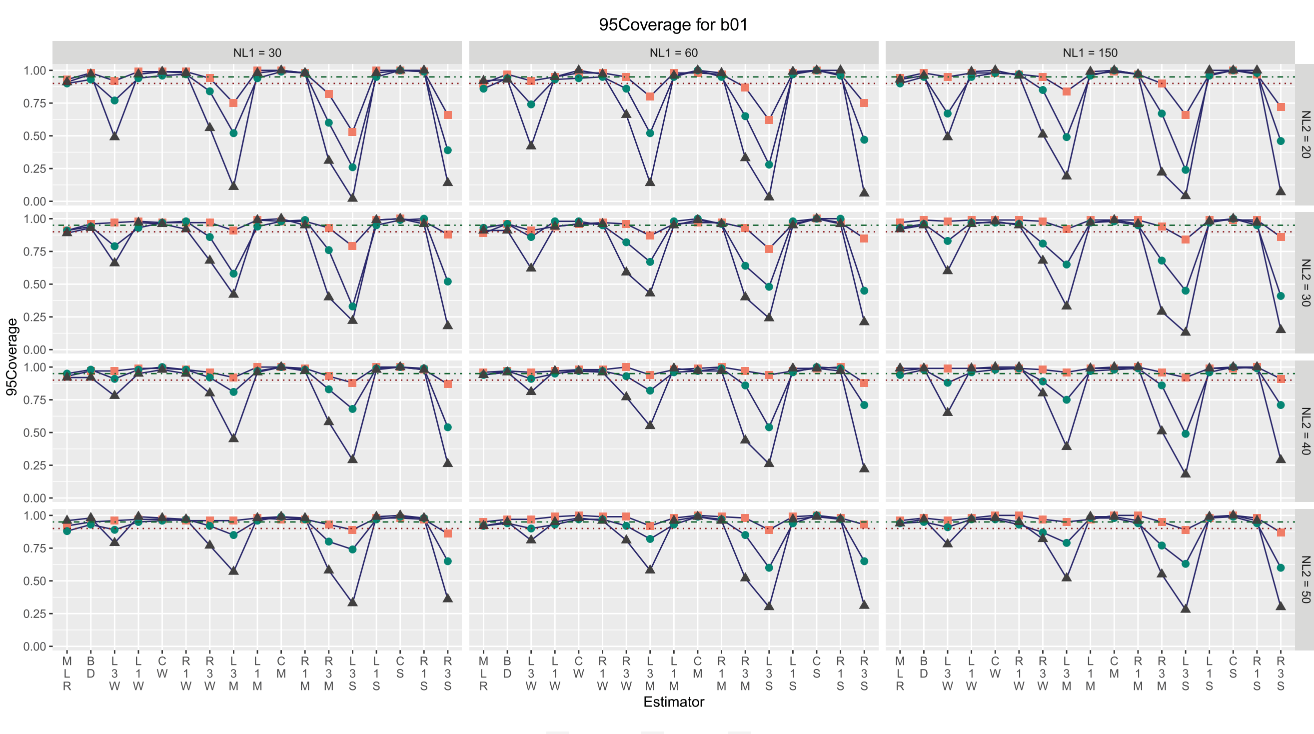

Figure 7. The 95% CI coverage for b_{01} in study 1. *Note*: MLR: robust maximum likelihood estimator; BD: Bayesian estimator with default priors. L3: Bayesian estimator with the prior means deviating 3SD from the true value to the left; L1: Bayesian estimator with the prior means deviating 1SD from the true value to the left; C: Bayesian estimator with the prior means equaling the true value; R1: Bayesian estimator with the prior means deviating 1SD from the true value to the right; R3: Bayesian estimator with the prior means deviating 3SD from the true value to the right. W: weakly informative priors (prior variance = population value*50%); M: mediumly informative priors (prior variance = population value*20%); S: strongly informative priors (prior variance = population value*10%). Dotted lines: 95% CI coverage = 0.9; dot dashed lines: 95% CI coverage = 0.95.

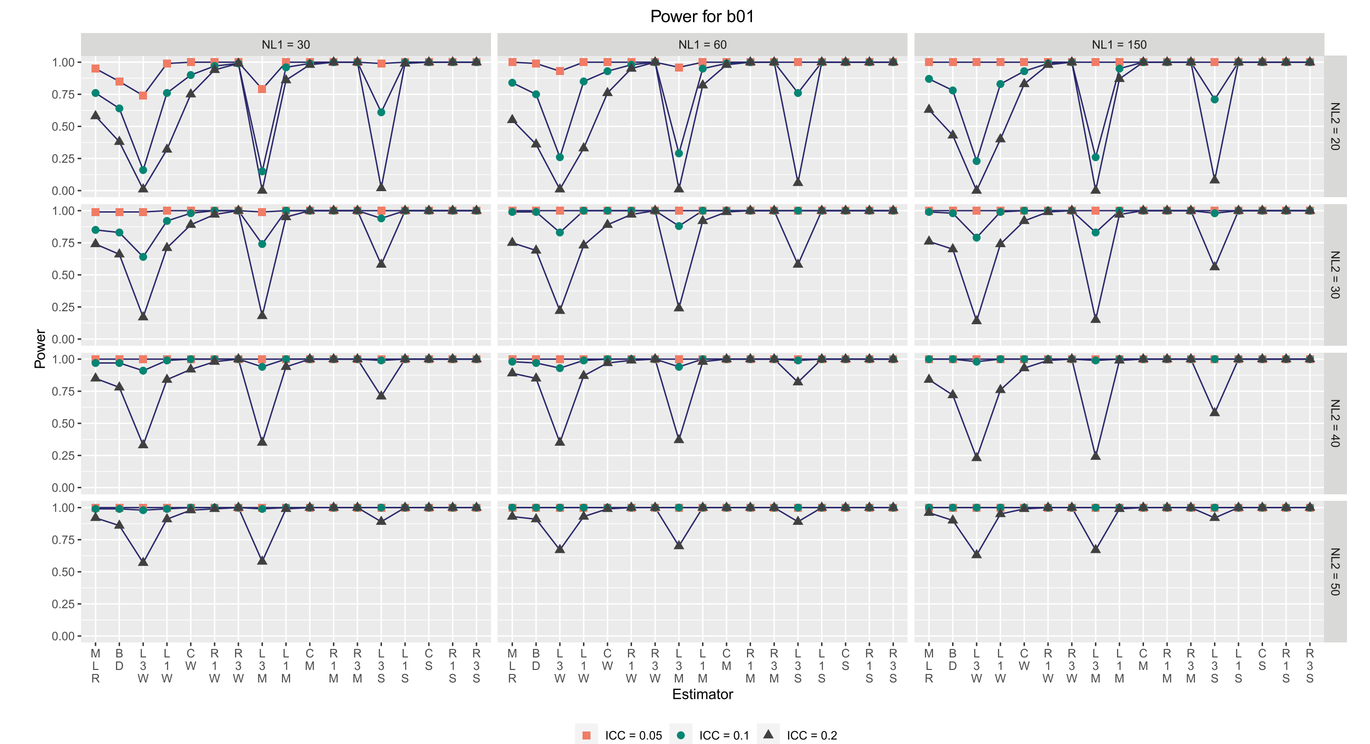

Figure 8. The power for b_{01} in study 1. *Note*: MLR: robust maximum likelihood estimator; BD: Bayesian estimator with default priors. L3: Bayesian estimator with the prior means deviating 3SD from the true value to the left; L1: Bayesian estimator with the prior means deviating 1SD from the true value to the left; C: Bayesian estimator with the prior means equaling the true value; R1: Bayesian estimator with the prior means deviating 1SD from the true value to the right; R3: Bayesian estimator with the prior means deviating 3SD from the true value to the right. W: weakly informative priors (prior variance = population value*50%); M: mediumly informative priors (prior variance = population value*20%); S: strongly informative priors (prior variance = population value*10%).

increases, but only increases slightly as the sample size at level 1 increases. For Bayesian analysis with a slightly deviated informative prior (magnitude of deviation = 1 SD), the 95% CI coverage of the parameter estimation was acceptable (>90%). However, if the prior mean deviates by 3 *SD* from the population value of the parameter, the 95% CI coverage rate declines considerably. Large sample sizes at level 2 and large ICC tend to buffer the negative impacts of incorrect prior information.

