Model Selection and Model Averaging for Mixed-Effects Models with Crossed Random Effects for Subjects and Items

José Á. Martínez-Huertas[a], Ricardo Olmos[a], & Emilio Ferrer[b]

a=Universidad Autónoma de Madrid; b=University of California, Davis

Multivariate Behavioral Research

This paper is not the copy of record and may not exactly replicate the final, authoritative version of the article. The final article will be available, upon publication, via its DOI: https://doi.org/10.1080/00273171.2021.1889946

ABSTRACT

A good deal of experimental research is characterized by the presence of random effects on subjects and items. A standard modeling approach that includes such sources of variability is the mixed-effects models (MEMs) with crossed random effects. However, under-parameterizing or over-parameterizing the random structure of MEMs bias the estimations of the *Standard Errors* (SEs) of fixed effects. In this simulation study, we examined two different but complementary perspectives: model selection with likelihood-ratio tests, AIC, and BIC; and model averaging with Akaike weights. Results showed that true model selection was constant across the different strategies examined (including ML and REML estimators). However, sample size and variance of random slopes were found to explain true model selection and SE bias of fixed effects. No relevant differences in SE bias were found for model selection and model averaging. Sample size and variance of random slopes interacted with the estimator to explain SE bias. Only the within-subjects effect showed significant underestimation of SEs with smaller number of items and larger item random slopes. SE bias was higher for ML than REML, but the variability of SE bias was the opposite. Such variability can be translated into high rates of unacceptable bias in many replications.

Keywords: mixed-effects models; crossed random effects; random slopes; model selection; model averaging; ML; REML.

INTRODUCTION

Hierarchical data can have complex random structures when, for example, sampled students are nested within neighborhoods and schools, but schools can have students from different neighborhoods and students from the same neighborhood can go to different schools. Both neighborhoods and schools could reflect part of the variability of the responses of sampled students. Usually, these complex random structures require crossed random effects models to properly analyze data (e.g., Raudenbush, 1993). A good deal of experimental research in the social sciences involves collecting measures that results in such random structures. As an example, consider a hypothetical experiment on reaction time where some individuals make decisions about the semantic similarity of different word pairs. In this experiment, differences across people could be modeled using subject random effects (e.g., individuals could systematically differ in their response to the same experimental items). Similarly, differences across items could be modeled using random effects because each stimulus has its own idiosyncrasy (e.g., an item could elicit systematically faster responses than another item). As such, much experimental research can be characterized by the presence of random effects on both subjects and items.

This fact has been embraced by psycholinguistic researchers for decades (see Baayen, Davidson, & Bates, 2008). Both psycholinguistics and other experimental researchers are usually interested in specific task effects, and model random components of both subjects and items to avoid potential bias in the fixed effects when their variances are ignored (see a detailed explanation below; see also Hoffman, 2015; Hox, Moerbeek, & Van de Schoot, 2018; Meyers & Beretvas, 2006). They are often times interested in estimating such random components as informative parameters of the psychological processes of interest (e.g., testing if experimental effects present variability between subjects or items; see also Barr, 2013 for a similar rationale on the use of random effects as a confirmatory hypothesis testing approach).

A standard modeling approach to capture random effects in the data are the mixed-effects models (MEMs). Here, fixed effects identify the systematic relations between independent and dependent variables, whereas random effects quantify the heterogeneity or variability in intercepts or slopes of different clusters[1] (e.g., Raudenbush & Bryk, 2002; Hoffman, 2015; Pardo & Ruiz, 2012). In psycholinguistics and experimental research, different strategies have been proposed to analyze subjects/items variances using Analysis of Variance (ANOVA; see for example the F_1 X F_2 criterion, Clark, 1973, Raaijmakers, Schrijnemakers, & Gremmen, 1999), but these techniques are not able to properly analyze subjects and items random slope variances simultaneously in the presence of incomplete or unbalanced data (see Hoffman, 2015 for a complete rationale on the difference between least squares and likelihood-based estimations). For this reason, MEMs with crossed random effects have become a standard method to analyze experimental data that include random slopes for subjects and items simultaneously using likelihood-based estimations (e.g., Baayen et al., 2008; Hoffman & Rovine, 2007; Quené & van den Bergh, 2004; Bates, Kliegl, Vasishth, & Baayen, 2015).

Despite the flexibility and wide use of MEMs, a model specification is needed that is optimal for the data in order to avoid errors of statistical inference (McNeish & Kelley, 2019; McNeish, Stapleton, & Silverman, 2017). For example, an important consequence of under-estimating or over-estimating the random structure of MEMs is biasing the estimation of the *Standard Errors* (SEs) and *p-values* of the fixed effects. Estimated fixed effects in MEMs are similar across different random structures, but the SEs and the *p-values* of those fixed effects are a function of which random effect variances are estimated (see, for example, Hoffman, 2015; Hox et al., 2018; Meyers & Beretvas, 2006). Because of this, improving the estimation

[1] In some contexts, fixed and random effects of MEMs are called models for the mean and variance, respectively.

of the SEs and *p-values* of fixed effects is critical for making correct decisions about analyses of experimental data. Corrections for SEs of fixed effects have been proposed, for example, for situations when one crossed random effect is ignored (Lai, 2019). However, the true model is unknown and, thus, it is unclear which are the correct SEs estimations.

In the present study we focus on true model selection and estimation of SEs of fixed effects in situations in which random effects are assumed to exist but the need to include crossed random effects (e.g., random effects for subjects and items) is not known. This is a common situation, for example, in experimental research such as psycholinguistics where variation across items and subjects is key to understand the psychological processes underlying the data. In particular, we study true model selection and the bias of SEs of fixed effects in MEMs with crossed random effects in conditions when the true variance components are not known. For this, we use two different but complementary perspectives: model selection based on likelihood-ratio tests and model averaging based on Akaike weights. Model selection strategies try to find the model that best fits the data. Model averaging approaches combine information from the competing models to provide an optimal set of estimates (see for example: Burnham & Anderson, 2002, 2004; Kaplan & Lee, 2018; Konishi & Kitagawa, 2008).

Model Selection: Bottom-Up vs. Top-Down

Although, in theory, random effects should be included as justified by the experimental design[2] (Barr, Levy, Scheepers, & Tily, 2013), the true population model is almost never known. Thus, selecting the right model can be difficult when facing with various

[2] Note that random effects are relevant for different reasons. One of them is their relevance to estimate SEs and *p-values* of fixed effects. Other more common reason is they provide information about variance in the data. Not only the presence of relevant random effects can be useful, but can be the focus of hypothesis testing. This aspect is particularly relevant because the parameterization of random effects should be considered a confirmatory hypothesis testing itself (Barr, 2013).

competing models. One of the most common criterion for model selection is the likelihood-ratio test. Using this test, two standard strategies for model selection typically used are the bottom-up and the top-down model selection scheme. The bottom-up strategy starts with the simplest model, and likelihood-ratio tests are used to decide if adding a random effect improves the model fit. The top-down strategy starts with the more complex model, and likelihood-ratio tests are used to decide if deleting a random effect worsens the model fit. Both the bottom-up and the top-down strategies are based on the deviance (−2LL) and the likelihood-ratio test as a goodness-of-fit measure for model selection. However, they have different starting points. Top-down consists in reducing the complexity of random effects (Barr et al., 2013), whereas bottom-up consists in increasing the complexity of random effects (Matuschek, Kliegl, Vasishth, Baayen, & Bates, 2017). Each model selection strategy requires the same number of model comparisons, but they have different starting points, which make the order of comparisons different. Although, intuitively it may appear as if both strategies would converge on a similar model, it is not clear if each strategy favors simpler or more complex models due to their different starting points. See the *Data analysis* section as an example of how we applied model selection with both bottom-up and top-down strategies in this study.

