How to Politely Re-dip Your Chip!

On the Use of Data-Based Informative Priors in Linear Mixed Models

Zhigang Zhang

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Education

University of Washington

2025

Committee:

Elizabeth Sanders

Chun Wang

Program Authorized to Offer Degree:

Education

University of Washington

Abstract

How to Politely Re-dip Your Chip!

On the Use of Data-Based Informative Priors in Linear Mixed Models

Zhigang Zhang

Chair of the Supervisory Committee:

Elizabeth Sanders

Education

The use of Bayesian analyses in social science research has been on the rise, yet the issue of prior specification still poses theoretical controversies and practical challenges. In educational psychology, the prevalence Bayesian analysis and choice of priors is currently unknown, and the impact of using sample-model-based informative priors for multilevel models has yet to be evaluated. The current study therefore investigates: 1) the use of Bayesian analyses and prior specification choices in recent applied educational psychology research, and 2) the consequences of using increasingly informative sample-model-based priors ("double-dipping") on fixed effect coefficient parameter recovery for 2-level hierarchical linear models. Our results show that, first, applied researchers tend to rely on software default priors (i.e., noninformative or weakly informative priors), and on rare occasions where informative priors are used, about one-third rely

on sample-related values. Second, our simulation results show that posterior standard errors are progressively underestimated (leading to over-credibility) as fixed effect coefficient sample-model-based prior informativeness increases, particularly for conditions involving a larger number of clusters (L2 sample size). Third, the best approach for obtaining unbiased fixed effect coefficient credible intervals is to use weakly informative priors; the next-best alternative is to use a cross-validation method whereby a random half of the data is modeled using uninformative priors to obtain sample-model-based priors for a subsequent model for the other half of the data. These findings are consistent with previous methodological work warning that data-based priors require careful implementation. Limitations and future directions are discussed.

Keywords: Multilevel modeling, Bayesian estimation, sample-based priors, data-based priors

How to Politely Re-dip Your Chip!

On the Use of Data-Based Informative Priors in Linear Mixed Models

The development and use of Bayesian statistical analyses has grown substantially over the past decade across the social sciences, including the fields of psychology (van de Schoot et al., 2017), sociology (Lynch & Bartlett, 2019), and education (König & Van De Schoot, 2018). A typical Bayesian analysis involves three steps: 1) specifying prior distributions that capture existing knowledge about parameters before data collection; 2) determining the likelihood function based on observed data; 3) combining both sources of information to compute the posterior distribution for balanced inference. Researchers may be attracted to Bayesian analyses for theoretical and practical benefits, including its capability to incorporate existing knowledge or beliefs with observed data, provide intuitive probability statements about parameters, and its ability to estimate complex models with relatively small samples (Wagenmakers et al., 2018).

Among these benefits, the Bayesian analysis integration of priors with likelihood information using Markov chain Monte Carlo (MCMC) is often promoted as better suited for modeling small-sample data compared to frequentist maximum likelihood (ML) estimation (McNeish, 2016). Without reliance on large-sample asymptotic properties (i.e., as in ML estimation), the quality of inference is not controlled by the sample size, but rather by the effective sample size drawn during MCMC estimation. Additionally, Bayesian MCMC estimation avoids the need for higher order numerical integration that ML requires, which can improve the efficiency of estimating complex models (Depaoli & Clifton, 2015). Last, by utilizing additional information in the form of *informative* priors, Bayesian analysis has the potential to provide more stable and accurate parameter estimates. However, when small samples are used, the prior distribution has a much greater weight on the posterior distribution compared

to the likelihood, relative to when the sample size is large. In other words, inaccurate informative priors with small samples can yield estimates that are more biased than frequentist estimates (Depaoli, 2014; van de Schoot et al., 2014).

Clustered Data and Multilevel Models

Hierarchical or "nested" data structures are prevalent in educational and psychological research, where observations (micro units) are clustered or "nested" within higher-level (macro) units (Snijders & Bosker, 2012), such as students within classrooms, employees within organizations, and repeated measurements within individuals. Ignoring such dependencies that often naturally arise among observations within clusters can lead to distorted fixed effect parameter estimates and their standard errors (Snijders & Bosker, 2012). Multilevel modeling (MLM), also known as random effects modeling or mixed modeling, is arguably the most flexible approach for analyzing clustered data because it decomposes the variation in the dependent variable into respective levels that can then be modeled by level-specific regressors. A simple 2-level hierarchical linear MLM in which the intercept (conditional mean of the dependent variable) and the L1 predictor slope (X-Y relationship) are free to vary across clusters (i.e., modeling intercept *variance* and slope *variance*) can be defined as follows:

$$Y_{ij} = \gamma_{00} + \gamma_{10}\left(X_{1ij} - \bar{X}_{1.j}\right) + \gamma_{01}\left(\bar{X}_{1.j} - \bar{X}_{1..}\right) + u_{0j} + u_{1j}\left(X_{1ij} - \bar{X}_{1.j}\right) + r_{ij}, \qquad (1)$$

where Y_{ij} is the *i*th observation in the *j*th cluster, modeled as a function of the sum of the intercept (conditional mean across clusters, γ_{00}), the within-cluster effect of the regressor ($X_{1ij} - \bar{X}_{1.j}$), the between-cluster effect of the regressor ($\bar{X}_{1.j} - \bar{X}_{1..}$), the between-cluster L2 residual for the intercept (u_{0j}), the between-cluster L2 residual for the L1 regressor slope (u_{1j}), and the within-cluster L1 residual error (r_{ij}). The two cluster-level (L2) residuals are typically specified to assume multivariate normality (although constraints can be used to restrict the random effects

to be normally distributed and uncorrelated), and the L1 residual is assumed normally distributed and uncorrelated with the L2 residuals.

Bayesian Estimation of Multilevel Models with Small Samples

The frequentist approach for estimating a multilevel model relies on large-sample theory. When the number of clusters (L2 units) is relatively small (e.g., Dedrick et al., 2009), restricted ML (REML) is recommended for frequentist MLM estimation (Snijders & Bosker, 2012), but REML-estimated models cannot be compared with likelihood ratio tests. As a result, even with small L2 sample sizes, researchers may turn to full information ML (FIML, or just ML), which can in turn produce biased standard errors due to non-normal L2 residuals or underestimated variance components (e.g., Maas & Hox, 2005). In contrast, Bayesian estimation treats parameters (θ) as random variables rather than fixed quantities. Through an application of Bayes' theorem, the probability of θ conditional on the observed data (y) is referred to as the posterior distribution $p(\theta|y)$:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} \propto p(y|\theta)p(\theta). \quad (2)$$

Where, the posterior distribution combines information about the parameters available in the observed data $p(y|\theta)$ or likelihood function with prior knowledge about θ before data collection $p(\theta)$, normalized by the marginal likelihood $p(y)$. Point estimates and credible intervals can be obtained the posterior distribution to describe likely parameter values based on a fusion of prior knowledge and observed data.

Bayesian-estimated MLMs rely on sampling algorithms that provide accurate parameter values even when the number of clusters is small – but only if *appropriate priors* are carefully selected (McNeish & Stapleton, 2016). As McNeish (2016) noted, with small samples, even

using diffuse priors (which are often specified as software defaults, see e.g., van Erp et al., 2018) can result in problematic posterior parameter estimates that are more biased than frequentist methods since the effect of sampling extreme values cannot be cancelled out by the likelihood and the prior impact is stronger in small samples than with large samples. Not surprisingly, the problem of inaccurate prior effects for small-sample Bayesian model estimation only gets worse when researchers use priors that are strongly informative (Holtmann et al., 2016).

