Comparing Three Predictive Information Criteria for Linear Multilevel Bayesian Regression Selection

Sean Devine

Department of Psychology, McGill University

PSYC 749: Statistical Foundations for Research in Quantitative Psychology

Dr. Carl Falk

December 1st, 2022

All results, figures, and scripts are available at

https://github.com/seandamiandevine/PSYC749_Final

Introduction

From 2007 to 2017 there has been a threefold increase in published psychology articles that utilize multilevel modeling techniques (Huang, 2018). Popular packages for estimating multilevel models such as *lme4* (Bates et al., 2014) have doubtlessly accelerated this trend. However, *lme4* only provides the most likely point estimates of parameters—i.e., the maximum likelihood estimates (MLE). Often however, researchers want to make claims not only about a single value of a given parameter, but also about the uncertainty surrounding that estimate. Accordingly, another popular package that offers many of the same user-features as *lme4*, but estimates multilevel models in a Bayesian framework, is *brms* (Bürkner, 2017), which estimates posterior distributions for each parameter using MCMC sampling as implemented in the probabilistic programming language, Stan.

While the Bayesian approach yields some advantages over the frequentist approach (e.g., direct examination of posterior uncertainty, the ability to include prior beliefs into estimates), multilevel model selection in the Bayesian framework can be less straightforward than in its frequentist counterpart. This is because, in the absence of significance-based model comparison tests—such as likelihood ratio tests—or point parameter estimates needed to compute overall fit indices—like AIC or BIC—it can be unclear which, of a set of candidate models, is the "best model" who's parameters should be interpreted—i.e., the most likely given the data at hand.

To address this issue, researchers have proposed several information criteria for Bayesian multilevel model selection (cf. Gelman, Hwang, & Vehtari, 2014), including the deviance information criterion (DIC; Spiegelhalter, et al., 2002), the widely applicable information criterion (WAIC; Watanabe, 2010), and leave-one-out cross-validation (LOO-CV; Vehtari, Gelman, & Gabry, 2017). Much like AIC and BIC, which rely on the MLE, each of these criteria

summarize the posterior likelihood of the data by a single value, which can be used to disambiguate which model, of a set of candidate models, should be selected as the most likely model given the data. In the case of all three measures are on the deviance scale and thus lower values suggest better model fit.

Past simulation work has shown that the DIC may be biased to select overfitted models and that the WAIC and LOO-CV may be better alternatives (Ando, 2011; Vehtari et al., 2017). However, this work has not systematically explored these criteria's performance in a multilevel context, where level-1 (i.e., cluster) and level-2 (i.e., observation) sample sizes may be small and/or asymmetrical, and the degree of clustering may influence criteria performance. In other words, it remains unclear whether and how differences in information criteria performance for Bayesian multilevel model selection emerge as a function of 1) the number of groups in the data, 2) the number of observations present in each group, and 3) the degree of clustering present in the dataset (i.e., whether most groups are fully distinct or overlap heavily in the outcome in question). Furthermore, the relationship between the *nature* of model misspecification—that is, which features of a given model deviate from the generative (i.e., "true" population) model—and the accuracy of a given information criterion has not been systematically examined. This last point is important, as evidence suggests that reporting practices for random effects structure remains poor (Luo et al., 2021), raising the risk that many models in the literature may suffer from a misspecification in terms of random effects specifically.

To this end, the present simulation study explores the performance of three predictive information criteria for Bayesian multilevel model selection—DIC, WAIC, and LOO-CV—across cluster size, observation size, degree of clustering, and type of model misspecification. Five models—four of which suffer from various misspecifications (described below)—are

estimated at different level-1 sample sizes, level-2 sample sizes, and degrees of clustering (operationalized as the magnitude of the random intercept variance).

Below, I provide a summary of how each criterion is calculated from a posterior distribution (here estimated using MCMC sampling as implemented by *brms*). I then provide details about the simulations themselves and the estimated models at each cell, followed by the results of these simulations. I conclude with a short discussion on the limited scope of this study in the context of this course and suggestions for how it could be extended into a larger treatment of this question.

Information Criteria[1]

Deviance Information Criterion (DIC)

The DIC is a Bayesian and hierarchical extension of the Akaike information criterion (AIC). The AIC estimates the predicted out-of-sample accuracy as:

$$\text{AIC} = -2 \ln p\left(D|\hat{\theta}_{MLE}\right) + 2k \qquad \text{Eq. 1}$$

where $\ln p\left(D\middle|\hat{\theta}_{MLE}\right)$ is the log-likelihood of the data, D, at the MLE, $\hat{\theta}_{MLE}$, and k is the number of parameters in the model. Smaller AIC values indicate better model fit. AIC penalizes overfitting by adding a constant factor (2) for each new parameter added to the model. However, a key feature of hierarchical models is that information about some parameters (the random effects) is tied to information about others (the random variance terms). The same is true for Bayesian models with informative priors, where information about a given parameter is contained within the hyperparameters and form of the prior distribution. For example, if it is

[1] Much of the summaries provided in this section are inspired from Gelman et al., 2014, with minor changes made to the equations for consistency throughout the present paper.

assumed a priori that a given parameter is distributed according to a Gamma distribution, then information about its value—e.g., that is positive—is contained within the prior. As such, in models with a hierarchical structure and/or models that are influenced by informative priors, the number of effective parameters will be less than the number of total parameters, k.

This latter insight is the key improvement the DIC makes over the AIC. That is, the DIC makes two changes to Eq. 1: 1) $\hat{\theta}_{MLE}$ is replace by $\hat{\theta}_{MAP}$, which is the mean posterior value, rather than the MLE, and 2) $2k$ is replaced by a data-driven bias correction term, p_D. Using an information-theoretic approach, Spiegelhalter et al. (2002) showed that p_D could be computed from an estimated posterior as:

$$p_D = 2\left(\ln p\left(D|\hat{\theta}_{MAP}\right) - \frac{1}{S}\sum_{j=1}^{S}\ln p\left(D|\theta_j\right)\right) \quad \text{Eq. 2}$$

where S is the number of samples in the posterior and θ_j is the vector of parameters at a given MCMC sample. As Spiegelhalter et al. (2002) put it, p_D is the "mean deviance minus the deviance of the means"—that is, it captures the degree to which the likelihood of the data under $\hat{\theta}_{MAP}$ deviates from the average likelihood of the data under all posterior draws. If this value is large, there is a large discrepancy between the average likelihood and the likelihood of the data under $\hat{\theta}_{MAP}$, which may suggest overfitting, and the penalty is thus larger. Conversely, if p_D is small, it suggests that the posterior is precisely centred around $\hat{\theta}_{MAP}$ and thus the number of effective parameters is small, as is the penalty. To compute the DIC, p_D is substituted for $2k$ in Eq. 1 and $\hat{\theta}_{MLE}$ is replaced with $\hat{\theta}_{MAP}$:

$$\text{DIC} = -2\ln p\left(D|\hat{\theta}_{MAP}\right) + 2p_D \quad \text{Eq. 3}$$

Apart from criticisms about its efficacy as a good metric for model selection (Ando, 2011), the DIC has also been criticized as not being "fully Bayesian", because its estimate of

likelihood relies solely on a point estimate, $\hat{\theta}_{MAP}$, rather than pooling information across the entire posterior, which is to many Bayesians a key advantage of the Bayesian framework. As we will see next, the WAIC is formulated precisely to address this concern.