Model Averaging: Akaike Weights

As an alternative to selecting a model based on its fit to the data, model averaging proposals attempt to use all the available information in the competing models to increase the precision of the model estimations (e.g., Burnham & Anderson, 2002, 2004; Konishi & Kitagawa, 2008). In the present study, we use Akaike weights (Burnham & Anderson, 2002; Steele, Ferrer, & Nesselroade, 2014) for model averaging, but other approaches exist that account for model uncertainty using *Bayesian model averaging* (e.g., Kaplan & Chen, 2014; Kaplan & Lee, 2018). Model averaging using Akaike weights involves fitting the various

competing models and using information relative to the estimation of the parameters and the fit indices (see Burnham & Anderson, 2002). This procedure can be summarized into three consecutive steps: First, all relevant models are fitted. Second, *Akaike weights* are computed using the Akaike information criterion (AIC; Akaike, 1973, 1974) for each competing model as a relative evidence in favor of each model among all the competing models (Burnham & Anderson, 2002). Finally, the target parameters of the model (fixed effects, standard errors, variances, etc.) are estimated using Akaike weights. Different derivations of the AIC fit index like the AIC with correction for small sample sizes (AICc) can be used within the same procedure. See the *Data analysis* section as an example of how we applied model averaging based on Akaike weights in this study.

On the difference between ML and REML estimators

Restricted Maximum Likelihood (REML) is recommended for estimating variance and covariance parameters, while *Maximum Likelihood* (ML) is suggested for estimating fixed effects (West, Welch, & Galecki, 2014; see also Morrell, 1998). This is the case because ML and REML estimators guarantee the generation of, at least, positive semi-definite estimations of variance components, but they are not necessarily full rank estimations (e.g., Anderson, Anderson, & Olkin, 1986; Vasdekis & Vlachonikolis, 2005). This means that the rank of such estimations will depend on some conditions of the data, where the ML rank will always be less or equal than the one from REML (Vasdekis & Vlachonikolis, 2005).

In the context of longitudinal data analysis, Vasdekis & Vlachonikolis (2005) showed that incorrect models of the variances generate differences in terms of efficiency in favor of REML, relative to ML estimators. These results reinforce the assumption that REML produce more accurate estimates of the random effects than ML (e.g., Thompson, 1962; Jiang, 1996). The consequences of these differences have also been observed in terms of better balances of

type I and II error rates in REML than in ML (Luke, 2017), but also in an anti-conservative tendency of ML when evaluating the significance of fixed effects with likelihood ratio tests (Pinheiro & Bates, 2000; c.f., Barr et al., 2013). Furthermore, as described in Hoffman (2015), REML maximizes the likelihood of the data treating the fixed effects as known, while ML only maximizes the residuals treating the fixed effects as unknown. Such differences in how the likelihood of ML and REML estimators is computed determine what aspects of the model would be indexed in its model fit indicators (Hoffman, 2015). Thus, as we fit different random structures for the same fixed effects in this study, we expect to find relevant differences between both estimators in true model selection and bias of standard errors of fixed effects (SE bias).

The Present Study

Given that the under-parameterization and the over-parameterization of the random structure of MEMs is supposed to increase SE bias, a correct specification of the random effects is needed when fitting MEMs with crossed random effects. Unfortunately, the true structure of the variance components is almost never known, leading to notable uncertainty about the various possible models. In the present study, we examine the SE bias of fixed effects in MEMs with crossed random effects using two different but complementary perspectives: model selection based on likelihood-ratio tests and model averaging based on Akaike weights.

Specifically, we are going to compare the performance of two different model selection strategies based on likelihood-ratio tests that are commonly applied in empirical research. Both strategies have the same model comparisons, but they have different decision pathways (i.e., bottom-up vs. top-down). Moreover, previous research reports some differences between different likelihood ratio tests when selecting multilevel random

coefficient models (LaHuis & Ferguson, 2009). LaHuis & Ferguson (2009) found small differences in favor of the one-tailed likelihood ratio test balancing type I and II error rates in front of the two-tailed likelihood ratio test and the likelihood ratio test with a mixture chi-square distribution (mixture likelihood ratio test). The latter one is an alternative version of likelihood ratio tests that assumes a mixture of different chi-square distributions with one and two degrees of freedom (LaHuis & Ferguson, 2009; Stoel, Garre, Dolan, & van den Wittenboer, 2006; Stram & Lee, 1994). Thus, we analyzed these versions of likelihood ratio test with different cut-off points for χ^2 differences using α=0.01 and α=0.05 in all the conditions.

All the MEMs of the present study are going to be estimated using ML and REML to analyze differences between them, as we expect to find that REML performs better in true model selection and SE bias because it is supposed to produce more accurate estimations of random effects. Furthermore, these strategies are compared with model selection of AIC and BIC indices. Also, we expect to find appropriate estimations of SEs of fixed effects using model averaging with Akaike weights because of using all the relevant information of competing models. To the best of our knowledge, this is the first time that model averaging has been applied to MEMs with crossed random effects for subjects and items in order to obtain appropriate estimations for SEs of fixed effects.

METHODS

Simulation Study

We conducted a simulation study that emulates an experimental design in psycholinguistics where multiple items were answered twice in two different conditions (control vs. experimental) by two groups of participants (expert vs. novice) and used response time (milliseconds -ms-) as the outcome. While subjects and items random effects were

crossed (i.e., all subjects answered all items), the control vs. experimental conditions represent the within-subject effect and the expert vs. novice participants represent the between-subject effect. Let Y_{tsi} be the outcome variable (ms) of subject *s* and item *i* where the *t* subscript indexes the within condition. The data generating model is presented in *Equation 1* based on Locker, Hoffman & Bovaird (2007) formulation:

$$Y_{tsi} = \gamma_{000} + \gamma_{010}(B_s) + \gamma_{100}(W_{tsi}) + \gamma_{110}(B_s)(W_{tsi}) + U_{0s0} + U_{00i} + U_{1s0}(W_{tsi}) + U_{10i}(W_{tsi}) + e_{tsi} \qquad [1]$$

where B_s (novice vs. expert groups) and W_{tsi} (control vs. experimental conditions) are the between and within factors, respectively. B_s and W_{tsi} factors were equally distributed in the study, that is, half of the participants were novice (B_s=0) and the other half were expert (B_s=1), and half of the items were for the control condition (W_{tsi}=0) and the other half were for the experimental condition (W_{tsi}=1). Fixed effects are represented with gamma (γ) letters and random effects are represented with *U* letters.