3.3.1.5. Power. The power results for b_{01} are presented as an example in Figure 8. When the sample sizes are small and the ICC is large, MLR and Bayesian estimation with non-informative priors obtain low power. For the Bayesian estimation with correct prior information, the power of the parameter estimation is acceptable (>0.8), even in the case of small samples. If the prior tends to underestimate the magnitude of effects (that is, the absolute values of the parameters), the corresponding power of the parameter estimation is lower than that of the correct priors. If the prior overestimates the magnitude of effects, the Bayesian estimator will obtain a higher power. However, in this situation, the high power of parameter estimation may increase Type I error rates. With the increase in the sample size at level 2 and the decrease in the ICC, the negative impacts of incorrect prior information on power decrease accordingly.

3.3.1.6. Conclusions. In general, Bayesian estimation with correct informative priors performs best, especially in terms of the ARB of parameter estimation. The performance of Bayesian estimation with non-informative priors and MLR is similar in parameter estimation. Both MLR and Bayesian estimation with non-informative priors perform poorly in terms of MSE and power, especially in small samples and large ICCs. For Bayesian analysis with incorrect informative priors, a large ICC and a small sample size at level 2 may increase the negative impacts of incorrect prior information. Stronger informative priors also lead to a more negative influence of incorrect prior information.

3.3.2. Study 2: Non-Normally Distributed Outcomes

To further explore the impacts of incorrect informative priors in the situation of non-normally distributed outcomes, we generated non-normally distributed dependent variables in Study 2. We compared the results of negatively and positively skewed data and found that the impact of the direction of skewness was negligible. In addition, the impacts of the ICC, sample sizes at both levels, the prior variance, and the magnitude of deviation of prior means in Study 2 are roughly consistent with those in Study 1. Therefore, to save space, we plotted the results of study 2 in the Supplemental online material.

3.3.2.1. ARB. In study 2, MLR and Bayesian estimators with non-informative priors obtain relatively robust ARBs, regardless of the non-normality of the dependent variables. However, in Bayesian estimators with incorrect informative priors, non-normality tends to increase the ARB for parameter estimation when the level of the ICC is high.

3.3.2.2. SE/SD. In general, the ratio of SE/SD is larger in non-normally distributed data (Study 2) than in normally distributed data (Study 1). The non-normality of the data significantly increases the ratio of SE/SD for Bayesian estimators with (both correctly and incorrectly) informative priors. In non-normally distributed data, SE/SD mainly depends on the strength of the prior information.

3.3.2.3. MSE. As for the results of MSE, MLR and non-informative Bayesian estimators behave worse for non-normal data than normal ones. For informative Bayesian, non-normality dramatically increases the negative impacts of incorrect prior information. In addition, non-normality tends to increase the adverse effects of ICC for b_{01}. However, the effects of ICC are not consistent with other coefficients.

3.3.2.4. Coverage. As for the 95% CI coverage of parameter estimation for b_{01}, MLR and non-informative Bayesian estimators were hardly affected by the normality of the data. For informative Bayesian estimators, the impact of non-normality differs among different levels of the ICC. If the ICC is high, non-normality tends to decrease the coverage of parameters. However, if the ICC is low (0.05 or 0.1), the non-normality slightly increases the coverage of parameters.

3.3.2.5. Power. Similarly, the impacts of non-normality differ among the different levels of the ICC for MLR and Bayesian estimators. Non-normality leads to low power when the level of ICC is high, but this is reversed when the level of the ICC is low. When the ICC is low, the non-normal data obtain slightly higher power than the normal data, especially when the sample size at level 2 is small.

3.3.2.6. Conclusions. Study 2 generated data from negatively and positively skewed distributions separately. Generally, MLR and Bayesian estimators perform worse in Study 2 (non-normal situations) than in Study 1 (normal-data conditions). The results indicate that the difference between the specific patterns of skewness (negatively or positively skewed) is negligible. The increase in the sample size at level 2 and prior variance decreases the negative impacts of incorrect prior information. In addition, non-normality tends to increase the adverse effects of incorrect prior information regarding parameter estimation bias.

3.3.3. Study 3.1: Binary Outcomes with Random Slopes

Both Study 3.1 and Study 3.2 analyzed binary dependent variables. In Study 3.1, we conducted multilevel modeling with random slopes, while in Study 3.2, we considered multilevel models only with random intercepts. In Study 3.1, all the models converged successfully with Bayesian estimation, while the convergence rate of MLR ranges between 0.62 and 1. To save space, the results of study 3.1 and study 3.2 are shown in the Supplemental online material.

3.3.3.1. ARB. Generally, models with binary dependent variables obtain larger ARBs than those with continuous and normally distributed variables (Study 1). MLR performs similarly to the non-informative Bayesian estimator, yielding a small parameter estimation bias. The ARB of all estimators decreases as the sample size at level 2 increases. A large sample size at level 2 buffers the negative impacts of incorrect prior information. In addition, a large prior variance weakens the strength of priors and the adverse effects of incorrect prior information.

3.3.3.2. SE/SD. The results of the SE/SD ratio for b_{01} indicated that MLR is unstable in binary variables and may obtain a large SE/SD. For Bayesian estimators, informative priors generally obtain larger ratios of SE/SD than non-informative priors. A smaller prior variance leads to larger ratios of SE/SD, and incorrect prior information tends to increase the SE/SD. This trend decreases as the sample size at level 2 increases.

3.3.3.3. MSE. With the increase in the sample size at level 2, the negative impacts of incorrect prior information on the MSE decrease accordingly. The decrease in ICC also tends to buffer the adverse effects of incorrect informative priors.

3.3.3.4. Coverage. As the results of 95% CI coverage for b_{01} show, the coverage of the parameter estimation increases as the sample size at level 2 increases. However, the coverage of MLR tends to decrease with an increase in the sample size at level 1. For Bayesian estimators, large sample sizes at level 2 buffer the negative impacts of incorrect prior information on coverage. In addition, a large prior variance tends to reduce the effects of incorrect informative priors.

3.3.3.5. Power. Generally, the power of MLR and Bayesian estimation with non-informative priors is relatively low, especially for small samples. The increase in sample sizes at both levels 1 and 2 increases the power of the parameters efficiently. The rise of ICC tends to decrease the power of parameter estimation.

3.3.3.6. Conclusions. Similar to the patterns of the continuous data settings, as the ICC decreases and the sample sizes increase, the impacts of incorrect informative priors decrease accordingly. However, compared to Study 1, the performance of parameter estimation is more sensitive to the specifications of the priors. In addition, the buffering effects of the sample size at level 2 are more evident.

In Study 3.1, the MLR estimator requires a large amount of multidimensional integration, thus leading to problems in terms of convergence. In the replications that MLR converges successfully, the parameter estimation and the ratio of SE/SD may be seriously biased. The power of the parameters was low in the small sample cases. In contrast, Bayesian estimators can successfully reach to convergence and perform well in terms of parameter estimation even in small samples.

3.3.4. Study 3.2: Binary Outcomes without Random Slopes

To compare the performance of Bayesian estimators to the WLSMV (which is more commonly used in binary data), we additionally conducted Study 3.2 based on Study 3.1. In Study 3.2, we simplified the models and omitted random slopes in the models. Thus, the only treatment effect (b_{01}) was the main focus of the research.

3.3.4.1. ARB. The model in Study 3.2 is simpler than that of Study 3.1 and results in a lower estimation bias in general. Similar to Study 3.1, the estimation of parameters is sensitive to the specifications of priors. In addition, the impacts of priors on Bayesian analysis show a similar pattern to that of Study 3.1. The ARB of the WLSMV was relatively small across the conditions, even in small samples. However, Bayesian estimation with non-informative priors may lead to more severe ARB than those with slightly incorrect prior information in small samples.

3.3.4.2. SE/SD. Likewise, informative Bayesian estimators yield higher ratios of SE and SD than Bayesian estimation with non-informative priors and WLSMV in general. The decreasing of the prior variance and the increasing of the prior mean deviation lead to increased ratios of SE and SD, especially when the level of ICC is high. When the level-2 sample size and ICC are small, the WLSMV obtains a large ratio of SE and SD, which may even surpass those of informative Bayesian estimators.

3.3.4.3. MSE. Across all estimators, a large sample size at level 2 decreases the MSE more significantly than the sample size at level 1. Bayesian estimation with non-informative priors is highly sensitive to the sample sizes and obtains an overly large MSE when the sample sizes at both levels are small. Likewise, seriously incorrect prior information significantly increases the MSE of parameter estimation. The MSE of the WLSMV was slightly smaller than that of the non-informative Bayesian. However, when the sample size at level 2 is small, the WLSMV obtains an even larger MSE than the Bayesian estimator with slightly incorrect informative priors.