Watanabe-Akaike information criterion (WAIC)

The WAIC is a more "fully-Bayesian" approach to estimating out-of-sample predictive accuracy, because it computes model fit using the pointwise posterior predictive density, rather than a point estimate of predictive likelihood, as in the case of the DIC. By pointwise posterior predictive density, $lppd$, I mean that the WAIC evaluates the (log) probability of each data point $D_i, i = 1, \dots, N$, where N is the total number of data points in the sample, at each $\hat{\theta}_j, j = 1, \dots S$, where S is the total number of posterior draws, such that:

$$lppd = \sum_{i=1}^{N} \ln \frac{1}{S} \sum_{j=1}^{S} p(D_i|\theta_j) \quad \text{Eq. 4}$$

In this sense, the WAIC evaluates the likelihood of each data point using information from the entire posterior distribution, rather than only its central tendency. Similarly, the correction term applied to the WAIC, p_W, mirrors that of the DIC, except that it is evaluated pointwise, for each data point, D_i in the sample, as:

$$p_W = 2 \sum_{i=1}^{N} \left(\ln \frac{1}{S} \sum_{j=1}^{S} p(D_i|\theta_j) - \frac{1}{S} \sum_{j=1}^{S} \ln p(D_i|\theta_j) \right) \quad \text{Eq. 4}$$

These terms are then combined and multiplied by -2 in order to put the WAIC on a deviance scale, rendering it comparable to AIC and DIC:

$$\text{WAIC} = -2(llpd - p_W) \quad \text{Eq. 5}$$

Leave-One-Out Information Criterion (LOOIC)

The goal of all information criteria is to estimate the predictive accuracy of a model in a new sample—that is, when applied to new data, what is the expected likelihood of the new data under this model? A practical way to do this would be to divide the data into two sets, a train and a test set, estimate the model on the train set, and predict the data in the test set, noting the discrepancy directly.

Leave-one-out cross-validation (LOO-CV) implements a version of this procedure. For a given data point, D_i, LOO-CV will train the model on all other data points in the sample, D_{i-1}, and the predictive likelihood will be computed on D_i. This is done iteratively for all $i, .., N$, and for all posterior samples, $j, \dots, S$, as:

$$\text{LOO-CV} = \sum_{i=1}^{N} \ln \frac{1}{S} \sum_{j=1}^{S} p\left(D_i | D_{i-1}, \theta_j\right) \quad \text{Eq. 6}$$

However, actually computing $p\left(D_i \middle| D_{i-1}, \theta_j\right)$ in this way can be very computationally intensive for data with many observations and is not appropriate when observations are not independent, as is the case in hierarchical models. Accordingly, Vehtari et al. (2017) proposed a method for approximating LOO-CV from the observed posterior draws, using an important sampling approach. The details of how this procedure is done is outside of the scope of this paper, and elaborated on by Vehtari and colleagues (2017), but in short, an importance score for the left-out data point i (for a given posterior draw, j) can be approximated as $r_{i,j} = \frac{1}{p\left(D_i | \theta_j\right)}$, which can be used to evaluate the log predictive density at the left-out data point, as:

$$p\left(D_i \middle| D_{i-1}, \theta_j\right) \approx \frac{1}{\frac{1}{S} \sum_{j=1}^{S} \frac{1}{p\left(D_i | \theta_j\right)}} \quad \text{Eq. 7}$$

More specifically, the value produced by Eq. 7, and specifically $r_{i,j}$, is unstable and requires some refinement to properly estimate the predictive density—namely, $p\left(D_i \middle| D_{i-1}, \theta_j\right)$ must be weighted by a vector of weights, $w_{i,j}$ obtained through Pareto smoothed importance sampling (Vehtari & Gelman, 2015). Again, details are presented in Vehtari et al. (2017), but once these weights are calculated, the LOO expected log pointwise predictive density is computed as:

$$elpd = \sum_{i=1}^{N} \ln\left(\frac{\sum_{j=1}^{S} w_{i,j} p\left(D_i \middle| \theta_j\right)}{\sum_{j=1}^{S} w_{i,j}}\right) \quad \text{Eq. 7}$$

which when multiplied by -2 yields the LOO-IC on the deviance scale, which can be compared to the DIC and WAIC presented above:

$$\text{LOO-IC} = -2elpd \quad \text{Eq. 7}$$

Method

To test which of these information criteria is best able to select the generative model among a set of candidate models, and moreover how this discriminability is influenced by level-1 sample size, level-2 sample size, the degree of clustering, and the nature of model misspecification, I simulated data from a known multilevel linear model, estimated four models at various levels of misspecification using MCMC sampling as implemented in *brms*, and computed the information criteria as described above for each model.

True Data Simulation

Data were sample from a true generative model with known parameters, which had the following form:

$$y_{ij} = \gamma_{00} + U_{0j} + (\gamma_{10} + U_{1j})X_{1_{ij}} + \gamma_{20}X_{2_{ij}} + R_{ij} \quad \text{Eq. 8}$$

where, y_{ij} is an observation i of a continuous outcome variable for cluster j. γ_{00} is the fixed intercept (i.e., the grand mean value across clusters and observations) and U_{0j} is the cluster-level deviation from γ_{00}. γ_{10} is the fixed slope for X_1, which is itself a standard normal random continuous variable, and U_{1j} is the cluster-level deviation from γ_{10}. γ_{20} is the fixed slope for X_2, which is normally distributed random variable, $X_2 \sim N(2,1)$. U_{0j} and U_{1j} were sampled from a multivariate normal distribution, $U_{kj} \sim MVN(0,\Sigma)$, where Σ is a covariance matrix of random variance terms for the intercept and slopes of X_1, τ_0^2 and τ_1^2 respectively, and the covariance between these terms was set to 0, such that $\Sigma = \begin{bmatrix} \tau_0^2 & 0 \\ 0 & \tau_1^2 \end{bmatrix}$. γ_{20} was assumed to hold for all clusters, such that $\tau_2^2 = 0$. Finally, R is the residual error, distributed as $R \sim N(0,\sigma^2)$, where σ^2 is the residual variance. Each parameter's population value was fixed (except for τ_0^2) and is displayed in Table 1.

To simulate data from this model, a hand-coded function was used (see https://github.com/seandamiandevine/PSYC749_Final/blob/main/fx.R). To ensure that this simulation code was well-implemented, a small parameter recovery study was conducted, assessing whether a given set of input parameters could be recovered from simulated data using *lme4* (Bates et al., 2015). The results of this procedure are visualized Figure S1 in the Supplemental Materials, where it is shown that parameters were well-recovered.

Comparison Models

To assess the performance of DIC, WAIC, and LOO-IC, I fit four models with different misspecifications to data generated from Eq. 8, as well as a model that has the form presented in Eq. 8 (the "true" model, Model E, not shown below). Using the same notation as described in the previous section, these models had the following form:

Model A: $$y_{ij} = \gamma_{00} + U_{0j} + \gamma_{10}X_{1_{ij}} + \gamma_{20}X_{2_{ij}} + R_{ij}$$ Eq. 9

Model B: $$y_{ij} = \gamma_{00} + U_{0j} + (\gamma_{10} + U_{1j})X_{1_{ij}} + R_{ij}$$ Eq. 10

Model C: $$y_{ij} = \gamma_{00} + U_{0j} + (\gamma_{10} + U_{1j})X_{1_{ij}} + (\gamma_{20} + U_{2j})X_{2_{ij}} + R_{ij}$$ Eq. 11

Model D: $$y_{ij} = \gamma_{00} + U_{0j} + \gamma_{10}X_{1_{ij}} + \gamma_{20}X_{2_{ij}} + \gamma_{30}X_{3_{ij}} + R_{ij}$$ Eq. 12

In words, Model A wrongly assumes that $\tau_1^2 = 0$, Model B wrongly assumes $\gamma_{20} = 0$, Model C wrongly assumes $\tau_2^2 > 0$, and Model D assumes that both $\tau_1^2 = 0$ and γ_{30}, the fixed slope for a new predictor X_3, which is distributed as $X_3 \sim N(0,1.41)$, is not zero. To be sure, these models do not cover the full space that misspecification can take (a point I will return to in the Discussion), but they represent how common models in the literature might be misspecified relative to the generative model, by, for instance, ignoring important cluster-level variable (Model A and D), ignoring predictors that may be important (Model B and D), and/or overfitting random effects' structures in a "keep it maximal" approach (Barr et al., 2013; Model C).