In our simulation study, we used similar fixed and random effects population parameters as those in previous related work (Matuschek et al., 2017). *Table 1* presents the simulation parameters. The main effects were set to 0 in all simulation conditions, so the interaction effect γ_{110} can be interpreted as the difference in control vs. experimental conditions between expert and novice groups. Three effect sizes were simulated (00, 25 and 50 ms). The cluster-specific random effects for subjects and items intercepts are represented by U_{0s0} and U_{00i}, respectively, and their variances (τ_{0s0} and τ_{00i}, respectively) were set to 10,000 ms in all conditions. The subject cluster-specific random slopes are represented by U_{1s0} and the item cluster-specific random slopes by U_{10i}. Three variances (τ_{1s0} and τ_{10i}, respectively) were simulated for random slopes (0, 3,600 and 14,400 ms). An intercept-slope covariance ($\tau_{U_{0s0},U_{1s0}}$ and $\tau_{U_{00i},U_{10i}}$ for subjects and items, respectively) –equivalent to an

intercept-slope correlation = 0.60– was imposed for all simulated conditions, which means that random slopes are positively correlated with random intercepts. This value was used for both subjects and items when random slopes were simulated. Given that the residual level-1 variance (σ^2_e) was 90,000 ms, the effect sizes can be standardized in terms of Cohen's criteria as d=0.00 (γ_{110}=0.00), d=0.083 (γ_{110}=25) and d=0.167 (γ_{110}=50), in the conditions where no random variance was imposed for the slopes[3]. Although the effect sizes appear small in terms of Cohen's criteria, all fixed and random effects parameters were generated based on prior empirical studies (similar simulation parameters can be found in Baayen et al., 2008; Barr et al., 2013; or Matuschek et al., 2017). The random effects and the error term were normally distributed. Furthermore, different sample sizes were considered for subjects and items given that the number of available observations is supposed to influence MEMs (Baayen et al., 2008; Bell et al., 2010, July; Maas & Hox, 2005; Vasishth & Nicenboim, 2016). Specifically, different number of subjects (sample size for subjects: 30, 50, 100, 200, and 500 subjects) and items (sample size for items: 12, 24, 48, and 96 items) were simulated. A total of 540 simulation conditions were considered in the present study, and 1,000 replications were generated for each simulation condition using MATLAB 2017b software.

TABLE 1 HERE

Data analysis

In the present study, we consider that every MEM with crossed random effects will naturally have random intercepts for both subjects and items (i.e., *minimal MEM*[4]). Also, we

[3] Given that the effect size can be influenced by random slopes variances, two different versions of the effect size were tested in the present study. The first one was the simulated effect size (*Table 1*). The second one was an effect size that was corrected by the sampling variance considering the error term and the size of the random slopes of subjects and items. Willett (1989) presents two different illustrations for longitudinal modeling, and Judd, Westfall, & Kenny (2017) present a similar approach for experimental research. In the present study, no relevant differences were found between these two approaches when evaluating the influence of effect size in true model selection and SE bias. Thus, we report the results of the original effect size from *Table 1*.

[4] We called it *minimal model* following Matuschek et al. (2017).

consider that the most complex MEM will be one with crossed random effects with random intercepts and random slopes for both subjects and items (i.e., *maximal MEM*[5]). Then, the same fixed effects can be analyzed with four different statistical models depending on their random structure: a *minimal MEM*, a MEM with random intercepts for both subjects and items and random slopes for subjects (*subject random slopes*), a MEM with random intercepts for both subjects and items and random slopes for items (*item random slopes*), and a *maximal MEM*. Each of these models adds different random effects relative to the minimal model, as all of them share random intercepts for both subjects and items. Then, each of the replications of the simulation study was analyzed using these four MEMs with crossed random effects, and both ML and REML estimators, for a total of 4,320,000 analyses (540,000 replications x 4 MEMs x 2 estimators). All MEMs were fitted with the *lme4* package (Bates, Mächler, Bolker, & Walker, 2014) in R software (R Development Core Team, 2019). Approximately, 99.92% and 99.81% of the estimated models converged for ML and REML, respectively.

To recap, we fitted the four different MEMs with crossed random effects. We then implemented model selection with the two (bottom-up and top-down) strategies, using the different likelihood-ratio tests based on the deviance or −2LL as the criterion. Each likelihood-ratio test had two degrees of freedom for model comparisons because each additional random slope entails the random slope itself and its intercept-slope covariance. For each model selection strategy, we made four comparisons (explained in detail below) and examined the convergence between strategies using different fit indices. We used different likelihood ratio tests (one-tailed, two-tailed, and mixture likelihood ratio tests) with different cut-off points for χ^2 differences of deviance comparisons using α=0.01 and α=0.05.

[5] We called it *maximal model* following Barr et al. (2013).

In the bottom-up strategy, we used a likelihood-ratio test to compare the −2LL of the *minimal MEM* against a model with subject random slopes and a model with item random slopes. If none of the comparisons were statistically significant, the *minimal MEM* was selected. If only one of those comparison was statistically significant, either the subject or the item random slopes MEM was selected. If both comparisons were significant, the −2LLs of the intermediate models were compared against a model with all random effects, the so-called *maximal MEM*. If these two new comparisons were statistically significant, the *maximal MEM* was selected. In the cases where one or both of those last comparisons were not statistically significant but both intermediate models (MEMs with crossed random slopes for subjects or items) obtained a better fit than the *minimal MEM*, the *maximal MEM* was selected. The latter scenario was very unlikely (less than 1% of the analyzed cases).

Whereas the bottom-up strategy was based on improvement of model fit, the top-down strategy was based worsening of model fit. Here, the likelihood-ratio test was used to compare the −2LL of the *maximal MEM* against that of subject random slopes and item random slopes. If both comparisons were statistically significant, the *maximal MEM* was selected. If only one comparison was non-statistically significant, either the subject random slopes or the item random slopes was selected. Then, the −2LL of the selected model was compared against the −2LL of the *minimal MEM*. If this new comparison was statistically significant, the intermediate MEM was selected. Otherwise, the *minimal MEM* was selected. When both the subject or the item random slopes MEMs provided a better fit than the *minimal* and the *maximal MEMs*, the *maximal MEM* was selected. The latter scenario was very unlikely (less than 1% of the analyzed cases).

For model averaging, we used Akaike weights (Burnham & Anderson, 2002) on the estimations of SEs of the fixed effects. Specifically, we computed the SEs of fixed effects weighting the estimations of each competing model using its corresponding Akaike weight

(ω_i). In this case, no model was selected. Instead, a weighted estimation was computed using the information of all the competing models. In this study, we fit four models to the simulated data (i.e., minimal MEM, subject random slopes, item random slopes, and maximal MEM). For all *R* competing models (*i* = 1, …, *R*), the AIC (Akaike, 1973, 1974) of each *i* competing model is obtained. Then, the fits of all *R* models are ranked as following:

$$\Delta_i = AIC_i - AIC_{min} \qquad [2]$$

where Δ_i represents the difference between the AIC of each *i* competing model and that of the best fitting model (AIC_{min}). Akaike weights are then calculated as:

$$\omega_i = \frac{\exp(-\Delta i/2)}{\sum_{r=1}^{R} exp(-\Delta r/2)} \qquad [3]$$

where ω_i is the resulting Akaike weight for each competing model based on Δ_i of all *R* competing models. In this way, ω_i is considered as relative evidence in favor of model *i* among the *R* competing models (Burnham & Anderson, 2002). Next, a weighted estimation of the parameters of interest (in this study, the SEs of fixed effects) is obtained weighting the estimations of each model *i* using its corresponding ω_i.