The Practice of Bayesian Analyses in Recent Psychology and Education Research

While existing systematic reviews have documented developments and practices of Bayesian methods in psychology and education up to 2015 (König & Van De Schoot, 2018; van de Schoot et al., 2017), the use of Bayesian analyses in social science research is likely to have progressed considerably since those reviews were conducted. The development of user-friendly, open-source software like the 'brms' package in *R*, which is based on *Stan* (Bürkner, 2017), along with a proliferation of online forums such as *Stack Exchange* (n.d.) and educational texts like *Statistical Rethinking* (McElreath, 2020) have lowered the barriers to implementing Bayesian analyses, likely increasing their adoption among applied researchers with varying levels of experience in Bayesian statistics.

To gain insight into the prevalence of applied Bayesian analysis in educational psychology research, a systematic review of original articles appearing in five high-impact journals published between 2014 and 2024 was conducted. These journals included: *Computers & Education* (CAED; Impact Factor: 8.9), *Child Development* (CD; Impact Factor: 3.9), *Developmental Psychology* (DP; Impact Factor: 3.1), *Journal of Educational Psychology* (JEP; Impact Factor: 5.6), and *Journal of Youth and Adolescence* (JYA; Impact Factor: 3.7). The search was conducted within each journal's website using the term "Bayesian" for original

research articles only in the target year range. The initial search identified 1,109 articles out of 8,415 published. All identified articles were screened with full text to only include those that implemented Bayesian estimation through the MCMC estimation algorithm. The exclusion process left 115 eligible articles, which is approximately 1.4% of all the publications.

Articles were then coded for the software used, type of Bayesian priors that were employed, and the extent to which Bayesian prior checks were performed. For $n = 31$ articles in which the authors did not discuss their prior specifications in the main text or supplemental materials (e.g., analysis code), we assumed that the authors used their chosen software's default prior settings. For these articles as well as those that explicitly mentioned the use of software default settings, articles were classified as either the noninformative or weakly informative categories (depending on the known prior settings of the software).

Table 1 presents the review results by each journal and combined. Multilevel models were used in $n = 41$ of the total articles using Bayesian analysis (36%). Notably, over half of the articles that used Bayesian analysis ($n = 64$, 56%) used noninformative priors for, the majority of which were software defaults ($n = 59$). Weakly informative priors were used in $n = 40$ studies (35%), with 28 specifying their own weakly informative priors (variance component estimates often relied on software defaults). Moderate or strongly informative priors (for fixed effects coefficients only), were only used in $n = 11$ articles (10%). Interestingly, among these 11 articles, three constructed their informative priors with an explicitly data-based approach, where hyperparameters for the prior are derived from sample data through frequentist model estimates or descriptive statistics. At its most extreme, the practice of using data-based priors is known as "double-dipping" when the same data is used twice: first to inform the prior beliefs and then again to obtain the posterior. The practice of double-dipping is criticized because it

fundamentally violates the likelihood principle. Specifically, a sample-based prior will be highly correlated with its likelihood and the integration of the two will cause the posterior distribution to be narrower in shape around sample values than it should be.

The three articles that used sample-based priors did so in different degrees. One used regression coefficient estimates and standard errors from previous year's data as priors for their current year data analysis. Another plugged in estimates from their regional sample as the priors for another sample. The third used sample-based bivariate correlation estimates as factor loading starting values in their (same-sample) factor models. Although none of these three studies used data-model-based double-dipping (which would be estimating the model with uninformative priors first and then using model-based estimates as priors in a second stage of analysis), these authors' use of directly accessible information for specifying priors can shed light on the logic model about priors that applied researchers may be using , even if it is inconsistent with recommended methodological procedures (see e.g., Zondervan-Zwijnenburg et al., 2017).

Data-Dependent Priors and Empirical Bayes

Setting prior distributions can be challenging when reliable external information is inadequate, or when the hyperparameters are difficult to interpret. To address this, some researchers use data-dependent priors. These approaches determine the hyperparameters based on either sample descriptive statistics or baseline model estimates obtained through frequentist methods or noninformative prior analysis (Carlin & Louis, 2000; Wasserman, 2000). By using the data to both construct priors and update them via the likelihood, researchers effectively use the information twice, potentially leading to overconfidence in parameter estimates and underestimated posterior uncertainty (Berger, 2006). Empirical Bayes (EB) methods typically combine initial estimates with large variances through certain procedures or adjustments to

account for the uncertainty in estimating hyperparameters (Darnieder, 2011). For instance, van Erp et al. (2018) proposed an EB prior for location parameters in Bayesian Structural Equation Modeling (SEM). Instead of centering around the potentially unstable ML estimates, the prior is centered around a reference value. Then the prior variance was set equal to the sum of the estimated error variance and the square of the estimated effect, ensuring the prior contains minimal information less than or equal to one observation.

The use of data-dependent priors in MLM has received limited attention. In their systematic review of simulation studies on the performance of Bayesian estimation in small samples, Smid et al. (2020) identified only one study (Browne & Draper, 2000) that specifically investigated data-dependent priors in MLM. Existing studies that compared data-dependent priors with other default noninformative priors or frequentist methods in different statistical models have yielded inconsistent results and recommendations.

While methodologists have cautioned against naive implementation of data-dependent priors, our systematic review revealed that researchers in applied setting may adopt informal and intuitive approaches by using the initial estimates to construct informative priors. This direct "plug-in" approach uses the data twice and does not account for the uncertainty in the initial estimates, leading to problematic posterior inferences. Recent work by Konold et al. (2025) demonstrated that when data-based priors are derived from a biased baseline model (in the context of biased L2 predictor coefficient due to the use of doubly manifest aggregation), more informative data-based priors resulted in tightened credible intervals around the biased estimate while producing better model fit indices. This creates a situation where researchers might have increased confidence in results that are less likely to capture the true parameter values.

Current Study

The current study extends the earlier methodological research by examining how varying degrees of informativeness in data-model-based priors (i.e., "double-dipping") affects fixed effects parameter recovery performance across a range of plausible multilevel data conditions, focusing on circumstances in which a correctly specified hierarchical linear model is applied. In addition, we investigate a split-sample cross-validation informative prior setting approach for reducing overfitting in the context of predictive inference (Picard & Berk, 1990). By randomly dividing the dataset and using one portion to derive priors and the remaining portion for Bayesian updating, independence between prior information and the likelihood can be preserved. Specifically, we aim to answer the following research questions:

1. What is the impact of different sample-based prior specifications (ranging from noninformative to highly informative) on the estimation performance of multilevel linear model regression coefficient parameters at both L1 and L2 across varying data conditions, including different numbers of clusters, cluster sizes, intraclass correlations, and magnitude of fixed effects?
2. Does a random split-sample approach help mitigate potential problems with using the same dataset to derive and apply priors?
3. To what extent do the consequences of using sample-based priors differ for within-cluster and between-cluster parameter recovery?

Method

To answer the research questions, a Monte Carlo simulation study was conducted using *Mplus* within *R*. Sample data were generated from a population two-level linear random intercept

model under varying conditions and then were analyzed with the correctly specified model using different prior specifications.