Simulation Design Matrix

The models above were fit to simulated data using Eq. 8. Three features of these data were changed iteratively across simulations. First, level-1 sample size—that is, the number of clusters in the dataset—took on values of 10, 50, or 100. Second, level-2 sample size—that is, the number of observations per cluster—took on values of 10, 50, or 100. Third, the degree of

clustering[2], operationalized as the value of τ_0^2, took on values of 0.05, 1, or 5. All together then, the simulation design matrix was a $3 \times 3 \times 3$ design, or 27 total cells. One-hundred simulations were conducted at each cell, yielding a total of 2700 simulations for this study.

Model Fitting and Computation of the Information Criteria

Models A-E were estimated using *brms*. One chain of 3000 draws were made from the posterior, with 1000 draws acting as burn-in and no thinning applied. Only one chain was used due to computational and time restrictions (see Discussion section).

brms does not offer an in-built way to compute DIC from a fitted model's posterior, nor could I find any package that did so for a *brms* object in R. Accordingly, I programmed my own function to compute the DIC, as specified in Eq. 2 and 3 (see https://github.com/seandamiandevine/PSYC749_Final/blob/main/fx.R). Conversely, functions to compute both WAIC and LOO-IC are included in *brms* by default (*WAIC()* and *LOO()* respectively), and these functions were used to compute these values for the current study.

In summary, at each of 2700 iterations, five models (A-E) were fit using *brms* to randomly sampled data from Eq. 8 that varied in terms of level-1 sample size, level-2 sample size, and the degree of clustering. DIC, WAIC, and LOO-IC values were computed from each model at each iteration and were stored in an outcome matrix for analysis.

The selected model was defined as the model with the lowest value on each criterion and accuracy was assessed as the proportions of samples where the generative model (E) was selected. Performance was then modeled using a logistic regression, predicting model selection

---

[2] The degree of clustering can be computed by the ICC as ICC $= \frac{\tau_0^2}{\tau_0^2 + \sigma^2}$. Thus, holding σ^2 constant, the degree of clustering depends on τ_0^2, where large values indicate strong clustering and low values indicate weak clustering.

accuracy from level-1 sample size (centred at 50), level-2 sample size (centred at 50), and random intercept variance (centred at 1), the results of which are summarized in Table 2. Coefficients are reported on the log-odds scale and confidence intervals (CI) are at the 95% level.

Results

Model Selection Accuracy

Information criteria performance is summarized in Figures 1-3. In line with past findings (Ando, 2011; Vehtari et al., 2017), DIC performed the worse of the three, correctly selecting model E in 57% of samples—notably still well above chance-level (here, 20%; $p < .0001$). WAIC correctly selected the generative model in 68% of samples and LOO-IC ($p < .0001$) in 71% of samples ($p < .0001$)

DIC. Interestingly, level-1 sample size exerted a significant negative effect on DIC accuracy ($b = -0.003$, CI = [-0.0054 – -0.0008], $p = .0085$), such that DIC failed to identify model E as the best fitting model more often when there was a large number of groups (Figure 1). Additionally, I observed a small, but significant, interaction between level-2 sample size and the degree of clustering, such that at high levels of clustering, higher level-2 sample size improved DIC performance, but as the degree of clustering became smaller, this improvement was lost ($b = 0.01$, CI = [0.0001 – 0.002], $p = .0389$). No other effects achieved statistical significance (see Table 2).

WAIC. As seen in Figure 2, larger level-2 ($b = 0.06$, CI = [0.004 – 0.009], $p < .0001$), but not level-1 ($p = .1340$), sample sizes improved WAIC's discriminative accuracy on average. This was not only the case overall, but as both the number of groups increased (level-1 sample

size got larger; b = 0.0001, CI = [0.000 – 0.0002], p = .0106), and with more clustering (b = 0.002, CI = [0.0009 – 0.0031], p = .0003), the benefit of larger level-2 sample sizes increased significantly. Finally, I observed a three-way interaction between level-1 sample size, level-2 sample size, and random intercept variance magnitude (b = 0.00003, CI = [-0.0001, -0.0000], p = .0186), such that the aforementioned benefits of larger level-2 sample sizes at larger level-1 sample sizes were strongest at high levels of clustering. However, this effect was very small and, in a design with only three levels per condition per predictor, should be interpreted with caution.

LOO-IC. Again, level-2 sample size positively predicted LOO-IC's model selection accuracy (b = 0.005, CI = [0.0025 – 0.0077], p < .0001; Figure 3). This effect was more pronounced at higher levels of clustering (b = 0.001, CI = [0.0003 – 0.0025], p = .0122). As for the WAIC, a significant three-way interaction was observed (b = 0.00003, CI = [-0.0001, -0.0000], p = .0186), such that the benefits of larger level-2 sample sizes at larger level-1 sample sizes were strongest at high levels of clustering.

Influence of Model Misspecification

The above analysis tested the influence of level-1 sample size, level-2 sample size, and degree of clustering on model selection accuracy for DIC, WAIC, and LOO-IC. None of these metrics were perfectly identified the correct model in all cases. Thus, the question remains: *which* models were selected?

Figures 4-6 visualize the relative selection rates of each model (see Eq. 9-12) as a function of each predictor value for each information criteria. As can be clearly seen from these figures, apart from at low level-1 and level-2 sample sizes, all three information criteria consistently chose either model E (the generative model) or model C, which overfits the data by

assuming that $\tau_2^2 > 0$. This was confirmed by an inspection of the cumulative raw scores for each criterion, which (unsurprisingly) reveal the same pattern: DIC, WAIC, and LOO-CV for models C and E are comparably low, relative to the other models under consideration (Supplemental Figures 2-4).

Discussion

The present simulation study examined the performance of three popular information criteria for Bayesian multilevel model selection—DIC, WAIC, and LOO-IC—for selecting among five models, four of which suffered from various misspecifications, across datasets with different numbers of groups (level-1 sample size), observations (level-2 sample size), and where the magnitude of clustering differed (random intercept variance).

Taken together, these results reveal two important features common to all three criteria under considerations. First, level-2 sample size—that is, the number of observations within each group—had the greatest impact on model selection accuracy, where the degree of clustering seems to have mainly accentuated or attenuated the benefits of larger level-2 sample size. This is of practical importance, because researchers often aim to maximize power by either increasing level-1 or level-2 sample size—a decision that comes with various financial (e.g., "how many participants can I recruit for my experiment?") and temporal (e.g., "how many trials can I administer to a single participant?") considerations. If the aim of a given study is to disambiguate between two competing hypotheses using DIC, WAIC, or LOO-IC as metrics, this finding may highlight the importance of maximizing level-2 sample size where possible.

Second, though all three criteria most often correctly identified model E as the generative model, model C was identified as the winning model in a substantial proportion of samples.

Recall, model C contained a random slope variance term for predictor X_2, which, under the generative model, did not vary per cluster. Furthermore, this proportion of incorrectly identified subsamples was not insignificant, resting between 20-30% of samples depending on the condition and criteria used. If we are to take the traditional 5% error rate as acceptable, this is a worrying number of incorrectly identified samples, especially since best practices currently suggest fitting models with maximal random effects structures (Barr et al., 2013), which, at least in the present case, would result in a Type I error (claiming $\tau_3^2 > 0$ when $\tau_3^2 = 0$) at a non-negligible frequency.

Limitations and Future Directions

While earlier, I described the study as consisting of 2700 iterations, it is more accurate to account for the sub-iterations incurred from the need to use MCMC sampling across 5 models. Together then, this study was a 3 (level-1 sample sizes) × 3 (level-2 sample sizes) × 3 (random variance intercepts) × 3000 (MCMC samples) × 5 (models) design, yielding a total of 405 000 iterations. Accordingly, the scope of the present study was limited by time and computational restrictions.

With a longer time-horizon and more computational resources (as well as perhaps more efficient coding), there are a number of changes and extensions I believe would prove insightful for better understanding Bayesian multilevel model selection criteria. I describe three such modifications below, in order of feasibility and conceptual distance from the current study.

Small changes. In the current design, criteria performance was only assessed at three levels of each predictor. The logic behind this decision—beyond limiting computational costs—was to gain a "low, medium, high" interpretation of the results. However, this is suboptimal and a simple change that could be made to the current design would be to increase the number

conditions. As mentioned above, this could be done by improving the efficiency of the simulation code by, for instance, parallelizing processes and splitting simulation across multiple machines.