To examine the results of the simulation, we considered two dependent variables. First, we looked at the performance of model selection strategies for true model selection. We computed this as the proportion of times the data generating model was correctly selected. This was determined by the presence of random slopes for subjects and items in the population model: the *minimal MEM* was considered the true model in conditions without random slopes and the *maximal MEM* was considered the true model in conditions with both random slopes, while the intermediate models were considered the true model when only one of the random slopes were different from zero. We also considered the performance of AIC

and BIC fit indices[6], and their agreement with the bottom-up and top-down model selection strategies. Second, we considered the bias of *SEs* of the fixed effects. SE bias was computed using the following formula: $SE\ bias = 100 * \left(SE\left(\widehat{\Theta}_i\right) - SD(\widehat{\Theta}_i)\right) / SD(\widehat{\Theta}_i)$, where $SE\left(\widehat{\Theta}_i\right)$ is the estimation of the SE of fixed effects of each approach (the estimation of the selected model, or the averaged estimation of model averaging), and $SD(\widehat{\Theta}_i)$ is the standard deviation of the distribution of the estimated fixed effects in the 1,000 replications (the estimations of the selected model, or the averaged estimation of model averaging). Then, SE bias is the percentage of difference between the estimated SEs of each approach (i.e., model selection or model averaging) and the standard deviation of the distribution of the estimated fixed effects in the 1,000 replications.

RESULTS

In the first set of analyses, we examined the proportion of accurate decisions (true model selection) and the agreements between bottom-up and top-down likelihood ratio tests and different fit indices (AIC and BIC). We also compared the performance of the likelihood-ratio tests with AIC and BIC model selection, and explored which models were incorrectly selected in each strategy. Second, we investigated factors of the simulation study that affected true model selection. Third, we analyzed the bias of SEs of the estimated fixed effects using model selection and model averaging.

Performance of model selection strategies

Table 2 includes the overall proportion of true model selection (correct model selection) in all scenarios and the proportions for the different simulated scenarios (when the correct model is the minimal MEM, the maximal MEM, or the subject or the item random

[6] We also analyzed the performance of AICc correcting for the sample size (number of subjects) and the number of observations (number of subjects multiplied by number of items), but no relevant differences in true model selection were obtained with AIC index. For the sake of brevity, we decided to focus on AIC and BIC indices.

slopes). The proportions are shown as a function of the strategy (bottom-up and top-down) comparing the estimators (ML and REML) and their AIC and BIC fit indices. *Table 2* also includes the three likelihood ratio tests (one-tailed vs. two-tailed vs. mixture) for two significance levels (α=0.01 vs. α=0.05). As it can be seen, the proportion of true model selection was virtually the same in all simulation conditions for bottom-up and top-down strategies and, as well, for both ML and REML estimators. True model selection, however, was generally medium (approximately, 0.80), indicating that the true model was correctly selected in only 80% of the replications.

A first ANOVA was conducted to test the effects of the estimator (ML vs. REML), the likelihood ratio tests (one-tailed vs. two-tailed vs. mixture), the significance levels (α=0.01 vs. α=0.05), and the strategy (bottom-up vs. top-down). No interaction effect was relevant to explain model selection performance, but a main effect of significance level was found ($\eta_p^2 = 0.027$), favoring the correct selection of models with α=0.05 versus α=0.01. In addition, a main effect of likelihood ratio test was found ($\eta_p^2 = 0.018$), favoring the one-tailed and the mixture likelihood ratio tests in front of the two-tailed one by a small proportion difference of 0.015.

A second ANOVA was conducted to test potential differences between the performance of likelihood ratio tests comparing AIC and BIC, considering the ML and REML estimators (in this case, we fixed the likelihood ratio test to the one-tailed with α=0.05). Results showed no relevant interaction effect between the estimator and the approach, but a relevant main effect of the approach was observed with an $\eta_p^2 = 0.090$ where a mean proportion difference of, approximately, 0.12 was found in favor of the one-tailed likelihood ratio test and AIC in front of BIC.

TABLE 2 HERE

Table 3 shows the agreements in selected models between the different likelihood ratio tests, combined with the strategy (bottom-up and top-down), and its derived fit indices (AIC and BIC) for both ML and REML estimators. Overall, the agreement between the likelihood tests and AIC (approximately 0.80) was higher than the one with BIC (approximately 0.70), and no differences were found between REML and ML estimators. In addition, an analysis on the agreements in incorrectly selected models revealed that bottom-up and top-down strategies select the same incorrect model in 95% of the replications (regardless of the likelihood ratio tests). Seldom one strategy selected the right model and the other did not (this occurred only in 4.1% of all replications analyzed). Again, no relevant differences were observed between REML and ML estimators (the incorrect decisions were very similar). In conclusion, the correct and incorrect decisions were very similar for the conditions examined and thus it is useful to know which models tend to be incorrectly selected.

TABLE 3 HERE

Table 4 presents the relative proportion of incorrect model selections of each strategy. Specifically, the relative proportion of incorrect selections of each strategy in each simulated scenario (true model) was analyzed. As it can be seen, all the strategies tend to favor the simpler models (i.e., the minimal MEM). In fact, a strong association was found between the incorrect model selection of the bottom-up and the top-down strategies with both ML and REML estimators (e.g., the one-tailed likelihood ratio test with α=0.05 obtained a Cramer's V test = 0.813 between ML and REML in bottom-up strategy which means that they coincide in 86.2% in the incorrect model selections). These results mean that both strategies and both estimators usually select the same incorrect model (this occurs regardless of the likelihood ratio test). However, their small differences can be explained due to the top-down strategy slightly tends to incorrectly select the maximal MEM more frequently than the bottom-up strategy (consequently, bottom-up tends to incorrectly select the minimal MEM on more

occasions). Regarding AIC and BIC indices, AIC showed a similar pattern than likelihood ratio tests, but BIC tends to select the simpler model more frequently (approximately, minimal MEM was selected in the 0.60 of the replications whose model selection was incorrect).

TABLE 4 HERE

Up to this point, it can be noted that the strategy (bottom-up and top-down), the estimator (ML and REML) and the AIC index derived from these estimators reach a very similar success rates of true model selection. The likelihood ratio test that favors more the true model selection is the one-tailed at a significance level of 0.05. BIC performance was markedly worse than the one of the likelihood ratio tests and the AIC index.

Factors affecting true model selection

In the next analyses, we examined the simulation conditions that most affected the true model selection using the partial η^2 of ANOVAs. We analyzed the performance for true model selection of bottom-up strategy with one-tailed likelihood ratio test and α=0.05, and AIC from both ML and REML estimators (this within-factor was called strategy). This decision was made following an additional univariate ANOVA that showed no relevant effect sizes (partial η^2) about the moderation of the differences among previous findings by means of the simulation conditions. The simulation conditions (between-factors) were the sample size for subjects and items, the size of random slope variances for subjects and items, and the effect size for the fixed interaction effect. The BIC index was discarded from these analyses because its performance was much worse than the likelihood ratio tests and the AIC in both ML and REML estimators.

