Behavior Research Methods (2019) 51:440–450
https://doi.org/10.3758/s13428-018-1188-3

Value of sample size for computation of the Bayesian information criterion (BIC) in multilevel modeling

Julie Lorah[1] · Andrew Womack[2]

Published online: 25 January 2019

Abstract

The Bayesian information criterion (BIC) can be useful for model selection within multilevel-modeling studies. However, the formula for the BIC requires a value for sample size, which is unclear in multilevel models, since sample size is observed for at least two levels. In the present study, we used simulated data to evaluate the rate of false positives and the power when the level 1 sample size, the effective sample size, and the level 2 sample size were used as the sample size value, under various levels of sample size and intraclass correlation coefficient values. The results indicated that the appropriate value for sample size depends on the model and test being conducted. On the basis of the scenarios investigated, we recommend using a BIC that has different penalty terms for each level of the model, based on the number of fixed effects at each level and the level-based sample sizes.

Keywords Multilevel modeling · Bayesian information criterion · BIC · Monte Carlo study · Model comparison · Hierarchical linear modeling

The Bayesian information criterion (BIC; Raftery, 1995; Schwarz, 1978) represents a useful measure for the comparison of multilevel models (McCoach & Black, 2008). The BIC offers several advantages over traditional hypothesis-testing procedures, including the abilities to compare nonnested models (McCoach & Black, 2008), make selection among several competing models (Raftery, 1995), and "show that a smaller model is better than a larger model" (Weakliem, 2004, p. 179). The BIC is a simplification of an approximation to minus twice the log of the marginal density of the observed data (Kass & Raftery, 1995). Specifically, the marginal density is approximated by taking the joint distribution of the data and the parameters, dividing by the Laplace approximation of the posterior of the parameters, and evaluating at the posterior mode. The BIC is obtained from this approximation by plugging in the maximum likelihood estimator instead of the posterior mode and dropping terms that are constant as a function of the sample size and model complexity.

The BIC is a function of model deviance (where the deviance is − 2 times the log-likelihood of the maximum-likelihood estimator) and is given as

$$\mathrm{BIC} = \mathrm{deviance} + k \times \ln(N), \tag{1}$$

where k is the number of parameters estimated and N is the sample size (Hox, 2010). The BICs for two or more models can be compared, and the lower BIC indicates a better fit (McCoach & Black, 2008). The difference between the BICs for two models (ΔBIC) can be computed, where guidelines indicate that a difference greater than 10 indicates very strong evidence; 6–10 indicates strong evidence; 2–6 indicates positive evidence; and 0–2 indicates weak evidence for the more complex model (Raftery, 1995). To compare fixed effects using the BIC, full maximum-likelihood estimation (as opposed to restricted maximum-likelihood estimation) should be used (Hox, 2010). The comparison of nested models using BICs as a testing procedure (with, say, $\Delta\mathrm{BIC} > c$ for fixed c as a testing rule) results in a test whose Type I and Type II error rates go to zero asymptotically as the sample size goes to infinity under suitable regularity conditions. The $\ln(N)$ value in the BIC causes the Type I error rate to decrease to zero as N increases. The tradeoff is a decrease in the statistical power of the test (as compared to a test with a fixed Type I error rate), although the power still increases to 1.

Use of the BIC may aid data analysts beginning to explore a dataset, or BIC values may be reported in a final multilevel-

✉ Julie Lorah
jlorah@iu.edu

[1] Department of Counseling and Educational Psychology, Indiana University School of Education, Bloomington, IN, USA

[2] Department of Statistics, Indiana University, Bloomington, IN, USA

modeling analysis. For example, a study examining a music ensemble participation outcome reports the results from three multilevel logistic regression models, including a BIC value for each model (Lorah, Sanders, & Morrison, 2014). The researchers may have preferred to report BIC in order to guard against some of the weaknesses of hypothesis-testing procedures, such as arbitrary selection of *p* values or inflation of the Type I error rate when multiple models are proposed for explaining variation in a dataset (Cohen, 1990). The change in BIC provides a route to the approximate Bayes factors, but it does not contain a multiplicity correction, which is typically handled using the model space prior distribution (Hoeting, Madigan, Raftery, & Volinsky, 1999; Scott & Berger 2010). In classical testing, multiplicity correction is obtained by spreading the significance over multiple tests (such as through a Bonferroni correction). In the Bayesian context, multiplicity correction is best viewed as a penalization of model complexity through the specification of a prior distribution on the space of models. For instance, one can decrease the prior probability of a model as the dimension of the model's parameter increases. When the number of models is small and the sample size is large, the prior over the model space is essentially inconsequential.

The BIC may be used with many models, including the multilevel model, which is a model with a random effect for group membership added. However, a complication arises because the BIC requires that the sample size, *N*, be specified for the computation, and this value is ambiguous (Hox, 2010; McCoach & Black, 2008; Raftery, 1995) because sample size is given at two or more levels. Researchers argue that the value for sample size somehow should be reduced (Raftery, 1995), and in practice, different software packages deal with the issue differently (Hox, 2010; McCoach & Black, 2008). This choice could impact the rate of false positives and the power for testing effects, indicating that if an inappropriate value for sample size is chosen in computation of the BIC, researchers could be more likely to make erroneous claims or to miss important substantive findings.