Moderate changes. Despite the focus of this study being Bayesian models, the effect of prior information on model selection was not examined, and default (uninformative) priors were used for all models. This is an important addition that could be made, as changes in priors can bias not only parameter estimation, but influence the effective number of parameters, perhaps exacerbating performance differences between information criteria. For instance, since the DIC relies only on a point summary of the posterior, $\hat{\theta}_{MAP}$, we might expect that strong priors would reduce its precision, thus increasing its penalty term, p_D. This should be especially true at smaller sample sizes, where the prior holds more weight. To my knowledge, the influence of prior form and parameterization on model selection using DIC, WAIC, or LOO-CV has not been previously investigated, and as such represents a clear avenue for future research.

Larger changes. Beyond basic linear multilevel regressions, a host of computational "process" models have taken advantage of hierarchical Bayesian techniques in recent years (e.g., Wiecki and colleagues' (2013) *HDDM* package for fitting hierarchical Bayesian drift diffusion models). Owing in part to the novelty of these methods, virtually no work exists which assesses different metrics of model fit and model selection for this non-linear class of process model. However, this is a critically important area of research, as parameters from these models are often given substantive weight in scientific inference as reflecting the underlying mechanisms of human cognition. Without an investigation of the means used to disambiguate one hypothesis form another in these models, some of these inferences may be faulty. The simulation approach employed here could be easily extended to address this issue.

References

Ando, T. (2011). Predictive Bayesian model selection. *American Journal of Mathematical and Management Sciences*, *31*(1-2), 13-38. https://doi.org/10.1080/01966324.2011.10737798

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language*, *68*(3), 255-278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.

Bürkner, P. C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of statistical software, 80*, 1-28. https://doi.org/10.18637/jss.v080.i01

Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and computing*, *24*(6), 997-1016. https://doi.org/10.1007/s11222-013-9416-2

Huang, F. L. (2018). Multilevel modeling myths. *School Psychology Quarterly*, *33*(3), 492. https://doi.org/10.1037/spq0000272

Luo, W., Li, H., Baek, E., Chen, S., Lam, K. H., & Semma, B. (2021). Reporting practice in multilevel modeling: A revisit after 10 years. *Review of Educational Research*, *91*(3), 311-355. https://doi.org/10.3102/0034654321991229

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the royal statistical society: Series b (statistical methodology)*, *64*(4), 583-639. https://doi.org/10.1111/1467-9868.00353

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and computing*, *27*(5), 1413-1432. https://doi.org/10.1007/s11222-016-9696-4

Vehtari, A., Simpson, D., Gelman, A., Yao, Y., & Gabry, J. (2015). Pareto smoothed importance sampling. *arXiv preprint arXiv:1507.02646*.

Watanabe, S., & Opper, M. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of machine learning research*, *11*(12).

Wiecki, T. V., Sofer, I., & Frank, M. J. (2013). HDDM: Hierarchical Bayesian estimation of the drift-diffusion model in Python. *Frontiers in neuroinformatics*, 14. https://doi.org/10.3389/fninf.2013.00014

Tables

Table 1. Population parameter values for true generative model, described in Eq. 8

Parameter	Population Value
γ_{00}	1.10
γ_{10}	1.66
γ_{20}	1.20
τ_1^2	0.75
τ_2^2	0.00
σ^2	3.00

Table 2. Results from logistic regression predicting model selection accuracy.

	b	*SE*	*p*
	DIC		
Intercept	3.21E-01	4.37E-02	1.99E-13
N	-3.10E-03	1.18E-03	0.00852
J	-1.77E-03	1.18E-03	0.13303
τ_0^2	-9.39E-03	1.84E-02	0.60957
$N \times J$	1.71E-05	3.19E-05	0.59045
$N \times \tau_0^2$	-4.58E-04	4.97E-04	0.35633
$J \times \tau_0^2$	1.03E-03	4.98E-04	0.03892
$N \times J \times \tau_0^2$	-3.04E-06	1.34E-05	0.82128
	WAIC		
Intercept	7.98E-01	4.68E-02	< 2e-16
N	-1.90E-03	1.26E-03	0.133981
J	6.35E-03	1.28E-03	7.49E-07
τ_0^2	-1.17E-02	2.00E-02	0.55786
$N \times J$	8.89E-05	3.48E-05	0.010641
$N \times \tau_0^2$	-1.03E-03	5.37E-04	0.054038
$J \times \tau_0^2$	2.01E-03	5.59E-04	0.000323
$N \times J \times \tau_0^2$	-3.52E-05	1.50E-05	0.018638
	LOO-IC		
Intercept	9.09E-01	4.76E-02	< 2e-16
N	-7.85E-04	1.29E-03	0.5418
J	5.09E-03	1.30E-03	9.25E-05
τ_0^2	1.02E-02	2.05E-02	0.6185
$N \times J$	6.15E-05	3.53E-05	0.0819
$N \times \tau_0^2$	-9.07E-04	5.48E-04	0.098
$J \times \tau_0^2$	1.42E-03	5.68E-04	0.0122
$N \times J \times \tau_0^2$	-3.81E-05	1.52E-05	0.0121

Note: N refers to level-1 sample size, J refers to level-2 sample size

Figures

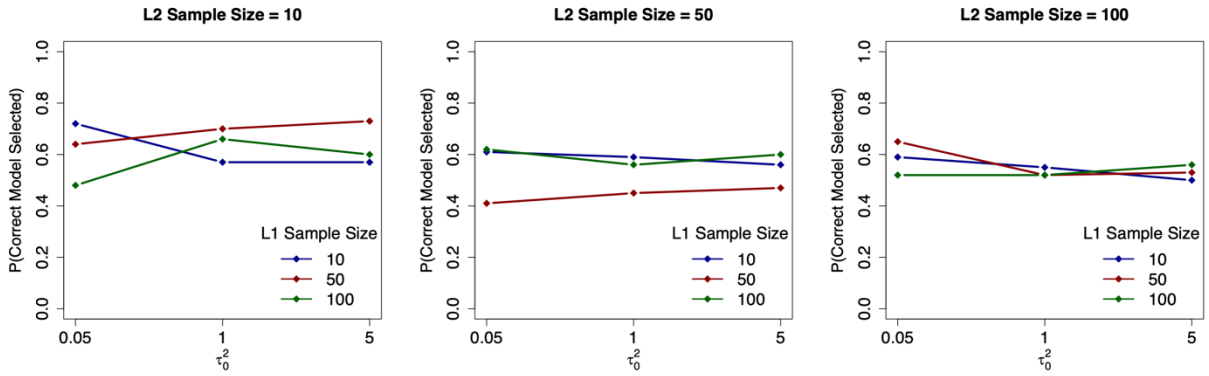

Figure 1. Model selection accuracy for DIC. The y-axis shows the probability of correctly identifying model E as having the lowest DIC value. The x-axis represents the magnitude of the random intercept variance, τ_0^2. Colours represent level-1 sample size. Panels represent level-2 sample size, as indicated by the title.

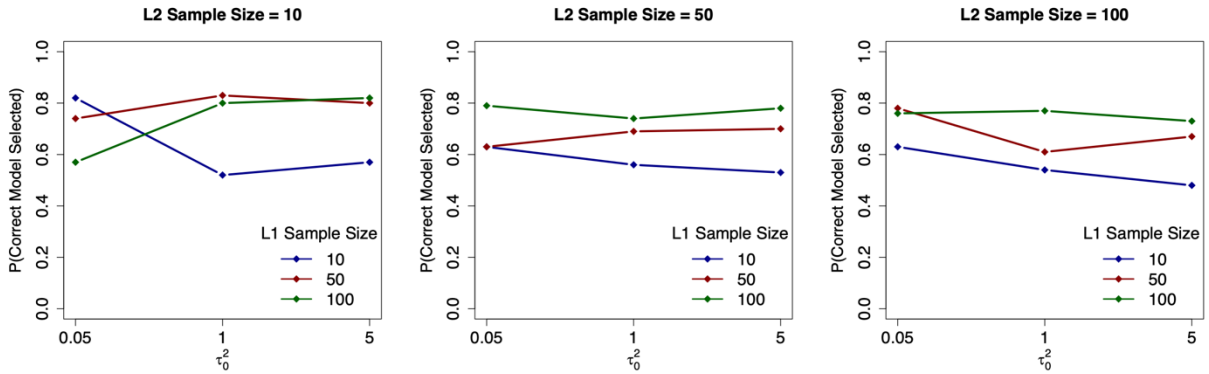

Figure 2. Model selection accuracy for WAIC. The y-axis shows the probability of correctly identifying model E as having the lowest WAIC value. The x-axis represents the magnitude of the random intercept variance, τ_0^2. Colours represent level-1 sample size. Panels represent level-2 sample size, as indicated by the title.