Data Generation

Data were generated using *Mplus 8* (Muthén & Muthén, 1998-2017) within the *MplusAutomation* package in *R* (Hallquist & Wiley, 2018). Our population generating model was a two-level random-intercept model with three predictors, cluster-mean centered at L1 and grand-mean centered at L2, as follows.

$$Y_{ij} = \gamma_{00} + \gamma_{10}\left(X_{1ij} - \bar{X}_{1.j}\right) + \cdots + \gamma_{30}\left(X_{3ij} - \bar{X}_{3.j}\right) + \gamma_{01}\left(\bar{X}_{1.j} - \bar{X}_{1..}\right) + \cdots + \gamma_{03}\left(\bar{X}_{3.j} - \bar{X}_{3..}\right) + u_{0j} + r_{ij} \quad (3)$$

In (2) we assumed multivariate normality for the fixed effects and normality for the two random effects. The data-generating parameters were set a priori using a pre-specified correlation matrix among the outcome and predictors to reflect different effect sizes (all predictors were assumed to have intercorrelations of .10 to be realistic), desired total population R^2 at each level, and the desired intraclass correlation coefficient (ICC). The ICCs were set equal for the outcome and predictors, which set the L2 variance for the outcome and each predictor to be the same. As a result, the L2 regression coefficients were equivalent to the L1 regression coefficients.

Varied Conditions

Five population conditions were varied to reflect realistic educational psychology research scenarios as well as for consistency with previous multilevel model methodological research, as follows.

1. Three numbers of clusters (J = 10, 30, 100) were chosen to reflect both methodological research as well as realistic applied studies. Specifically, J = 10 is well below methodological recommendations for using maximum likelihood

estimation, but yet represents the traditionally lowest threshold for treating a clustering effect as a random effect and is also a circumstance where Bayesian estimation may be preferred. The $J = 30$ condition was used because it is typically considered a minimum threshold for assuming normality of L2 residuals and can easily represent a sample of classrooms or schools in an educational psychology study. Lastly, $J = 100$ was selected because it represents a large sample that would be expected to provide estimates closer to the population level (relative to the low number of parameters being estimated in our model).

2. Two cluster sizes ($M = 5, 30$) were selected to represent small and large within-cluster sample sizes. A cluster size of 5 can represent scenarios common in longitudinal studies where measurement occasions form the lower level, while 30 is typical in educational research where students are nested within classrooms.
3. Three ICC levels ($ICC = .10, .20, .50$) were used to represent realistic levels of non-independence in educational research, with .10 being a small clustering effect (10% of variability explained by clustering) and .50 representing strong clustering that would be expected in longer or more intensive intervention studies.
4. Two types of regression coefficient sizes were used: null and non-null. In the null condition, all standardized regression coefficients were set to 0. In the non-null condition, the regression coefficients were set at 0.11, 0.33, and 0.56 for the three predictors, respectively (predictor-outcome correlations were set at .20, .40, and .60, respectively, with intercorrelations of .10). (We note that the standardized and unstandardized values were identical due to how the population data generation was constructed.)

For each of the 3*2*3*4 = 72 conditions (fully crossed), 1,000 samples were drawn. Each of these 72,000 samples were then modeled using the (data-generating) correctly specified model shown in (2) and estimated with Bayesian MCMC (TYPE = BAYES) with seven different fixed effects coefficient prior specification approaches (variance parameter prior distributions were set to be diffuse noninformative per the *Mplus* default, which uses an improper inverse gamma, *IG*(-1, 0)). The seven Bayesian prior approaches were as follows.

1. Diffuse noninformative priors were specified as $N(0, 10000)$. Posterior point estimate and variance from this baseline analysis were used to construct sample-based priors.
2. Point estimate as the prior mean with noninformative variance $N(\hat{\mu}, 10000)$, where $\hat{\mu}$ represents the posterior point estimate from the initial analysis with diffuse prior specification.
3. Point estimate as the prior mean with a variance 4 times that of the posterior variance $N(\hat{\mu}, \hat{\sigma}^2 * 4)$, a weakly informative prior.
4. Point estimate as the prior mean with a variance 2 times that of the posterior variance $N(\hat{\mu}, \hat{\sigma}^2 * 2)$, a moderately informative prior.
5. Point estimate as the prior mean with a variance equal to the posterior variance $N(\hat{\mu}, \hat{\sigma}^2 * 1)$, a strongly informative prior.
6. Diffuse noninformative prior on a split dataset $N(0, 10000)$ Split. For each replication, the sample data were randomly split in half, with the first half analyzed using the diffuse noninformative prior.
7. Point estimate as the prior mean with a variance twice that of the posterior variance on a split dataset $N(\hat{\mu}, \hat{\sigma}^2 * 2)$ Split. The resulting posterior estimates from the default

prior analysis on the first half of the data were used to construct a moderately informative prior for analyzing the second half of the sample.

Notably, *Mplus* invokes an automatic convergence criterion based on the potential scale reduction (PSR; Brooks & Gelman, 1998) that is monitored at every 100th iteration (Asparouhov & Muthén, 2010) such that all model parameters must reach PSR values of < 1.1 for iterations to stop. This setting was used for the current study as well, and as such, all models reached convergence.

Simulation Results Analysis

Bayesian analysis prior approach performance for the L1 and L2 fixed effects coefficient parameters was evaluated using three metrics: 1) posterior parameter median raw bias, 2) posterior parameter coverage based on 95% Bayesian highest posterior density (HPD) credible intervals, and 3) empirical bias of the posterior parameter standard deviation (i.e., standard error) relative to the population-generating parameter standard deviation (standard error).

Raw bias was computed as the difference between the Bayesian-estimated parameter median value and the true parameter value. Positive bias values indicate overestimation, while negative values indicate underestimation. Raw bias values within ±0.05 were considered acceptable, as this represents a relatively small deviation in the context of standardized coefficients that were being estimated.

Coverage was calculated as the proportion of replications in which the model-generating true coefficient value was within the bounds of the HPD credible interval. Coverage rates below 95% indicate overly narrow intervals that are overconfident (under-coverage), while coverage rates above 95% indicate too conservative intervals that are underconfident (over-coverage). Following Bradley's (1978) liberal criterion, coverage rates between 92.5% and 97.5% (i.e., 95%

± 2.5%) were considered reasonably close to the nominal rate, although we note that Hoogland and Boomsma (1998) have suggested a more relaxed criterion of 95% ± 5.0%.

Last, to assess the accuracy of Bayesian-estimated parameter posterior variance, **empirical bias** for each condition cell was calculated as the ratio of the mean Bayesian-estimated parameter posterior standard deviation to the empirical standard deviation of the coefficient point estimates across the 1,000 replications. Values of 1.0 indicate no bias, meaning the model's uncertainty estimates perfectly match the actual sampling variability of the estimates. Values below 1.0 indicate underestimated posterior standard deviations (narrow credible intervals), while values above 1.0 indicate overestimated posterior standard deviations. Following Hoogland and Boomsma (1998), values between 0.90 and 1.10 (±10% bias) were considered within an acceptable range.

