METHODOLOGY
European Journal of Research Methods for the Behavioral and Social Sciences

PsychOpen GOLD

Original Article

Check for updates

A Comparison of Methods for Specifying Optimal Random Effects Structures

Wen Luo [1], Haoran Li [2]

[1] *Department of Educational Psychology, Texas A&M University, College Station, TX, USA.* **[2]** *Department of Educational Psychology, University of Minnesota, Minneapolis, MN, USA.*

Methodology, 2023, Vol. 19(4), 365–386, https://doi.org/10.5964/meth.9601

Received: 2022-05-31 • **Accepted:** 2023-09-18 • **Published (VoR):** 2023-12-22

Handling Editor: Katrijn Van Deun, Tilburg University, Tilburg, The Netherlands

Corresponding Author: Wen Luo, 4225 Department of Educational Psychology, Texas A&M University, College Station, TX, 77843, USA. E-mail: wluo@tamu.edu

Abstract

Using Monte Carlo simulations, this study compared the performance of various approaches to the specification of random effects structures in linear mixed effects models (LMMs), including the minimal approach, the maximal approach, the forward search, the backward search, and the all-possible structures approach. The results showed that if the predictor of interest is at the within-cluster level or involves a cross-level interaction, the maximal approach, the best-path forward search, and the best-path backward search are all desirable methods. If the predictor of interest is at the cluster level, it is not essential to specify random slopes of Level-1 predictors. In addition, it is important to specify random slopes of within-cluster control variables, as they can increase the statistical power for testing the main within-cluster variables, especially when the sample size is small and the variance of the random slope of the control variable is large.

Keywords

linear mixed effects models (LMMs), random effects structures, model specification, model selection, mathematics subject

Linear mixed effects models (LMMs) are commonly used to analyze clustered data in educational, social, and behavioral sciences. The general form of LMMs is given by $Y = X\beta + Zu + e$, where Y represents the outcome variable, X the predictors, β the fixed effects, Z the design matrix for random effects, u the random effects, and e the errors. The random effects (u) can be associated with either a random intercept or a random slope.

The specification of LMMs poses a challenge as researchers must not only specify the fixed effects but also the random effects. A systematic review of applied multilevel modeling research in the past decade revealed that information regarding the specification of random effects in models is frequently missing or incomplete (Luo et al., 2021). Applied researchers often fail to provide clear rationales for the selection of a particular random effect structure and tend to default to using random intercept models. However, this practice is problematic as research has demonstrated that constraining a randomly varying slope of a Level-1 predictor to be constant across Level-2 clusters can lead to inflated Type I error in the test of the fixed effect (e.g., Barr et al., 2013).

One challenge in specifying random effects is the lack of sufficient details in substantive theories and previous studies to guide *a priori* specifications. Therefore, Barr et al. (2013) argued for the maximal model approach, which involves using the full variance-covariance structure of random effects. However, some researchers argue that the maximal approach may lead to overfitting when population variance components are close to zero, potentially resulting in reduced power, particularly with small sample sizes (Matuschek et al., 2017). Therefore, a model selection approach is recommended, where a certain criterion is used to select the random effect structure that is most supported by the data (e.g., Matuschek et al., 2017).

Facing the challenges of specifying random effects in LMMs, there is a lack of methodological investigations that evaluate the different approaches and provide evidence-based guidelines for applied researchers. Hence, the purpose of this study is to compare various approaches to random effects specification in terms of their impact on the statistical inferences of fixed effects. In the remainder of the paper, we first provide a brief review of the maximal and model selection approach. Next we introduce the design of the simulation study and report the results. Finally, we discuss the findings and provide recommendations for applied researchers.

Literature Review

The Maximal Approach

For confirmatory hypothesis testing, Barr et al. (2013) argued that LMMs should include the maximal random effect structure that is justified by the design. Specifically, it is recommended to include random slopes for all within-cluster factors. When a within-cluster variable is involved in a cross-level interaction, a random slope should always be included for the within-cluster variable (Heisig & Schaeffer, 2019). For control variables, however, it remains unclear whether including random slopes is essential (Barr et al., 2013).

The rationales behind the maximal approach are twofold. First, random effects are crucial for capturing the measurement dependances in the design (Bolker et al., 2009). Second, a data-driven approach could incorrectly exclude random effects due to insufficient power, which is particularly relevant for small sample sizes. Through a simulation

study, Barr et al. (2013) showed that the maximal model can control Type I error rate without a significant loss of power, whereas the model selection approach tended to be anticonservative with minimal power gains. However, it is important to note that the model used for data generation in Barr et al. (2013) was based on a two-way random effects ANOVA design (i.e., involving items and subjects) with a single within-subject treatment factor. Hence it remains unclear whether their findings can be generalized to models with more predictors at both the within-cluster and between-cluster levels.

The Model Selection Approach

In contrast, other researchers have recommended using model selection approaches to determine an optimal random effect structure (e.g., Stroup, 2012). To investigate the potential power loss associated with the maximal model, Matuschek et al. (2017) compared the maximal approach with two model selection approaches, namely backward selection based on likelihood ratio test (LRT) and AIC-based selection from all possible subset models. The results showed that the backward selection based on LRT can achieve higher power without inflated Type I error rates. The AIC-based all subset model selection can achieve similar power to the LRT-based backward selection, but may suffer from slightly inflated Type I error rates in conditions with small sample sizes. However, the maximal approach leads to a significant loss of statistical power when the true value of the variance components is small.

Although Matuschek et al. (2017) used similar data generation models and sample size conditions as Barr et al. (2013), the conclusions drawn from the two studies were quite different. Two potential reasons may explain the contradictory findings. First, the relative magnitudes of the variance components in the two studies differed significantly. Barr et al. (2013) selected all variance components from a uniform distribution ranging from 0 to 3, allowing for a wide range of intraclass correlation coefficients (ICCs) from 0 to 1. In contrast, Matuschek et al. (2017) had a narrower range of ICC from 0 to 0.35. Secondly, the empirical power was calculated differently between the two studies. Barr et al. (2013) corrected the empirical power to account for anticonservativeness, while Matuschek et al. (2017) directly compared the power of different approaches that varied in their Type I error rates.