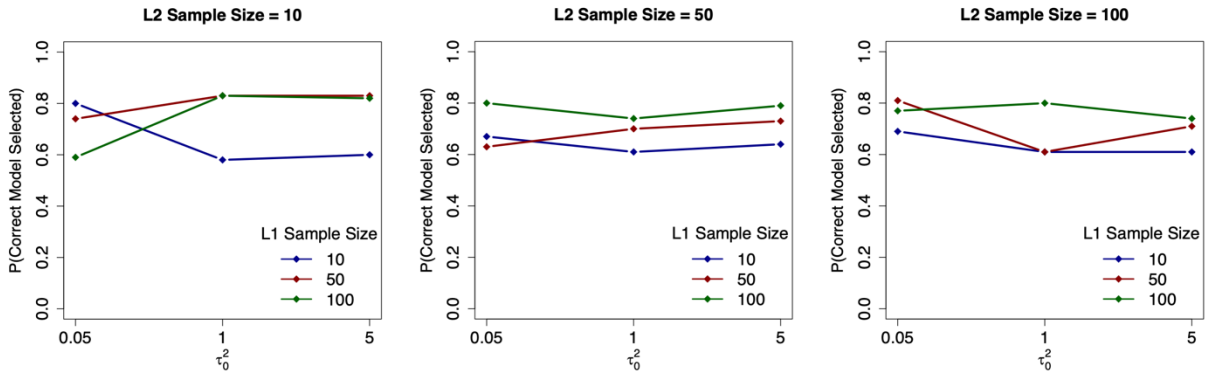

Figure 3. Model selection accuracy for LOO-IC. The y-axis shows the probability of correctly identifying model E as having the lowest LOO-IC value. The x-axis represents the magnitude of the random intercept variance, τ_0^2. Colours represent level-1 sample size. Panels represent level-2 sample size, as indicated by the title.

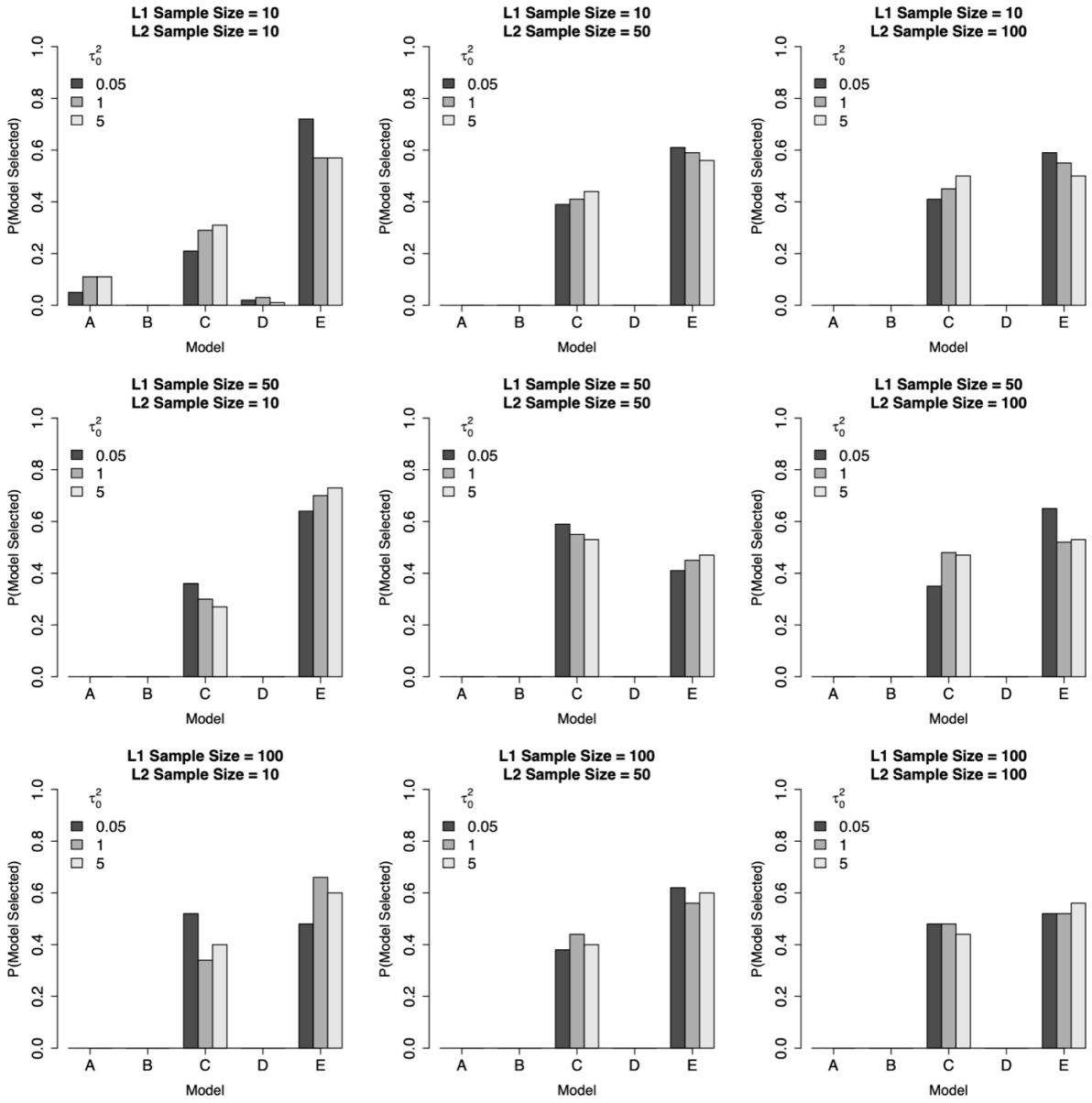

Figure 4. Relative model selection for DIC. The y-axis shows the relative proportion of samples where a given model—shown on the x-axis (see Eq. 9-12)—was selected. Bar colours represent random intercept variance, τ_0^2. Panels represent configurations of level-1 and level-2 sample sizes, as indicated by panel titles.

