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

PsychOpen GOLD

Original Article

Check for updates

Specifying the Random Effect Structure in Linear Mixed Effect Models for Analyzing Psycholinguistic Data

Jungkyu Park[a], Ramsey Cardwell[b], Hsiu-Ting Yu[cd]

[a] *Department of Psycholgy, Kyungpook National University, Daegu, South Korea.* **[b]** *Department of Educational Research Methodology, University of North Carolina at Greensboro, Greensboro, USA.* **[c]** *Department of Psychology, National Chengchi University, Taipei, Taiwan.* **[d]** *Research Center for Mind, Brain & Learning, National Chengchi University, Taipei, Taiwan.*

Methodology, 2020, Vol. 16(2), 92–111, https://doi.org/10.5964/meth.2809

Received: 2018-07-08 • **Accepted:** 2019-09-14 • **Published (VoR):** 2020-06-18

Corresponding Author: Hsiu-Ting Yu, Department of Psychology, National Chengchi University, No. 64 Sec 2, Zhinan Rd., Wenshan Dist., Taipei City 11605, Taiwan. Phone: (0) + 886-2-2938-7384, E-mail: hsiutingyu@gmail.com, htyu@nccu.edu.tw

Abstract

Linear Mixed Effect Models (LMEM) have become a popular method for analyzing nested experimental data, which are often encountered in psycholinguistics and other fields. This approach allows experimental results to be generalized to the greater population of both subjects and experimental stimuli. In an influential paper Bar and his colleagues (2013; https://doi.org/10.1016/j.jml.2012.11.001) recommend specifying the maximal random effect structure allowed by the experimental design, which includes random intercepts and random slopes for all within-subjects and within-items experimental factors, as well as correlations between the random effects components. The goal of this paper is to formally investigate whether their recommendations can be generalized to wider variety of experimental conditions. The simulation results revealed that complex models (i.e., with more parameters) lead to a dramatic increase in the non-convergence rate. Furthermore, AIC and BIC were found to select the true model in the majority of cases, although selection accuracy varied by LMEM random effect structure.

Keywords

linear mixed-effect models, psycholinguistic data, random effect structure, model specification, random effects

In psycholinguistic studies, a common outcome measure is reaction time (RT). For example, subjects might judge whether strings of letters are words or non-words, indicating their decision by pressing a button. The real words represent different categories, constituting the experimental manipulation. The nature of psycholinguistic studies necessitates accounting for variability in the outcome variable caused by particular subjects and items (i.e., by-subject and by-item random effects).

Classical methods for analyzing psycholinguistic data are by-subjects or by-items analysis of variance (ANOVA), known as F_1 and F_2 (Clark, 1973). The distinct feature of these methods is that subjects and items are treated as fixed factors. However, this choice might be inappropriate, as the interest of most studies is generalizing conclusions beyond the particular individuals and stimuli used in the experiment. Thus, both subjects and stimuli should be treated as random factors when analyzing data from such experiments.

The quasi *F*-ratio (*F'*; Clark, 1973) was an early attempt at solving this problem. The *F'* and min-*F* statistics account for both item and subject variability (e.g., Forster & Dickinson, 1976; Santa, Miller, & Shaw, 1979; Wickens & Keppel, 1983). However, this method requires balanced data with no missing responses and computing by-subject and by-item effects separately.

The recently-popularized linear mixed effects model (LMEM) enables simultaneous modeling of by-subject and by-item random effects while handling missing data better than previous ANOVA methods (Baayen, Davidson, & Bates, 2008). Such random effects include random intercepts, reflecting differences in the overall level of the dependent variable across experimental units (e.g., subjects or items) and/or random slopes, reflecting differences in the effects of predictors across the experimental units.

Although LMEMs are receiving increased attention in psycholinguistics, there is no widely-accepted rule for determining the random effects to include in an LMEM, causing confusion and inconsistent use of LMEMs. To provide practical recommendations concerning the choice of random effect structure, Barr, Levy, Scheepers, and Tily (2013) conducted a simulation study. They manipulated the number of items, whether treatment was within- or between-items, and the presence of a treatment effect as the experimental factors, resulting in eight experimental conditions. The simulated data sets from the eight conditions were then analyzed with min-*F'*, F_1, $F_1 \times F_2$, and several LMEMs differing in the complexity of the random effects component.

Based on the results, the authors suggested that LMEMs with the most complex random effect structure justified by the design (i.e., maximal LMEM) should be implemented, providing the model converges. Their results showed that maximal LMEMs performed well in terms of type I error rate, power, and model selection accuracy, while random-intercept-only models performed worse on all criteria, and usually even worse than separate F_1 and F_2 tests. On the other hand, Matuschek, Kliegl, Vasishth, Baayen, and Bates (2017) demonstrated that fitting the maximal model successfully controls the Type I error, but leads to a significant loss of power. They noted that higher power can be achieved without inflating Type I error if information criteria is used to select the optimal model.

Although innovative and informative, Barr et al. (2013) was limited to relatively simple conditions, including only one predictor with a continuous outcome variable in the model. It is unclear whether their recommendations regarding choice of random effects generalize to more complex experimental designs common in contemporary psy-

cholinguistic studies. When fitting LMEMs under complex experimental designs, a model non-convergence issue may arise. Several papers noted the increasing likelihood of non-convergence as models become more complex (e.g., Baayen, Vasishth, Kliegl, & Bates, 2017; Bates, Kliegl, Vasishth, & Baayen, 2015). This problem has also been encountered in a number of empirical applications of LMEMs. For example, Nixon, Chen, and Schiller (2015) investigated the processing of tones in Chinese. In this study, the maximal model with reaction time as the response variable and three predictors (trial number, stimulus onset asynchrony, and experimental condition) failed to converge. After simplifying the model by removing random effect correlations, the model still did not converge.