Model Selection Criteria

In model selection, researchers must strike a balance between goodness of fit and parsimony. Many model comparison methods aim to address the trade-off between goodness of fit and model complexity (Kuha, 2004; Vandekerckhove et al., 2015). Among these methods, the most popular model selection criteria are the Akaike information criterion (AIC; Akaike, 1973, 1974), the Bayesian information criterion (BIC; Schwarz, 1978) and likelihood ratio tests (LRT).

Information Criterion (IC) Indexes — AIC and BIC incorporate a goodness-of-fit term along with a penalty term for overfitting. Hence, each criterion selects the model with best penalized log-likelihood, given by $-2\ln p\left(y|\widehat{\theta}\right) + A_n k$. The first term, $-2\ln p\left(y|\widehat{\theta}\right)$, quantifies the goodness of fit, where y is the dataset and $\widehat{\theta}$ is the maximum likelihood estimate. The second term, $A_n k$, is a penalty for model complexity, where A_n is a constant or a function of the sample size (n) and k is the number of parameters. The information criteria, such as AIC and BIC, differ in the choice of A_n, indicating different degrees of penalty for model complexity.

For AIC, $A_n = 2$, and thus it is expressed as $-2\ln p\left(y|\widehat{\theta}\right) + 2k$, which estimates the degree of information loss due to the probability distribution associated with the true data generation mechanism approximated by the probability distribution with the fitted model. Therefore, the model with the lowest AIC is considered most likely to provide a data generation mechanism close to the true mechanism in the population.

For BIC, the information criterion is expressed as $-2\ln p\left(y|\widehat{\theta}\right) + kln(n)$, with $A_n = ln(n)$, where n is the number of observations. BIC is an estimate of a function of the posterior probability of a model being true with the unit information prior, so a lower BIC means that a model is considered closer to the true model (Dziak et al., 2020). When $n \geq 8$, BIC places more weight on the penalty for model complexity compared to AIC. There is evidence that AIC tends to select complex models that over fit the data, whereas BIC tends to select simpler models that underfit data (Burnham & Anderson, 2002; O'Hagan & Forster, 2004). Matuschek et al. (2017) showed that the model selection approach based on AIC was slightly anticonservative, with empirical Type I error rates exceeding the nominal level. However, the performance of BIC was not investigated in the study.

Likelihood Ratio Test (LRT) — To select an optimal random effect structure, the LRT is another promising approach. Because the regular LRT is overconservative when testing a variance component (Ryoo, 2011). To address this issue, Crainiceanu and Ruppert (2004) derived a spectral representation of the exact finite null distribution of the restricted likelihood ratio test ($RLRT_{CR}$). Using spectral decomposition, the null distribution of the $RLRT_{CR}$ statistic can be simulated rapidly because the distribution only depends on the design matrix of the fixed and random effects. This procedure can be implemented via the *RLRsim* package in R (Scheipl et al., 2008).

Model Search Algorithms

In the context of selecting random effect structures in LMMs, three search algorithms have been used in prior research: forward selection, backward elimination, and all possible structures. The all-possible structures technique involves examining LMMs with every conceivable random effects structure. The best model can be determined based on AIC or BIC. However, if the number of potential random effects is large, this algorithm

could rapidly become expensive to compute and difficult to evaluate. For example, when there are two within-cluster predictors, the number of all possible random effects structures would increase to 14 (see Table 1 for all the possible structures).

Table 1

Candidate Random Effect Structure

M1: $\begin{bmatrix} \tau_{00} & & \\ 0 & 0 & \\ 0 & 0 & 0 \end{bmatrix}$	M2: $\begin{bmatrix} 0 & & \\ 0 & \tau_{11} & \\ 0 & 0 & 0 \end{bmatrix}$	M3: $\begin{bmatrix} 0 & & \\ 0 & 0 & \\ 0 & 0 & \tau_{22} \end{bmatrix}$
M4: $\begin{bmatrix} \tau_{00} & & \\ 0 & \tau_{11} & \\ 0 & 0 & 0 \end{bmatrix}$	M5: $\begin{bmatrix} \tau_{00} & & \\ 0 & 0 & \\ 0 & 0 & \tau_{22} \end{bmatrix}$	M6: $\begin{bmatrix} 0 & & \\ 0 & \tau_{11} & \\ 0 & 0 & \tau_{22} \end{bmatrix}$
M7: $\begin{bmatrix} \tau_{00} & & \\ \tau_{10} & \tau_{11} & \\ 0 & 0 & 0 \end{bmatrix}$	M8: $\begin{bmatrix} \tau_{00} & & \\ 0 & 0 & \\ \tau_{20} & 0 & \tau_{22} \end{bmatrix}$	M9: $\begin{bmatrix} 0 & & \\ 0 & \tau_{11} & \\ 0 & \tau_{21} & \tau_{22} \end{bmatrix}$
M10: $\begin{bmatrix} \tau_{00} & & \\ 0 & \tau_{11} & \\ 0 & 0 & \tau_{22} \end{bmatrix}$	M11: $\begin{bmatrix} \tau_{00} & & \\ \tau_{10} & \tau_{11} & \\ 0 & 0 & \tau_{22} \end{bmatrix}$	M12: $\begin{bmatrix} \tau_{00} & & \\ 0 & \tau_{11} & \\ \tau_{20} & 0 & \tau_{22} \end{bmatrix}$
M13: $\begin{bmatrix} \tau_{00} & & \\ 0 & \tau_{11} & \\ 0 & \tau_{21} & \tau_{22} \end{bmatrix}$	M14: $\begin{bmatrix} \tau_{00} & & \\ \tau_{10} & \tau_{11} & \\ \tau_{20} & \tau_{21} & \tau_{22} \end{bmatrix}$	

In the forward selection method, variances and covariances of random effects are added to a model sequentially based on statistical significance of LRTs. In the context of model selection, the significance level should be interpreted as the relative weight of goodness of fit and model complexity (Matuschek et al., 2017), similar to the trade-off between goodness of fit and model complexity with information criteria. Both Barr et al. (2013) and Matuschek et al. (2017) used the significance level of 0.2 for LRT in model selection.