3.3.4.4. Coverage. Apart from the Bayesian estimator with seriously incorrect prior information, all other estimators perform well across different situations (regarding 95% CI coverage). When the ICC was low, 95% CI coverage was generally acceptable. However, a large ICC increases the impact of incorrect prior information, especially when the sample sizes and prior variance are small.

3.3.4.5. Power. In Study 3.2, the power of different estimators (including Bayesian estimation with serious incorrect prior information) is generally high when the sample size at level 2 is large. However, when the sample size at level 2 is small, the power of the Bayesian estimation is significantly influenced by the corresponding prior information. In addition, WLSMV obtains low power in small samples. Bayesian estimation with non-informative priors obtains higher power than that of the WLSMV and performs similarly to Bayesian estimation with weak prior information.

3.3.4.6. Conclusions. In summary, when the level-2 sample size and ICC are small, the WLSMV obtains large ratios of SE and SD, large MSE, and low power. Bayesian estimators with relatively strong informative priors generally perform well. However, specifying non-informative priors may lead to seriously biased parameter estimations in small samples. In addition, similar to the results of Study 3.1, the negative impacts of incorrect informative priors are more significant than those under the continuous data conditions.

4. Empirical Analysis: Democratic School Health Education

4.1. Research Background

A publicly available dataset from a cRCT research[4] (Boberova et al., 2017) were used to further illustrate the conclusions of simulations. In the cRCT research, Boberova et al. (2017) evaluated the effect of the democratic school health education intervention, which is based on the investigation-vision-action-change (IVAC) model. Participants were 180 Slovakia fourth-grade pupils from 10 classes (52% male and the average age was 10.29). To minimize the effect of treatment contamination, the intervention was conducted in class units. Researchers randomly selected 5 classes as experimental groups and the remaining 5 classes were control groups. In the experimental classes, students were instructed to attend the intervention program of democratic school health education. In control classes, however, students just took their normal courses as usual.

Specifically, the intervention program for the experimental classes included 4 phases. Firstly, students were instructed to pay close attention to, think about, and describe their health condition. For example, students should continuously record their activities and corresponding feelings for a week. Secondly, they were encouraged to think about the impacts of the campus environment on their physical and mental health, and have extensive discussions on factors that can improve their well-being in campus. Next, based on the discussion, students were inspired to develop specific plans for health improvement, and present the plan directly to the principal of the campus. Finally, students were guided to critically review their efforts during the program. In both experimental and control classes, all students were required to answer some questionnaires before (pre-measurement) and after the intervention (post-measurement). Key outcome variables included perception of school, subjective well-being, and violent behavior in school.

[4]Retrieved from osf.io/m5cth.

4.1.1. Data Analysis

Two multilevel analysis were conducted based on the empirical data. In empirical example 1, post-measured well-being was selected as the outcome, corresponding to simulation study 1 (in which the outcome is normally distributed). In empirical example 2, the post-measured violent behavior was the outcome instead, corresponding to simulation study 2 (in which the outcome is non-normally distributed). In empirical studies, it is generally more difficult to specify appropriate prior information than in simulations. As previous research suggested (Brown, 2008; Holtmann et al., 2016; van der Linden, 2008), due to the lack of prior knowledge, we conducted ML estimation first and assumed that the parameter estimation of the ML is true. Next, as the simulations, we imposed priors with different mean and variance hyperparameters based on the ML estimation (that is, data-dependent priors, DDP). For the prior distribution $N(\mu, \sigma^2)$, the mean hyperparameter μ would deviate -3SD, -1SD, 0SD, +1SD, or +3SD from the corresponding estimates of ML. σ^2, the variance hyperparameter, would be 10%, 20%, or 50% of the corresponding estimates of ML.

4.2. Results

4.2.1. Empirical Example 1: Normally Distributed Outcomes

In the first empirical example, post-measured well-being was the dependent variable. Age, gender, and pre-measured well-being were independent variables at level 1, while the intervention condition (experimental vs. control group) was the independent variable at level 2 (class level). We deleted missing data listwise and obtained 158 complete answers based on the variables that we were interested in. The ICC of the dependent variable (the post-measured well-being) was 0.208, so multilevel modeling was generally suggested (Cohen, 1988). Since the effect of the intervention was the research focus, we grand-mean centered the independent variables at level 1 according to previous suggestions (Enders & Tofighi, 2007; McNeish, 2016a). Like the simulation studies, we considered random slope effects in the model to investigate the cross-level effects of the intervention (Figure 9). As the skewness and kurtosis of the dependent variable (post-measured well-being) were −0.482 and 1.571 respectively, and less than the rule-of-thumb cutoff values ($|\text{skewness}| \leq 2$ and $|\text{kurtosis}| \leq 7$; Curran et al., 1996) we assumed the pattern of results were similar to those of the simulation study 1, in which the dependent variable is normally distributed.

Table 3 presents the results of ML estimation. Compared to pupils in control classes, pupils in experimental classes develop a significantly higher level of well-being after the intervention of democratic school health education ($b_{01} = 3.342$, $p < .001$). Based on the results of ML estimation, we further conducted Bayesian estimation with informational priors. Specifically, we set the prior mean hyperparameter based on the ML parameter estimation, and the prior variance hyperparameter based on the standard error of the ML estimation. Like the ML, the result of

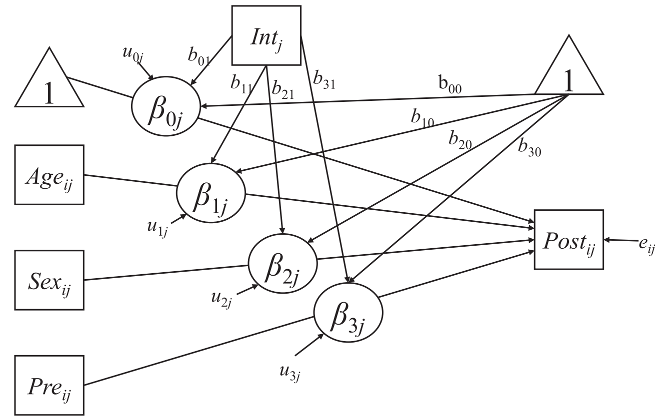

Figure 9. The path diagram of the empirical examples. *Note*: Int_j: the democratic school health education intervention at level 2; Pre_{ij}: the pre-measured well-being (in empirical example 1) or violent behavior (in empirical example 2); $Post_{ij}$: the post-measured well-being (in empirical example 1) or violent behavior (in empirical example 2). e_{ij}: the residual terms at level 1; u_{0j} : the residual term of the random intercept (β_{0j}); $u_{1j}-u_{3j}$: residual terms of the random slopes ($\beta_{1j}-\beta_{3j}$).

Table 3. Empirical example 1: the results of MLR analysis.

Para	est	SE	est/SE	*p*
b11	−1.409	2.539	−0.555	0.579
b21	1.977	1.503	1.315	0.189
b31	0.139	0.180	0.775	0.438
b01	3.342**	0.947	3.528	<.001
b10	−0.621	0.800	−0.776	0.438
b20	−1.105	1.268	−0.872	0.383
b30	0.097	0.128	0.763	0.445

Note: est: the estimation of parameters; SE: the standard error; $^{**}p<.001$.

Table 4. Empirical example 1: the results of Bayesian analysis with DDP.

			95% CI	
Para	est	SD	Lower 2.5%	Upper 2.5%
b11	−1.570	1.673	−4.883	1.748
b21	2.172	1.025	0.072	4.184
b31	0.127	0.132	−0.125	0.392
b01	3.362	0.725	1.894	4.782
b10	−0.587	0.691	−1.886	0.812
b20	−1.023	0.801	−2.654	0.486
b30	0.090	0.091	−0.082	0.269

Note: DDP: data − dependent priors; est: the estimation of parameters; sd: the standard deviation of the posterior distribution; 95% CI: the 95% credible intervals of parameters.

Bayesian estimation (Table 4) indicates significant intervention effects on the well-being ($b_{01} = 3.362$, 95% CI = [1.894, 4.782]). In addition, the invention significantly mitigates the sex difference in well-being ($b_{21} = 2.172$, 95% CI = [0.072, 4.184]). That is, boys may benefit more than girls in the program, and the gap in well-being between boys and girls may be smaller after the intervention of democratic school health education.

To further explore the effect of priors on Bayesian estimation, we imposed different priors on the regression parameters (as the simulations), and the pattern of results (Tables 5 and 6) is similar to that of the simulations. When the prior mean hyperparameters seriously deviate from the actual effects (i.e., the parameter estimation in the ML), Bayesian estimators obtain seriously biased parameter estimation and may fail to detect the effect of the intervention. Bayesian estimators with slightly deviated priors, however, obtain parameter estimation that is relatively close to the actual effects.