Results

Across conditions, Bayesian-estimated coefficient parameter estimates exhibited minimal bias at both L1 (Table 2) and L2 (Table 3) across all design conditions, with values consistently lower than ±0.05. This lack of bias aligns with established findings that fixed-effect estimates in properly specified multilevel models are generally unbiased (Maas & Hox, 2004). Given this (expected) result, the foregoing results primarily focus on the second and third evaluation metrics of coverage and standard deviation/error empirical bias.

Level 1 Results

Coefficient Parameter Estimate Coverage Rates

Table 4 presents the mean L1 coverage rates for coefficient parameter estimates across sample size conditions and prior types, averaged over ICC levels and true coefficient values. (Results were averaged across ICC conditions because an analysis of variance of the coverage

rates using design conditions as factors revealed that ICC had no significant effect, with an $\hat{\eta}^2$ value close to zero.) The differences across true coefficient values are presented in Figure 1. The results demonstrate a clear pattern wherein increased sample-based prior informativeness significantly reduces coverage rates below the nominal .95 level, except for the split-data approach which maintained satisfying coverage. This pattern held across sample size conditions but appeared to be moderated by the number of clusters (*J*). The default noninformative prior, the split-data approach with the default prior, and the point estimate with default variance prior performed similarly and produced adequate coverage rates (.91 – .95). As priors became more informative, coverage rates declined substantially. The most informative prior, which directly used the posterior point estimate and standard deviation obtained using the default prior, exhibited severe undercoverage (.76 – .85) well below acceptable levels. However, the informative split-data prior $N(\hat{\mu}, \hat{\sigma}^2 * 2)$ Split produced excellent coverage rates close to or slightly above the nominal (.93 – .97) that often outperformed the default priors. Additionally, *J* = 30 conditions generally produced the best coverage rates across prior types. The undercoverage with more informative priors was most pronounced for small numbers of clusters (*J* = 10). Figure 1 illustrates these patterns in greater detail by showing the coverage rates of each true coefficient value (0, 0.11, 0.33, 0.56). Coverage rates were generally better for null and large effect sizes compared to moderate effect sizes.

Coefficient Standard Error Empirical Bias

Given the minimal parameter estimate raw bias across conditions, the observed coverage issues should be driven by biased standard error estimates. Table 5 presents mean L1 coefficient standard error empirical bias across sample size conditions and prior types. Figure 2 adds details across true coefficient values. The empirical standard error bias results show a clear pattern that

aligns directly with the coverage deficits observed previously. The noninformative priors produced empirical bias values close to 1 (0.95 – 1.06), explaining their adequate coverage rates. As prior informativeness increased, standard errors became increasingly underestimated, which corresponds to the declining coverage rates. The informative split-data prior consistently overestimated standard errors but mostly within the acceptable range (1.03 – 1.16), which explains its adequate coverage rates. Again, sample size conditions moderated these effects. With $J = 10$, standard error underestimation was most severe for informative priors, particularly for moderate coefficient sizes ($\beta = 0.11$, $\beta = 0.33$).

Level 2 Results

Coefficient Parameter Estimate Coverage Rates

Table 6 presents the mean L2 coverage rates across sample size conditions and prior types, aggregated across ICC levels and true coefficient values. The overall pattern of increasingly informative data-based priors lead to more severe undercoverage persists. The most informative data-based prior showed severe undercoverage well below the acceptable boundary (.77 – .87). Moreover, as the number of clusters and cluster size increases, the undercoverage of highly informative priors was more severe, with the largest sample size condition ($J = 100$, $M = 30$) having the largest undercoverage. The informative split-data prior performed relatively well but showed more variability than at L1, with coverage rates ranging from .88 to .98. When cluster size is smaller ($M = 5$), the informative split prior produced over-coverage (.98) or coverage rates close to the nominal level (.94 – .95), whereas when cluster size is larger ($M = 30$), it still exhibited undercoverage that is equivalent to the weakly informative data-based prior with four times variance (.88 – .93). Figure 3 illustrates these coverage patterns with the additional separation of true coefficient sizes. The undercoverage effect of more informative

priors consistently occurred for all coefficient sizes. Differences between coefficient sizes mostly occurred in underperforming informative priors (most obvious for moderate coefficient sizes in $J = 100$, $M = 30$), while for priors producing coverage rates close to the nominal level, coefficient sizes didn't exhibit obvious differences.

Coefficient Standard Error Empirical Bias

Table 7 presents the mean L2 standard error empirical bias across sample sizes and prior types. Compared to that at L1, the L2 empirical bias results show more extreme patterns, with substantial overestimation and underestimation of standard errors depending on conditions. The default noninformative prior and the point estimate with default variance prior showed considerable overestimation of standard errors in the small-sample conditions (bias = 1.10 – 1.33), whereas for conditions with many clusters ($J = 100$) the standard error estimates were quite accurate (bias = 0.97 – 1.01). The split-data default prior demonstrated substantial standard error overestimation in the small-J conditions, especially for the smallest total sample size condition ($J = 10$, $M = 5$), with an empirical bias of 2.11. Consistent with the declining coverage rates, as prior informativeness increased, standard errors became increasingly underestimated. The underestimation was further exacerbated by an increasing number of clusters and cluster sizes. The informative split-data prior showed variable performance at L2, producing fairly unbiased standard error estimates for most sample size conditions (0.99 – 1.09) except for the two extreme conditions with substantial overestimation (1.46) for $J = 10$, $M = 5$ and underestimation (0.86) for $J = 100$, $M = 30$. Figure 4 illustrates these patterns with added comparisons across coefficient values. The empirical bias was generally consistent across coefficient values, though the downward bias of informative sample-based priors for $J = 30$ was

less pronounced for the moderate coefficient sizes ($\beta = 0.11$, $\beta = 0.33$) and for $J = 100$ was less pronounced for the large coefficient size of $\beta = 0.56$.

Discussion

Prior specification is arguably the most important step in Bayesian MCMC analysis, yet it can be complicated and challenging. This study's systematic review showed that, while most researchers use software defaults, a few may take the intuitive approach of constructing informative priors with an explicitly data-model-based approach. It is likely that many applied researchers may not be aware of the procedures needed to account for the uncertainty in using priors derived from the data to estimate their model parameters, and the ripple effects doing so can cause.

While Konold et al. (2025) demonstrated how informative data-based priors derived from misspecified multilevel models can lead to severely biased credible intervals (even as model fit indices were seemingly good), the current study extends this work by investigating the consequences of using informative data-model-based priors on fixed effects coefficient parameters with varying degrees of informativeness in *correctly specified* multilevel models across a wide range of realistic research conditions. These results found that, while coefficient parameter values themselves were largely unbiased, the parameter posterior distribution variability was progressively underestimated as sample-based priors became more informative. This pattern was particularly pronounced with a larger number of clusters, especially for L2 parameters. With increasing L2 units, the underestimation of posterior uncertainty was amplified by using overly confident L2 parameter values to create informative priors.

Another and perhaps equally important contribution of this study is the evaluation of a split-sample approach as a potential remedy to the double-dipping problem, which used one

portion of the data to derive priors and the remaining portion for Bayesian updating. With a moderate variance, the split-sample approach outperformed other data-based prior specifications in most cases at both levels of the analysis model. This said, the standard errors from the split-sample approach were consistently larger than those from the full-sample approach with the same level of informativeness because only half the data was used for likelihood estimation; as such, future research might examine the use of a stratified split sample (stratified by cluster) and/or a smaller portion for the prior derivation phase. In any case, by preserving the independence between the prior and likelihood information, the split-sample approach allows more accurate uncertainty quantification.