When using the forward selection method, researchers also need to determine the sequence of adding variances or covariances. For example, when there are two within-cluster predictors (X1 and X2), starting from a random intercept model, one could first add the random slope of X1 or the random slope of X2. If the random slope of X1 is included in the model, in the subsequent step, one can add the covariance between the random intercept and random slope of X1, or add the random slope of X2. However, Barr et al. (2013) found that using an arbitrary sequence led to poor results and, therefore, suggested the "best path" algorithm, which involves testing all possible random effects that are not currently in the model and including any that pass the LRT.

Finally, in the backward elimination method, one starts with the maximal random effects structure, and sequentially eliminates variances and covariances based on LRT.

Similar to the forward selection method, there are multiple sequences that can be used in the backward elimination method. The backward "best-path" algorithm tests all random effects in the maximal model and excludes the statistically non-significant ones. Barr et al. (2013) found that when using the "best path" algorithm, the backward elimination method performed equally well as the forward selection method, but outperformed the latter when using arbitrary sequences.

The Current Study

Existing studies have only included a few approaches for comparison and have used different evaluation criteria, making it challenging to draw definitive conclusions about the best approach. In addition, previous studies have used overly simplistic data generation models, failing to consider Level-2 predictors and control variables. As a result, the generalizability of their findings to more complex models is questionable. Therefore, the current study aims to address the gap by comparing a broader range of approaches and generating data using more complex models.

First, we compare the maximal approach, forward selection, backward elimination, and the all-possible structure approach in terms of their performance in achieving optimal empirical Type I error and power when testing fixed effects associated with Level-1 predictors, Level-2 predictors, and cross-level interactions. Second, we examine whether it is essential to correctly model the random slope of a Level-1 control variable.

Method

Data Generation and Design Factors

We generated data using the following LMM to simulate cross sectional clustering, such as students nested within schools. The model includes five predictors, consisting of two Level-1 predictors, two Level-2 predictors, and one cross-level interaction term. Additionally, multiple random effects were included to represent the complexity of models observed in empirical studies.

Level-1:

$$Y_{ij} = \beta_{0j} + \beta_{1j}X1_j + \beta_{2j}X2_j + e_{ij}$$

Level-2:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}W1_j + \gamma_{02}W2_j + \mu_{0j}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}W1_j + \mu_{1j}$$

$$\beta_{2j} = \gamma_{20} + \mu_{2j}$$

$$\begin{bmatrix}\mu_{0j}\\ \mu_{1j}\\ \mu_{2j}\end{bmatrix} \sim MVN\left(\begin{bmatrix}0\\0\\0\end{bmatrix}, \begin{bmatrix}\tau_{00} & \tau_{01} & \tau_{02}\\ \tau_{10} & \tau_{11} & \tau_{12}\\ \tau_{20} & \tau_{21} & \tau_{22}\end{bmatrix}\right) \quad e_{ij} \sim N(0, \sigma_e^2),$$

where i and j are indexes of participants and clusters, respectively; X1 is the main Level-1 predictor, X2 is the Level-1 control variable, and W1 and W2 are the Level-2 predictors; μ_{0j}, μ_{1j}, and μ_{2j} are the random effects associated with the intercept, X1, and X2 for cluster j, respectively; and e_{ij} is the within-cluster error term. X1 and X2 are generated from a multivariate normal distribution with a mean vector of zeros and a variance of 1.0. The bivariate correlation between X1 and X2 is set to 0.3. The same specifications apply to the Level-2 predictors W1 and W2. The random effects μ_{0j}, μ_{1j}, and μ_{2j} are assumed to follow a multivariate normal distribution with a mean vector of zeros. The variances of the random effects are indicated by τ_{00}, τ_{11}, and τ_{22}, and the covariances by the off-diagonal components τ_{10}, τ_{20} and τ_{21}. The Level-1 error term e_{ij} is assumed to follow an independent and identically normally distributed (i.i.d.) distribution with a mean of zero and a variance of σ_e^2.

The number of clusters (J) was set to 10, 50 or 100, and the cluster size (I) was set to 10, 25, or 50. These values represent the typical range of sample sizes at Level-1 and Level-2 as found in the review of empirical studies using MLMs (Luo et al., 2021). The intercept (γ_{00}) was fixed at 0 since it is typically not of interest. All the other fixed effects (γ_{10}, γ_{20}, γ_{01}, γ_{02}, and γ_{11}) were assigned values of either 0 (no effect) or 0.5 (medium effect). The Level-1 error variance was set to 1, and the variance of the random intercept (τ_{00}) was set to 0.25. We chose to fix τ_{00} at a specific value because the impact of misspecifying the random intercept is not the focus of the investigation, given that researchers typically include a random intercept in their model specifications (Luo et al., 2021).

On the other hand, the variance components associated with random slopes of X1 and X2 (τ_{11} and τ_{22}) were set to 0.05, 0.25 or 0.5, representing conditions where the variance component was trivial, medium, or large. The correlation among random effects was set to 0.5. These specific variance-covariance values were selected to ensure that the intraclass correlations ranged between 0.25 and 0.55, as found to be common in cross-sectional clustered studies in behavioral science by Hedges and Hedberg (2007). Taking into account all the design factors, a total of 162 conditions were considered. In each condition, 1000 independent data sets (i.e., replications) were simulated using R 4.0.3 (R Core Team, 2021).

Model Selection Algorithms

In the all-possible structure approach, we fitted 14 possible random effects structures as shown in Table 1, and the model with the smallest value of AIC or BIC was adopted as the final model.