One possibility for reducing the value of sample size is to use the value for the effective sample size, which can be computed as the total sample size divided by the design effect (Kish, 1965; Snijders & Bosker, 2012). This reduction is explored in the present study and is compared with use of the level 1 sample size (i.e., the total sample size) and the level 2 sample size (i.e., the number of groups), while varying group size and number of groups. Additionally, the intraclass correlation coefficient (ICC), given as

$$\mathrm{ICC} = \tau^2/\left(\tau^2+\sigma^2\right) \tag{2}$$

(Snijders & Bosker, 2012), will be varied in the present study, where τ^2 represents the variance at level 2 and σ^2 represents the variance at level 1. The ICC represents the proportion of variance in the outcome that is explained by group membership (Snijders & Bosker, 2012). Note that as it is used in the present study, this is a *conditional* definition of the ICC. It is defined as the correlation in the outcome variable among the group members in a two-level model conditioned on the fixed-effect design matrix. An alternative definition of the ICC is the *marginal* ICC, which views the design itself as random and incorporates the variance of the design within and between groups into the definition of the ICC.

In this study, we explored different sample size calculations for *N* in the formula for the BIC through a simulation study. In the next section, we discuss the connection between the BIC and hypothesis testing. Then we discuss the theoretical justification for different sample size calculations. In the section after that, we describe the research questions and hypotheses. Next, we describe the simulation study. Finally, we present the results of the simulation study and discuss its implications.

BIC and hypothesis testing

The BIC represents an information criterion approach, and as such, the concept of statistical significance does not strictly apply. Model selection using information criteria is usually achieved by viewing a criterion as a loss function and selecting the model that minimizes the loss. The criterion and action rule are set, and their properties, such as frequentist risk or error rates, can be evaluated. Because it approximates the log marginal, the BIC can be considered a "Bayesian approach to hypothesis testing, model selection, and accounting for model uncertainty" (Raftery, 1995, p. 111). We discuss the connection between the BIC and hypothesis testing, as well as various evaluation schemes, in this section.

To use the BIC for hypothesis testing, it is useful first to conceptualize a null hypothesis test of a fixed effect as the model selection question when there are only two models under consideration and one is nested in the other. The test will consist of failing to reject the smaller model whenever $\mathrm{BIC}_0 - \mathrm{BIC}_1 = \mathrm{BIC} > 0$. The difference in BICs between the smaller and larger models is the difference in log-likelihood minus twice the difference in model dimension times the log of the sample size. Under suitable regularity conditions, the difference in log-likelihood converges in distribution to a chi-square random variable whose number of degrees of the freedom is the difference in the model dimensions. Thus, the BIC-based test produces both Type I and Type II error rates that decrease to zero as the sample size increases. This is in contrast to the test based on the Akaike information criterion, whose decision rule provides a Type I error rate that, for each difference in model dimensions, converges to a nonzero constant as the sample size increases (Yang, 2005).

In the general case, a hypothesis might not be uniquely described by one particular model, but rather by a collection of models. For example, when a researcher wants to perform a null-hypothesis test of a particular fixed effect, there are two model classes under consideration. One model class contains all of the models in which the fixed effect is assumed to be zero, and the other class contains all of the models in which it is not. These classes might be able to be disaggregated into finer hypotheses (e.g., also considering a hypothesis regarding a different parameter), but an entire class of models represents the single hypothesis being tested.

There are multiple ways to make Bayesian decision rules for addressing the hypothesis test in this setting. One approach is to consider class-wide minimization of the BIC—that is, to choose the class that contains the model that achieves the minimal BIC. This is akin to the thresholding decision rule discussed above, though its frequentist properties are more nuanced and are tied to the particular model classes under consideration. A more traditional Bayesian approach would be to choose the class of models that attains the maximum posterior probability. This can be achieved by converting the BICs for the models to approximate posterior probabilities and computing the posterior probability of each model class. Both testing procedures are asymptotically consistent for independent and identically distributed data when the true data-generating process is among the considered models.

The false positive rate (equivalent to the Type I error rate in hypothesis testing) is the likelihood of claiming evidence for an effect that does not exist. Within hypothesis-testing procedures, this rate is defined as α and is typically set at .05 (Cohen, 1992; Hox, 2010). However, since computation of the BIC takes into account the sample size, this rate is expected to vary on the basis of the number of groups and the group size. Specifically, as the sample size increases, the equivalent of a smaller *p* value would be required in order to find evidence for a given effect with the BIC (Raftery, 1995). Furthermore, although at the smallest sample sizes (about 30–50) these corresponding *p* values are slightly higher than .05, for any larger sample size the corresponding *p* value for the BIC is much lower than .05, and this value continues to decrease as sample size increases (see Table 9 of Raftery, 1995).

Power can be considered the likelihood of correctly claiming evidence for a real effect, and typically researchers desire power of approximately .80 or higher (Cohen, 1992; Spybrook, 2008). Increasing either the group size or the number of groups is expected to increase power, although the number of groups has been shown to have a bigger influence on power than does group size (Spybrook, 2008). Furthermore, increases in ICC are related to decreases in power (Spybrook, 2008), because of a decrease in the effective sample size.

Because Bayesian methods typically have both the Type I and Type II error rates going to zero, this results in a tradeoff between reductions in the number of false positives and in statistical power. The cost of decreasing the probability of a false positive to zero is a decrease in statistical power. This can be seen intuitively by once again considering a simple null-hypothesis test with two nested models. When the smaller model is true, the ΔBIC increases as a logarithmic function of the sample size. When the larger model is true, the ΔBIC decreases as a linear function of the sample size. Thus, the cost of a reduction in power is not burdensome when sample sizes are large, because the logarithmic function grows much more slowly than a linear function.

BIC and sample size

For the study in this article, three distinct BIC values were computed using different values for the number of observations: the level 1 sample size, level 2 sample size, and effective sample size (i.e., the total level 1 sample size divided by the design effect). The design effect (Kish, 1965) is defined as

$$\text{Design effect} = 1 + (n{-}1) \times \text{ICC}, \qquad (3)$$

where n represents the group size (Snijders & Bosker, 2012). This definition of the design effect uses balanced groups of size n, and we also used balanced groups in our simulation study. However, for the discussion in this section, let n_j be the size of group j, $J = N_{\text{grp}}$ be the number of groups (level 2 sample size), and N_{tot} be total sample size, given by the sum of the n_j.