As expected from previous analysis, no substantive differences were found for the within-factor strategy (i.e., the one that compares the performance between likelihood ratio

test and AIC using both ML and REML): η_p^2=0.001 was found for the main within-factor effect and none of the partial η^2 involving the interactions of this within factor with the rest exceeded η_p^2=0.014. Instead, four η_p^2>0.06 were found for the (between-factors) simulation conditions. The highest value was for the main effect of the number or items (η_p^2=0.191), followed by the main effect of the number of subjects (η_p^2=0.099), the interaction effect between number of items and size of random slope variance for subjects (η_p^2=0.080), and the main effect for the random slope variance for subjects (η_p^2=0.062). The random slope variance for items was less important (a main effect with η_p^2=0.024, and an interaction effect with number of subjects of η_p^2=0.025, were found). The effect size for the interaction fixed effect handled in this study was irrelevant (all η_p^2 involving this factor were less than 0.001). *Figure 1* presents a graphical summary of the results.

FIGURE 1 HERE

As it can be seen in *Figure 1*, when the number of items is small (especially for 12 and 24 items) the model selection has a very low performance, unless there is no variance of slopes neither in subjects nor in items (upper left graph). With 24 or more items, the model selection performance exceeds 0.90 as long as there are more than 200 subjects. On the other hand, when the random slope variances (both for the subjects and items) is medium (i.e., 3,600) the model selection is invariably bad unless there are 96 items (regardless of the number of subjects) or with 48 items together with 100 or more subjects. Otherwise, that is, with fewer items or few subjects, the incorrect selection of the minimal model is being favored, presumably because the variances of the slopes are not detected. Considering partial η^2 in our simulated design, the performance is more affected by the random slope variance for subjects than by the random slope variance for the items.

Bias in Standard Errors (SEs) of fixed effects in model selection and model averaging

In the next analyses, we examined the bias in SEs of the estimated fixed effects using model selection and model averaging. In the light of the previous results, the following analyses were conducted with the one-tailed likelihood ratio test with α=0.05 using the bottom-up model selection strategy. The standard deviation of the estimated fixed effects of all the replications of each condition was considered the true (population) values per condition. Then, the percentage of SE bias was computed as the difference of each SE estimation with that standard deviation of fixed effects per condition divided by that standard deviation of fixed effects per condition. *Table 5* presents different descriptive analysis (mean and standard deviation of SE bias) for some representative conditions of the simulation study. We analyzed the influence of simulation conditions on the SE bias using the partial η^2 of ANOVA (reporting $\eta_p^2 \geq 0.06$ effects, medium effect sizes according to Cohen, 1988) for within-subject, between-subject and interaction fixed effects to test the influence of the estimator (ML vs. REML), the approach (model selection vs. model averaging), and the simulation conditions (number of subjects and items, effect size, and size of subjects and items random slopes).

TABLE 5 HERE

First, we analyzed the influence of simulation conditions on the bias of SEs of the within-subjects main effect (control vs. experimental conditions for items that was set to 0 in all simulation conditions). An interaction effect was found between the estimator, the number of items and the size of the item random slopes (η_p^2=0.152). Similarly, two different interaction effects were found between the estimator and the number of items (η_p^2=0.252) and between the estimator and the size of the item random slopes (η_p^2=0.173). A principal effect of the estimator was also found (η_p^2=0.565). Different minor interaction effects were found, but no interaction effect was found for the approach (model selection vs. model averaging).

Second, we considered the influence of simulation conditions on the bias of SEs of the between-subjects main effect (expert vs. novice, also set to 0 in all simulation conditions). A general interaction effect was found for all the variables included in the analysis -except the estimator- (η_p^2=0.144). Also, an interaction effect between the estimator and the number of subjects was found (η_p^2=0.225), and a principal effect of the estimator (η_p^2=0.286). Medium effect sizes were found in different interaction effects concerning the influence of the size of the subject random slopes in SE bias.

Third, we examined the influence of simulation conditions on the bias of SEs for the interaction effect. Similarly, a general interaction effect was found for all the variables included in the analysis -except the estimator- (η_p^2=0.131). Also, an interaction effect between the estimator and the number of subjects was found (η_p^2=0.110), and a principal effect of the estimator (η_p^2=0.158). Medium effect sizes were found in different interaction effects concerning the influence of the size of the subject random slopes in SE bias.

Table 5 also reports the standard deviation of SE bias as a measure of the variability of the SE bias of the estimations. As it can be observed, there is a direct relation between the number of subjects and items and the variability of SE bias. While this result was expected, the standard deviation of many conditions shows that, although the SE bias does not present directionality (i.e., its mean tends to zero), many replications would present important SE bias. These results are problematic as some conditions present standard deviations >20% of SE bias. In this way, model averaging with Akaike weights seems to be a less risky option considering the central tendency and the variability of the SE bias.

In light of the results regarding SE bias, we examined the main interaction effects in *Figure 2*. This figure presents a graphical summary of the results using bottom-up model selection with one-tailed likelihood ratio test (α=0.05) and model averaging (Akaike weights)

for ML and REML (please note that results were very similar for both model selection and model averaging approaches). *Figure 2* shows that SE bias was higher for the within-subject effects, relative to the between-subjects or the interaction effects, when there were fewer items. It also shows that SEs of fixed effects are underestimated in the presence of random slopes, and that ML tends to be more biased than REML. A decay of SE bias that depends on sample sizes can also be observed when there are random slopes (although the influence of sample size is significantly higher for the within-subject effect, relative to the between-subjects or the interaction effects). This means that larger sample sizes and larger random slope variances are both related to lower levels of SE bias. SE bias was acceptable for between-subjects and interaction effects. In sum, complex interactions in the patterns of SE bias were apparent, but the general trend was that the lack of items/subjects increased SE bias when there are larger variances in their respective random slopes. This tendency affected ML estimator more negatively than REML. A similar pattern of results was observed for top-down model selection, AIC index, and model averaging strategies.

FIGURE 2 HERE

DISCUSSION

Summary of findings

The aim of the present study was to examine model selection strategies as well as SE bias of fixed effects when using MEMs with crossed random effects for subjects and items. Specifically, we tested two different model selection strategies based on likelihood-ratio tests and model averaging using Akaike weights. Top-down had been hypothesized in prior research as reducing the complexity of random effects (Barr et al., 2013), whereas bottom-up had been proposed as increasing the complexity of random effects (Matuschek et al., 2017). In our results, true model selection was very similar for all the strategies considered (bottom-

up and top-down likelihood ratio tests, and AIC model selection), except for the BIC fit index whose performance was lower. True model selection was approximately 0.80. This unexpected performance can be related to the high demands of the model selection task that was conducted in this simulation study. Such model selection task (where we evaluated if random slopes should be added or deleted to an already complex crossed random effects model with a partial random structure) is more demanding than evaluating the presence of a full crossed random effect such as subjects or items. In light of these results, these model selection tasks leave room for important improvements.