For the forward selection method, we started with the random intercept model (M1) and used the "best path" algorithm (FWB) as well as 9 arbitrary sequences (FW1-FW9) to construct the model. In the backward elimination method, we started with the maximal model (M14) and used the "best path" algorithm (BWB) as well as 5 arbitrary sequences

(BW1-BW5) to trim the model. The flow charts for all the arbitrary sequences in the forward selection and backward elimination methods are presented in the Supplementary Materials.

All the models were estimated in the R package lme4 (Bates et al., 2015) through restricted maximum likelihood estimation. In the forward and backward search algorithms, the $RLRT_{CR}$ approach was adopted for the significance test of variance components to address the boundary issue. The test was implemented using the R package *RLRsim* (Scheipl et al., 2008). To test covariance components, the regular LRT was performed using the R package *lmtest* (Zeileis & Hothorn, 2002). The significance level for these tests was set at $\alpha = .20$.

The minimum approach (i.e., random-intercept model, M1) and the maximum approach (i.e., full variance-covariance structure, M14) were also examined. The corresponding R codes for all the models (M1 to M14), along with the best-path forward and backward search, are provided in the Supplementary Materials.

Empirical Type I Error Rate and Power

Significance tests for all five predictors were performed using the t-test with Satterthwaite approximation (Satterthwaite, 1946) to correct the degrees of freedom, ensuring more accurate inferential results in conditions with small sample sizes. To evaluate the performance, empirical Type I error rate and power were calculated. With 1000 replications, we expected the empirical Type I error rate to range from .037 to .063 ($.05 \pm 1.96 \times \sqrt{0.05 * 0.95 / 1000}$).

Because the empirical power of an anticonservative approach is inflated, directly comparing the empirical power of approaches that differ in Type I error rate can be misleading. Hence, we adopted the method used in Barr et al. (2013) to compute the corrected empirical power.

Results

Arbitrary Forward Search and Backward Search (FWA and BWA)

The arbitrary sequences used in the forward search approach exhibited similar empirical Type I error rates and corrected power for the significance tests of all the fixed effects. Similarly, minimal differences were found among the sequences used in the arbitrary backward search approach. Therefore, the mean Type I error rate and corrected power across the sequences within the arbitrary forward and backward search approaches were computed and used in the subsequent comparisons with the other approaches.

Empirical Type I Error Rates

Overall Performance

For X1 and X2, the MAX, FWB, BWB, and AIC approaches showed acceptable Type I error rate overall. The FWA, BWA, and BIC approaches exhibited slightly inflated Type I error rates, while the random intercept model (RI) resulted in severely inflated Type I error rate (ranging between 0.08 and 0.75). Regarding W1 and W2, all methods demonstrated slight inflation in the Type I error rate, except for RI. For X1W1, RI showed a significant inflation in the Type I error rate, while the other methods exhibited only slight inflations. The specific observed Type I error rates under different data generation conditions are elaborated upon and presented in Table 2, Table 3, Table 4, and Table 5[1]. More detailed results of the Type I error rate and power for each combination of the design factors were available in Supplementary Materials.

Effect of Sample Size

For the slopes of X1 and X2, FWA, BWA, and BIC all had a slightly inflated Type I error rate when the number of clusters was 10. However, the Type I error rate became acceptable when the number of clusters increased to 25 or above. When the number of clusters increased to 50, all of the approaches (except for RI) had similar Type I error rates.

For W1 and W2, only RI had an acceptable Type I error rate across all conditions. For the other approaches, when the number of clusters was 10, they all had unacceptably high Type I error rates, especially when the approach was MAX, FWB, BWB, AIC, or BIC. However, the Type I error rate became acceptable when the number of clusters increased to 25 or above, and the differences among the approaches diminished when the number of clusters increased to 50.

For X1W1, all of the methods had unacceptably high Type I error rates when the number of clusters was 10, among which RI, AIC, and BIC were the worst. However, the Type I error rate decreased and became similar among the approaches (except for RI) when the number of clusters increased to 25 or above.

On the other hand, the cluster size (CS) had little impact on the Type I error rate of the coefficients of X1 and X2, except when the approach was RI or BIC. For RI, the Type I error rate notably increased when CS increased. For BIC, the Type I error rate was unacceptably high when CS = 10, and decreased to an acceptable range when CS increased to 50 or above.

For W1 and W2, regardless of the cluster size, RI had the best performance, followed by FWA and BWA which had slightly inflated but acceptable Type I error rates. The other approaches all had unacceptably high Type I error rate, which tended to increase as

1) Because the results for W1 and W2 were similar, we only presented those for W1.

CS increased. The only exception was BIC, which had a smaller Type I error rate for W1 and W2 when the CS increased.

For X1W1, FWA and BWA had the best performance, with an acceptable Type I error rate regardless of the cluster size. MAX, FWB, and BWB had an acceptable Type I error rate only when CS was 50 or below. AIC and BIC all had unacceptably high Type I error rates regardless of CS.

Effect of Variance Components (τ11 and τ22)

The magnitude of τ11 and τ22 had little impact on the Type I error rate across all the methods, except for BIC and RI. Specifically, BIC had an unacceptably high Type I error rate for X1 when τ11 was small (0.05), and similarly for X2 when τ22 was small. For RI, which had the highest Type I error rate among all methods, the Type I error rate for a level-1 predictor (e.g., X1) was positively related to the magnitude of its variance component (e.g., τ11) and negatively related to the variance component of the slope of the other level-1 predictor (e.g., τ22).

For X1W1, only FWA and BWA had an acceptable Type I error rate regardless of the magnitude of τ11 and τ22. MAX, FWB and BWB had an acceptable Type I error rate when τ11 and τ22 were small. AIC and BIC had an unacceptably high Type I error rate, especially when τ11 was small and τ22 was large. RI had an unacceptably high Type I error rate in all conditions, which was exacerbated as τ11 increased.