The total sample size can be derived for testing level 1 fixed effects when the covariance structure is assumed to be known (Dellatre, Lavielle, & Poursat, 2014). Similarly, the BIC can be derived by using a unit information prior approach (Kass & Raftery, 1995; Kass & Wasserman 1995). Let the covariance of the data be given by Σ, the covariate design matrix be X, and the prior precision of the fixed effects be $\Omega\Omega = X^{T}\Sigma\ ^{1}X/N_{\text{tot}}$, which is a unit information prior precision. The BIC conditioned on the maximum-likelihood estimator (MLE) of Σ uses the level 1 sample size. We note that in the two-level case, Ω can be derived using group-specific covariate matrices, X_j, and covariance structures, Σ_j. The prior precision is obtained by

$$\Omega = \sum_{j=1}^{J} w_j\Omega_j, \text{where } \Omega_j = \frac{X_j^T\Sigma_j^{-1}X_j}{n_j} \text{ and } w_j = n_j/N_{\text{tot}} \qquad (4)$$

Thus, the unit information precision can be viewed as a convex combination of the unit information precision from each group, where the weights are determined by the relative sample sizes. The insight here is that the groupings only affect the testing of a level 1 effect through a change in the contribution of each group to the precision of the estimate of the level 1 effect.

The effective sample size presented here is a special case of counting the effective number of observations in a random-intercepts model (Jones, 2011). The number of effective samples is counted by computing the sum of the elements of the inverse correlation matrix for each group, which is related to the group precision matrix. In the particular case of compound symmetry, the number of effective observations (*ne*) in group *j* is defined as

$$ne_j = \frac{n_j}{1+\left(n_j-1\right)\mathrm{ICC}} \tag{5}$$

and the total number of effective observations, N_{icc}, is given as the sum of the ne_j. When the group sizes are all the same, this results in the definition of the effective sample size being the same as that described at the beginning of this section.

The effective number of observations in multilevel models with more than two levels, random slopes, or nonnested groupings can be calculated by using a general approach (Jones, 2011). Specifically, one computes the MLE for Σ, converts it to a correlation matrix, inverts the correlation matrix, and sums the elements of the resulting matrix. This effective sample size could then be used analogously for computation of the BICs for a variety of models (Jones, 2011), but demonstration of these topics is beyond the scope of the present study.

Using the number of level 2 groups *(J)* can be motivated by the case in which the ICC is near 1. Then each group has approximately one effective observation, and N_{icc} is approximately *J*. Similarly, when the ICC is near 0, then N_{icc} is approximately N_{tot}. When the number of samples in each group is large relative to 1/ICC (e.g., when the ICC is bounded away from 0 and group sizes are large), then the effective number of observations in each group is approximately 1/ICC, so that N_{icc} is approximately *J*/ICC. Using the number of level 2 groups (*J*) can also be theoretically justified when testing random effects and not fixed effects (Dellatre et al., 2014), although this case is not considered here.

Research questions and hypotheses

The present study examines the following research questions and offers the following hypotheses:

Research Question 1 What is the likelihood of choosing the correct null model when assessing a fixed effect in a multilevel model with three different values of sample size for computation of the BIC—level 1 sample size, effective sample size, and level 2 sample size—while varying the number of groups, group size, and ICC? This research question is related to the Type I error rate of hypothesis tests and is also explored through a Bayesian model selection rule.

Hypothesis As the number of groups and the group size increases, the false positive rate will decrease (Raftery, 1995). As the ICC decreases, the effective information in the data will increase, resulting in lower false positive rates. Since the level 1 sample size will be larger than the effective sample size or the level 2 sample size, the rate of false positives will be lowest for level 1 sample size, then for effective sample size, and highest for level 2 sample size.

Research Question 2 What is the likelihood of correctly selecting the model with a nonzero fixed effect in a multilevel model using the same three values of sample size for computation of the BIC under the same varying conditions? This research question is related to the power of hypothesis tests and is similarly investigated using a Bayesian hypothesis selection rule.

Hypothesis As the number of groups and the group size increases, power will increase. As the ICC decreases, the effective information in the data will increase, resulting in higher power. Since the BIC implicitly makes it harder to find an effect as the sample size increases, the computation using level 2 sample size will show the highest power, followed by effective sample size, followed in turn by level 1 sample size.

Method

A simulation study was conducted to answer the research questions. For each simulated dataset, the following four multilevel models were estimated:

$$Y_{ij} = B_0 + u_j + e_{ij}, \tag{6}$$

$$Y_{ij} = B_0 + B_1 X_{ij} + u_j + e_{ij}, \tag{7}$$

$$Y_{ij} = B_0 + B_2 Z_j + u_j + e_{ij}, \tag{8}$$

$$Y_{ij} = B_0 + B_1 X_{ij} + B_2 Z_j + u_j + e_{ij}, \tag{9}$$

where Y_{ij} is the outcome for individual *i* within group *j*, B_0 is the intercept, B_1 is the slope coefficient for the individual-level predictor variable X_{ij}, and B_2 is the slope for the group-level predictor Z_j. The u_j term is the random effect for clusters and is normally distributed with mean of zero and variance of τ^2, and

e_{ij} is the random error term, which is also normally distributed with mean of zero and variance of σ^2. The independent variables, X and Z, were generated as independent standard normal random variables.