4.2.2. Empirical Example 2: Non-Normally Distributed Outcomes

In empirical example 2, we selected post-measured violent behavior as the dependent variable. Likewise, age, gender, and pre-measured violent behavior were selected as level-1 independent variables, and the intervention condition (experimental vs. control group) was the level-2 independent variable. After deleting the missing data listwise (based on the dependent and independent variables), we obtained 157 complete samples in total. The ICC of the dependent variable (the post-measured violent behavior) was.084, so we'd better analyze the data with multilevel models (Cohen, 1988). Likewise, we grand-mean centered the independent variables at level 1, and considered random slope effects in the model. As the skewness and kurtosis of the dependent variable are relatively large in empirical example 2 (skewness $=$ 2.045, kurtosis $=$ 4.130), we assumed the pattern of results might be similar to that of the simulation study 2, in which the dependent variable is non-normally distributed.

As Table 7 shows, the intervention significantly decreases participants' violent behavior after the program ($b_{01} = -0.858$, p =.036). Pupils with more (vs. less) pre-measured violent behavior tend to report more post-measured violent behavior ($b_{30} = 0.599$, p =.002), but the intervention significantly mitigates this tendency ($b_{31} = -0.515$, p =.008). However, the gender differences of post-measured violent behavior (boys appear more violent than girls, $b_{20} =$ 1.238, p <.001) may not be mitigated by the intervention.

We further conducted Bayesian estimation with informational priors based on the ML results. As presented in Table 8, the informational Bayesian estimation appear more powerful than the ML estimation. In addition to supporting the significant effects in the ML estimation, the Bayesian estimator further indicates that the intervention significantly

Table 5. Empirical example 1: the parameter estimation of different estimators.

Esitmator	b11	b21	b31	b01	b10	b20	b30
MLR	−1.409	1.977	0.139	3.342	−0.621	−1.105	0.097
BD	−2.100	2.411	0.108	3.270	−0.200	−1.110	0.085
L3W	−3.494	0.293	−0.131	1.132	−1.952	−2.508	−0.067
L1W	−2.033	1.735	0.073	2.810	−0.999	−1.416	0.060
CW	−1.429	2.111	0.125	3.362	−0.594	−1.052	0.089
R1W	−0.814	2.483	0.178	3.909	−0.187	−0.674	0.117
R3W	0.594	3.723	0.370	5.467	0.760	0.294	0.236
L3M	−2.858	0.660	−0.125	1.582	−1.560	−2.166	−0.087
L1M	−1.884	1.679	0.071	2.849	−0.915	−1.397	0.053
CM	−1.422	2.064	0.129	3.365	−0.613	−1.074	0.091
R1M	−0.965	2.438	0.190	3.877	−0.309	−0.740	0.126
R3M	−0.001	3.369	0.394	5.086	0.328	−0.013	0.280
L3S	−2.484	0.936	−0.088	1.934	−1.318	−1.930	−0.072
L1S	−1.763	1.699	0.075	2.927	−0.847	−1.353	0.051
CS	−1.417	2.030	0.134	3.364	−0.616	−1.086	0.091
R1S	−1.068	2.364	0.191	3.797	−0.385	−0.816	0.132
R3S	−0.365	3.083	0.365	4.737	0.084	−0.253	0.268

Note: BD: Bayesian estimator with default priors. L3: Bayesian estimator with the prior means deviating 3SD from the true value to the left; L1: Bayesian estimator with the prior means deviating 1SD from the true value to the left; C: Bayesian estimator with the prior means equaling the true value; R1: Bayesian estimator with the prior means deviating 1SD from the true value to the right; R3: Bayesian estimator with the prior means deviating 3SD from the true value to the right. W: weakly informative priors (prior variance = population value*50%); M: mediumly informative priors (prior variance = population value*20%); S: strongly informative priors (prior variance = population value*10%).

Table 7. Empirical example 2: the results of MLR analysis.

Para	est	SE	est/SE	p
b11	−0.174	0.527	−0.33	0.741
b21	−0.639	0.393	−1.625	0.104
b31	−0.515*	0.196	−2.634	0.008
b01	−0.858*	0.410	−2.093	0.036
b10	−0.017	0.498	−0.034	0.973
b20	1.238**	0.308	4.020	<.001
b30	0.599*	0.197	3.037	0.002

Note: est: the estimation of parameters; SE: the standard error; *p<.05; **p<.001.

Table 6. Empirical example 1: the 95% CIs of different estimators.

Esitmator	b11	b21	b31	b01	b10	b20	b30
MLR	[−6.387, 3.568]	[−0.970, 4.923]	[−0.213, 0.492]	[1.485, 5.199]	[−2.190, 0.948]	[−3.590, 1.380]	[−0.153, 0.348]
BD	[−9.446, 4.423]	[−1.415, 6.343]	[−0.387, 0.656]	[0.294, 5.979]	[−4.785, 4.778]	[−3.939, 1.494]	[−0.257, 0.448]
L3W	[−5.095, −1.856]	[−1.738, 2.061]	[−0.665, 0.254]	[−1.181, 3.084]	[−2.973, −0.927]	[−4.009, −1.174]	[−0.489, 0.223]
L1W	[−3.595, −0.434]	[0.028, 3.241]	[−0.249, 0.402]	[1.004, 4.452]	[−1.972, −0.003]	[−2.647, −0.253]	[−0.174, 0.298]
CW	[−2.944, 0.145]	[0.457, 3.624]	[−0.184, 0.467]	[1.699, 5.013]	[−1.540, 0.393]	[−2.264, 0.100]	[−0.134, 0.328]
R1W	[−2.367, 0.730]	[0.839, 4.081]	[−0.135, 0.531]	[2.185, 5.645]	[−1.133, 0.829]	[−1.870, 0.478]	[−0.100, 0.360]
R3W	[−0.990, 2.214]	[1.968, 5.630]	[0.014, 0.940]	[3.478, 7.898]	[−0.266, 1.822]	[−1.027, 1.664]	[−0.043, 0.699]
L3M	[−3.864, −1.794]	[−0.543, 1.795]	[−0.463, 0.157]	[0.096, 2.943]	[−2.202, −0.889]	[−3.086, −1.306]	[−0.361, 0.145]
L1M	[−2.900, −0.843]	[0.521, 2.717]	[−0.183, 0.325]	[1.555, 4.139]	[−1.557, −0.248]	[−2.270, −0.581]	[−0.132, 0.236]
CM	[−2.443, −0.379]	[0.920, 3.107]	[−0.117, 0.382]	[2.057, 4.658]	[−1.240, 0.053]	[−1.919, −0.257]	[−0.089, 0.275]
R1M	[−1.976, 0.067]	[1.326, 3.513]	[−0.056, 0.460]	[2.571, 5.167]	[−0.936, 0.353]	[−1.572, 0.069]	[−0.052, 0.323]
R3M	[−1.023, 1.021]	[2.246, 4.553]	[0.103, 0.737]	[3.669, 6.584]	[−0.318, 1.014]	[−0.900, 0.863]	[0.046, 0.569]
L3S	[−3.205, −1.730]	[0.096, 1.742]	[−0.318, 0.121]	[0.881, 2.971]	[−1.784, −0.843]	[−2.566, −1.304]	[−0.260, 0.104]
L1S	[−2.496, −1.023]	[0.858, 2.465]	[−0.122, 0.270]	[1.919, 3.922]	[−1.311, −0.367]	[−1.993, −0.753]	[−0.098, 0.199]
CS	[−2.136,−0.677]	[1.200,2.809]	[−0.062,0.332]	[2.367,4.339]	[−1.074,−0.137]	[−1.703,−0.486]	[−0.054,0.242]
R1S	[−1.793,−0.336]	[1.548,3.140]	[0.001,0.404]	[2.804,4.792]	[−0.844,0.091]	[−1.435,−0.209]	[−0.014,0.291]
R3S	[−1.090,0.375]	[2.278,3.905]	[0.145,0.603]	[3.698,5.800]	[−0.385,0.564]	[−0.889,0.380]	[0.092,0.466]

Note: BD: Bayesian estimator with default priors. L3: Bayesian estimator with the prior means deviating 3SD from the true value to the left; L1: Bayesian estimator with the prior means deviating 1SD from the true value to the left; C: Bayesian estimator with the prior means equaling the true value; R1: Bayesian estimator with the prior means deviating 1SD from the true value to the right; R3: Bayesian estimator with the prior means deviating 3SD from the true value to the right. W: weakly informative priors (prior variance = population value*50%); M: mediumly informative priors (prior variance = population value*20%); S: strongly informative priors (prior variance = population value*10%).