Methodology
2023, Vol. 19(4), 365–386
https://doi.org/10.5964/meth.9601

Table 2

Type I Error and Power for X1 Slope

	RI		MAX		FWA		FWB		BWA		BWB		AIC		BIC	
Condition	**Type I Error**	**Corr. Power**	**Type I Error**	**Corr. Power**	**Type I Error**	**Corr. Power**	**Type I Error**	**Corr. Power**	**Type I Error**	**Corr. Power**	**Type I Error**	**Corr. Power**	**Type I Error**	**Corr. Power**	**Type I Error**	**Corr. Power**
Number of Clusters (*N*)																
10	0.414	0.712	0.050	0.671	0.071	0.666	0.053	0.669	0.073	0.669	0.051	0.671	0.061	0.652	0.077	0.624
25	0.404	0.948	0.049	0.950	0.057	0.950	0.049	0.950	0.057	0.950	0.049	0.950	0.050	0.950	0.053	0.947
50	0.404	0.997	0.052	0.998	0.054	0.998	0.053	0.998	0.054	0.998	0.052	0.998	0.052	0.998	0.053	0.998
Cluster Size (*n*)																
10	0.207	0.838	0.051	0.829	0.059	0.825	0.054	0.827	0.059	0.827	0.052	0.829	0.060	0.814	0.071	0.798
50	0.450	0.905	0.051	0.891	0.061	0.890	0.051	0.891	0.062	0.891	0.051	0.891	0.054	0.889	0.060	0.876
100	0.566	0.914	0.049	0.899	0.062	0.899	0.049	0.899	0.063	0.898	0.049	0.899	0.050	0.897	0.052	0.896
$\tau 11$																
0.05	0.236	0.984	0.049	0.977	0.059	0.977	0.050	0.977	0.060	0.977	0.049	0.977	0.058	0.975	0.075	0.965
0.25	0.454	0.895	0.052	0.880	0.061	0.877	0.053	0.878	0.062	0.879	0.052	0.880	0.054	0.867	0.057	0.858
0.50	0.533	0.779	0.051	0.762	0.061	0.761	0.051	0.761	0.063	0.761	0.051	0.762	0.051	0.758	0.051	0.746
$\tau 22$																
0.05	0.430	0.895	0.051	0.874	0.059	0.872	0.052	0.873	0.060	0.873	0.051	0.874	0.054	0.871	0.057	0.862
0.25	0.405	0.885	0.050	0.873	0.061	0.873	0.051	0.873	0.063	0.873	0.051	0.874	0.055	0.864	0.061	0.855
0.50	0.387	0.877	0.050	0.872	0.061	0.869	0.051	0.870	0.062	0.871	0.050	0.871	0.055	0.866	0.064	0.853

Methodology
2023, Vol. 19(4), 365–386
https://doi.org/10.5964/meth.9601

Table 3

Type I Error and Power for X2 Slope

	RI		MAX		FWA		FWB		BWA		BWB		AIC		BIC	
Condition	**Type I Error**	**Corr. Power**	**Type I Error**	**Corr. Power**	**Type I Error**	**Corr. Power**	**Type I Error**	**Corr. Power**	**Type I Error**	**Corr. Power**	**Type I Error**	**Corr. Power**	**Type I Error**	**Corr. Power**	**Type I Error**	**Corr. Power**
Number of Clusters (*N*)																
10	0.406	0.755	0.049	0.708	0.071	0.704	0.052	0.705	0.073	0.706	0.051	0.708	0.058	0.694	0.073	0.669
25	0.402	0.956	0.049	0.957	0.057	0.956	0.050	0.957	0.057	0.957	0.049	0.957	0.050	0.956	0.054	0.955
50	0.402	0.997	0.050	0.998	0.054	0.998	0.050	0.998	0.054	0.998	0.050	0.998	0.051	0.998	0.051	0.998
Cluster Size (*n*)																
10	0.201	0.859	0.048	0.852	0.059	0.847	0.053	0.848	0.059	0.849	0.051	0.851	0.058	0.837	0.069	0.819
50	0.450	0.920	0.050	0.905	0.061	0.905	0.050	0.905	0.062	0.905	0.050	0.905	0.051	0.905	0.058	0.897
100	0.560	0.928	0.050	0.906	0.062	0.906	0.050	0.906	0.063	0.906	0.050	0.906	0.050	0.906	0.051	0.906
τ11																
0.05	0.425	0.907	0.050	0.886	0.059	0.884	0.052	0.886	0.060	0.884	0.051	0.886	0.053	0.881	0.059	0.873
0.25	0.402	0.901	0.048	0.888	0.061	0.886	0.050	0.887	0.062	0.888	0.049	0.888	0.052	0.883	0.059	0.876
0.50	0.383	0.900	0.049	0.889	0.061	0.888	0.051	0.887	0.062	0.889	0.051	0.889	0.053	0.884	0.061	0.873
τ22																
0.05	0.229	0.992	0.045	0.988	0.059	0.987	0.049	0.987	0.060	0.987	0.047	0.988	0.054	0.986	0.070	0.982
0.25	0.443	0.914	0.053	0.895	0.061	0.892	0.054	0.893	0.063	0.894	0.053	0.895	0.055	0.884	0.058	0.872
0.50	0.538	0.802	0.050	0.780	0.061	0.779	0.050	0.780	0.062	0.779	0.050	0.780	0.050	0.778	0.051	0.768