For cases in which a maximal model does not converge, Barr et al. (2013) proposed a simple alternative: keeping all random slopes for the predictor of interest and fixing correlations among all random effect components to zero. In the follow-up paper, Barr (2013) suggested that the "keeping maximal model with zero correlations" strategy can also be applicable when higher-order interaction terms are included in LMEMs.

The purpose of this paper is to systematically investigate the applicability of the Barr et al. (2013) recommendation under complex experimental conditions, and to provide practical recommendations for selecting the random effect structure when analyzing psycholinguistic data with LMEMs. The remainder of this section briefly introduces essential concepts of LMEMs and explains the hypotheses and approach of the present study. The details of the simulation study are provided, followed by the results. Finally, the results are discussed, including recommendations and future directions.

Linear Mixed Effects Models

A LMEM can be formally specified as follows,

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{b} + \varepsilon \tag{1}$$

In Equation (1), **X** is a design matrix encoding all factor contrasts and predictors, which is multiplied by the vector of population coefficients (β) containing the overall intercept and main-effect slope(s). What differentiates a LMEM from linear regression is the term **Zb**, where **Z** is another design matrix and **b** is a vector of subject and item random effects assumed to be normally distributed with mean 0 and variance-covariance of the two random effects. These random effects adjust the intercept and/or slope for each subject and each item. This might be clarified by an example using different notation. The term ε is the residual errors that are assumed to follow a normal distribution with mean 0 and variance of σ^2.

If a hypothetical study were both within-subjects (i.e., each subject sees all items) and within-item (i.e., each item occurs in all experimental conditions), the corresponding LMEM could be expressed as follows:

$$Y_{si} = \beta_0 + S_{0s} + I_{0i} + (\beta_1 + S_{1s} + I_{1i})\, X_i + e_{si} \tag{2}$$

In Equation (2), the response of a subject to an item is modeled as the fixed-effect intercept (β_0) and fixed-effect slope (β_1) plus random effects, where "fixed effect" means that the value does not change for different subjects and items. The fixed-effect intercept represents the average response value under one of the experimental conditions, while the fixed-effect slope represents the mean difference between the two experimental conditions (i.e., the overall treatment effect). The random intercepts adjust the baseline response value for each subject (S_{0s}) and item (I_{0i}). The random slopes adjust the treatment effect for each subject (S_{1s}) and item (I_{1i}). Consequently, a different response is predicted for each unique subject–item combination. While this is the conceptual basis of LMEM, the actual model-fitting procedure does not estimate individual subject and item random effects, instead estimating the population variance-covariance matrices of random subject effects (Equation 3) and random item effects (Equation 4), where τ_{00}^2 and ω_{00}^2 represent the variances of the by-subject and by-item intercept distributions respectively. The parameters τ_{11}^2 and ω_{11}^2 likewise represent the by-subject and by-item slope distribution variances. The parameters ρ_s and ρ_i represent the correlation between subject random slopes and intercepts, and item random slopes and intercepts respectively. Formally, these parameters are specified as:

$$T = \begin{pmatrix} \tau_{00}^2 & \rho_s \tau_{00} \tau_{11} \\ \rho_s \tau_{00} \tau_{11} & \tau_{11}^2 \end{pmatrix} \tag{3}$$

$$\Omega = \begin{pmatrix} \omega_{00}^2 & \rho_i \omega_{00} \omega_{11} \\ \rho_i \omega_{00} \omega_{11} & \omega_{11}^2 \end{pmatrix} \tag{4}$$

Multiple software packages can implement LMEMs, including SAS (e.g., Yu, 2015), and R (e.g., Bates, Maechler, Bolker, & Walker, 2015). Several books and papers (e.g., Baayen et al., 2008) are often credited with popularizing the use of the R package *lme4* (Bates, 2005; Bates, Maechler, Bolker, & Walker, 2015) for applying LMEMs to psycholinguistic data. More recently, promising new tools for fitting LMEMs have become available. For example, *MixedModels.jl*, an LMEM package coded in the Julia programming language (https://github.com/dmbates/MixedModels.jl). Additionally, LMEMs can be implemented in a Bayesian framework (Sorensen, Hohenstein, & Vasishth, 2016) using the programming language *Stan* (Stan Development Team, 2016).

Evaluation Criteria

Three evaluation criteria – parameter estimation accuracy, non-convergence rate, and model selection accuracy – were considered to evaluate different random effect structures in LMEMs. We first assessed parameter estimate accuracy of both fixed and random effects under the random effect structures included in the simulation study.

The second criterion was non-convergence rate. Non-convergence occurs when the model-fitting algorithm fails to reach a solution within the specified number of iterations or stopping criteria. In real-world experiments, it is closely related to model complexity, especially concerning the random effects structure; non-convergence is more likely to occur as models become more complex, such as when additional parameters are estimated, including item order effects, extra random components.

The third evaluation criterion is model selection accuracy. Model selection means selecting a statistical model from candidate models for interpretation of the results. The candidate models should ideally be grounded in sound theory, and thus researchers should develop several theory-based candidate models for comparison using objective model-selection techniques (Vallejo, Tuero-Herrero, Núñez, & Rosário, 2014). In the LMEM context, model selection aims to determine the fixed and random effects to include in the model. However, the best method for selecting the random component structure is particularly unclear (Barr et al., 2013; Yu, 2015).

Simulation Study

Design Factors

A simulation study was designed to investigate non-convergence rates and model selection accuracy in LMEMs. Two factors were varied to generate datasets from nine models (Table 1) with various LMEM structures: (1) number of predictors and (2) random component complexity. The numbers in Table 1 indicate the number of binary predictors (one, two, or three) in the model.