Table 8. Empirical example 2: the results of Bayesian analysis with DDP.

Para	est	SD	95% CI Lower 2.5%	95% CI Upper 2.5%
b11	−0.183	0.429	−1.022	0.677
b21	−0.630	0.290	−1.209	−0.061
b31	−0.516	0.132	−0.766	−0.246
b01	−0.836	0.272	−1.368	−0.302
b10	−0.082	0.338	−0.739	0.604
b20	1.225	0.223	0.773	1.646
b30	0.588	0.108	0.382	0.801

Note: DDP: data − dependent priors; est: the estimation of parameters; SD: the standard deviation of the posterior distribution; 95% CI: the 95% credible intervals of parameters.

Table 9. Empirical example 2: the parameter estimation of different estimators.

Estimator	b11	b21	b31	b01	b10	b20	b30
MLR	−0.174	−0.639	−0.515	−0.858	−0.017	1.238	0.599
BD	−0.100	−0.612	−0.517	−0.818	−0.176	1.215	0.585
L3W	−0.429	−1.273	−0.652	−1.239	−0.924	1.122	0.593
L1W	−0.088	−0.925	−0.533	−0.795	−0.423	1.298	0.576
CW	0.067	−0.845	−0.490	−0.615	−0.247	1.360	0.569
R1W	0.219	−0.777	−0.449	−0.433	−0.086	1.424	0.560
R3W	0.553	−0.401	−0.355	0.107	0.371	1.625	0.559
L3M	−0.264	−1.451	−0.699	−1.109	−0.779	1.053	0.546
L1M	−0.039	−1.025	−0.533	−0.713	−0.413	1.330	0.569
CM	0.069	−0.914	−0.486	−0.542	−0.256	1.411	0.564
R1M	0.176	−0.797	−0.440	−0.370	−0.102	1.501	0.559
R3M	0.397	−0.331	−0.281	0.110	0.246	1.835	0.602
L3S	−0.172	−1.438	−0.723	−0.978	−0.666	1.048	0.473
L1S	−0.010	−1.066	−0.533	−0.650	−0.386	1.339	0.559
CS	0.070	−0.938	−0.485	−0.504	−0.255	1.444	0.560
R1S	0.149	−0.807	−0.435	−0.350	−0.127	1.554	0.563
R3S	0.310	−0.377	−0.232	0.024	0.143	1.945	0.665

Note: BD: Bayesian estimator with default priors. L3: Bayesian estimator with the prior means deviating 3SD from the true value to the left; L1: Bayesian estimator with the prior means deviating 1SD from the true value to the left; C: Bayesian estimator with the prior means equaling the true value; R1: Bayesian estimator with the prior means deviating 1SD from the true value to the right; R3: Bayesian estimator with the prior means deviating 3SD from the true value to the right. W: weakly informative priors (prior variance = population value*50%); M: mediumly informative priors (prior variance = population value*20%); S: strongly informative priors (prior variance = population value*10%).

reduces the sex differences in violent behavior ($b_{21} = -0.63$, 95% CI = [-1.209, -0.061]).

To further explore the effect of prior specification on Bayesian estimation in non-normal situations, we imposed different settings of priors on the regression parameters like the simulations. Tables 9 and 10 present the results of different estimators in empirical example 2. Bayesian estimator with seriously deviated prior information obtained terrifically biased parameter estimation. However, when the prior mean deviation is minor, the corresponding results of parameter estimation are much closer to those of the ML estimation.

4.2.3. Summary and Discussion

The results of the two empirical examples jointly indicated the effectiveness of the democratic school health education intervention. Pupils in the experimental classes develop better states of well-being and less violent behavior than those in the control classes. In addition, pupils with worse baseline conditions seem to benefit even more from the intervention. Specifically, boys, who probably develop less well-being and more violent behavior than girls, may gain more benefits from the intervention in improving well-being and reducing violent behavior. Likewise, for pupils with more (vs. less) pre-measured violent behavior, the intervention reduces the level of post-measured violent behavior more significantly.

In sum, the empirical examples show similar pattern of results to the simulation studies. With the decrease of the prior variance, the influence of prior information on Bayesian estimation increases rapidly, and the width of the corresponding confidence interval becomes narrower. When the prior is strongly informative and the prior deviation is serious, the Bayesian estimator obtains detrimentally biased parameter estimation, and may even fail to detect the intervention effects. Bayesian estimation with slightly incorrect prior information obtains similar results to those of the ML

Table 10. Empirical example 2: the 95% CIs of different estimators.

Estimator	b11	b21	b31	b01	b10	b20	b30
MLR	[−1.207, 0.859]	[−1.409, 0.132]	[−0.899, −0.132]	[−1.661, −0.054]	[−0.993, 0.960]	[0.635, 1.842]	[0.212, 0.986]
BD	[−2.534, 2.038]	[−1.769, 0.650]	[−0.972, −0.028]	[−1.779, 0.017]	[−1.752, 1.369]	[0.329, 2.012]	[0.233, 0.942]
L3W	[−0.794, −0.058]	[−2.610, −0.320]	[−1.244, −0.235]	[−2.002, −0.617]	[−1.617, −0.311]	[0.111, 1.880]	[0.209, 0.909]
L1W	[−0.436, 0.273]	[−1.786, −0.144]	[−0.932, −0.139]	[−1.404, −0.228]	[−0.982, 0.099]	[0.624, 1.959]	[0.276, 0.883]
CW	[−0.281, 0.420]	[−1.673, −0.035]	[−0.865, −0.100]	[−1.224, −0.023]	[−0.768, 0.284]	[0.704, 2.018]	[0.273, 0.878]
R1W	[−0.122, 0.573]	[−1.597, 0.088]	[−0.815, −0.028]	[−1.051, 0.209]	[−0.598, 0.467]	[0.771, 2.094]	[0.258, 0.881]
R3W	[0.202, 0.916]	[−1.320, 0.956]	[−0.742, 0.310]	[−0.637, 1.122]	[−0.192, 1.043]	[0.834, 2.967]	[0.232, 0.994]
L3M	[−0.495, −0.024]	[−2.312, −0.741]	[−1.275, −0.315]	[−1.635, −0.634]	[−1.201, −0.380]	[0.189, 1.680]	[0.029, 0.832]
L1M	[−0.264, 0.192]	[−1.695, −0.429]	[−0.878, −0.199]	[−1.191, −0.261]	[−0.793, −0.023]	[0.761, 1.880]	[0.293, 0.829]
CM	[−0.156, 0.297]	[−1.545, −0.283]	[−0.805, −0.143]	[−1.020, −0.098]	[−0.627, 0.127]	[0.851, 1.977]	[0.305, 0.829]
R1M	[−0.050, 0.405]	[−1.433, −0.142]	[−0.751, −0.083]	[−0.852, 0.112]	[−0.468, 0.287]	[0.926, 2.073]	[0.297, 0.841]
R3M	[0.174, 0.629]	[−1.069, 0.554]	[−0.661, 0.328]	[−0.415, 0.702]	[−0.140, 0.662]	[1.100, 2.944]	[0.293, 1.191]
L3S	[−0.338, −0.003]	[−2.046, −0.930]	[−1.193, −0.387]	[−1.358, −0.620]	[−0.960, −0.374]	[0.400, 1.562]	[0.014, 0.772]
L1S	[−0.172, 0.156]	[−1.602, −0.581]	[−0.823, −0.252]	[−1.022, −0.299]	[−0.667, −0.094]	[0.850, 1.816]	[0.315, 0.779]
CS	[−0.090, 0.231]	[−1.447, −0.446]	[−0.750, −0.196]	[−0.872, −0.155]	[−0.537, 0.032]	[0.957, 1.932]	[0.334, 0.789]
R1S	[−0.013, 0.311]	[−1.322, −0.301]	[−0.701, −0.137]	[−0.720, 0.003]	[−0.399, 0.159]	[1.067, 2.058]	[0.342, 0.813]
R3S	[0.151, 0.474]	[−0.924, 0.216]	[−0.591, 0.238]	[−0.365, 0.426]	[−0.135, 0.436]	[1.308, 2.777]	[0.358, 1.180]

Note: BD: Bayesian estimator with default priors. L3: Bayesian estimator with the prior means deviating 3SD from the true value to the left; L1: Bayesian estimator with the prior means deviating 1SD from the true value to the left; C: Bayesian estimator with the prior means equaling the true value; R1: Bayesian estimator with the prior means deviating 1SD from the true value to the right; R3: Bayesian estimator with the prior means deviating 3SD from the true value to the right. W: weakly informative priors (prior variance = population value*50%); M: mediumly informative priors (prior variance = population value*20%); S: strongly informative priors (prior variance = population value*10%).

estimation, which is consistent with the conclusions of the simulations.