Methodology
2023, Vol. 19(4), 365–386
https://doi.org/10.5964/meth.9601

Table 4

Type I Error and Power for W1 Slope

	RI		MAX		FWA		FWB		BWA		BWB		AIC		BIC	
Condition	**Type I Error**	**Corr. Power**	**Type I Error**	**Corr. Power**	**Type I Error**	**Corr. Power**	**Type I Error**	**Corr. Power**	**Type I Error**	**Corr. Power**	**Type I Error**	**Corr. Power**	**Type I Error**	**Corr. Power**	**Type I Error**	**Corr. Power**
Number of Clusters (*N*)																
10	0.050	0.517	0.089	0.540	0.071	0.533	0.087	0.539	0.075	0.535	0.087	0.540	0.095	0.499	0.091	0.473
25	0.049	0.943	0.062	0.970	0.058	0.968	0.062	0.969	0.058	0.969	0.062	0.968	0.064	0.967	0.064	0.952
50	0.049	0.999	0.054	1.000	0.054	1.000	0.054	1.000	0.054	1.000	0.054	1.000	0.055	1.000	0.057	1.000
Cluster Size (*n*)																
10	0.049	0.772	0.063	0.793	0.059	0.790	0.062	0.792	0.059	0.791	0.062	0.792	0.070	0.768	0.081	0.753
50	0.048	0.842	0.070	0.854	0.061	0.851	0.071	0.853	0.063	0.852	0.070	0.853	0.072	0.844	0.065	0.828
100	0.051	0.845	0.072	0.862	0.062	0.859	0.071	0.862	0.064	0.859	0.071	0.862	0.071	0.854	0.066	0.844
$\tau 11$																
0.05	0.049	0.824	0.067	0.837	0.060	0.834	0.067	0.835	0.061	0.834	0.068	0.835	0.070	0.823	0.068	0.814
0.25	0.050	0.819	0.068	0.838	0.061	0.835	0.067	0.839	0.062	0.837	0.067	0.840	0.071	0.821	0.072	0.805
0.50	0.049	0.816	0.070	0.834	0.062	0.831	0.069	0.820	0.063	0.832	0.069	0.834	0.072	0.822	0.072	0.806
$\tau 22$																
0.05	0.051	0.825	0.067	0.835	0.060	0.834	0.066	0.834	0.060	0.835	0.067	0.834	0.070	0.822	0.069	0.809
0.25	0.050	0.818	0.069	0.820	0.062	0.837	0.068	0.841	0.063	0.839	0.068	0.841	0.070	0.808	0.070	0.813
0.50	0.048	0.816	0.070	0.824	0.061	0.829	0.069	0.833	0.063	0.829	0.069	0.833	0.073	0.810	0.073	0.802

Methodology
2023, Vol. 19(4), 365–386
https://doi.org/10.5964/meth.9601

Table 5

Type I Error and Power for X1W1 Slope

	RI		MAX		FWA		FWB		BWA		BWB		AIC		BIC	
Condition	**Type I Error**	**Corr. Power**	**Type I Error**	**Corr. Power**	**Type I Error**	**Corr. Power**	**Type I Error**	**Corr. Power**	**Type I Error**	**Corr. Power**	**Type I Error**	**Corr. Power**	**Type I Error**	**Corr. Power**	**Type I Error**	**Corr. Power**
Number of Clusters (*N*)																
10	0.478	0.628	0.074	0.649	0.071	0.636	0.075	0.640	0.074	0.641	0.075	0.641	0.081	0.621	0.089	0.579
25	0.482	0.906	0.060	0.952	0.058	0.949	0.060	0.951	0.058	0.950	0.060	0.950	0.063	0.949	0.065	0.937
50	0.474	0.988	0.055	0.998	0.054	0.998	0.055	0.998	0.054	1.000	0.054	0.998	0.056	0.998	0.060	0.997
Cluster Size (*n*)																
10	0.256	0.793	0.060	0.815	0.059	0.806	0.061	0.808	0.059	0.810	0.061	0.810	0.068	0.797	0.079	0.780
50	0.530	0.861	0.063	0.889	0.061	0.886	0.063	0.887	0.063	0.887	0.063	0.887	0.065	0.882	0.069	0.859
100	0.650	0.869	0.066	0.895	0.062	0.892	0.066	0.893	0.064	0.893	0.066	0.846	0.066	0.890	0.066	0.874
$\tau 11$																
0.05	0.336	0.958	0.061	0.966	0.059	0.964	0.062	0.964	0.060	0.964	0.062	0.964	0.068	0.961	0.082	0.948
0.25	0.514	0.843	0.064	0.867	0.061	0.859	0.064	0.862	0.062	0.861	0.064	0.862	0.067	0.852	0.069	0.831
0.50	0.584	0.721	0.064	0.766	0.062	0.760	0.088	0.762	0.063	0.763	0.064	0.763	0.065	0.756	0.064	0.734
$\tau 22$																
0.05	0.467	0.870	0.060	0.867	0.060	0.862	0.061	0.863	0.060	0.864	0.060	0.864	0.063	0.856	0.067	0.840
0.25	0.482	0.836	0.065	0.863	0.062	0.858	0.066	0.861	0.063	0.859	0.066	0.861	0.069	0.855	0.073	0.834
0.50	0.484	0.817	0.064	0.868	0.061	0.863	0.064	0.865	0.062	0.866	0.064	0.866	0.067	0.858	0.074	0.840

Corrected Power

The Slope of X1 and X2

Because the Type I error rate of the RI approach was unacceptably high when testing the slopes of X1 and X2 across the board, it was not considered when examining the corrected power. Among the remaining methods, in general, MAX had the best performance across all conditions because it achieved higher corrected power while maintaining the Type I error rate at the nominal level. Although the other methods did not perform as well as MAX, the differences were less than 0.10 among the conditions with an acceptable Type I error rate. It was noted that BIC had the worst performance as it tended to have lower corrected power as well as unacceptably high Type I error rate when the sample size was small.

The Slopes of W1 and W2

When the number of clusters was small (i.e., 10 clusters), the RI approach performed the best because it had the highest corrected power among all the methods with acceptable Type I error rates. Under this condition, BIC generally had the worst performance because it had an unacceptably high Type I error rate and the lowest corrected power. When the number of clusters increased to 25, MAX showed small advantages over RI. Although MAX had the highest corrected power most frequently, while RI had the lowest corrected power most frequently, the differences between the two approaches were small, ranging from 0.010 to 0.065 with an average of 0.029. Finally, when the number of clusters increased to 50, all methods reached the maximum power.