Table 1

Summary of Examined Models

	Number of predictors		
Random structure complexity	**1X**	**2X**	**3X**
Random-intercept only (A)	A1	A2	A3
No random correlations (B)	B1	B2	B3
Maximal (C)	C1	C2	C3

Additionally, the three random effects structures considered in this study – random-intercept-only model, random intercepts and slopes without random correlations in the covariance matrix, and the maximal random effects model – are designated A, B, and C, respectively. The letter designations follow the order of increasing model complexity, i.e., the intercept-only model A was extended to create the no-correlation model B by

adding by-subject and by-item random slopes. The introduction of correlations between random intercepts and random slopes gives the maximal model C. These correlation coefficients represent conceptually the pairwise relatedness between random intercepts and slopes. They are part of the off-diagonal terms in the random effect covariance matrices. In the three-predictor maximal (C3) model, the random correlation coefficients add 56 parameters compared to the three-predictor no-correlation (B3) model.

The combination of number of predictors and random effect component letter creates unique identifiers for each model. For example, the random-intercept-only model with two predictors is labeled A2. In each of the nine simulation conditions and generating models, both the number of subjects and items was 24, equivalent to the larger-sample conditions in Barr et al. (2013).

Analysis

The simulation design included nine combinations of the levels of the two manipulated factors. As in Barr et al. (2013), all predictors were binary and deviation coded (-0.5, 0.5). Generated models with multiple predictors also included all possible two- and three-way interactions in both fixed and random components. As in Barr et al. (2013), all fixed-effect parameters were set to 0.8. The parameter sampling ranges from Barr et al. (2013) were used to maximize comparability of results. However, for maximal models with more than one fixed effect, independent random sampling of random effect variances and correlation coefficients can produce covariance matrices that are not positive definite. To circumvent this problem, we instead randomly sampled the eigenvalues of the random component covariance matrix, then simulated the covariance matrix from these eigenvalues (Varadhan, 2008). The eigenvalues were restricted such that the random effect variances did not exceed 3, the upper limit of the variance sampling range used in the non-maximal models. The parameters used in data simulation are summarized in Table 2.

Table 2

Parameter Values for Fixed Effect and Parameter Sampling Ranges for Random Effects

Parameter	Description	Value
β_0	Grand-average intercept	~*U*(-3,3)
β	Grand slopes X1, X2, etc	0.8
τ^2	By-subject random effect variances	~*U*(0,3)
ω^2	By-item random effect variances	~*U*(0,3)
λ	Random effect matrix eigenvalues	~*U*(0,4)
ρ	Correlation between by-subject and by-item random effects	~*U*(-1,1)
σ^2	Residual error	~*U*(0,3)

After generating 1,000 datasets from each of the nine models A1 through C3, each dataset was fit with four models differing in complexity of the random component. The simplest model had no random component, i.e., a fixed-effects-only regression model. This basic model was included to provide an additional candidate for model selection, and so that all LMEMs considered would have a simpler alternate model. In all fitted models, the fixed-effect structure was specified such that the fitted model included the same number of predictors and interactions as the model from which the data were generated.

From each fitted model, we recorded model convergence and the model's AIC (Akaike, 1974) and BIC (Schwarz, 1978). A model was recorded as not converging if the model-fitting process returned a convergence warning message based on lme4's default convergence criteria; otherwise, the model was considered to have converged. To formally test the effects of our experimental factors on convergence, we performed logistic regression with model non-convergence as the outcome variable (1 = non-convergence, 0 = convergence). The number of predictors and classifications of fitted models (i.e., underfit/true model/overfit) were also included as predictors to capture the possible effect of the number of predictors and model mismatch on non-convergence.[1]

We also performed logistic regression to evaluate the effects of our experimental factors on whether a fitted model was "selected" using AIC and BIC from the candidate models (i.e., the model had the lowest IC value). In each regression analysis, the outcome variable was a binary variable in which 1 represented selection using the particular IC. For the analysis, a model was first fit containing model match, number of predictors (as two dummy-coded variables), and generated model (as two dummy-coded variables) as predictors.

All LMEMs were fit using the *lmer* function of the *R* package *lme4* version 1.1-7 in *R* version 3.1.3 (R Core Team, 2013). All default settings of the function were used, including the REML estimation method, the optimizer BOBYQA with the default tolerance value of 0.02, and default starting values.

Results

Accuracy of Parameter Estimation

Table 3 presents mean fixed-effects estimates and corresponding standard errors (*SE*) for all nine generated models. The results show that the mean fixed-effects estimates were close to the true values (0.8) under most conditions, indicating that they are robust to changes in the random effects specification. This result is consistent with previous stud-

1) The number of predictors in the simulation was included as two dummy-coded binary predictors, one-predictor (1 = model with 1 predictor) and three-predictors (1 = model with 3 predictors). Likewise, two dummy-coded binary variables - Underfit (1 = fitted model less complex than generated model) and Overfit (1 = fitted model more complex than generated model) - were also included as predictors.

ies demonstrating robust fixed-effects estimates when the outcome variable is continuous and normally distributed (McCulloch, Searle, & Neuhaus, 2008; Verbeke & Lessafre, 1997).