Limitations and Future Research

Several limitations of the current study provide directions for future investigation. First, this study's results are limited to correctly specified models with a fairly simple 2-level random intercept structure, which might be seen as the best-case scenario. Future attention can be paid to more complex structures, such as models with random slopes or cross-level interactions. Investigating sample-based priors for misspecified models or models with interactions would provide insights about how the patterns observed here may unfold in more practical scenarios. Second, these results are limited to continuous outcomes. While it is unlikely that the major patterns observed would be markedly different when estimating binary or count outcomes, for example, it may be worthwhile to investigate whether sample-based priors might need to be adjusted using the standard deviation hyperparameter in particular. Third, while this study provides an evaluation of the effectiveness of a split-sample approach, it used a simple random split with equal proportions; alternative splitting strategies like stratified sampling or other splitting proportions warrant examination. Further, expanding the comparison to include other

data-based prior methods that incorporate uncertainty adjustments (e.g., van Erp et al., 2018) could further our understanding of the relative severity of the double-dipping problem and the extent to which methodological adjustments can mitigate the reuse of information.

Conclusion

The current study sheds light on two major ways: first, it provides a systematic review of the current state of the use of Bayesian MCMC analysis in educational psychology across five top-tiered journals spanning the most recent decade. Those results indicate that applied researchers are using Bayesian analyses but are largely doing so using software defaults, which may mean that researchers are not aware of how to properly construct priors to take advantage of a key advantage in using a Bayesian analysis. Although only a tiny portion of studies used informative priors, there is a sense that, in the absence of good prior study information, it may be tempting to use priors derived from the same data in a 2-step fashion.

As this study's simulation results show – and which is consistent with earlier methodological work – using double-dipped highly informative priors for multilevel model fixed effects is largely a bad idea because the coefficients' posterior distribution variabilities will be underestimated. Moreover, this underestimation will not be alleviated by larger L2 sample sizes; to the contrary, it can be exacerbated as the number of clusters increases. Instead, the use of weakly informative or a split-sample technique is recommended. Future research can provide guidance around the best split-sample approach.

References

Asparouhov, T., & Muthén, B. (2010). *Bayesian analysis using Mplus: Technical implementation*. https://www.statmodel.com

Berger, J. (2006). The case for objective Bayesian analysis. *Bayesian Analysis*, *1*(3). https://doi.org/10.1214/06-BA115

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

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, *7*(4), 434–455. https://doi.org/10.1080/10618600.1998.10474787

Browne, W. J., & Draper, D. (2000). Implementation and performance issues in the Bayesian and likelihood fitting of multilevel models. *Computational Statistics*, *15*(3), 391–420. https://doi.org/10.1007/s001800000041

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1). https://doi.org/10.18637/jss.v080.i01

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

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

Dedrick, R. F., Ferron, J. M., Hess, M. R., Hogarty, K. Y., Kromrey, J. D., Lang, T. R., Niles, J. D., & Lee, R. S. (2009). Multilevel modeling: A review of methodological issues and applications. *Review of Educational Research*, *79*(1), 69–102. https://doi.org/10.3102/0034654308325581

Depaoli, S. (2014). The impact of inaccurate "informative" priors for growth parameters in Bayesian growth mixture modeling. *Structural Equation Modeling: A Multidisciplinary Journal*. https://www.tandfonline.com/doi/abs/10.1080/10705511.2014.882686

Depaoli, S., & Clifton, J. P. (2015). A Bayesian approach to multilevel structural equation modeling with continuous and dichotomous outcomes. *Structural Equation Modeling: A Multidisciplinary Journal*, *22*(3), 327–351. https://doi.org/10.1080/10705511.2014.937849

Hallquist, M. N., & Wiley, J. F. (2018). MplusAutomation: An R package for facilitating large-scale latent variable analyses in Mplus. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(4), 621–638. https://doi.org/10.1080/10705511.2017.1402334

Holtmann, J., Koch, T., Lochner, K., & Eid, M. (2016). A comparison of ML, WLSMV, and Bayesian methods for multilevel structural equation models in small samples: A simulation study. *Multivariate Behavioral Research*, *51*(5), 661–680. https://doi.org/10.1080/00273171.2016.1208074

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and a meta-analysis. *Sociological Methods & Research*, *26*(3), 329–367. https://doi.org/10.1177/0049124198026003003

König, C., & Van De Schoot, R. (2018). Bayesian statistics in educational research: A look at the current state of affairs. *Educational Review*, *70*(4), 486–509. https://doi.org/10.1080/00131911.2017.1350636

Konold, T. R., Sanders, E. A., & Afolabi, K. (2025). From Seinfeld to statistics: On the dangers of double dipping with Bayesian inference. *Structural Equation Modeling: A Multidisciplinary Journal*, 1–10. https://doi.org/10.1080/10705511.2025.2487053

Lynch, S. M., & Bartlett, B. (2019). Bayesian statistics in sociology: Past, present, and future. *Annual Review of Sociology*, *45*(1), 47–68. https://doi.org/10.1146/annurev-soc-073018-022457

Maas, C. J. M., & Hox, J. J. (2004). Robustness issues in multilevel regression analysis. *Statistica Neerlandica*, *58*(2), 127–137. https://doi.org/10.1046/j.0039-0402.2003.00252.x

Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *1*(3), 86–92. https://doi.org/10.1027/1614-2241.1.3.86

McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan* (Second edition). CRC Press.

McNeish, D. (2016). On using Bayesian methods to address small sample problems. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(5), 750–773. https://doi.org/10.1080/10705511.2016.1186549

McNeish, D., & and Stapleton, L. M. (2016). Modeling clustered data with very few clusters. *Multivariate Behavioral Research*, *51*(4), 495–518. https://doi.org/10.1080/00273171.2016.1167008

Muthén, L. K., & Muthén, B. O. (1998). *Mplus user's guide* (8th ed.). Muthén & Muthén.

Picard, R. R., & Berk, K. N. (1990). Data splitting. *The American Statistician*, *44*(2), 140–147. https://doi.org/10.1080/00031305.1990.10475704

Smid, S. C., McNeish, D., Miočević, M., & van de Schoot, R. (2020). Bayesian versus frequentist estimation for structural equation models in small sample contexts: A systematic review. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(1), 131–161. https://doi.org/10.1080/10705511.2019.1577140

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed). Sage.