One of the most relevant conditions for true model selection was the significance levels of the likelihood ratio tests. Using $\alpha=0.05$ lead to better performances than $\alpha=0.01$. Moreover, the sample sizes and the sizes of random slopes were a strong determinant for true model selection. True model selection was considerably low for conditions with smaller random slopes, so larger sample sizes are required to obtain adequate proportions of true model selection. On the contrary, true model selection was better for conditions with null or large random slopes, although higher sample sizes also favored true model selection. These results reinforce the differences between previous findings about the likelihood ratio tests being anti-conservative when testing the statistical significance of fixed effects (Pinheiro & Bates, 2000; Luke, 2017), and other papers that did not find them so anti-conservative for typical experimental conditions (e.g., Barr et al., 2013). Also, our results support the conclusions of LaHuis & Ferguson (2009) because we also found a slight difference in favor of the one-tailed in front of the two-tailed and the mixture likelihood ratio tests (although the differences with the later were negligible). No relevant differences were found between the bottom-up and the top-down model selection strategies in their true model selection performance.

The analysis of the incorrect model selections revealed that all the strategies tend to under-parameterize the random structure of the models (approximately, the minimal MEM was incorrectly selected in the 0.40 of the replications). This under-parameterization of the model was considerably higher for BIC (approximately, the minimal MEM was selected in the 0.60 of the replications whose models were incorrectly selected). These results can explain why true model selection was significantly higher for minimal MEMs, relative to the rest of the simulation conditions (those with random slopes for one or two crossed random effects). In this way, although the bottom-up and the top-down strategies presented considerable agreements (approximately, 0.95) and their correct and incorrect model selections were very similar, it can be seen than the bottom-up tend to favor the election of the minimal MEM more frequently than the top-down one that tends to favor the maximal MEM.

We also examined the SE bias of fixed effects using model averaging with Akaike weights (Burnham & Anderson, 2002). Our results indicate that the SE bias of fixed effects were similar between model selection and model averaging. However, we found important differences across the various simulated fixed effects (within-subject, between-subject, and interaction effects). Specifically, in some conditions involving fewer items, the underestimation (SE bias) of the within-subject effect was significantly higher than the one of the other two effects. It is worth to mention that between-subject and interaction effects did not present significant bias even for small sample sizes. SE bias of fixed effects was related to the interaction between the estimator, the number of items and the size of the item random slopes. In this way, within-subject SE bias was higher in those conditions with lower item sample sizes and higher random slope variances, especially for ML estimator. On the contrary, the variability of SE bias led to unacceptable SE bias for some conditions (e.g., presenting standard deviations >20%). The differences between the experimental effects (within-subject vs. between-subject and interaction effects) and estimators (ML vs. REML)

are probably related to the relatively small number of items that were used to simulate the within-subject effect as the sample size of items would be limiting the available information necessary to estimate the parameters comparing to the between subjects and the interaction effects that would have more available information. These results are in accordance with the properties of ML and REML estimators (e.g., Hoffman, 2015).

Overall, these findings suggest that, in designs with interactions involving between- and within-subject effects, the bias of SEs of fixed effects is acceptable (small), but it is related to the interaction between the number of subjects and the sizes of subjects random slopes. A similar pattern can be observed for between-subject effects. However, the bias of SEs of within-subject effects, such as control and experimental conditions for items, is larger and is related to the interaction between the number of items and the size of random slope variance for items. These results are in line with those found in previous studies (Barr, 2013). However, in contrast to previous recommendations advocating the use of REML to estimate covariance parameters and using ML to estimate fixed effects (e.g., West et al., 2014), we found that, in our simulated conditions, ML leads to the same proportion of true model selection (although ML yields more SE bias than REML). ML showed less variability of SE bias than REML. This was expected based on past research (Hoffman, 2015). In our case, however, this was probably associated with the under-estimation of random effects and thus the SEs of fixed effects. Such differences were especially relevant for conditions with small sample sizes and larger random slopes, being the only simulation conditions that were determinant for true model selection.

Theoretical and methodological considerations

When selecting a correct model among nested alternatives, using a bottom-up strategy (Matuschek et al., 2017), a top-down strategy (Barr et al., 2013), different likelihood ratio

tests (LaHuis & Ferguson, 2009), or a model averaging approach like Akaike weights (Burnham & Anderson, 2002) lead to similar results. Not only do these strategies share a similar performance in selecting the true model but also they obtain similar estimations of SEs of fixed effects whose bias depends on sample sizes and the variance of random slopes. Although model selection and model averaging are two convergent and useful solutions for reducing bias of SEs of fixed effects, we also showed that model averaging with Akaike weights could be a less risky option, as its variability of SE bias was smaller than the one in model selection.

A general conclusion from our analyses is that both subjects and items sample sizes affect bias of SEs of fixed effects despite the presence of random slopes. Given that MEMs with crossed random effects are complex models and, in the conditions studied here, have two sources of variability (i.e., subjects and items), an increase in subjects and items leads to an increase in true model selection and a reduction in potential bias. Our findings agree with previous proposals suggested to reduce the negative effects of the random effects of subjects and items in experimental contexts (Judd et al., 2012, 2017): (1) use homogeneous subject or item samples, or (2) increment sample sizes if they have high variability. Then, given that both random intercepts and random slopes effects can be naturally found in representative samples, a solution to preserve the validity of empirical studies is to increase the number of subjects in order to gain representativeness of the entire population. Similarly, improving the experimental control of the items would be a good strategy to reduce the potential negative effects of items random effects when few items are used. This is especially important as we found higher SE bias for within-subject effects that usually depend on items from different experimental conditions.

These methodological considerations could have important theoretical implications for empirical research because they can avoid errors of statistical inference. For example,

previous research showed that MEMs with crossed random effects can overcome the so-called *language-as-fixed-effect-fallacy* (i.e., generalizing findings beyond the specific sample of chosen materials; Clark, 1973), because item random effects are conceived as a random sample from a larger population of those items (Quené & van den Bergh, 2008). However, as the true variance model is not always evident in empirical studies, deciding the extent to which adding random slopes for subjects and items in our experimental research is an important decision. Similarly, research has pointed out that all item random variation should be included in order to take into account all relevant dimensions in experimental conditions for ecological validity (Hoffman, 2015). This means that MEMs with crossed random effects enable to test item exchangeability, which implies to control item variation for experimental control and to test if there are differential experimental effects for items. As long as random slopes for subjects or items are available, researchers ought to clarify theoretically why the experimental effects can depend on idiosyncratic differences. More importantly, as suggested by Barr et al. (2013), even the parameterization of random effects should be considered a confirmatory hypothesis itself. In this vein, the importance of model selection and model averaging strategies to determine random effects is emphasized.

Limitations and future directions

The present study examined true model selection and SE bias of fixed effects in a simulated experimental design with two interacting effects (a between- and a within-subject effect) where subjects and items random effects were crossed (i.e., all subjects answered all items). However, experimental designs strongly vary depending on the researcher's hypothesis. For example, further research should determine if these findings are generalizable to designs with missing data and or incomplete experimental designs where subjects only answer a subset of items. Given that experimental designs are closely related to the random structure of the statistical models (Judd et al., 2017), future studies should investigate the

influence of idiosyncratic experimental designs on the bias in the SEs of fixed effects. Among the many possibilities, one interesting experimental design to explore is the single-case experimental designs (e.g., Smith, 2012).