The following values were all fully crossed, for a total of 1,600 separate conditions: group sizes of 5, 15, 25, 35, and 45; numbers of groups of 10, 20, 35, 50, and 100; ICCs of .1, .2, .3, and .5; and individual- and group-level predictor slopes (B_1 and B_2) of 0, .1, .2, and .3. This rich set of conditions was guided by commonly encountered datasets within the field of education. Previous simulation studies had used similar (Lorah, 2018; Maas & Hox, 2004) sample size conditions (group size and number of groups). Maas and Hox chose 30 clusters as a minimum specified in the literature, 100 clusters as a sufficient size specified in the literature, and 50 clusters as a common size found in research; they chose a cluster size of 5 to represent a common condition in family research, 30 to represent a common condition for educational research, and 50 based on the literature. Furthermore, guidance regarding the minimum sample size for multilevel models was considered. The "30/30 rule" specifies a minimum of 30 clusters with 30 individuals per cluster (Hox, 2010), whereas others have suggested that a minimum of 10 clusters should be used (Snijders & Bosker, 2012). Exploring simulated datasets at and around these minima should be particularly relevant.

Previous research has assessed typical ICC conditions: for K–12 academic achievement, the average ICCs were .22 overall and .09 for low-achieving schools (Hedges & Hedberg, 2007), and another study reported average achievement in the United States between .10 and .20 (Spybrook, 2008). A previous simulation study had used similar ICC values: .10, .20, and .30 (Maas & Hox, 2004). The present values were chosen to represent the range of typically observed ICC values in educational research, as well as a high ICC value. Note that the actual observed ICC values based on simulated data are expected to differ slightly, due to a known bias in ICC (Atenafu et al., 2012), as well as a known bias in variance component estimation with full maximum-likelihood estimation (Hox, 2010) and an omitted-variable bias when a meaningful group-level predictor is omitted from the estimated model.

For each of the 1,600 conditions, a total of 200 simulation replications were run. All data were generated and analyzed in R (R Core Team, 2017), and multilevel models were estimated with the lmer() function within the lme4 package (Bates, Mächler, Bolker, & Walker, 2015) using full maximum-likelihood methods. For each scenario, the models were evaluated in two ways. First we considered the simple selection rule, according to which the model with the smallest BIC was selected. The results of such a rule were summarized by considering the rank of the true model amongst the models under consideration. Second, we considered hypothesis-testing questions by using the class-wide minimization selection rule. The hypothesis for testing the level 1 (and similarly the level 2) fixed effect was represented by a model class containing two models. The model class containing the model that had the smallest BIC was selected.

Results and discussion

Tables 1, 2, 3, 4, 5 and 6 in Appendix A suggest that using the BIC with N_{grp} provided the best ranking for the true model across the range of conditions. However, this is misleading, because of aggregation over all the different models that were simulated. In the rest of this section, we discuss the nuances of testing level 1 and level 2 fixed effects using the different methods for calculating the BIC.

Research Question 1 Tables 7, 8, 9 and 10 in Appendix B show the results for testing a level 1 fixed effect using the three different BIC methods. Table 7 shows the expected behavior that the decrease in penalty from going from N_{tot} to N_{icc} to N_{grp} increases the Type I error rate. Table 8 shows that the Type I error rate only changes with the ICC for the penalty using N_{icc}. Table 9 shows that increasing the number of observations by increasing the number of groups decreases the Type I error rate for all three BIC types, with the smallest false positive rates coming from N_{tot}.

The most interesting result for testing level 1 fixed effects is in Table 10, which shows that increasing the number of observations by increasing group size renders the penalties from using N_{grp} and N_{icc} useless for controlling false positives. Considering the model in Eq. 7 or 9, Appendix D shows that increasing the group size while keeping the number of groups fixed is asymptotically equivalent to mean-centering X_{ij} by groups when considering a level 1 fixed effect.

As was discussed in Delattre et al. (2014), the concentration rate for the variance of level 1 fixed effects is $1/N_{\text{tot}}$. The other methods for computing the BIC only provide asymptotic protection from false positives when the number of groups increases. Thus, requiring in any manner that the Type I error rate decrease to 0 as the sample size increases restricts us to the use of N_{tot} when testing level 1 fixed effects. This restriction is reasonable for testing a level 1 fixed effect, because the consistency of the Bayesian testing procedure needs to be maintained even when there is only a single group.

Similar empirical results hold for the false positive rate when testing a level 2 fixed effect, as is evidenced in Tables 11, 12, 13 and 14 in Appendix C. Note that the penalty from N_{tot} is a sum of the penalties from the number of groups and the average group size. Thus, although increasing the group size does decrease the Type I error rate under the BIC computed using N_{tot}, it is not immediately clear that this is desirable for level 2 fixed effects. Intuitively, when the number of groups is fixed, it does not make sense that increasing group size should provide overwhelming information about a level 2 fixed effect.

Research Question 2 First, consider the power for testing a level 1 fixed effect. Although Appendix B does show the power increasing to 1 for the penalties using N_{icc} and N_{grp}, we restrict our attention to N_{tot}, because its Type I error rate decreases to 0 even when the number of groups is fixed. We observe empirically the expected increases in power as the sample size increases. However, we do not observe the expected decrease in power as the ICC increases in Table 8. This is due to the fact that the level 1 design matrix X was simulated with the theoretical group means set at 0. For instance, the Fisher information for B_1 in Eq. 7 with $B_0 = 0$ is given by

$$I(B_1) = \sum_j \frac{n_j}{\sigma^2} \left(\overline{x_j^2} - \bar{x}_j^2 \times \text{ICC} \times ne_j \right) \tag{10}$$

and thus the effect of an increase in ICC causing a decrease in power was not observed in the simulation study.