Stack Exchange. (n.d.). Stack Exchange. Retrieved June 1, 2025, from https://stackexchange.com

van de Schoot, R., Kaplan, D., Denissen, J., Asendorpf, J. B., Neyer, F. J., & van Aken, M. A. G. (2014). A gentle introduction to Bayesian analysis: Applications to developmental research. *Child Development*, *85*(3), 842–860. https://doi.org/10.1111/cdev.12169

van de Schoot, R., Winter, S. D., Ryan, O., Zondervan-Zwijnenburg, M., & Depaoli, S. (2017). A systematic review of Bayesian articles in psychology: The last 25 years. *Psychological Methods*, *22*(2), 217–239. https://doi.org/10.1037/met0000100

van Erp, S., Mulder, J., & Oberski, D. L. (2018). Prior sensitivity analysis in default Bayesian structural equation modeling. *Psychological Methods*, *23*(2), 363–388. https://doi.org/10.1037/met0000162

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., Selker, R., Gronau, Q. F., Šmíra, M., Epskamp, S., Matzke, D., Rouder, J. N., & Morey, R. D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review*, *25*(1), 35–57. https://doi.org/10.3758/s13423-017-1343-3

Wasserman, L. (2000). Asymptotic inference for mixture models using data-dependent priors. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *62*(1), 159–181. https://doi.org/10.1111/1467-9868.00226

Zondervan-Zwijnenburg, M., Peeters ,Margot, Depaoli ,Sarah, & and Van de Schoot, R. (2017). Where do priors come from? Applying guidelines to construct informative priors in small sample research. *Research in Human Development*, *14*(4), 305–320. https://doi.org/10.1080/15427609.2017.1370966

Table 1

Characteristics of Bayesian Analyses Published in Education and Developmental Research Journals, 2014-2024

Characteristic		CAED	CD	DP	JEP	JYA	*Total*
Total Articles		2041	1761	2006	861	1746	8415
Bayesian Estimation		9 (0.5%)	38 (2%)	28 (1%)	20 (2%)	20 (1%)	115 (1%)
Multilevel Model		2 (22%)	16 (42%)	10 (36%)	8 (40%)	5 (25%)	41 (36%)
	Total	3 (33%)	21 (53%)	14 (50%)	12 (60%)	14 (80%)	64 (56%)
Noninformative Priors	Default	2 (22%)	19 (50%)	13 (46%)	12 (60%)	13 (65%)	59 (51%)
	User-Specified	1 (11%)	2 (5%)	1 (4%)	0 (0%)	1 (5%)	5 (4%)
	Total	4 (44%)	13 (34%)	12 (43%)	7 (35%)	4 (20%)	40 (35%)
Weakly Informative Priors	Default	1 (11%)	4 (11%)	3 (11%)	2 (10%)	2 (10%)	12 (10%)
	User-Specified	3 (33%)	9 (24%)	9 (32%)	5 (25%)	2 (10%)	28 (24%)
Informative Priors		2 (22%)	4 (11%)	2 (7%)	1 (5%)	2 (10%)	11 (10%)
MCMC Diagnostics		6 (67%)	25 (66%)	13 (46%)	11 (55%)	13 (65%)	68 (59%)
Posterior Predictive Checks		5 (56%)	11 (29%)	9 (32%)	1 (5%)	2 (10%)	28 (24%)
Prior Sensitivity Analysis		4 (44%)	3 (11%)	1 (4%)	2 (10%)	0 (0%)	10 (9%)

Note. CAED = *Computers and Education*, CD = *Child Development*, DP = *Developmental Psychology*, JEP = *Journal of Educational Psychology*, JYA = *Journal of Youth and Adolescence*.

Table 2

Mean L1 Coefficient Parameter Estimate Raw Bias by Sample Sizes and Prior Types

	Fixed Effect Coefficient Prior						
Sample Sizes	$0, 10000$	$0, 10000$ (Split)	$\hat{\mu}, 10000$	$\hat{\mu}, \hat{\sigma}^2 * 4$	$\hat{\mu}, \hat{\sigma}^2 * 2$	$\hat{\mu}, \hat{\sigma}^2 * 2$ (Split)	$\hat{\mu}, \hat{\sigma}^2 * 1$
$J = 10$							
$M = 5$	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$M = 30$	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$J = 30$							
$M = 5$	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$M = 30$	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$J = 100$							
$M = 5$	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$M = 30$	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Note. Each cell is averaged across all ICC levels (.10, .20, .50) and true coefficient values (0, 0.11, 0.33, 0.56). Values closer to 0 are better.

Table 3

Mean L2 Coefficient Parameter Estimate Raw Bias by Sample Sizes and Prior Types

	Fixed Effect Coefficient Prior						
Sample Sizes	$0, 10000$	$0, 10000$ (Split)	$\hat{\mu}, 10000$	$\hat{\mu}, \hat{\sigma}^2 * 4$	$\hat{\mu}, \hat{\sigma}^2 * 2$	$\hat{\mu}, \hat{\sigma}^2 * 2$ (Split)	$\hat{\mu}, \hat{\sigma}^2 * 1$
$J = 10$							
$M = 5$	-0.03	-0.01	-0.03	-0.03	-0.03	-0.01	-0.03
$M = 30$	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03
$J = 30$							
$M = 5$	-0.01	0.00	-0.01	-0.01	-0.01	0.00	-0.02
$M = 30$	-0.02	-0.02	-0.02	-0.03	-0.03	-0.03	-0.03
$J = 100$							
$M = 5$	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$M = 30$	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Note. Each cell is averaged across all ICC levels (.10, .20, .50) and true coefficient values (0, 0.11, 0.33, 0.56). Values closer to 0 are better.

Table 4

Mean L1 Coefficient Parameter Estimate Coverage Rates by Sample Sizes and Prior Types

	Fixed Effect Coefficient Prior						
Sample Sizes	$0, 10000$	$0, 10000$ (Split)	$\hat{\mu}, 10000$	$\hat{\mu}, \hat{\sigma}^2 * 4$	$\hat{\mu}, \hat{\sigma}^2 * 2$	$\hat{\mu}, \hat{\sigma}^2 * 2$ (Split)	$\hat{\mu}, \hat{\sigma}^2 * 1$
$J = 10$							
$M = 5$	.91	.94	.91	.87	.82	.94	.76
$M = 30$	.91	.91	.91	.86	.82	.93	.76
$J = 30$							
$M = 5$	.95	.94	.95	.92	.89	.96	.84
$M = 30$	.95	.95	.95	.93	.90	.97	.84
$J = 100$							
$M = 5$	.93	.93	.93	.90	.86	.95	.80
$M = 30$	.91	.92	.91	.88	.84	.94	.79

Note. Each cell is averaged across all ICC levels (.10, .20, .50) and true coefficient values (0, 0.11, 0.33, 0.56). Values closer to .95 are better.

Table 5

Mean L1 Coefficient Standard Error Empirical Bias by Sample Sizes and Prior Types

	Fixed Effect Coefficient Prior						
Sample Sizes	$0, 10000$	$0, 10000$ (Split)	$\hat{\mu}, 10000$	$\hat{\mu}, \hat{\sigma}^2 * 4$	$\hat{\mu}, \hat{\sigma}^2 * 2$	$\hat{\mu}, \hat{\sigma}^2 * 2$ (Split)	$\hat{\mu}, \hat{\sigma}^2 * 1$
$J = 10$							
$M = 5$	0.98	1.04	0.98	0.85	0.77	1.06	0.66
$M = 30$	0.95	0.95	0.95	0.84	0.77	1.03	0.66
$J = 30$							
$M = 5$	1.06	1.01	1.06	0.95	0.87	1.09	0.76
$M = 30$	1.06	1.05	1.06	0.95	0.88	1.16	0.77
$J = 100$							
$M = 5$	1.01	0.98	1.01	0.90	0.82	1.05	0.71
$M = 30$	1.00	1.00	1.00	0.89	0.81	1.10	0.71

Note. Each cell is averaged across all ICC levels (.10, .20, .50) and true coefficient values (0, 0.11, 0.33, 0.56). Values closer to 1 are better.