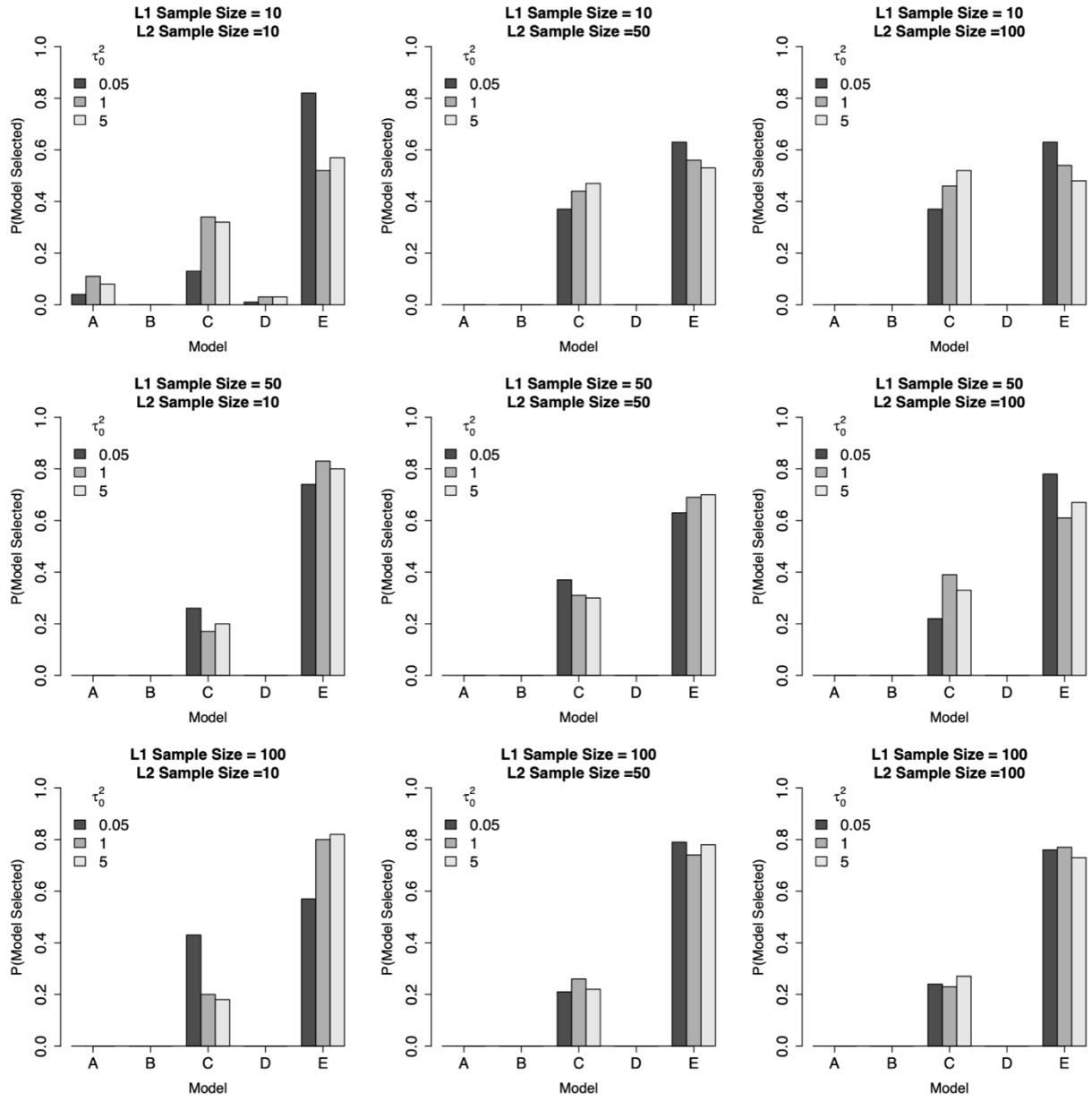

Figure 5. Relative model selection for WAIC. The y-axis shows the relative proportion of samples where a given model—shown on the x-axis (see Eq. 9-12)—was selected. Bar colours represent random intercept variance, τ_0^2. Panels represent configurations of level-1 and level-2 sample sizes, as indicated by panel titles.

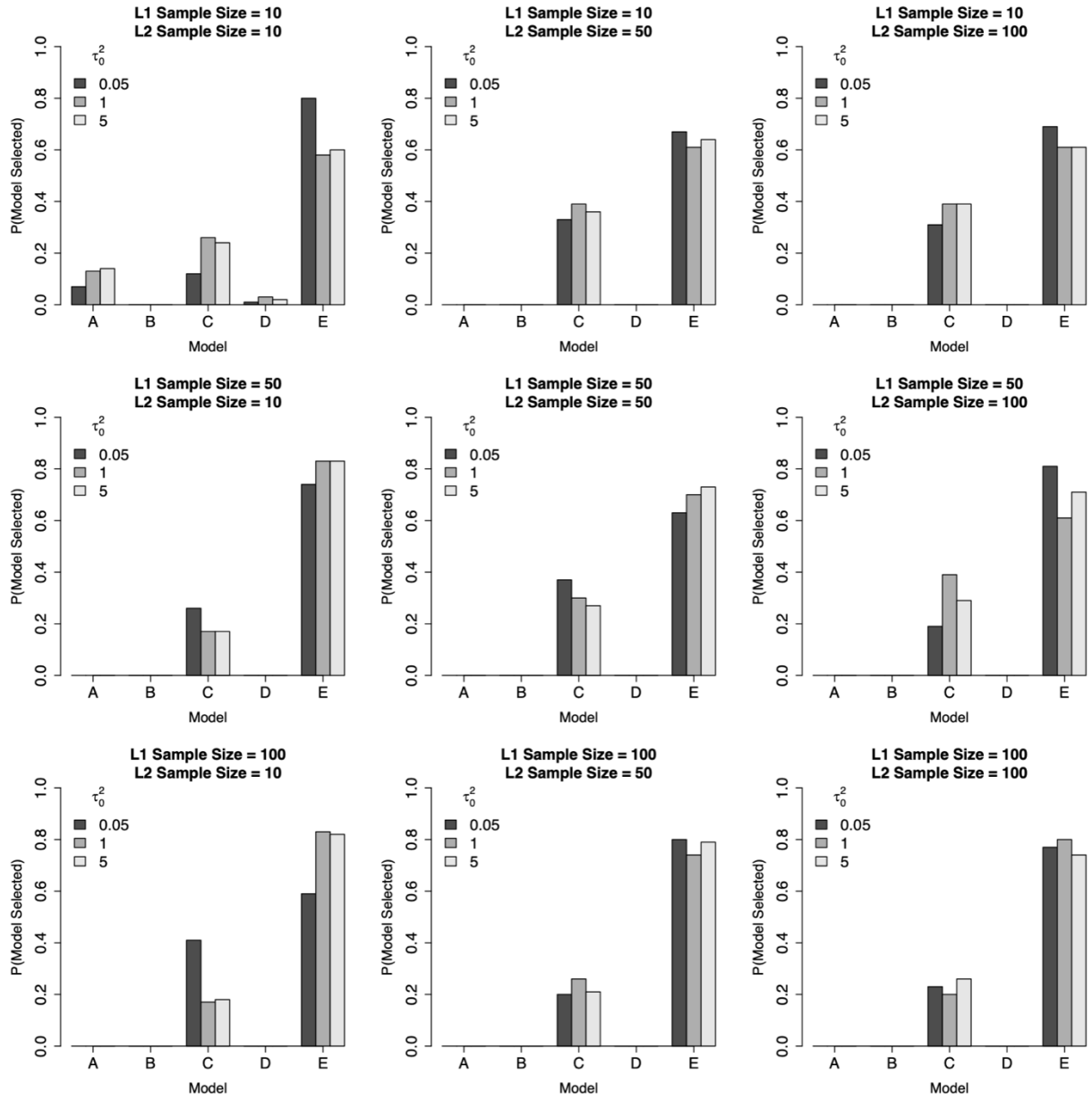

Figure 6. Relative model selection for DIC. The y-axis shows the relative proportion of samples where a given model—shown on the x-axis (see Eq. 9-12)—was selected. Bar colours represent random intercept variance, τ_0^2. Panels represent configurations of level-1 and level-2 sample sizes, as indicated by panel titles.

Supplemental Materials

Parameter Recovery

To ensure that my hand-coded simulation code produced data that corresponded to the desired parameters, I ran a very small parameter recovery study. Here, 100 samples of 250 observations from 100 groups was generated according to the model specified in Eq. 8 (model E). $\hat{\theta}_{MLE}$ were then estimated using *lme4* and estimated parameters were compared to true known parameters, which were randomly sampled each iteration. Fixed effects (γ) were sampled from standard normal distributions, $N(0,1)$, and random variance terms (τ) were sampled from a Gamma distribution, $G(2,2)$. Covariance between random variance terms were set to 0 for each iteration.

As can be seen in Figure S1, recovery was very good. Poorest recovery was seen for extreme random variance terms, as is common in multilevel estimation, but even so, correlations between true and estimated effects were no less than 0.95. These effects confirm that the simulation function I programmed properly produced data with the desired coefficients and was thus suitable for use in the larger-scale simulation study.