Barr et al. (2013) advocated a superiority of maximal MEMs in front of other alternative models due to not ignoring specific random slopes. In this line, future research should explore the consequences of ignoring specific random slopes when the true model is the maximal MEM and it has crossed random effects. This is because the SE bias of fixed effects that was observed in our simulation study was the consequence of a model decision task and we could expect that not using a model selection or a model averaging strategy would lead to a worse scenario for SE bias. Moreover, our simulation study fixed the intercept-slope covariance in all its conditions. But it is interesting to simulate different scenarios that manipulate that parameter to determine if there is a better true model selection in conditions with smaller intercept-slope covariances than conditions with larger ones. A new direction for future research could study the statistical power of detecting and estimating such parameters and its relation with samples sizes and sizes of random intercepts and slopes.

In our study, true model selection for both ML and REML estimators showed unlikely rates, presenting no differences in the strategies that were tested. It should be noted here that we used the classic AIC index (AIC = $2k - 2LL$; being k the number of parameters), but some corrections have been proposed for this estimator in MEMs due to AIC not being an asymptotically unbiased estimator (Greven & Kneib, 2010). One could expect improvements in true model selection and, maybe, in SE bias of model averaging with Akaike weights using such corrections for AIC. Similarly, although we found that REML estimator does not present important SE bias even for small sample sizes, the Kenward-Roger correction for degrees of freedom has been proposed by different authors as an appropriate correction to compute the SEs of fixed effects (e.g., Luke, 2017; McNeish, 2017). Thus, we could also expect some

improvement in SE bias of fixed effects. These aspects were not included in this study because they were computationally prohibitive for those conditions with larger sample sizes for subjects or items, but future research should try to research their performance in similar conditions.

Conclusion and recommendations

We examined model selection and model averaging strategies, and their effect on the SEs estimates of fixed effects in MEMs with crossed random effects. The selection of the true model was equivalent across the different strategies. Similarly, the SE bias of fixed effects was constant across the strategies. When comparing ML and REML estimators, true model selection using ML was very similar to REML, but the SE bias of fixed effects was lower for REML. True model selection of all the strategies showed approximately 0.80 of performance (except BIC that showed about 0.70 of performance). Model selection in this study evaluated if random slopes should be added or deleted to an already complex crossed random effects model with a partial random structure. This is a very demanding task and we believe that there is room for important improvements in such model selection.

Based on our findings, we endorse the combination of using the one-tailed likelihood ratio test with $\alpha=0.05$ for model selection, and model averaging based on Akaike weights. AIC index model selection was very similar to likelihood ratio tests while BIC was more divergent. Also, our results showed that true model selection was the same for bottom-up and top-down model selection strategies, and that both make almost the same incorrect decisions, although the top-down strategy could underestimate the random structure of the model less frequently than the bottom-up one. The only plausible solution against incorrect model selection was the reduction of subjects and items variability (which is not usually an option due to ecological validity), or the increase of their respective sample sizes. As expected,

REML tends to present less SE bias (underestimation) than ML. The variability of SE bias of the estimations also shows that SE bias can be unacceptable under certain conditions (such as lower number of subjects and items) and thus, although the central tendency is to present no SE bias, many empirical estimations could be considerably biased.

In general, we endorse the combination of model selection and model averaging as two different but complementary perspectives that can lead to better estimations of SEs of fixed effects in MEMs with crossed random effects. As such, these strategies can provide researchers with another tool against potential errors of statistical inference.

DECLARATIONS OF INTEREST

None.

REFERENCES

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

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

Anderson, B. M., Anderson, T. W., & Olkin, I. (1986). Maximum likelihood estimators and likelihood ratio criteria in multivariate components of variance. *The Annals of Statistics, 14*(2), 405-417. https://doi.org/10.1214/aos/1176349929.

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*(4), 390-412. https://dx.doi.org/10.1016/j.jml.2007.12.005.

Barr, D. J. (2013). Random effects structure for testing interactions in linear mixed-effects models. *Frontiers in Psychology, 4*(328), 1-2. 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*(3), 255-278. https://doi.org/10.1016/j.jml.2012.11.001.

Bates, D., Kliegl, R., Vasishth, S., & Baayen, R. H. (2015). *Parsimonious Mixed Models*. Retrieved from http://arxiv.org/abs/1506.04967.

Bates, D., Mächler, M., Bolker, B. & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*, 1–48. https://dx.doi.org/10.18637/jss.v067.i01.

Bell, B. A., Morgan, G. B., Schoeneberger, J. A., Loudermilk, B. L., Kromrey, J. D., & Ferron, J. M. (2010, July). *Dancing the sample size limbo with mixed models: How low can you go?* Poster presented at the SAS Global Forum 2010. Columbia, SC.

Burnham, K. P., & Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach, Second Edition*. New York: Springer.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: understanding AIC and BIC in model selection. *Sociological Methods & Research, 33*(2), 261-304. https://doi.org/10.1177/0049124104268644.

Clark, H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior, 12*(4), 335-359. https://doi.org/10.1016/S0022-5371(73)80014-3.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences (2nd Ed.)*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Greven, S., & Kneib, T. (2010). On the behaviour of marginal and conditional AIC in linear mixed models. *Biometrika, 97*(4), 773-789. https://doi.org/10.1093/biomet/asq042.

Hoffman, L. (2015). *Longitudinal Analysis: Modeling Within-Person Fluctuation and Change*. New York: Routledge.

Hoffman, L., & Rovine, M. J. (2007). Multilevel models for the experimental psychologist: Foundations and illustrative examples. *Behavior Research Methods, 39*(1), 101-117. https://doi.org/10.3758/BF03192848.

Hox, J. J., Moerbeek, M., & Van de Schoot, R. (2018). *Multilevel Analysis: Techniques and Applications*. New York, NY: Routledge.

Jiang, J. (1996). REML estimation: asymptotic behavior and related topics. *The Annals of Statistics, 24*(1), 255-286. http://dx.doi.org/10.1214/aos/1033066209.

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology, 103*(1), 54–69. https://doi.org/10.1037/a0028347.

Judd, C. M., Westfall, J., & Kenny, D. A. (2017). Experiments with more than one random factor: Designs, analytic models, and statistical power. *Annual Review of Psychology, 68*, 601-625. https://dx.doi.org/10.1146/annurev-psych-122414-033702.

Kaplan, D. & Chen, J. (2014). Bayesian model averaging for propensity score analysis. *Multivariate Behavioral Research, 49*(6), 505-517. https://doi.org/10.1080/00273171.2014.928492.

Kaplan, D. & Lee, C. (2018). Optimizing Prediction Using Bayesian Model Averaging: Examples Using Large-Scale Educational Assessments. *Evaluation Review, 42*(4), 423-457. https://doi.org/10.1177/0193841X18761421.

Konishi, S., & Kitagawa, G. (2008). *Information Criteria and Statistical Modeling*. New York: Springer Science & Business Media.

LaHuis, D. M., & Ferguson, M. W. (2009). The accuracy of significance tests for slope variance components in multilevel random coefficient models. *Organizational Research Methods, 12*(3), 418-435. https://doi.org/10.1177/1094428107308984.

Lai, M. H. C. (2019). Correcting Fixed Effect Standard Errors When a Crossed Random Effect Was Ignored for Balanced and Unbalanced Designs. *Journal of Educational and Behavioral Statistics, 44*(4), 448-472. https://doi.org/10.3102/1076998619843168.