Table 3

Mean Fixed-Effects Estimates and SES for all Nine Generated Models

Model	β_0 (*SE*)	β_1 (*SE*)	β_2 (*SE*)	β_3 (*SE*)	β_4 (*SE*)	β_5 (*SE*)	β_6 (*SE*)	β_7 (*SE*)
A1	0.795 (0.073)	0.798 (0.014)	-	-	-	-	-	-
B1	0.788 (0.072)	0.778 (0.068)	-	-	-	-	-	-
C1	0.791 (0.072)	0.805 (0.077)	-	-	-	-	-	-
A2	0.802 (0.070)	0.800 (0.011)	0.802 (0.010)	0.799 (0.020)	-	-	-	-
B2	0.799 (0.073)	0.783 (0.071)	0.784 (0.072)	0.806 (0.075)	-	-	-	-
C2	0.791 (0.073)	0.795 (0.076)	0.808 (0.073)	0.787 (0.077)	-	-	-	-
A3	0.809 (0.069)	0.799 (0.008)	0.799 (0.007)	0.800 (0.007)	0.801 (0.014)	0.800 (0.014)	0.802 (0.015)	0.796 (0.030)
B3	0.790 (0.074)	0.808 (0.071)	0.798 (0.073)	0.799 (0.071)	0.780 (0.071)	0.798 (0.073)	0.806 (0.074)	0.803 (0.076)
C3	0.790 (0.073)	0.795 (0.079)	0.806 (0.074)	0.786 (0.075)	0.820 (0.074)	0.813 (0.077)	0.808 (0.072)	0.817 (0.074)

Figures 1, 2, and 3 display the mean estimates of each random component variance in the nine generated models, with error bars representing *SE*. The expected value of all random effect variances is 1.5, as the variances were randomly sampled from $U(0,3)$. The Figures reveal that the random effect variances (τ^2 and ω^2) were also well recovered.

Figure 1

The Mean Estimates of Random Effect Variances for All 1-Predictor Generated Models (A1, B1, and C1), Error Bars Represent 1 Standard Error (SE)

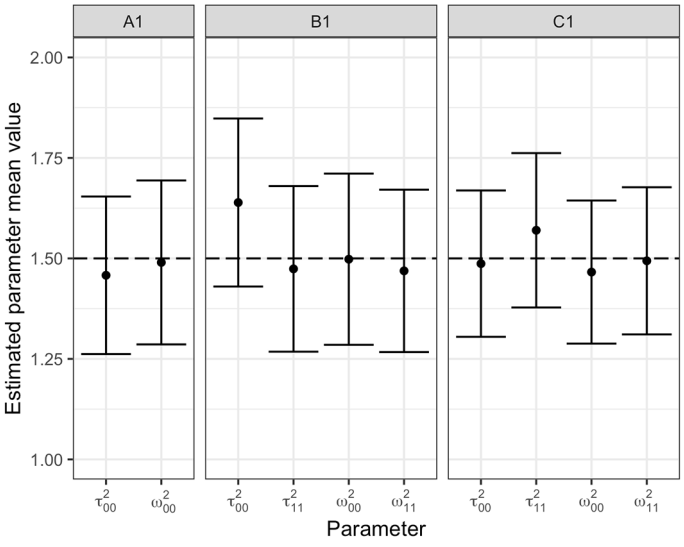


Figure 2

The Mean Estimates of Random Effect Variances for All 2-Predictor Generated Models (A2, B2, and C2), Error Bars Represent 1 Standard Error (SE)

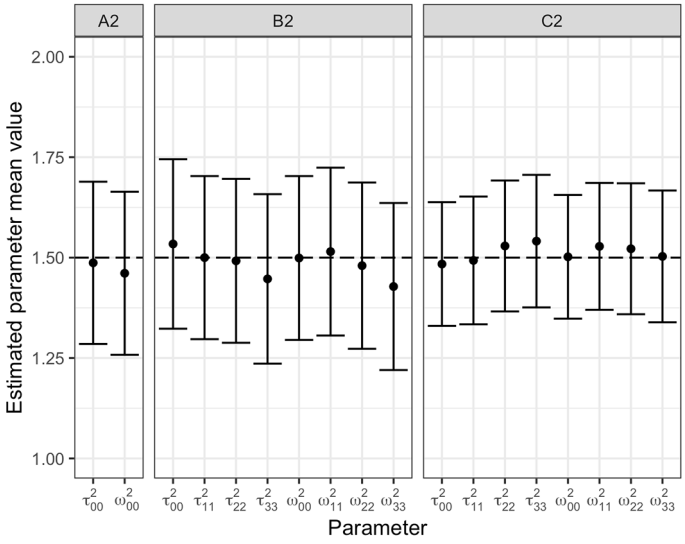


Figure 3

The Mean Estimates of Random Effect Variances for All 3-Predictor Generated Models (A3, B3, and C3), Error Bars Represent 1 Standard Error (SE)