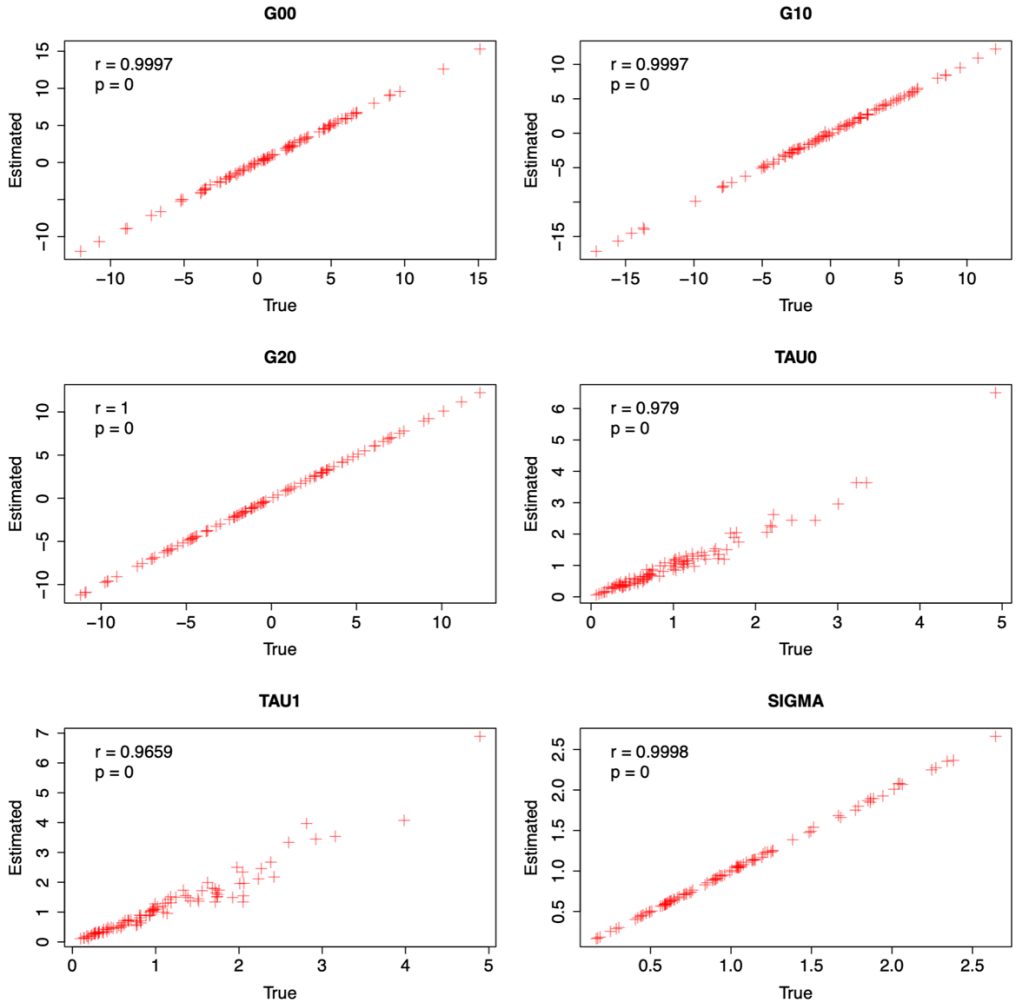

Figure S1. Parameter recovery results. Each panel shows a parameter under consideration, as indicated by the panel's title. The x-axis shows the true parameter value, and the y-axis shows the estimated parameter. Each cross represents a true, estimated parameter pair for a single iteration. Pearson's *r* and associated p-values are indicated in the top-left corner of each panel.

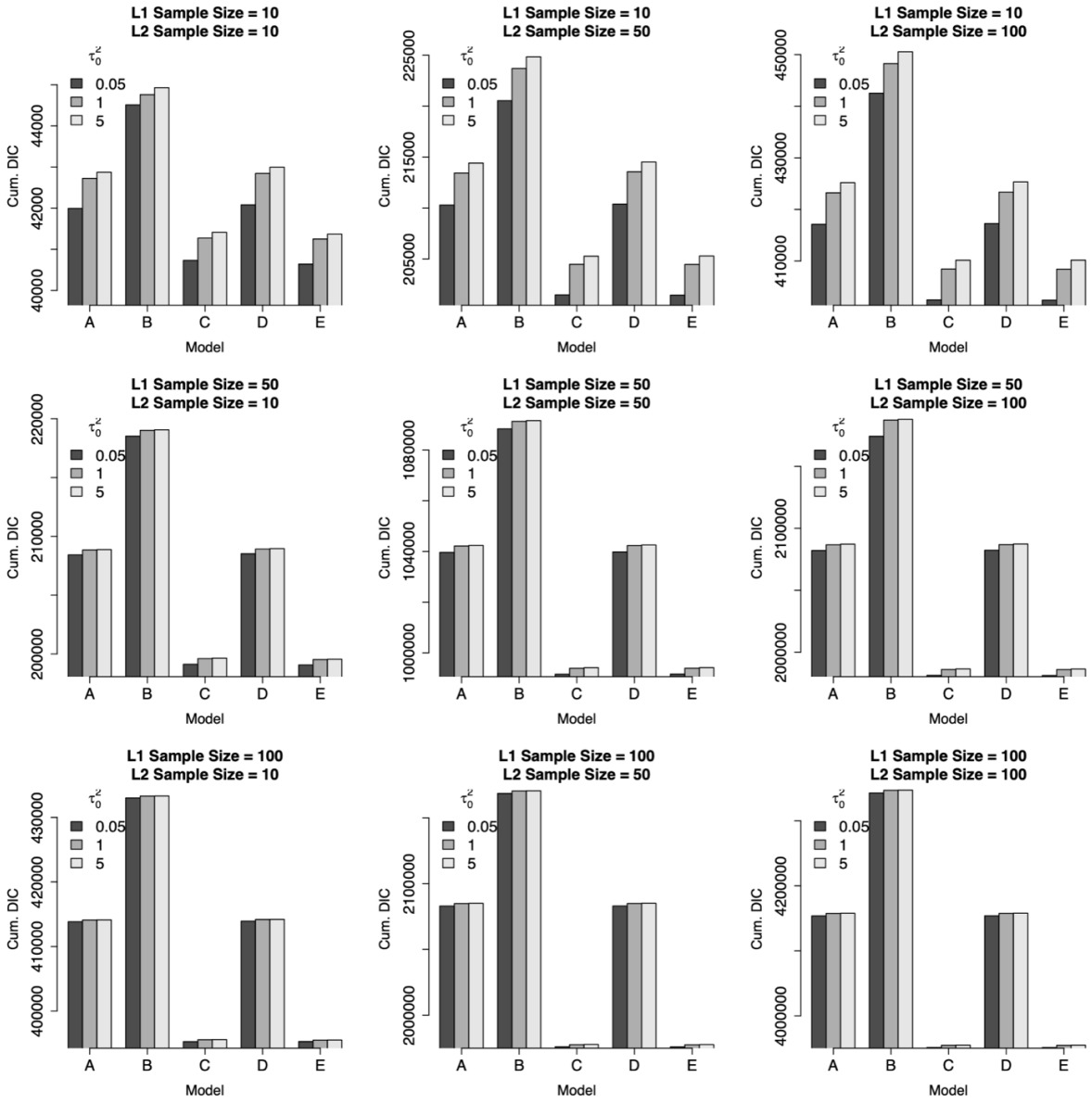

Figure S2. Cumulative DIC values. The y-axis shows the cumulative DIC values for a given model—shown on the x-axis (see Eq. 9-12). Bar colours represent random intercept variance, τ_0^2. Panels represent configurations of level-1 and level-2 sample sizes, as indicated by panel titles.