Power for level 2 fixed effects is a bit more subtle. We did observe the increase in power due to an increase in the number of groups, as expected. Table 12 shows the expected decrease in power due to an increase in ICC. Now we need to determine whether N_{tot} is a reasonable penalty for a level 2 fixed effect. From Table 14 it is not obvious, but the power of the test does decrease to zero as group size increases for a fixed number of groups when penalizing using N_{tot}. To see this, consider the model in Eq. 8, with $B_0 = 0$ and any $\tau > 0$. The Fisher information for B_2 is given by

$$I(B_2) = \text{ICC} \times \sum_j \frac{ne_j}{\tau^2} z_j^2 \tag{11}$$

So that the variance of the MLE for B_2 concentrates at a rate of $1/J$. Thus, N_{tot} penalizes too much as the group size increases, decreasing the power to 0. Again, consider the extreme case in which there are only two groups. As the group sizes increase, $\log(N_{\text{tot}})$ produces an infinite penalty, even though there are only two groups with which to estimate B_2 and $I(B_2)$ converges to $(z_1{}^2 + z_2{}^2)/\tau^2$.

Finally, Eq. 11 provides an additional justification for using the sum of the effective sample sizes as a penalty. $I(B_2)/N_{\text{icc}}$ converges to a positive number under suitable conditions as J increases, and thus $\log(N_{\text{icc}})$ could be an appropriate penalty for the BIC for level 2 fixed effects. However, we note that Eq. 11 itself does not really provide a reason to prefer N_{icc} to N_{grp}, or vice versa.

Conclusions and limitations

Through this study, we have been able to conclude something similar to the results of Delattre et al. (2014). We have demonstrated that a single computation for the sample size in the penalty for the BIC for a multilevel model is insufficient. Delattre et al. considered testing both fixed and random effects and used both group size and the total number of observations. We have shown that both are needed in order to properly penalize fixed effects at different levels of the data. To get both the Type I and Type II error rates to decrease to 0, both the level 1 sample size and the level 2 sample size need to be used to penalize fixed effects at their respective levels. Thus, the preferred form of the BIC should include a penalty term of the form $d_1 \log(N_{\text{tot}})$, where d_1 is the number of level 1 fixed effects, and a term of the form $d_2 \log(N_{\text{grp}})$, where d_2 is the number of level 2 fixed effects. This will produce appropriate model selection behavior from the BIC for testing fixed effects at both levels.

One limitation to this study is that the results do not necessarily generalize beyond the specific conditions examined in the simulation. In the present study we examined a two-level linear regression model with balanced groups, but future research should extend this to more complicated models, such as three-level models, nonnested groupings, random-slope models, nonlinear models, latent-variable models, and so forth. In particular, three-level models include sample size at three rather than two levels, making the value for sample size additionally ambiguous. We conjecture that an appropriate BIC would include a penalty that is the product of the number of fixed effects at each level times the log of the appropriate level-based sample size. Furthermore, in the present study we examined testing for a fixed effect, but future research should extend this inquiry to random effects. Additionally, although the penalty based on N_{icc} in the present study did not seem to show dramatically different behavior from that for N_{grp}, its use in analyzing unbalanced groups needs to be explored.

Tabulated output for proportions of ranks for the true model

Table 1 Proportions of ranks for the true model (overall average)

BIC Type	Rank True = 1	Rank True = 2	Rank True = 3	Rank True = 4
N_tot	.42	.39	.10	.09
N_icc	.47	.37	.09	.07
N_grp	.54	.33	.07	.05

Table 2 Proportions of ranks for the true model (average by level 1 fixed effects)

BIC Type	Rank True = 1	Rank True = 2	Rank True = 3	Rank True = 4	Rank True = 1	Rank True = 2	Rank True = 3	Rank True = 4
	L1_FE = 0				**L1_FE = .1**			
N_tot	.48	.36	.15	.01	.29	.29	.16	.26
N_icc	.52	.30	.15	.02	.35	.32	.13	.20
N_grp	.56	.26	.14	.03	.44	.32	.10	.14
	L1_FE = .2				**L1_FE = .3**			
N_tot	.44	.42	.07	.07	.47	.48	.03	.02
N_icc	.49	.40	.05	.06	.52	.44	.02	.02
N_grp	.57	.36	.03	.04	.59	.39	.01	.01

Table 3 Proportions of ranks for the true model (average by level 2 fixed effects)

BIC Type	Rank True = 1	Rank True = 2	Rank True = 3	Rank True = 4	Rank True = 1	Rank True = 2	Rank True = 3	Rank True = 4
	L2_FE = 0				**L2_FE = .1**			
N_tot	.82	.13	.04	.01	.08	.60	.17	.15
N_icc	.84	.12	.04	.01	.12	.61	.15	.12
N_grp	.81	.13	.05	.01	.21	.57	.12	.09
	L2_FE = .2				**L2_FE = .3**			
N_tot	.29	.47	.12	.12	.48	.34	.08	.09
N_icc	.37	.44	.10	.10	.57	.30	.07	.07
N_grp	.48	.38	.08	.07	.65	.25	.05	.05

Table 4 Proportions of ranks for the true model (average by ICC)

BIC Type	Rank True = 1	Rank True = 2	Rank True = 3	Rank True = 4	Rank True = 1	Rank True = 2	Rank True = 3	Rank True = 4
	ICC = .1				**ICC = .2**			
N_tot	.55	.30	.08	.07	.45	.36	.10	.09
N_icc	.58	.29	.07	.07	.50	.34	.08	.07
N_grp	.67	.24	.05	.04	.58	.30	.07	.05
	ICC = .3				**ICC = .5**			
N_tot	.38	.41	.11	.10	.29	.47	.13	.11
N_icc	.45	.39	.09	.08	.37	.45	.11	.08
N_grp	.51	.36	.08	.06	.40	.43	.10	.07