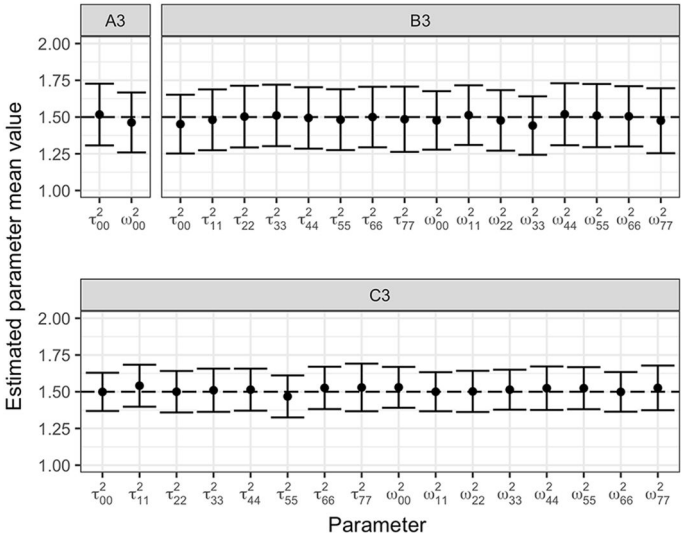
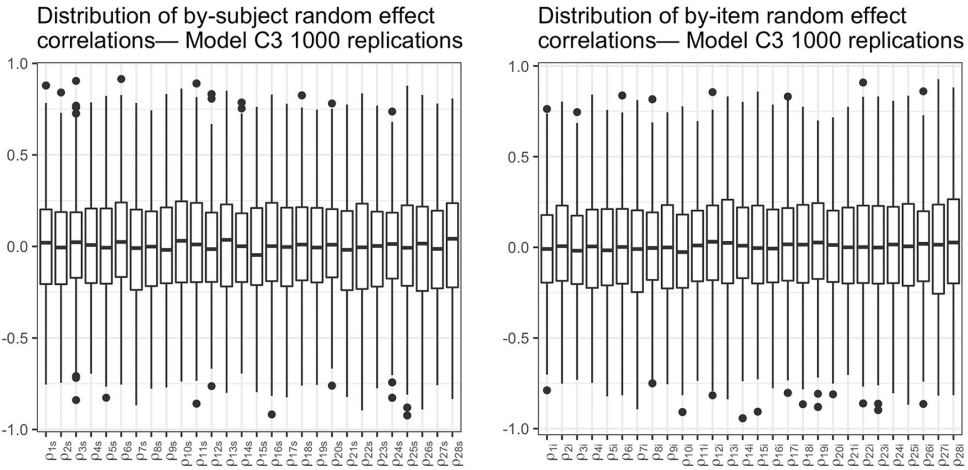


Figure 4 shows that the average values of correlations between random intercepts and random slopes are all close to zero and the distributions nearly span the full range of possible values (-1 to 1), indicating that the data was successfully generated with random effect correlations.

Figure 4

Distribution of Model Estimates for All 28 by-Subject (Left) and by-Item (Right) Random Correlation Coefficients



Model Non-Convergence Rates

Table 4 summarizes the model non-convergence rate for each combination of generated model, fitted model, and number of predictors. Each number represents the proportion of simulation runs in the corresponding category that did not reach convergence. None of the random-intercept-only models (A1, A2, and A3) fit to data experienced non-convergence. For no-correlation models (B1, B2, and B3), non-convergence rates were relatively low, ranging from 0 to .023. However, there is a steady increase in the non-convergence rate of B models fit to A datasets with increasing number of predictors, from .002 in the one-predictor condition to .023 in the three-predictor condition. Some non-convergence occurred in all conditions fitting the maximal (C) model, with a dramatic increase in non-convergence for C models fit to A datasets as the number of predictors increased.

Table 4

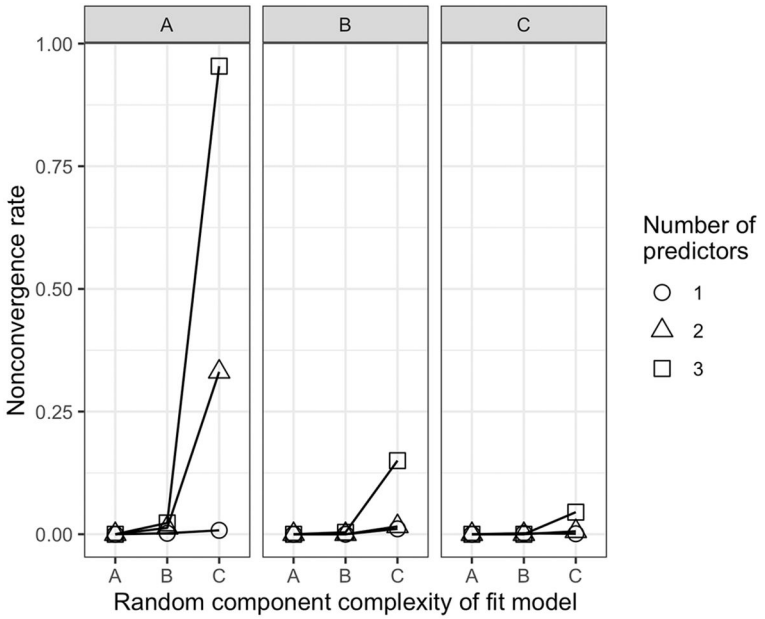
Model Non-Convergence Rates for All Combinations of Generated and Fitted

	Fitted Model								
Generated Model	**A1**	**B1**	**C1**	**A2**	**B2**	**C2**	**A3**	**B3**	**C3**
A1	0	0.002	0.008						
B1	0	0	0.011						
C1	0	0.002	0.001						
A2				0	0.013	0.331			
B2				0	0	0.016			
C2				0	0	0.006			
A3							0	0.023	0.954
B3							0	0.004	0.150
C3							0	0	0.045

Each pair of generated and fitted models can be classified as underfit, overfit, and true models. The diagonal of the table corresponds to fitting the "true model" (i.e., the fitted model is the same as the generated model). Non-convergence rates are relatively low in the true model scenarios, despite a slight increase as additional predictors are added. Below the diagonal, which represents when models were underfit (i.e., the model was less complex than the data), non-convergence only occurred twice, so the non-convergence rate is effectively zero regardless of the number of predictors. Almost all cases of non-convergence are above the diagonal, representing when models were overfit (i.e., the model was more complex than the data). Overfit models suffered from higher non-convergence rates than other model pairs with the same number of predictors, and also displayed a trend of dramatically increasing non-convergence rates with the inclusion of additional predictors. When displayed visually (Figure 5), it is even more apparent that non-convergence occurs most often for maximal models with multiple predictors, especially when overfitting.

Figure 5

The Model Non-Convergence Rates of All 27 Combinations of Generated and Fitted Models



Note. The letters in grey boxes at the top of each panel represent the generated models, while the letters along the X-axis correspond to the fitted models. The different lines represent the number of predictors in the model (1, 2, or 3).