In addition, the impacts of incorrect prior information on Bayesian estimation seem relatively more detrimental when the dependent variable is relatively non-normal, which is also consistent with the conclusion of the simulation studies. When specifying incorrect prior information, the bias of the Bayesian estimator is generally more serious in empirical example 2 than in empirical example 1. The results of both empirical examples further validate the conclusions of the simulations, reminding empirical researchers to be careful with the impacts of incorrect prior information, especially when the dependent variables are non-normally distributed.

5. Discussion

In many psychological experiments (e.g., cRCT), researchers must conduct multilevel modeling to analyze nested data. Multilevel data are prone to small-sample problems and may lead to biased estimations in traditional estimators. Bayesian methods can incorporate prior information for the estimation. However, due to the lack of empirical research and theoretical evidence, it is usually challenging to set priors in real studies correctly. Seriously incorrect informative priors may lead to destructively biased parameter estimation, but it is unclear how this impacts the analysis in more specific situations under the multilevel framework. To investigate the potentially detrimental effects of inaccurate prior information on Bayesian approaches, we conducted a series of simulations under the multilevel model framework with different dependent variable types, sample sizes, and ICCs. In summary, as expected, regardless of whether the dependent variables were continuous or binary, the overall performance was best for Bayesian estimation with correct informative priors. Seriously inaccurate informative priors have devastating impacts on Bayesian analysis, especially in the cases of larger ICC, smaller level 2 sample size, and smaller prior variance. When the dependent variable was non-normal or binary, these adverse effects were more prominent.

5.1. Continuous Data

First, the influence of incorrect informative priors on parameter estimation changes with the sample size. Consistent with Holtmann et al. (2016), the current study showed that larger sample sizes at both levels may lead to a more negligible influence of incorrect informative priors on parameter estimation. In Bayesian approach, when the sample sizes are small, the posterior distributions are primarily dependent on the specifications of the priors. With an increase in the sample size, the influence of the prior distribution gradually weakens, and the shape of the posterior distribution approximates that of the likelihood (Lynch, 2007).

Second, for the estimation of the treatment effects b_{01} (which is usually the focus of cRCT research), a large ICC may increase the negative impacts of incorrect informative priors. This tendency is consistent with previous studies, which concluded that the estimation bias of level 2 effects increased with ICCs, while power decreased with ICC (Cao et al., 2019; Finch & French, 2011). However, for the estimation of other regression parameters, the impact of ICC may be inconsistent across different situations in the current simulation. This is also consistent with the results of Cao et al. (2019), which indicated that ICC was highly associated with estimating the between-level interaction, but not the within-level or cross-level interaction effects. More research is warranted to explore the impacts of ICC on different effects at different levels, taking into consideration the potential impacts of incorrect informative priors.

In addition, the influence of incorrect informative priors is weaker when the prior variance is increased. When the prior variance is relatively large, the adverse effects of incorrect informative priors are primarily suppressed. When the prior variance is small, incorrect informative priors may negatively influence the Bayesian estimation, even with large sample sizes. Therefore, in general, it is usually more conservative to specify a larger prior variance.

The results of the standard error ratio were similar to those of previous studies. Generally, the standard error ratios are larger for Bayesian estimation with informative priors and increase with a decrease in prior variance, which indicates that the posterior SDs tend to overestimate the empirical SDs with small prior variance. However, this increase in the standard error ratio may also reveal a decrease in SD estimation due to informative priors (Holtmann et al., 2016). In addition, Bayesian estimation with incorrect informative priors tends to overestimate the standard error more seriously among all informative Bayesian estimators. Therefore, in Bayesian analysis, researchers often need to rely on the credibility interval to determine the significance of the parameters.

In the present study, we also considered situations in which the dependent variables were non-normal. In these cases, the overall performance of the parameter estimation is slightly worse than in the cases in which the dependent variables are normally distributed, and the impacts of incorrect informative priors are more detrimental. This is consistent with the results of previous studies, which indicated that compared to the conditions in which the dependent variables follow a normal distribution, informative priors tend to have a more significant impact on parameter estimation under non-normality conditions (Kelava & Nagengast, 2012). Therefore, when the dependent variables violate the assumption of normality, researchers should be more cautious about the specification of informative priors or use more robust estimation approaches to account for non-normality (e.g., Asparouhov & Muthén, 2016; Lai, 2019).

5.2. Binary Data

The present study conducted simulations for binary data, both with and without random slopes. Consistent with previous studies (Depaoli & Clifton, 2015; Muthén, 2010), in Study 3.1, ML became computationally demanding and might fail to converge when the dependent variable is binary with random slope effects. In Study 3.2, when analyzing data without random slope effects, WLSMV provided a biased estimation of the standard error in small samples, which is consistent with previous research (Muthén et al., 2015).

Similar to Asparouhov and Muthén's (2010) simulations, Bayesian estimation is more susceptible to prior information when the dependent variables are binary. The impact of incorrect informative priors is much more detrimental than under continuous conditions. The estimation of parameters is more dependent on the specification of informative priors when the dependent variables are categorical, thus increasing the negative impacts of incorrect informative priors. In Studies 3.1 and 3.2, increasing the between-level sample size can buffer the adverse effects of incorrect informative priors to a larger extent.

Overall, regardless of whether the dependent variables are continuous or binary, the impact of informative priors is similar. Bayesian estimation with correctly informative priors performed best among the different estimators, especially in terms of the MSE and power. Bayesian analysis with seriously incorrect prior information (i.e., the mean hyperparameters of priors deviated 3*SD* from the corresponding true values) led to seriously biased parameter estimation and low power. In addition, a large ICC, a small sample size at level 2, and a small prior variance may further increase the negative impacts of incorrect prior information.

5.3. Limitations and Prospects

Despite the novel contribution to incorrect informative priors, there are some limitations to the present research. First, the present study explored the impacts of inaccurate prior information on Bayesian approaches with different dependent variable types, sample sizes, and ICCs, but could not cover all the conditions in the real studies. We mainly focused on the sample sizes common in cRCT research, but the sample sizes may differ significantly from the current simulations in real studies. Moreover, to control the number of conditions, only three levels of sample sizes at the within-level (i.e., 30, 60, 150) are specified in the current simulations. Future studies should include a larger number of sample size conditions to better understand the impacts of incorrect informative priors. In addition, the present research assumed balanced sample sizes within groups, but the group sizes are usually unbalanced in actual situations. Further research is needed to explore whether incorrect informative priors will lead to more detrimental impacts if the group sizes are unbalanced.

Second, to simplify the simulations, we only investigated the effects of skewness (negatively or positively skewed) in the Study 2. However, the magnitude of skewness and kurtosis may also significantly impact model fitting and parameter power (Foss et al., 2011). Future studies are warranted to explore the influence of incorrect informative priors on parameter estimation under different magnitudes of skewness and kurtosis.

Considering that the target of cRCT research is generally the investigation of treatment effect b_{01}, we mainly focused on the estimation of b_{01} under different data conditions. However, for the analysis of other regression parameters, the impact of ICC may be inconsistent across various situations in the current simulation. Further research is warranted to explore the effects of incorrect informative priors under different levels of ICC.

5.4. Practical Recommendations

Traditional frequentist estimators rely on the asymptotic theory and require large sample sizes to obtain appropriate parameter estimation. When the sample size (especially the level-2 sample size) is relatively large, traditional frequentist methods (e.g., ML and WLSMV) perform well in the estimated models and, thus, are highly recommended.

However, in small sample sizes, traditional methods may yield biased estimations and low power. Bayesian approaches are advantageous in dealing with small sample sizes but are, at the same time, sensitive to the specification of priors. When the sample sizes are small, the software-default priors (whose distribution ranges are often very wide) may seriously influence the efficiency of MCMC estimation, and lead to biased parameter estimation in small samples (Erp et al., 2018; Smid & Winter, 2020). Therefore, researchers should not intuitively use default "non-informative" priors without careful consideration, and should search for as much information as possible (from previous research and knowledge, specialistic advice, and so on) to help specify proper priors. Note that if the dataset is considered sampled from a different population from previous studies in the current research, researchers should specify the priors with slightly larger prior variance (larger than that from the previous studies) to take the uncertainty of priors into consideration (Miocevic et al., 2017). Although facing some risks of inaccurate priors, informative priors with thoughtful consideration can narrow the distribution of parameter priors to a relatively reasonable range, which helps the MCMC algorithm reach to an accurate estimation.