Table 6

Mean L2 Coefficient Parameter Estimate Coverage Rates by Sample Sizes and Prior Types

	Fixed Effect Coefficient Prior						
Sample Sizes	$0, 10000$	$0, 10000$ (Split)	$\hat{\mu}, 10000$	$\hat{\mu}, \hat{\sigma}^2 * 4$	$\hat{\mu}, \hat{\sigma}^2 * 2$	$\hat{\mu}, \hat{\sigma}^2 * 2$ (Split)	$\hat{\mu}, \hat{\sigma}^2 * 1$
$J = 10$							
$M = 5$	.97	.99	.97	.93	.89	.98	.83
$M = 30$	.96	.96	.96	.91	.87	.91	.82
$J = 30$							
$M = 5$	.97	.95	.97	.94	.92	.95	.87
$M = 30$	.96	.96	.96	.93	.90	.93	.85
$J = 100$							
$M = 5$	.93	.94	.93	.89	.85	.94	.79
$M = 30$	.92	.92	.92	.87	.84	.88	.77

Note. Each cell is averaged across all ICC levels (.10, .20, .50) and true coefficient values (0, 0.11, 0.33, 0.56). Values closer to .95 are better.

Table 7

Mean L2 Coefficient Standard Error Empirical Bias across Sample Sizes and Prior Types

	Fixed Effect Coefficient Prior						
Sample Sizes	$0, 10000$	$0, 10000$ (Split)	$\hat{\mu}, 10000$	$\hat{\mu}, \hat{\sigma}^2 * 4$	$\hat{\mu}, \hat{\sigma}^2 * 2$	$\hat{\mu}, \hat{\sigma}^2 * 2$ (Split)	$\hat{\mu}, \hat{\sigma}^2 * 1$
$J = 10$							
$M = 5$	1.33	2.11	1.33	1.02	0.89	1.46	0.76
$M = 30$	1.31	1.31	1.31	0.99	0.87	0.99	0.74
$J = 30$							
$M = 5$	1.20	1.10	1.20	1.06	0.96	1.09	0.84
$M = 30$	1.17	1.17	1.17	1.02	0.93	1.03	0.81
$J = 100$							
$M = 5$	0.99	1.01	0.99	0.87	0.79	1.02	0.68
$M = 30$	0.97	0.98	0.97	0.86	0.78	0.86	0.67

Note. Each cell is averaged across all ICC levels (.10, .20, .50) and true coefficient values (0, 0.11, 0.33, 0.56). Values closer to 1 are better.

Figure 1

Illustration of Mean L1 Parameter Estimate Coverage Rates by Sample Sizes, Prior Types, and Ture Coefficient Sizes

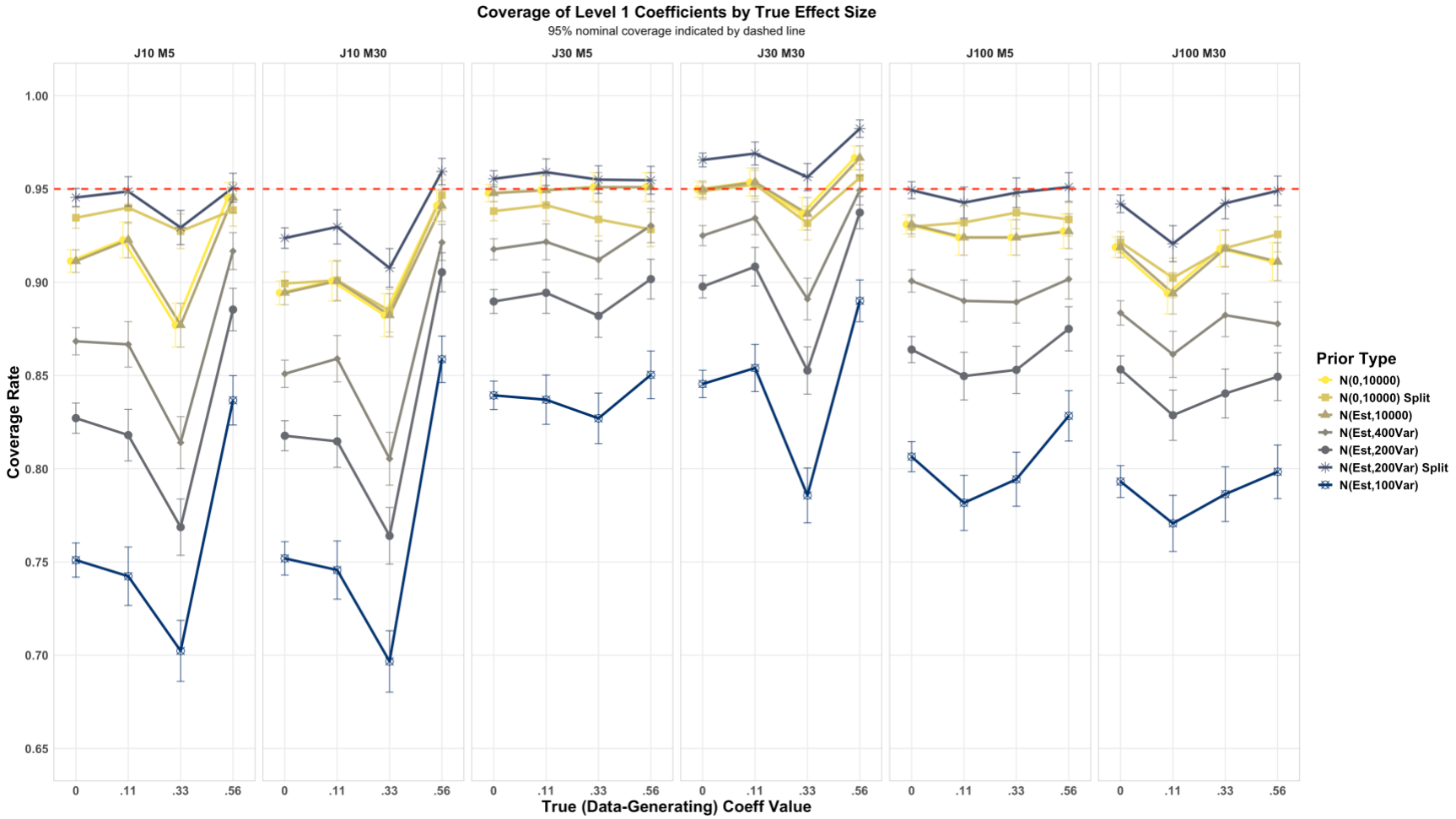

Figure 2

Illustration of Mean L1 Parameter Estimate Raw Bias and Empirical Bias by Sample Sizes, Prior Types, and Ture Coefficient Sizes