Table 5 Proportions of ranks for the true model (average by number of groups)

BIC Type	Rank True = 1	Rank True = 2	Rank True = 3	Rank True = 4	Rank True = 1	Rank True = 2	Rank True = 3	Rank True = 4
	# Groups = 10				**# Groups = 20**			
N_tot	.24	.38	.17	.20	.32	.42	.13	.13
N_icc	.28	.40	.15	.16	.37	.41	.11	.10
N_grp	.38	.38	.12	.12	.45	.38	.09	.07
	# Groups = 35				**# Groups = 50**			
N_tot	.41	.41	.10	.07	.49	.39	.07	.05
N_icc	.48	.39	.08	.06	.55	.35	.06	.04
N_grp	.54	.35	.07	.04	.60	.32	.05	.03
	# Groups = 100							
N_tot	.63	.32	.03	.02				
N_icc	.68	.27	.03	.01				
N_grp	.72	.24	.03	.01				

Table 6 Proportions of ranks for the true model (average by group size)

BIC Type	Rank True = 1	Rank True = 2	Rank True = 3	Rank True = 4	Rank True = 1	Rank True = 2	Rank True = 3	Rank True = 4
	Group Size = 5				**Group Size = 15**			
N_tot	.32	.32	.16	.20	.42	.37	.11	.10
N_icc	.34	.32	.15	.18	.46	.36	.10	.08
N_grp	.39	.33	.14	.14	.53	.34	.08	.06
	Group Size = 25				**Group Size = 35**			
N_tot	.44	.40	.09	.07	.46	.42	.08	.05
N_icc	.50	.37	.07	.05	.52	.38	.06	.03
N_grp	.57	.34	.06	.03	.59	.33	.05	.02
	Group Size = 45							
N_tot	.46	.43	.07	.04				
N_icc	.53	.39	.06	.02				
N_grp	.60	.34	.05	.02				

Tabulated output for tests of level 1 fixed effects

Table 7 Proportions rejecting H0: L1_FE = 0 (overall average)

BIC Type	L1_FE = 0	L1_FE = .1	L1_FE = .2	L1_FE = .3
N_tot	.014	.536	.874	.961
N_icc	.034	.623	.900	.968
N_grp	.071	.703	.926	.978

Table 8 Proportions rejecting H0: L1_FE = 0 (average by ICC)

BIC Type	ICC = .1	ICC = .2	ICC = .3	ICC = .5	ICC = .1	ICC = .2	ICC = .3	ICC = .5
	L1_FE = 0				**L1_FE = .1**			
N_tot	.014	.014	.014	.014	.542	.535	.534	.533
N_icc	.024	.030	.037	.045	.593	.615	.631	.655
N_grp	.073	.069	.072	.070	.708	.702	.704	.698
	L1_FE = .2				**L1_FE = .3**			
N_tot	.878	.875	.872	.870	.964	.959	.961	.959
N_icc	.891	.898	.904	.906	.967	.966	.969	.971
N_grp	.929	.926	.928	.921	.980	.979	.978	.976

Table 9 Proportions rejecting H0: L1_FE = 0 (average by number of groups)

BIC Type	# Groups = 10	# Groups = 20	# Groups = 35	# Groups = 50	# Groups = 100	# Groups = 10	# Groups = 20	# Groups = 35	# Groups = 50	# Groups = 100
	L1_FE = 0					**L1_FE = .1**				
N_tot	.025	.016	.012	.009	.007	.218	.380	.565	.682	.834
N_icc	.058	.042	.029	.023	.017	.348	.507	.656	.747	.859
N_grp	.130	.086	.061	.047	.033	.493	.615	.728	.795	.884
	L1_FE = .2					**L1_FE = .3**				
N_tot	.689	.837	.911	.941	.991	.877	.947	.984	.996	1.000
N_icc	.765	.865	.925	.951	.994	.899	.957	.988	.997	1.000
N_grp	.831	.897	.942	.964	.996	.930	.970	.992	.998	1.000

Table 10 Proportions rejecting H0: L1_FE = 0 (average by group size)

BIC Type	Group Size = 5	Group Size = 15	Group Size = 25	Group Size = 35	Group Size = 45	Group Size = 5	Group Size = 15	Group Size = 25	Group Size = 35	Group Size = 45
	L1_FE = 0					**L1_FE = .1**				
N_tot	.027	.015	.011	.009	.008	.180	.431	.594	.701	.774
N_icc	.040	.035	.032	.032	.031	.221	.533	.701	.800	.862
N_grp	.072	.072	.073	.068	.071	.294	.636	.791	.873	.920
	L1_FE = .2					**L1_FE = .3**				
N_tot	.580	.872	.948	.978	.991	.829	.978	.997	1.000	1.000
N_icc	.628	.911	.973	.990	.997	.855	.987	.999	1.000	1.000
N_grp	.700	.947	.987	.996	.999	.896	.995	1.000	1.000	1.000

Tabulated output for tests of level 2 fixed effects

Table 11 Proportions rejecting H0: L2_FE = 0 (overall average)

BIC Type	L2_FE = 0	L2_FE = .1	L2_FE = .2	L2_FE = .3
N_tot	.021	.091	.327	.550
N_icc	.037	.138	.408	.635
N_grp	.088	.237	.526	.726

Table 12 Proportions rejecting H0: L2_FE = 0 (average by ICC)