The Slope of X1W1

Among the methods that can maintain an acceptable Type I error rate, no single method stood out in terms of the corrected power as the differences between these methods were negligible (i.e., less than 0.03). However, it is worth noting that BIC had the worst performance because it not only had an unacceptably high Type I error rate but also the lowest corrected power, especially when the number of clusters was not large (i.e., 25 or less).

The Impact of Under-Specifying the Random Effects of the Level-1 Control Variable

When the random slope of the control variable (X2) was misspecified as fixed (i.e., Model 4 or Model 7), the Type I error rate of the slope of X1 was unacceptably low when the number of clusters was small (i.e., 10) and the magnitude of $\tau 22$ was large (i.e., $\tau 22 = 0.5$), which led to decreased power in those conditions. As shown in Table 6[2], the difference in the corrected power for the slope of X1 between Model 4 and the MAX approach (i.e., Model 14) ranged between 0.09 and 0.21 when the number of clusters was small and the

magnitude of $\tau 22$ was large. However, when the number of clusters increased to 25 or above, there was no noteworthy differences between the two models.

For the slope of X1W1, although the under-specification of the random effect of X2 did not negatively affect the Type I error rate, the corrected power was more than 10% lower than that under Model 14 when the number of clusters was small and the magnitude of $\tau 22$ was large.

2) The results based on Model 7 were almost identical to Model 4, hence only Model 4 results were presented in Table 6.

Methodology
2023, Vol. 19(4), 365–386
https://doi.org/10.5964/meth.9601

Table 6

Type I Error and Power for the Slope of X1 and X1W1 Under Model 4 and Model 14

			Type I Error				Corrected Power			
			X1		X1W1		X1		X1W1	
Cluster	Cluster Size	t22	M4	M14	M4	M14	M4	M14	M4	M14
10	10	.05	0.044	0.050	0.052	0.063	0.559	0.566	0.537	0.545
10	10	.25	0.033	0.049	0.051	0.067	0.489	0.567	0.463	0.533
10	10	.50	0.026	0.046	0.041	0.068	0.436	0.576	0.438	0.544
10	50	.05	0.037	0.046	0.045	0.064	0.694	0.726	0.658	0.711
10	50	.25	0.029	0.054	0.054	0.078	0.640	0.710	0.588	0.706
10	50	.50	0.026	0.055	0.049	0.089	0.581	0.696	0.545	0.681
10	100	.05	0.036	0.050	0.044	0.072	0.699	0.732	0.668	0.713
10	100	.25	0.029	0.051	0.058	0.094	0.652	0.729	0.604	0.677
10	100	.50	0.016	0.047	0.041	0.074	0.605	0.736	0.563	0.735
25	10	.05	0.047	0.056	0.049	0.057	0.914	0.916	0.898	0.905
25	10	.25	0.033	0.053	0.047	0.061	0.896	0.929	0.855	0.912
25	10	.50	0.026	0.052	0.052	0.058	0.860	0.921	0.830	0.916
25	50	.05	0.038	0.050	0.051	0.058	0.956	0.964	0.929	0.963
25	50	.25	0.021	0.047	0.046	0.060	0.946	0.966	0.915	0.969
25	50	.50	0.020	0.046	0.056	0.061	0.930	0.961	0.888	0.971
25	100	.05	0.030	0.040	0.046	0.060	0.962	0.970	0.944	0.975
25	100	.25	0.027	0.049	0.055	0.059	0.946	0.960	0.918	0.979
25	100	.50	0.017	0.049	0.058	0.062	0.924	0.961	0.893	0.975

			Type I Error				Corrected Power			
			X1		X1W1		X1		X1W1	
Cluster	Cluster Size	t22	M4	M14	M4	M14	M4	M14	M4	M14
50	10	.05	0.037	0.049	0.052	0.056	0.995	0.996	0.990	0.993
50	10	.25	0.028	0.052	0.054	0.055	0.990	0.995	0.983	0.995
50	10	.50	0.025	0.054	0.047	0.055	0.988	0.996	0.977	0.996
50	50	.05	0.042	0.059	0.047	0.053	0.999	1.000	0.997	0.999
50	50	.25	0.023	0.048	0.045	0.051	0.999	0.999	0.996	1.000
50	50	.50	0.013	0.051	0.049	0.051	0.997	0.999	0.989	0.999
50	100	.05	0.041	0.058	0.048	0.059	0.999	0.999	0.998	1.000
50	100	.25	0.019	0.049	0.054	0.061	0.998	0.999	0.990	1.000
50	100	.50	0.014	0.048	0.047	0.053	0.998	0.999	0.993	0.999

Methodology
2023, Vol. 19(4), 365–386
https://doi.org/10.5964/meth.9601

Discussion

The present study examined statistical inferences of fixed effects in MLM when the random effect structure was specified based on the minimal approach, the maximal approach, and the model selection approach. Our findings regarding the fixed effects of Level-1 predictors were consistent with Barr et al. (2013), which showed that (1) the maximal model can control the Type I error rate without a significant loss of power, and (2) the best-path algorithm in the forward search and backward elimination approach performed equally well as the maximal approach.

Our findings also corroborated Matuschek et al. (2017)'s finding regarding the performance of AIC-based all-subset model selection, which has been shown to suffer from slightly inflated Type I error rates in conditions with small sample sizes. Although Matuschek et al. (2017) did not examine BIC, our findings showed that BIC had similar and sometimes worse performance compared to AIC.

On the other hand, unlike Matuschek et al. (2017), we did not find noticeable advantages in the statistical power for testing the fixed effect of X1 and X2 when using the arbitrary backward elimination or forward selection approach. This is likely to be due to the use of corrected power in the evaluation.

Regarding the cross-level interaction effect, our findings were consistent with Heisig and Schaeffer (2019) in that the minimal approach led to severely inflated Type I error rates. However, we did not find clear evidence to show that the maximal approach significantly outperformed the model selection approach. Other than AIC- and BIC-based model selection, which had higher Type I error rates and lower corrected power, there was little difference among the other methods.