In Table 4, the upper left block summarizing the one-predictor models corresponds to the models used in Barr et al. (2013). While the authors did not formally investigate non-convergence, they mentioned that convergence rates exceeded 99% in all experimental conditions, with the lowest rate being 99.61% convergence. Thus, our observed non-convergence rates in the one-predictor models are consistent with those observed by Barr et al. (2013). It is only in the two- and three-predictor conditions, which go beyond the scope of Barr et al.'s study, that we observed a dramatic increase in model non-convergence. Specifically, when data generated from A models were fit with a C model, representing substantial model overfitting, the rates of non-convergence were 33.1% and 95.4% in the two- and three-predictor conditions respectively.

Logistic regression analysis reveals that both the number of predictors and the relationship between the generated and fitted models are independently significant predictors of model non-convergence. Models are more likely to result in non-convergence when overfit, as compared to the true model, by a factor of nearly 50 ($Z = 11.50$, $p < .001$). Non-convergence was also significantly more likely for three-predictor models

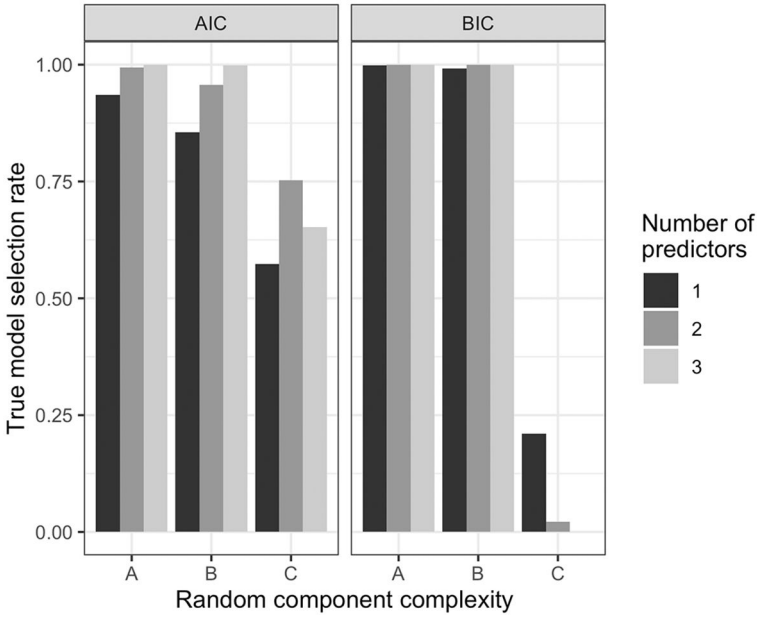
(Z = 4.85, p < .001) and significantly less likely for one-predictor models (Z = -13.38, p < .001) in comparison to two-predictor models, reflecting the observed trend of increasing non-convergence as the number of predictors increased.

Model Selection Using Information Criteria

In addition to model convergence, another research concern of this paper is using information criteria (AIC and BIC) to select the best-fitting model. In each replication, four models differing only in complexity of the random-effects component were fit to a dataset, one of which corresponded to the "true" model. Theoretically, the true model should be the best fitting of the four candidate models. Figure 6 displays the proportion of simulation runs in which the true model was selected (i.e., had the lowest IC value) among the four candidate models using AIC and BIC (hereafter "true model selection rate").

Figure 6

True Model Selection Rate Using AIC and BIC for All Nine Generated Models



Using AIC, the true model selection rate was generally high (> 85%) for random-intercept-only (A) and no-correlation (B) models, with the rate for models A3 and B3 reaching 100% and 99.9% respectively. For maximal (C) models, the rate was lower, ranging from 57.3% to 75.3%. The overall true model selection rate for AIC was 85.8%. Using BIC, the

true model selection rate was uniformly high (> 99%) for A and B models. However, the rate for C models was substantially lower, ranging from 21% (C1) to 0% (C3). When the true model was not selected in C model conditions (except two simulation runs in the C1 condition), the corresponding B model was selected by BIC. The overall true model selection rate for BIC was 69.1%. So overall, AIC and BIC selected the true model in the majority of cases, otherwise selecting the next simplest model.

Logistic regression was also implemented to predict model selection using AIC. A logistic regression model was first fitted containing model match, number of predictors (as two dummy-coded variables) and generated model (as two dummy-coded variables) as predictors. The saturated model containing model match and number of predictors as parameters was subjected to model simplification via stepwise Likelihood-ratio test (LRT), resulting in the final model with significant effects for Model Match, one-Predictor, and their interaction (Table 5).

Table 5

Results of the Logistic Regression Analysis Predicting Model Selection Using AIC

Variable	β (*SE*)	95% CI (Lower)	Odds	95% CI (Upper)
Model Match	4.98 (0.06)	130.54	146.16***	163.94
One-Predictor	0.74 (0.06)	1.87	2.09***	2.34
Model Match × One-Pred	-1.54 (0.08)	0.18	0.21***	0.25

Note. R^2 = .68 (Nagelkerke); R^2 =.49 (Cox & Snell).
***p < .001.

The final model includes model match and one-predictor and their interaction. Model Match is the primary variable that predicts AIC selection, with the true model being more likely to be selected by a factor of 146.16 (Z = 85.80, p < .001). The significant negative coefficient of the Model Match × One-Predictor interaction indicates that the true model was significantly less likely to be selected by AIC when the model only contains one predictor (Z = -18.28, p < .001). In other words, AIC was less accurate at selecting the true model in one-predictor conditions compared to two-predictor conditions.