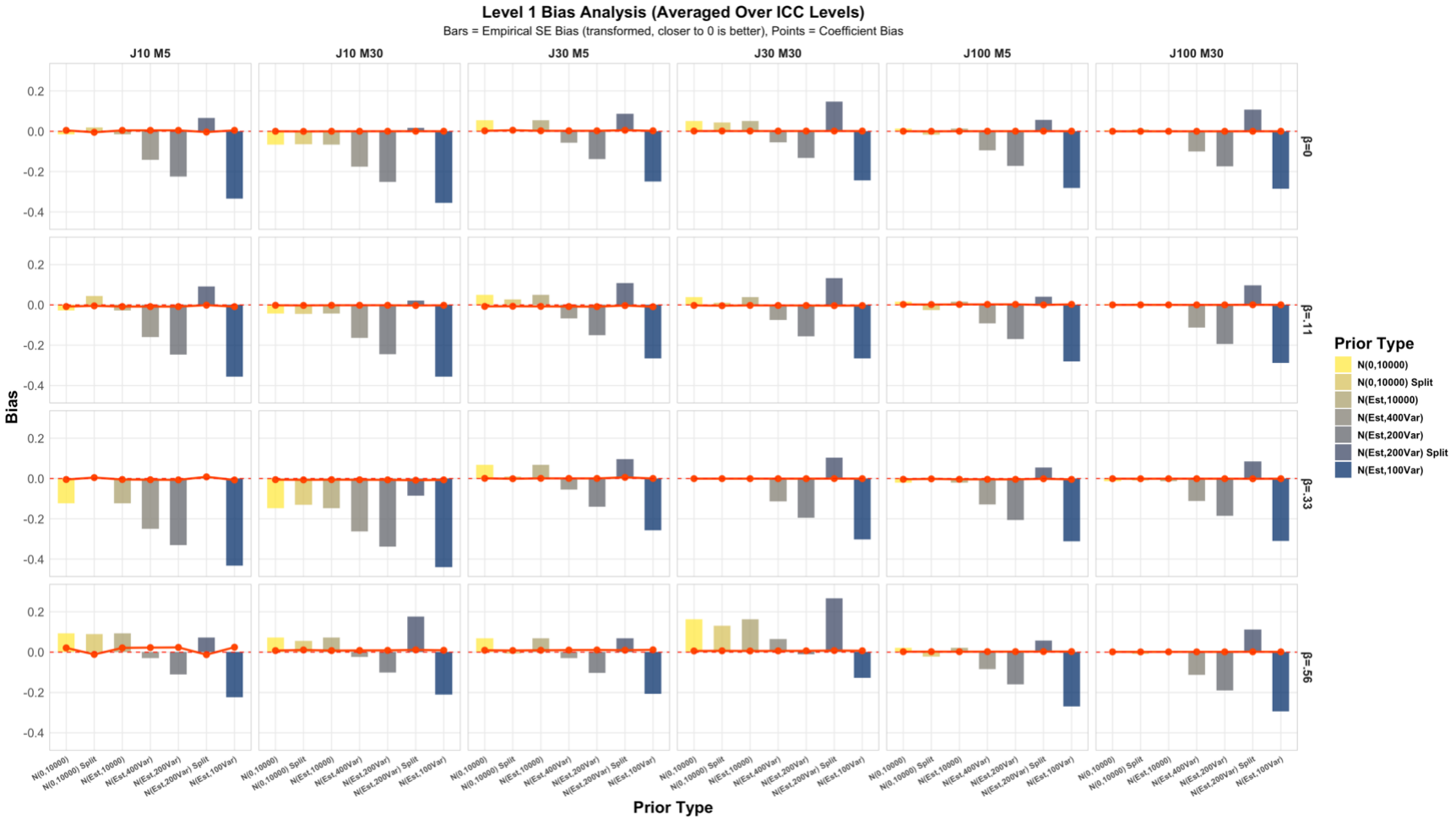

Figure 3

Illustration of Mean L2 Parameter Estimate Coverage Rates by Sample Sizes, Prior Types, and Ture Coefficient Sizes

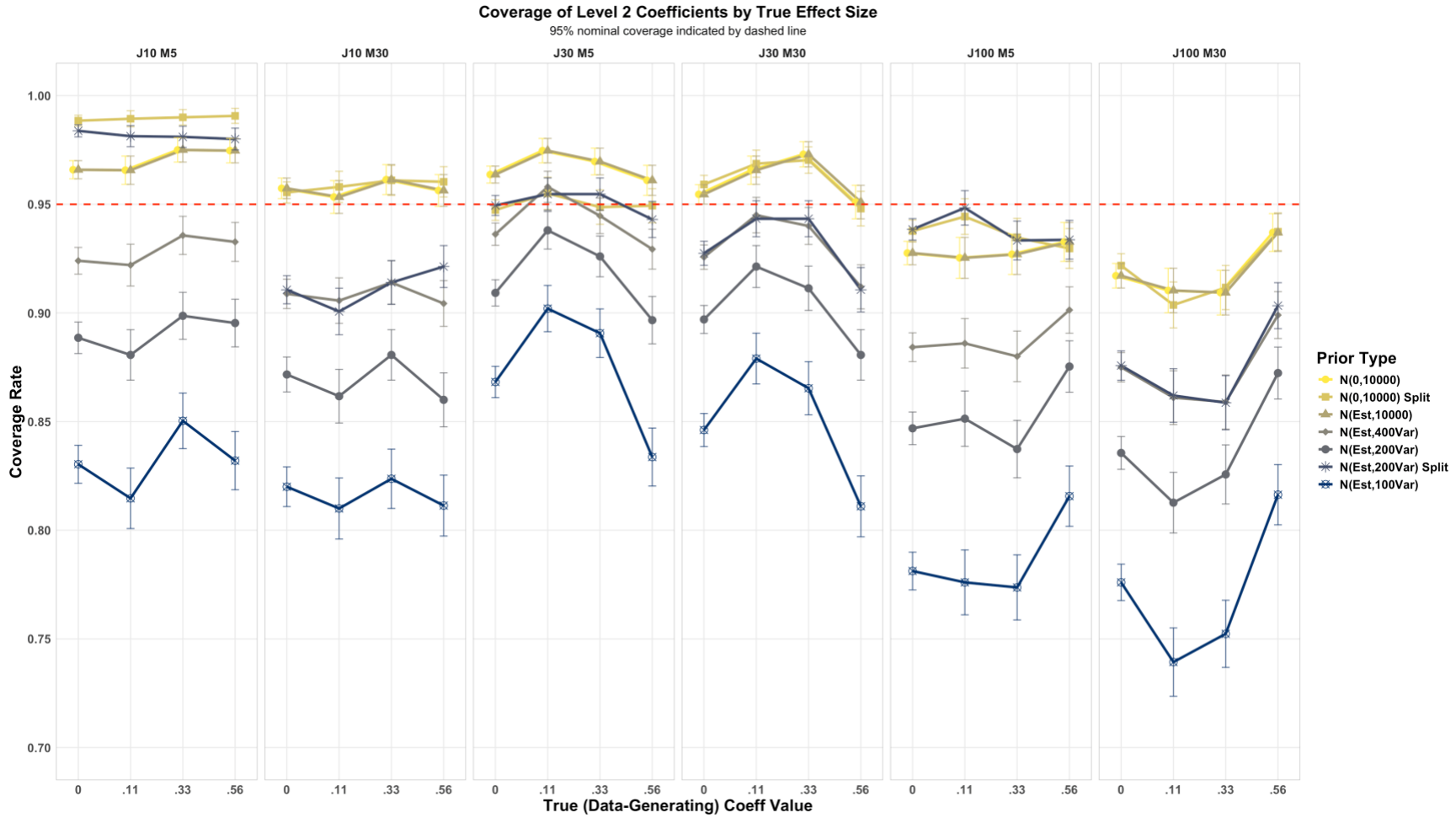

Figure 4

Illustration of Mean L2 Parameter Estimate Raw Bias and Empirical Bias by Sample Sizes, Prior Types, and True Coefficient Sizes