BIC Type	ICC = .1	ICC = .2	ICC = .3	ICC = .5	ICC = .1	ICC = .2	ICC = .3	ICC = .5
	L2_FE = 0				**L2_FE = .1**			
N_tot	.021	.019	.020	.022	.171	.093	.064	.038
N_icc	.024	.029	.040	.054	.207	.141	.111	.092
N_grp	.086	.086	.091	.090	.368	.250	.190	.138
	L2_FE = .2				**L2_FE = .3**			
N_tot	.574	.380	.245	.107	.805	.644	.496	.256
N_icc	.613	.458	.346	.214	.826	.709	.604	.401
N_grp	.762	.599	.464	.279	.910	.816	.702	.475

Table 13 Proportions rejecting H0: L2_FE = 0 (average by number of groups)

BIC Type	# Groups = 10	# Groups = 20	# Groups = 35	# Groups = 50	# Groups = 100	# Groups = 10	# Groups = 20	# Groups = 35	# Groups = 50	# Groups = 100
	L2_FE = 0					**L2_FE = .1**				
N_tot	.047	.024	.013	.011	.008	.068	.060	.075	.090	.164
N_icc	.070	.045	.029	.024	.016	.099	.097	.119	.142	.233
N_grp	.184	.106	.067	.052	.033	.234	.192	.204	.232	.322
	L2_FE = .2					**L2_FE = .3**				
N_tot	.138	.192	.293	.394	.615	.259	.390	.561	.675	.866
N_icc	.185	.266	.384	.495	.709	.330	.498	.664	.764	.918
N_grp	.362	.411	.503	.592	.762	.510	.624	.742	.816	.937

Table 14 Proportions rejecting H0: L2_FE = 0 (average by group size)

BIC Type	Group Size = 5	Group Size = 15	Group Size = 25	Group Size = 35	Group Size = 45	Group Size = 5	Group Size = 15	Group Size = 25	Group Size = 35	Group Size = 45
	L2_FE = 0					**L2_FE = .1**				
N_tot	.034	.022	.018	.014	.014	.101	.093	.093	.086	.083
N_icc	.042	.039	.035	.033	.036	.118	.136	.144	.145	.146
N_grp	.082	.093	.087	.089	.091	.189	.233	.247	.258	.257
	L2_FE = .2					**L2_FE = .3**				
N_tot	.304	.342	.332	.329	.326	.537	.562	.554	.551	.547
N_icc	.338	.408	.424	.434	.435	.572	.634	.650	.654	.665
N_grp	.429	.525	.550	.559	.566	.652	.727	.743	.747	.759

Score equations and concentration rates for the random intercepts model

We work with the model $Y_{ij} = B_0 + B_1X_{ij} + B_2Z_j + u_j + e_{ij}$, where the u_j are independent and identically distributed Gaussians with mean 0 and variance τ^2 and are independent of e_{ij}, which are Gaussians with mean 0 and variance σ^2. Here we assume that there are J groups with $j = 1, \ldots, J$ and that $i = 1, \ldots, n_j$ for group j. The score equation for estimating $B = (B_0, B_1, B_2)^T$ is given by

$$\sum_{j=1}^{J}\begin{pmatrix} n_j\overline{Y}_j/\left(n_j\tau^2+\sigma^2\right) \\ \frac{n_j}{\sigma^2}\left(\overline{X_jY_j}-n_j\tau^2\overline{X}_j\,\overline{Y}_j/\left(n_j\tau^2+\sigma^2\right)\right) \\ n_jZ_j\overline{Y}_j/\left(n_j\tau^2+\sigma^2\right) \end{pmatrix}$$

$$= \sum_{j=1}^{J}\begin{pmatrix} n_j/\left(n_j\tau^2+\sigma^2\right) & n_j\overline{X}_j/\left(n_j\tau^2+\sigma^2\right) & n_jZ_j/\left(n_j\tau^2+\sigma^2\right) \\ n_j\overline{X}_j/\left(n_j\tau^2+\sigma^2\right) & \frac{n_j}{\sigma^2}\left(\overline{X_j^2}-n_j\tau^2\,\overline{X}_j^{\,2}/\left(n_j\tau^2+\sigma^2\right)\right) & n_j\overline{X}_j\,Z_j/\left(n_j\tau^2+\sigma^2\right) \\ n_jZ_j/\left(n_j\tau^2+\sigma^2\right) & n_j\overline{X}_j\,Z_j/\left(n_j\tau^2+\sigma^2\right) & n_j\,Z_j^2/\left(n_j\tau^2+\sigma^2\right) \end{pmatrix}\begin{pmatrix} B_0 \\ B_1 \\ B_2 \end{pmatrix}$$

Consider the asymptotic score as group sizes grow. Letting each n_j be large relative to σ^2/τ^2, we obtain the asymptotic score approximation

$$\sum_{j=1}^{J}\begin{pmatrix} \bar{Y}_j/\tau^2 \\ \frac{n_j}{\sigma^2}\left(\overline{X_jY_j}-\bar{X}_j\,\bar{Y}_j\right) \\ Z_j\bar{Y}_j/\tau^2 \end{pmatrix} \approx \sum_{j=1}^{J}\begin{pmatrix} 1/\tau^2 & \bar{X}_j/\tau^2 & Z_j/\tau^2 \\ \bar{X}_j/\tau^2 & \frac{n_j}{\sigma^2}\left(\overline{X_j^2}-\bar{X}_j^{\,2}\right) & \bar{X}_j\,Z_j/\tau^2 \\ Z_j/\tau^2 & \bar{X}_j\,Z_j/\tau^2 & Z_j^2/\tau^2 \end{pmatrix}\begin{pmatrix} B_0 \\ B_1 \\ B_2 \end{pmatrix}.$$

Dividing each side by $\sum n_j$, we get the approximate equation for B_1,

$$\frac{\sum_{j=1}^{J} n_j\left(\overline{X_jY_j}-\bar{X}_j\,\bar{Y}_j\right)}{\sigma^2\sum_{j=1}^{J} n_j} \approx \left(\frac{\sum_{j=1}^{J} n_j\left(\overline{X_j^2}-\bar{X}_j^{\,2}\right)}{\sigma^2\sum_{j=1}^{J} n_j}\right) B_1.$$

This is the exactly the equation (without asymptotic approximation) for B_1 that is obtained when the design points have been mean centered by group ($X_{ij}\mapsto X_{ij}-\bar{X}_j$) before the analysis is performed. We note that, because this transformation depends on j, the MLE is not invariant to it. However, the MLE for B_1 computed without the transformation converges to that computed with the transformation.