The logistic regression results for BIC differed more starkly based on generated model than by number of predictors. Therefore, model match and generated model were used as predictors. The saturated model was subjected to model simplification via stepwise LRT, resulting in the final model containing Model Match, C-Generated (a binary dummy variable indicating whether the data were generated from a maximal model), and the Model Match × C-Generated interaction (Table 6).

Table 6

Results of the Logistic Regression Analysis Predicting Model Selection using BIC

Variable	β (*SE*)	95% CI (Lower)	Odds	95% CI (Upper)
Match	13.82 (0.49)	413688	1001297***	2826149
C-Generated	7.22 (0.44)	629.54	1364.12***	3560.81
B-Generated	0.12 (0.49)	0.430	1.13	3.00
Match × C-Gen	-16.14 (0.49)	0.000	0.000***	0.000

Note. R^2 = .82 (Nagelkerke); R^2 = .59 (Cox & Snell).
***p < .001.

The true model is extremely unlikely to be selected for C datasets (Z = -32.79, p < .001), reflecting the extremely low rates of true model selection observed for maximal (C) datasets. Since the model includes an interaction, coefficients for the single predictors can be interpreted as the effect when the variable with which it interacts is 0. Therefore, the coefficient of Model Match represents the effect of model match for random-intercept-only (A) and no-correlation (B) datasets, indicating that BIC is almost guaranteed to select the true model in the A and B model conditions (Z = 28.39, p < .001). Finally, the significant coefficient of C-Generated represents the very high odds that BIC will select a model other than the true model in the case of C datasets (Z = 16.49, p < .001).

Conclusion and Discussion

Barr et al. (2013) has been quite influential in the field of psycholinguistics. The authors conducted a simulation study generating data from a model with one binary predictor and a continuous outcome variable, concluding that the maximal model should be the "gold standard" when using LMEMs for confirmatory hypothesis testing because it achieved the Type I error rate closest to .05 under all conditions, while also having greater or similar corrected *Power* (a metric devised by the authors) compared to other methods.

The primary differences between our study and Barr et al. (2013) are the number of predictors in the generated model and the dependent variables of interest. While Barr et al. used a one-predictor model, we generated data from one-, two-, and three-predictor models. Furthermore, whereas Barr et al. (2013) compared fitted models based on Type I error and Power, we recorded model non-convergence, AIC, and BIC. Our results are therefore not directly comparable to those of Barr et al., and we cannot contest their findings concerning Type I error and Power in the models included in both studies. Rather, the goal of the present study is to investigate whether their recommendations can be generalized to more complex conditions by investigating a potential obstacle to their implementation (non-convergence) and a potential alternative (IC-based model selection) that minimizes non-convergence caused by overfitting.

The results showed that overall non-convergence rates increase with the addition of predictors. There was no clear pattern in non-convergence rates under the one-predictor conditions. However, when the simulation scenarios were expanded to include two and three predictors, two intertwined patterns of model non-convergence emerged. Furthermore, a significant effect of number of predictors was found in logistic regression analysis of the non-convergence data, with models with fewer predictors less likely to experience non-convergence.

Upon closer inspection of the non-convergence rates within models containing the same number of predictors, another pattern emerges: when a dataset is fit with its true model, non-convergence is very infrequent. In underfit models, there is essentially no non-convergence. In overfit models, however, non-convergence is considerably more common. This pattern of non-convergence is reflected in the logistic regression analysis by the significant effect of model match, showing that a true model is significantly less likely to suffer non-convergence.

The two patterns of non-convergence also form an interaction. There is no change in the non-convergence rate of underfit models as the number of predictors increases. For true models, there is a small increase in non-convergence with additional predictors. For overfit models, there is a more dramatic increase in non-convergence rates with increasing numbers of predictors. This significant interaction between model overfit and number of predictors is relevant to the recommendations of Barr et al. (2013) because the maximal model recommended therein cannot logically be underfit if it includes all relevant variables and the maximal random-effects structure. The maximal model likely overfits the data in many cases, and over-fitted models suffer the most from model non-convergence. Our findings thus support the application of the keeping maximal model with zero correlation strategy (Barr et al., 2013; Barr, 2013) in cases where a maximal model does not converge. Our results show that applying this strategy successfully controls the non-convergence rate even when a model is overfitted. Specifically, when fitting to data generated from the three-predictor random intercept model (A3), the suggested strategy (i.e., fitting B3 instead of C3 model) would drastically reduce the non-convergence rate from 95.4% to 2.3%.

Our results show that, overall, AIC selected the true model in 86% of simulated cases, while the success rate for BIC is 69%, supporting that AIC is overall more consistently accurate, with increasing accuracy as the number of predictors increased. However, AIC's true model selection rate was lower in maximal model conditions.

BIC's true model selection rate displayed a different pattern; BIC was extremely accurate at selecting the true model for data generated from random-intercept-only models and no-correlation models, with rates for these conditions all exceeding 99%. Only in the maximal model conditions did BIC perform poorly.

The issue of non-convergence presents a major obstacle to implementing the advice of Barr et al. (2013) to fit the LMEM with the maximal random-effects structure. Non-

convergence is especially problematic when the model overfits the data and includes multiple predictors. However, the back-up strategy proposed by Barr et al. (2013) provides a helpful solution when the maximal model fails to converge. As models with multiple predictors and higher-order interactions are utilized frequently in social science studies, the recommendation of constraining random-component correlations to zero while retaining all random slopes in the model will improve model convergence. Furthermore, recent software for LMEM such as *MixedModels.jl* and Bayesian LMEM using Stan also present alternative ways to circumvent the problem of non-convergence when fitting the maximal model.