Regarding Level-2 predictors, we found that it is sufficient to use the minimal approach (i.e., a random intercept model) to obtain a controlled Type I error rate, which is consistent with the finding from a recent study by Huang and Li (2022). On the contrary, the maximal approach and model selection approach tended to cause inflated Type I error rates, especially when the number of clusters is small.

Finally, our findings showed that it is important to consider random slopes of control variables. If the random slope of the Level-1 control variable (X2) was misspecified as fixed, the tests of the slope of the main Level-1 predictor (X1) and the cross-level interaction (X1W1) became too conservative, leading to decreased power, especially when the number of clusters was small (i.e., 10) and the magnitude of $\tau 22$ was large (i.e., $\tau 22 = 0.5$). Correctly specifying the random slope of a Level-1 control variable can reduce the error variance, thus enhancing the statistical power of the tests for the main Level-1 predictor. Hence, the working assumption proposed by Barr et al. (2013) that it is not essential to specify random effects for control variables to avoid anticonservative inference did not fully capture the role of control variables and should be revised based on the new findings.

Implications

First, if the predictor of interest is at the within-cluster level or involves a cross-level interaction, the maximal approach, the best-path forward search, and the best-path backward elimination are all desirable methods for specifying random effect structures. If there are convergence issues with estimating the maximal model, the best-path forward search is recommended. Second, if the predictor of interest is at the cluster level, it is not essential to specify random slopes of Level-1 predictors. Third, it is important to specify random slopes of within-cluster control variables, especially when the sample size is small and the variance of the random slope is large. Lastly, researchers should avoid using AIC- and BIC-based approaches, especially when the sample size is small.

Limitations and Future Directions

The findings of the current study should be interpreted in light of the limitations. First, our study focused on cross-sectional clustered data. It is not safe to directly apply our findings to studies with panel/longitudinal data, which may involve more complex Level-1 error structures (e.g., auto-correlated Level-1 errors). Second, we generated data in which the distributional assumptions of random effects were satisfied. It is unknown whether the performance of these approach would differ when there are violations of the normality assumptions of the random effects. These could be the directions for future studies.

Funding: The authors have no funding to report.

Acknowledgments: The authors have no additional (i.e., non-financial) support to report.

Competing Interests: The authors have declared that no competing interests exist.

Supplementary Materials

For this article, the materials provided are the arbitrary sequences flowcharts for the forward selection and backward elimination methods, and the corresponding R codes for the study's models and best-path forward and backward searches (see Luo & Li, 2023).

Index of Supplementary Materials

Luo, W., & Li, H. (2023). *Supplementary materials to "A comparison of methods for specifying optimal random effects structures"* [Flowcharts, R code]. PsychOpen GOLD. https://doi.org/10.23668/psycharchives.13976

References

Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Second International Symposium on Information Theory* (pp. 267–281). Akadémiai Kiadó.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

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. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., & White, J. S. S. (2009). Generalized linear mixed models: A practical guide for ecology and evolution. *Trends in Ecology & Evolution, 24*(3), 127–135. https://doi.org/10.1016/j.tree.2008.10.008

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information–theoretic approach* (2nd ed.). Springer.

Crainiceanu, C. M., & Ruppert, D. (2004). Likelihood ratio tests in linear mixed models with one variance component. *Journal of the Royal Statistical Society: Series B. Statistical Methodology, 66*(1), 165–185. https://doi.org/10.1111/j.1467-9868.2004.00438.x

Dziak, J. J., Coffman, D. L., Lanza, S. T., Li, R., & Jermiin, L. S. (2020). Sensitivity and specificity of information criteria. *Briefings in Bioinformatics, 21*(2), 553–565. https://doi.org/10.1093/bib/bbz016

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group randomized trials in education. *Educational Evaluation and Policy Analysis, 29*(1), 60–87. https://doi.org/10.3102/0162373707299706

Heisig, J. P., & Schaeffer, M. (2019). Why you should always include a random slope for the lower-level variable involved in a cross-level interaction. *European Sociological Review, 35*(2), 258–279. https://doi.org/10.1093/esr/jcy053

Huang, F. L., & Li, X. (2022). Using cluster-robust standard errors when analyzing group-randomized trials with few clusters. *Behavior Research Methods, 54*(3), 1181–1199. https://doi.org/10.3758/s13428-021-01627-0

Kuha, J. (2004). AIC and BIC: Comparisons of assumptions and performance. *Sociological Methods & Research, 33*(2), 188–229. https://doi.org/10.1177/0049124103262065

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

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language, 94*, 305–315. https://doi.org/10.1016/j.jml.2017.01.001

O'Hagan, A., & Forster, J. (2004). *Kendall's advanced theory of statistics, Vol. 2B: Bayesian inference* (2nd ed.). Arnold.

R Core Team. (2021). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Ryoo, J. H. (2011). Model selection with the linear mixed model for longitudinal data. *Multivariate Behavioral Research, 46*(4), 598–624. https://doi.org/10.1080/00273171.2011.589264

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin, 2*(6), 110–114. https://doi.org/10.2307/3002019

Scheipl, F., Greven, S., & Küchenhoff, H. (2008). Size and power of tests for a zero random effect variance or polynomial regression in additive and linear mixed models. *Computational Statistics & Data Analysis, 52*(7), 3283–3299. https://doi.org/10.1016/j.csda.2007.10.022

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Stroup, W. W. (2012). *Generalized linear mixed models: Modern concepts, methods and applications.* CRC Press.

Vandekerckhove, J., Matzke, D., & Wagenmakers, E.-J. (2015). Model comparison and the principle of parsimony. In J. R. Busemeyer, Z. Wang, J. T. Townsend, & A. Eidels (Eds.), *Oxford handbook of computational and mathematical psychology* (pp. 300–320). Oxford University Press.

Zeileis, A., & Hothorn, T. (2002). Diagnostic checking in regression relationships. *R News, 2*(3), 7–10. https://CRAN.R-project.org/doc/Rnews/