In practice, however, information about the prior specification is often rare, and the exact prior distributions of the parameters are unclear. In these situations, researchers often suggest specifying DDP (also known as empirical Bayes priors; Erp et al., 2018; McNeish, 2016b) or admissible-range-restricted priors (according to the admissible values of the parameters; McNeish, 2019) to the parameters. Compromising between classical frequentist and Bayesian approach, DDP adds more information about the data to the Bayesian analysis, and may improve the stability of the MCMC sampler (Carlin & Louis, 2000; Casella, 1992). It generally mitigates the small-sample bias in the models, and may obtain much better parameter estimation than frequentist or Bayesian analysis with non-informative priors (McNeish, 2016b). Although not so powerful as thoughtful informative priors, DDP may obtain low estimation bias for structural parameters (comparable to the thoughtful priors), and even relatively lower bias for variance parameters (Smid et al., 2019).

Some studies, however, indicated that DDP may result in poor performance when the sample size is small and the model is complex (e.g., growth mixture model; Depaoli, 2013), and suggested that only highly informative priors obtain accurate estimates. In addition, DDP may underestimate the standard errors in Bayesian analysis, as it ignores the variability of the hyperparameters in the random effect covariance matrix (Carlin & Louis, 2000). However, considering the detrimental influence of inaccurate informative priors, it may be safer to specify priors that are not so informative in empirical research. More research are needed to investigate the influence of DDP on Bayesian estimation in complex models (e.g., multilevel models), and the potential of using DDP to decrease the risks of specifying inaccurate prior information.

In sum, researchers should always be alert to the possibility of specifying inaccurate informative priors. For models estimated in the current simulations, the performance of Bayesian estimation with minorly incorrect informative priors (i.e., the prior mean deviates $\pm 1SD$ from the population value) is generally not too poor. However, when the ICC is large, or the sample size at level 2 is small, seriously incorrect informative priors (i.e., the prior mean deviates $\pm 3SD$ from the population value) may lead to seriously biased parameter estimation. In addition, non-normally distributed or binary dependent variables may further increase the impact of incorrect informative priors. Considering the severe impacts of incorrect prior information in these situations, we advise researchers to reduce the possible effects of priors by increasing the sample sizes at both levels (especially at level 2) as much as possible or increasing the prior variance to some extent. In addition, a prior sensitivity analysis is often suggested to check the robustness of the conclusions (Erp et al., 2018).

Disclosure Statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the National Natural Science Foundation of China under Grant [31871128]; Natural Science Foundation of Guangdong Province, China under Grant [2022A1515010367]; National Natural Science Foundation of China for Young Scholars under Grant [72104006].

ORCID

Lijin Zhang http://orcid.org/0000-0002-4222-8850
Zhehan Jiang http://orcid.org/0000-0002-1376-9439
Junhao Pan http://orcid.org/0000-0001-7156-6077

References

Aryee, S., Walumbwa, F. O., Seidu, E. Y., & Otaye, L. E. (2012). Impact of high-performance work systems on individual-and branch-level performance: Test of a multilevel model of intermediate linkages. *The Journal of Applied Psychology*, *97*, 287–300.

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

Asparouhov, T., & Muthén, B. (2012a). *Comparison of computational methods for high dimensional item factor analysis (Mplus technical report, Version 7).* http://www.statmodel.com/download/HighDimension11.pdf.

Asparouhov, T., & Muthén, B. (2012b). *General random effect latent variable modeling: Random subjects, items, contexts, and parameters.* http://www.statmodel.com/download/NCME12.pdf.

Asparouhov, T., & Muthén, B. (2013). Section on *statistics in epidemiology computationally efficient estimation of multilevel high-dimensional latent variable models.* https://www.statmodel.com/download/JSM2007000746.pdf.

Asparouhov, T., & Muthén, B. (2016). Structural equation models and mixture models with continuous nonnormal skewed distributions. *Structural Equation Modelling*, *23*, 1–19.

Barendse, M. T., & Rosseel, Y. (2020). Multilevel modelling in the "wide format" approach with discrete data: A solution for small cluster sizes. *Structural Equation Modeling*, *27*, 696–721.

Bauer, D. J., & Sterba, S. K. (2011). Fitting multilevel models with ordinal outcomes: performance of alternative specifications and methods of estimation. *Psychological Methods*, *16*, 373–390.

Boberova, Z., Paakkari, L., Ropovik, I., & Liba, J. (2017). Democratic school health education in a post-communist country. *Health Education*, *117*, 469–484.

Brown, L. D. (2008). In-season prediction of batting averages: A field test of empirical Bayes and Bayes methodologies. *Annals of Applied Statistics*, *2*, 113–152.

Campbell, M. J. (2019). Cluster randomised trials. *The Medical Journal of Australia*, *210*, 154–156.e1. https://doi.org/10.5694/mja2.13001

Campbell, M. J., & Walters, S. J. (2014). *How to design, analyse and report cluster randomised trials in medicine and health related research.* John Wiley & Sons.

Can, S., van de Schoot, R., & Hox, J. (2015). Collinear latent variables in multilevel confirmatory factor analysis: A comparison of maximum likelihood and Bayesian estimations. *Educational and Psychological Measurement*, *75*, 406–427.

Cao, C., Kim, E. S., Chen, Y.-H., Ferron, J., & Stark, S. (2019). Exploring the test of covariate moderation effects in multilevel mimic models. *Educational and Psychological Measurement*, *79*, 512–544.

Carlin, B. P., & Louis, T. A. (2000). Empirical Bayes: Past, present and future. *Journal of the American Statistical Association*, *95*, 1286–1289.

Casella, G. (1992). Illustrating empirical Bayes methods. *Chemometrics and Intelligent Laboratory Systems*, *16*, 107–125. https://doi.org/10.1016/0169-7439(92)80050-E

Cham, H., West, S. G., Ma, Y., & Aiken, L. S. (2012). Estimating latent variable interactions with nonnormal observed data: A comparison of four approaches. *Multivariate Behavioral Research*, *47*, 840–876.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates, Inc.

Curran, P. J., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods*, *1*, 16–29.

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

Depaoli, S. (2014). The impact of inaccurate "informative" priors for growth parameters in Bayesian growth mixture modeling. *Structural Equation Modelling*, *21*, 239–252. https://doi.org/10.1080/10705511.2014.882686

Depaoli, S., & Clifton, J. P. (2015). A Bayesian approach to multilevel structural equation modelling with continuous and dichotomous outcomes. *Structural Equation Modelling*, *22*, 327–351.

Depaoli, S., & Schoot, R. V. d (2017). Improving transparency and replication in Bayesian statistics: The WAMBS-checklist. *Psychological Methods*, *22*, 240–261.

Donner, A., & Zou, K. G. (2004). Methods for the statistical analysis of binary data in split-cluster designs. *Biometrics*, *60*, 919–925.

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods*, *12*, 121–138.

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

Fang, J., Wen, Z. L., & Hau, K. T. (2019). Mediation effects in 2-1-1 multilevel model: Evaluation of alternative estimation methods. *Structural Equation Modelling*, *26*, 591–606.

Finch, W. H., & French, B. F. (2011). Estimation of mimic model parameters with multilevel data. *Structural Equation Modeling*, *18*(2), 229–252. https://doi.org/10.1080/10705511.2011.557338

Finch, W. H., & Miller, J. E. (2019). The use of incorrect informative priors in the estimation of mimic model parameters with small sample sizes. *Structural Equation Modelling*, *26*, 497–508.

Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika*, *43*, 521–532. https://doi.org/10.1007/BF02293811

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, *9*, 466–491.

Foss, T., Joreskog, K. G., & Olsson, U. H. (2011). Testing structural equation models: The effect of kurtosis. *Computational Statistics & Data Analysis*, *55*, 2263–2275.

Gates, K. M., Lane, S. T., Varangis, E., Giovanello, K., & Guskiewicz, K. (2017). Unsupervised classification during time-series model building. *Multivariate Behavioral Research*, *52*, 129–148.

Gelfand, A. E., & Smith, A. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, *85*, 398–409. https://doi.org/10.1080/01621459.1990.10476213

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1996). *Markov Chain Monte Carlo methods in practice* (pp. 1–20). Information Age.

Ha, A. S., Lonsdale, C., Ng, J. Y. Y., & Lubans, D. R. (2017). A school-based rope skipping program for adolescents: Results of a randomized trial. *Preventive Medicine*, *101*, 188–194.

Heck, R. H., & Thomas, S. L. (2020). *An introduction to multilevel modelling techniques: MLM and SEM approaches*. Routledge.