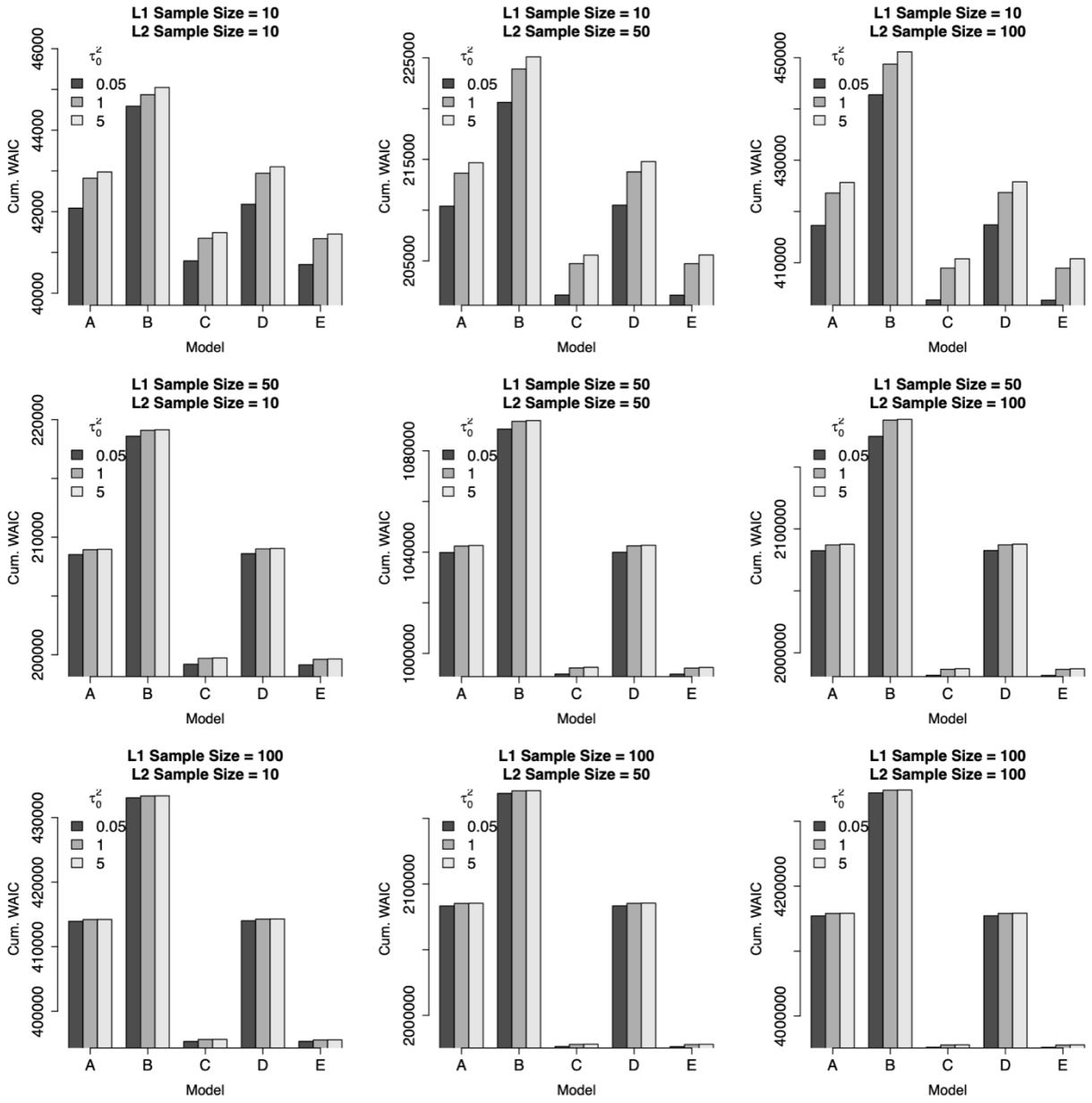

Figure S3. Cumulative WAIC values. The y-axis shows the cumulative WAIC values for a given model—shown on the x-axis (see Eq. 9-12). Bar colours represent random intercept variance, τ_0^2. Panels represent configurations of level-1 and level-2 sample sizes, as indicated by panel titles.

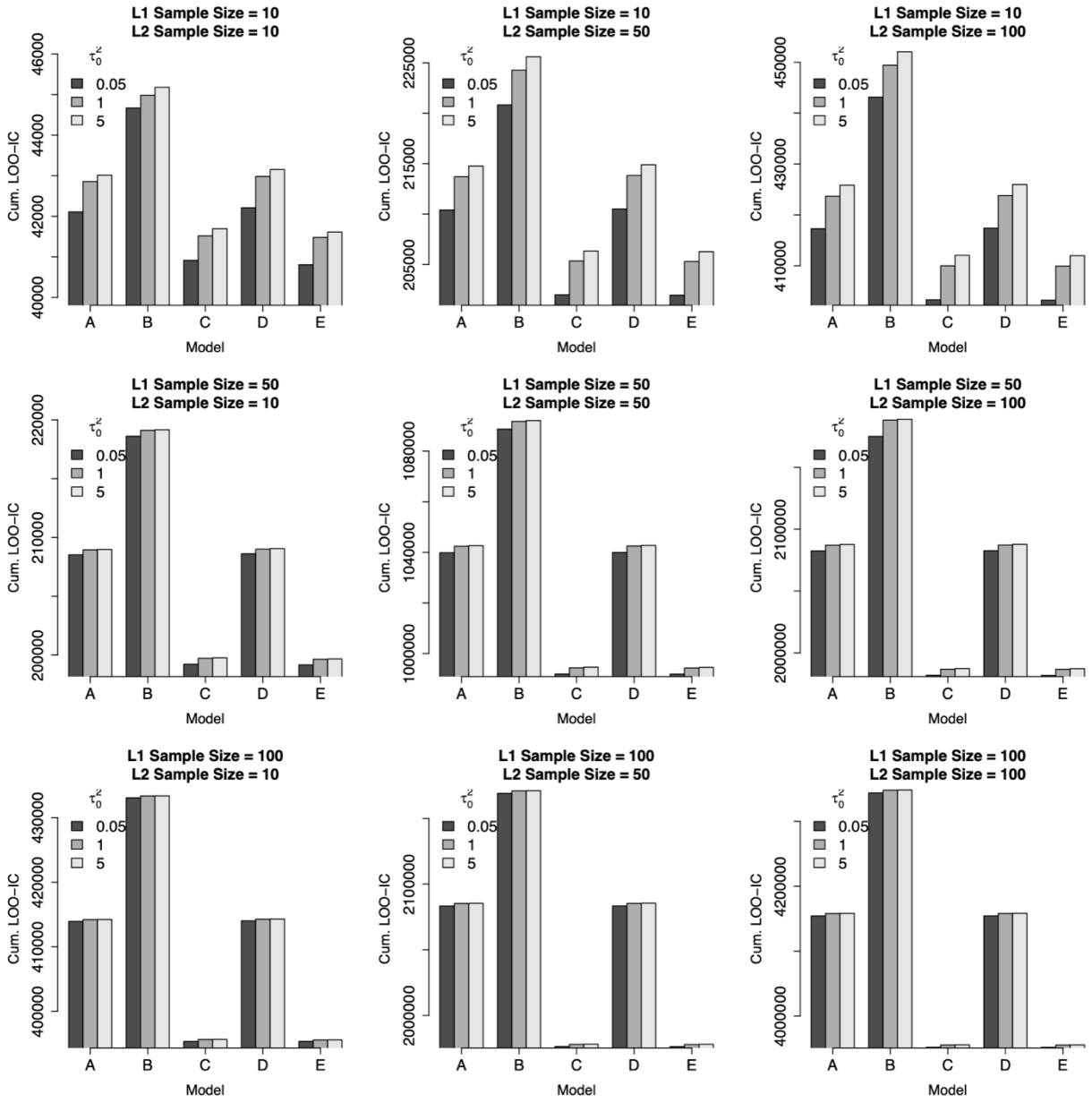

Figure S4. Cumulative LOO-IC values. The y-axis shows the cumulative LOO-IC values for a given model—shown on the x-axis (see Eq. 9-12). Bar colours represent random intercept variance, τ_0^2. Panels represent configurations of level-1 and level-2 sample sizes, as indicated by panel titles.