The Fisher information is given by

$$I = \sum_{j=1}^{J}\begin{pmatrix} n_j/\left(n_j\tau^2+\sigma^2\right) & n_j\bar{X}_j/\left(n_j\tau^2+\sigma^2\right) & n_jZ_j/\left(n_j\tau^2+\sigma^2\right) \\ n_j\bar{X}_j/\left(n_j\tau^2+\sigma^2\right) & \frac{n_j}{\sigma^2}\left(\overline{X_j^2}-n_j\tau^2\,\bar{X}_j^{\,2}/\left(n_j\tau^2+\sigma^2\right)\right) & n_j\bar{X}_j\,Z_j/\left(n_j\tau^2+\sigma^2\right) \\ n_jZ_j/\left(n_j\tau^2+\sigma^2\right) & n_j\bar{X}_j\,Z_j/\left(n_j\tau^2+\sigma^2\right) & n_j\,Z_j^2/\left(n_j\tau^2+\sigma^2\right) \end{pmatrix}$$

The asymptotic concentration rates for asymptotic variances are described by the growth rates of the diagonal terms. The diagonal term corresponding to B_0 grows at a rate that is bounded above by J/τ^2. Similarly, the diagonal term corresponding to B_2 grows at a rate that is bounded above by $\sum Z_j^2/\tau^2$. Assuming that $\sum Z_j^2/J$ and the MLE for τ^2 converge in probability to finite, positive numbers as J increases, the appropriate penalty for B_0 and B_2 in the BIC is log(J). In contrast, the diagonal term corresponding to B_1 is bounded below by $\sum\frac{n_j}{\sigma^2}\left(\overline{X_j^2}-\bar{X}_j^{\,2}\right)$ and above by $\sum\frac{n_j}{\sigma^2}\overline{X_j^2}$. Assuming that $\sum\left(\overline{X_j^2}-\bar{X}_j^{\,2}\right)/J$, $\sum\overline{X_j^2}/J$, and the MLE for σ^2 converge in probability to finite, positive numbers as J increases, then the appropriate penalty for B_2 in the BIC is $log\left(\sum_{j=1}^{J} n_j\right)$.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

Atenafu, E. G., Hamid, J. S., To, T., Willan, A. R., Felman, B. M., & Beyene, J. (2012). Bias-corrected estimator for intraclass correlation coefficient in the balanced one-way random effects model. *BCM Medical Research Methodology*, *12*, 1–8.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48. https://doi.org/10.18637/jss.v067.i01

Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, *45*, 1304–1312. https://doi.org/10.1037/0003-066X.45.12.1304

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Delattre, M., Lavielle, M., & Poursat, M. A. (2014). A note on BIC in mixed-effects models. *Electronic Journal of Statistics*, *8*, 456–475. https://doi.org/10.1214/14-EJS890

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, *29*, 60–87.

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, *14*, 382–417.

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications*. New York, NY: Routledge.

Jones, R. H. (2011). Bayesian information criterion for longitudinal and clustered data. *Statistics in Medicine*, *30*, 3050–3056. https://doi.org/10.1002/sim.4323

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795. https://doi.org/10.1080/01621459.1995.10476572

Kass, R. E., & Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association*, *90*, 928–934.

Kish, L. (1965). *Survey sampling*. New York, NY: Wiley.

Lorah, J. A. (2018). Estimating individual-level interaction effects in multilevel models: A Monte Carlo simulation study with application. *Journal of Applied Statistics*, *45*, 2238–2255. https://doi.org/10.1080/02664763.2017.1414163

Lorah, J. A., Sanders, E. A., & Morrison, S. J. (2014). The relationship between English language learner status and music ensemble participation. *Journal of Research in Music Education*, *62*, 234–244. https://doi.org/10.1177/0022429414542301

Maas, C. J. M., & Hox, J. J. (2004). Robustness issues in multilevel regression analysis. *Statistica Neerlandica*, *58*, 127–137.

McCoach, D. B., & Black, A. C. (2008). Evaluation of model fit and adequacy. In A. A. O'Connell, & D. B. McCoach (Eds.), *Multilevel modeling of educational data* (pp. 245–272). Charlotte, NC: Information Age.

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from https://www.R-project.org/

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, *25*, 111–163, disc. 165–195. https://doi.org/10.2307/271063

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

Scott, J. G., & Berger, J. O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *Annals of Statistics*, *38*, 2587–2619. https://doi.org/10.1214/10-AOS792

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

Spybrook, J. (2008). Power, sample size, and design. In A. A. O'Connell & D. B. McCoach (Eds.), *Multilevel modeling of educational data* (pp. 273–311). Charlotte, NC: Information Age.

Weakliem, D. L. (2004). Introduction to the special issue on model selection. *Sociological Methods and Research*, *33*, 167–187.

Yang, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation. *Biometrika*, *92*, 937–950.