In this study, AIC and BIC have proven to be useful tools for selecting the optimal random-effects structure under different conditions. Specifically, AIC performs well when the random-effects structure of the fitted model is more complex, while BIC is preferable under conditions when the fitted model is relatively simple. These IC are usually in the default output of most statistical software for LMEM, and are thus a realistic means for selecting the best-fitting model.

Crucially, when fitting LMEMs, studies suggest that a priori knowledge about the random-effects structure is important for gauging the potential risk of overfitting and non-convergence, even though the true random-effects structure is usually ''unknown.'' We suggest paying careful attention to the methodological literature on current LMEM best practices, substantive knowledge on the research topic, as well as information from visualization techniques (e.g., residuals pattern) and model criticism, as these would help in making a more confident decision concerning the appropriate random-effects structure.

There are limitations to the present study, which should be considered when weighing our recommendations. First, the generated simulation conditions may not reflect the wide array of scenarios in empirical studies, and thus researchers should interpret the results with caution and not overgeneralize the findings.

Second, we did not investigate Type I error rate and Power, so we cannot say if Barr et al.'s findings of the maximal model's superior performance on these metrics generalizes to models containing more predictors. Furthermore, it is unclear whether our method for generating positive definite random-effects covariance matrices was adequate in approximating the maximal model. Finally, we only considered model selection using AIC and BIC in isolation, without considering other ICs or using multiple ICs simultaneously to select a model. Future research could rectify these shortcomings.

Funding: Hsiu-Ting Yu's work is partially supported by the Ministry of Science and Technology, Taiwan. [MOST-108-2401-H-004-100].

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

Acknowledgments: The authors have no support to report.

Author Contribution: All authors contributed equally to the work.

References

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

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*, 390-412. https://doi.org/10.1016/j.jml.2007.12.005

Baayen, R. H., Vasishth, S., Kliegl, R., & Bates, D. (2017). The cave of shadows: Addressing the human factor with generalized additive mixed models. *Journal of Memory and Language, 94*, 206-234. https://doi.org/10.1016/j.jml.2016.11.006

Barr, D. J. (2013). Random effects structure for testing interactions in linear mixed-effects models. *Frontiers in Psychology, 4*, Article 328. https://doi.org/10.3389/fpsyg.2013.00328

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*, 255-278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D. M. (2005). Fitting linear mixed models in R. *R News, 5*, 27-30. https://cran.r-project.org/doc/Rnews/Rnews_2005-1.pdf

Bates, D., Kliegl, R., Vasishth, S., & Baayen, R. H. (2015). *Parsimonious mixed models* (arXiv:1506.04967 [stat.ME]). Retrieved from https://arxiv.org/abs/1506.04967

Bates, D., Maechler, 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

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior, 12*, 335-359. https://doi.org/10.1016/S0022-5371(73)80014-3

Forster, K. I., & Dickinson, R. G. (1976). More on the language-as-fixed-effect fallacy: Monte Carlo estimates of error rates for F1, F2, F′, and min F′. *Journal of Verbal Learning and Verbal Behavior, 15*, 135-142. https://doi.org/10.1016/0022-5371(76)90014-1

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, R. 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

McCulloch, C. E., Searle, S. R., & Neuhaus, J. M. (2008). *Generalized, linear, and mixed models.* Hoboken, NJ, USA: Wiley.

Nixon, J. S., Chen, Y., & Schiller, N. O. (2015). Multi-level processing of phonetic variants in speech production and visual word processing: Evidence from Mandarin lexical tones. *Language, Cognition and Neuroscience, 30*, 491-505. https://doi.org/10.1080/23273798.2014.942326

R Core Team. (2013). *R: A language and environment for statistical computing* [R Foundation for Statistical Computing, Vienna, Austria]. Retrieved from http://www.R-project.org/

Santa, J. L., Miller, J. J., & Shaw, M. L. (1979). Using Quasi *F* to prevent alpha inflation due to stimulus variation. *Psychological Bulletin, 86*, 37-46. https://doi.org/10.1037/0033-2909.86.1.37

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

Sorensen, T., Hohenstein, S., & Vasishth, S. (2016). Bayesian linear mixed models using Stan: A tutorial for psychologists, linguists, and cognitive scientists. *The Quantitative Methods for Psychology, 12*, 175-200. https://doi.org/10.20982/tqmp.12.3.p175

Stan Development Team. (2016). Stan modeling language Users Guide and Reference Manual (Version 2.12.0.). Retrieved from http://mc-stan.org/

Vallejo, G., Tuero-Herrero, E., Núñez, J. C., & Rosário, P. (2014). Performance evaluation of recent information criteria for selecting multilevel models in Behavioral and Social Sciences. *International Journal of Clinical and Health Psychology, 14*, 48-57. https://doi.org/10.1016/S1697-2600(14)70036-5

Varadhan, R. (2008). [R] how to randomly generate a n by n positive definite matrix in R? Retrieved from https://stat.ethz.ch/pipermail/r-help/2008-February/153708.html

Verbeke, G., & Lessafre, E. (1997). The effect of misspecifying the random effects distribution in linear mixed models for longitudinal data. *Computational Statistics & Data Analysis, 23*, 541-556. https://doi.org/10.1016/S0167-9473(96)00047-3

Wickens, T. D., & Keppel, G. (1983). On the choice of design and of test statistic in the analysis of experiments with sampled materials. *Journal of Verbal Learning and Verbal Behavior, 22*, 296-309. https://doi.org/10.1016/S0022-5371(83)90208-6

Yu, H.-T. (2015). Applying linear mixed effects models with crossed random effects to psycholinguistic data: Multilevel specification and model selection. *The Quantitative Methods for Psychology, 11*, 78-88. https://doi.org/10.20982/tqmp.11.2.p078

Methodology
2020, Vol.16(2), 92–111
https://doi.org/10.5964/meth.2809