Holtmann, J., Koch, T., Lochner, K., & Eid, M. (2016). A comparison of ml, WLSMV, and Bayesian methods for multilevel structural equation models in small samples: A simulation study. *Multivariate Behavioral Research*, *51*, 661–680.

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

Hox, J. J., Maas, C. J., & Brinkhuis, M. J. (2010). The effect of estimation method and sample size in multilevel structural equation modeling. *Statistica Neerlandica*, *64*, 157–170.

Hox, J. J., & Maas, C. J. M. (2001). The accuracy of multilevel structural equation modeling with pseudobalanced groups and small samples. *Structural Equation Modeling*, *8*, 157–174.

Hox, J. J., Moerbeek, M., & Schoot, R. V. d (2017). *Multilevel analysis: Techniques and applications*. Routledge.

Hsu, H.-Y., Lin, J.-H., Kwok, O.-M., Acosta, S., & Willson, V. (2017). The impact of intraclass correlation on the effectiveness of level-specific fit indices in multilevel structural equation modelling: A Monte Carlo study. *Educational and Psychological Measurement*, *77*, 5–31.

Jung, S., Seo, D. G., & Park, J. (2020). Regularized exploratory bifactor analysis with small sample sizes. *Frontiers in Psychology*, *11*, 507.

Kadane, J. B. (2015). Bayesian methods for prevention research. *Prevention Science*, *16*, 1017–1025. https://doi.org/10.1007/s11121-014-0531-x

Kelava, A., & Nagengast, B. (2012). A Bayesian model for the estimation of latent interaction and quadratic effects when latent variables are nonnormally distributed. *Multivariate Behavioral Research*, *47*, 717–742. https://doi.org/10.1080/00273171.2012.715560

Koch, T., Schultze, M., Burrus, J., Roberts, R. D., & Eid, M. (2015). A multilevel cfa-mtmm model for nested structurally different methods. *Journal of Educational and Behavioral Statistics*, *40*, 477–510.

Lai, K. (2019). More robust standard error and confidence interval for SEM parameters given incorrect model and nonnormal data. *Structural Equation Modeling*, *26*, 260–279. https://doi.org/10.1080/10705511.2018.1505522

Legare, F., Briere, N., Stacey, D., Bourassa, H., Desroches, S., Dumont, S., Fraser, K., Freitas, A., Rivest, L.-P., & Roy, L. (2015). Improving decision making on location of care with the frail elderly and their caregivers (the dolce study): Study protocol for a cluster randomized controlled trial. *Trials*, *16*, 50. https://doi.org/10.1186/s13063-015-0567-7

Levy, R., & Choi, J. (2013). Bayesian structural equation modelling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural Equation Modeling*. Information Age.

Li, X., & Jacobucci, R. (2021). Regularized structural equation modelling with stability selection. *Psychological Methods*, *27*, 497–518.

Lynch, S. M. (2007). *Introduction to applied Bayesian statistics and estimation for social scientists*. Springer Science & Business Media.

McNeish, D. (2016a). On using Bayesian methods to address small sample problems. *Structural Equation Modeling*, *23*, 750–773. https://doi.org/10.1080/10705511.2016.1186549

McNeish, D. (2016b). Using data-dependent priors to mitigate small sample bias in latent growth models. *Journal of Educational and Behavioral Statistics*, *41*, 27–56. https://doi.org/10.3102/1076998615621299

McNeish, D. (2019). Two-level dynamic structural equation models with small samples. *Structural Equation Modelling*, *26*, 948–966. https://doi.org/10.1080/10705511.2019.1578657

McNeish, D., & Stapleton, L. M. (2016). Modelling clustered data with very few clusters. *Multivariate Behavioral Research*, *51*, 495–518.

Miocevic, M., MacKinnon, D. P., & Levy, R. (2017). Power in Bayesian mediation analysis for small sample research. *Structural Equation Modelling*, *24*, 666–683.

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, *38*, 2074–2102. https://doi.org/10.1002/sim.8086

Muthén, B. (2010). Bayesian analysis in Mplus: A brief introduction. https://pdfs.semanticscholar.org/a941/cb99e52a8eafad701d4fb0391347232779f9.pdf?_ga=2.137417013.467066868.1585219703-884182655.1568515547.

Muthén, B., Muthén, L., & Asparouhov, T. (2015). Estimator choices with categorical outcomes. http://www.statmodel.com/download/EstimatorChoices.pdf.

Muthén, L. (2006). Re: TYPE = COMPLEX [Online forum comment]. http://www.statmodel.com/discussion/messages/12/1354.html?1153839055.

Muthén, L., & Muthén, B. (1998–2017). *Mplus user's guide*. Muthén & Muthén.

Newton, N. C., Teesson, M., Mather, M., Champion, K. E., Barrett, E. L., Stapinski, L., Carragher, N., Kelly, E., Conrod, P. J., & Slade, T. (2018). Universal cannabis outcomes from the climate and preventure (cap) study: A cluster randomised controlled trial. *Substance Abuse Treatment Prevention and Policy*, *13*, 34.

Preacher, K. J. (2011). Multilevel SEM strategies for evaluating mediation in three-level data. *Multivariate Behavioral Research*, *46*, 691–731.

R Core Team. (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Ribeiro, D. C., Milosavljevic, S., & Abbott, J. H. (2018). Sample size estimation for cluster randomized controlled trials. *Musculoskeletal Science & Practice*, *34*, 108–111..

Rutterford, C., Copas, A., & Eldridge, S. (2015). Methods for sample size determination in cluster randomized trials. *International Journal of Epidemiology*, *44*, 1051–1067. https://doi.org/10.1093/ije/dyv113

Ryu, E. (2015). The role of centering for interaction of level 1 variables in multilevel structural equation models. *Structural Equation Modelling*, *22*, 617–630. https://doi.org/10.1080/10705511.2014.936491

Schoeneberger, J. A. (2016). The impact of sample size and other factors when estimating multilevel logistic models. *The Journal of Experimental Education*, *84*, 373–397. https://doi.org/10.1080/00220973.2015.1027805

Shen, Y., Wang, T., Gao, M., Zhu, X., Zhang, X., He, C., Li, Y., & Sun, X. (2019). Effectiveness of low-cost reminder package combined with case-based health education to improve hypertensive patients' medication adherence: a clustered randomized controlled trial. *Patient Preference and Adherence*, *13*, 1083–1092. https://doi.org/10.2147/PPA.S194667

Shi, D., & Tong, X. (2017). The impact of prior information on Bayesian latent basis growth model estimation. *SAGE Open*, *7*, 1–14.

Smid, S. C., McNeish, D., Miocevic, M., & van de Schoot, R., (2019). Bayesian versus frequentist estimation for structural equation models in small sample contexts: A systematic review. *Structural Equation Modeling*, *27*, 1–31.

Smid, S. C., & Winter, S. D. (2020). Dangers of the defaults: A tutorial on the impact of default priors when using Bayesian SEM with small samples. *Frontiers in Psychology*, *11*, 611963.

Song, X.-Y., & Lee, S.-Y. (2004). Bayesian analysis of two-level nonlinear structural equation models with continuous and polytomous data. *The British Journal of Mathematical and Statistical Psychology*, *57*, 29–52.

Song, X.-Y., Pan, J.-H., Kwok, T., Vandenput, L., Ohlsson, C., & Leung, P.-C. (2010). A semiparametric Bayesian approach for structural equation models. *Biometrical Journal*, *52*, 314–332.

van Breukelen, G. J., & Candel, M. J. (2012). Calculating sample sizes for cluster randomized trials: We can keep it simple and efficient! *Journal of Clinical Epidemiology*, *65*, 1212–1218.

van der Linden, W. J. (2008). Using response times for item selection in adaptive testing. *Journal of Educational and Behavioral Statistics*, *33*, 5–20. https://doi.org/10.3102/1076998607302626

van de Schoot, R., Depaoli, S., King, R., Kramer, B., Märtens, K., Tadesse, M. G., Vannucci, M., Gelman, A., Veen, D., Willemsen, J., & Yau, C. (2021). Bayesian statistics and modelling. *Nature Reviews Methods Primers*, *1*, 1–26. https://doi.org/10.1038/s43586-020-00001-2

Van de Schoot, R., Kaplan, D., Denissen, J., Asendorpf, J. B., Neyer, F. J., & Van Aken, M. A. (2014). A gentle introduction to bayesian analysis: Applications to developmental research. *Child Development*, *85*, 842–860.

van Erp, S., & Browne, W. J. (2021). Bayesian multilevel structural equation modelling: An investigation into robust prior distributions for the doubly latent categorical model. *Structural Equation Modelling*, *28*, 875–893.