Locker, L., Hoffman, L., & Bovaird, J. A. (2007). On the use of multilevel modeling as an alternative to items analysis in psycholinguistic research. *Behavior Research Methods, 39*(4), 723-730. https://doi.org/10.3758/BF03192962.

Luke, S. G. (2017). Evaluating significance in linear mixed-effects models in R. *Behavior Research Methods, 49*(4), 1494-1502. https://doi.org/10.3758/s13428-016-0809-y.

Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 1*, 86–92. https://doi.org/10.1027/1614-1881.1.3.86.

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.

McNeish, D. (2017). Small sample methods for multilevel modeling: A colloquial elucidation of REML and the Kenward-Roger correction. *Multivariate Behavioral Research, 52*(5), 661-670. https://doi.org/10.1080/00273171.2017.1344538.

McNeish, D., & Kelley, K. (2019). Fixed effects models versus mixed effects models for clustered data: Reviewing the approaches, disentangling the differences, and making recommendations. *Psychological Methods, 24*(1), 20–35. https://dx.doi.org/10.1037/met0000182.

McNeish, D., Stapleton, L. M., & Silverman, R. D. (2017). On the unnecessary ubiquity of hierarchical linear modeling. *Psychological Methods, 22*(1), 114-140. https://dx.doi.org/10.1037/met0000078.

Meyers, J. L., & Beretvas, S. N. (2006). The impact of inappropriate modeling of cross-classified data structures. *Multivariate Behavioral Research, 41*(4), 473–497. https://doi.org/10.1207/s15327906mbr4104_3.

Morrell, C. (1998). Likelihood ratio testing of variance components in the linear mixed-effects model using restricted maximum likelihood. *Biometrics, 54*(4), 1560–1568. https://doi.org/10.2307/2533680.

Pardo, A., & Ruiz, M. Á. (2012). *Análisis de Datos en Ciencias Sociales y de la Salud III [Data Analysis in Health and Social Sciences III]*. Madrid: Editorial Síntesis.

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-Effects Models in S and S-Plus*. New York: Springer.

Quené, H., & van den Bergh, H. (2004). On multi-level modeling of data from repeated measures designs: A tutorial. *Speech Communication, 43*(1-2), 103-121. https://doi.org/10.1016/j.specom.2004.02.004.

Quené, H., & van den Bergh, H. (2008). Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language, 59*(4), 413-425. https://doi.org/10.1016/j.jml.2008.02.002.

Raaijmakers, J. G., Schrijnemakers, J. M., & Gremmen, F. (1999). How to deal with "the language-as-fixed-effect fallacy": Common misconceptions and alternative solutions. *Journal of Memory and Language, 41*(3), 416-426. https://doi.org/10.1006/jmla.1999.2650.

Raudenbush, S. W. (1993). A crossed random effects model for unbalanced data with applications in cross-sectional and longitudinal research. *Journal of Educational Statistics, 18*(4), 321-349. https://doi.org/10.3102/10769986018004321.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods (Vol. 1)*. London: Sage Publications.

R Development Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. URL http://www.R-project.org/.

Smith, J. D. (2012). Single-case experimental designs: a systematic review of published research and current standards. *Psychological Methods, 17*(4), 510–550. https://doi.org/10.1037/a0029312.

Steele, J. S., Ferrer, E., & Nesselroade, J. R. (2014). An idiographic approach to estimating models of dyadic interactions with differential equations. *Psychometrika, 79*(4), 675-700. https://doi.org/10.1007/s11336-013-9366-9.

Stoel, R. D., Garre, F. G., Dolan, C., & van den Wittenboer, G. (2006). On the likelihood ratio test in structural equation modeling when parameters are subject to boundary constraints. *Psychological Methods, 11*(4), 439–455. https://doi.org/10.1037/1082-989X.11.4.439.

Stram, D. O., & Lee, J. W. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics, 50,* 1171-1177. https://doi.org/10.2307/2533455.

Thompson, W. A. (1962). The problem of negative estimates of variance components. *The Annals of Mathematical Statistics, 33*(1), 273-289. http://dx.doi.org/10.1214/aoms/1177704731.

Vasdekis, V. G., & Vlachonikolis, I. G. (2005). On the difference between ML and REML estimators in the modelling of multivariate longitudinal data. *Journal of Statistical Planning and Inference, 134*(1), 194-205. https://doi.org/10.1016/j.jspi.2004.01.020.

Vasishth, S., & Nicenboim, B. (2016). Statistical methods for linguistic research: Foundational ideas – Part I. *Language and Linguistics Compass, 10*(8), 349-369. https://doi.org/10.1111/lnc3.12201.

West, B. T., Welch, K. B., & Galecki, A. T. (2014). *Linear Mixed Models: A Practical Guide Using Statistical Software, 2nd. Ed*. Boca Raton, FL: Chapman and Hall/CRC.

Willett, J. B. (1989). Some results on reliability for the longitudinal measurement of change: Implications for the design of studies of individual growth. *Educational and Psychological Measurement, 49*(3), 587-602. https://doi.org/10.1177/001316448904900309.

TABLE 1

Table 1. Simulation Parameters

	Parameter		Values
Fixed conditions	Intercept	γ_{000}	2,000 ms
	Between-subject effect	γ_{010}	0 ms
	Within-subject effect	γ_{100}	0 ms
	Random intercepts for subjects	τ_{0s0}	10,000 ms
	Random intercepts for items	τ_{00i}	10,000 ms
	Residual level 1 variance	σ^2_e	90,000 ms
	Intercept-slope correlation for subjects and items	r_{01}	0.60
Manipulated conditions	Interaction effect	γ_{110}	0 / 25 / 50 ms
	Random slopes for subjects	τ_{1s0}	0 / 3,600 / 14,400 ms
	Random slopes for items	τ_{10i}	0 / 3,600 / 14,400 ms
	Sample size for subjects	N_2	30 / 50 / 100 / 200 / 500 subjects
	Sample size for items	N_1	12 / 24 / 48 / 96 items

Note: Random effects are presented as variances. r_{01} represents the standardized intercept-slope covariance ($\tau_{U_{0s0},U_{1s0}}$ and $\tau_{U_{00i},U_{10i}}$ for subjects and items, respectively).

TABLE 2

Table 2. Proportion of true model selection of bottom-up and top-down likelihood ratio tests, and fit indices (AIC, BIC) model selection for ML and REML.

		True model	All scenarios		Minimal		Subject slopes		Item slopes		Maximal	
		Sig. level (α)	0.01	0.05	0.01	0.05	0.01	0.05	0.01	0.05	0.01	0.05
ML	One-tailed LRT	Bottom-up	0.783	0.819	0.985	0.816	0.797	0.823	0.830	0.854	0.701	0.774
		Top-down	0.795	0.826	0.968	0.895	0.781	0.795	0.820	0.834	0.746	0.820
	Two-tailed LRT	Bottom-up	0.763	0.806	0.993	0.961	0.778	0.817	0.811	0.850	0.673	0.741
		Top-down	0.776	0.817	0.976	0.942	0.766	0.796	0.804	0.835	0.717	0.787
	Mixture LRT	Bottom-up	0.776	0.815	0.988	0.937	0.791	0.822	0.824	0.855	0.692	0.760
		Top-down	0.789	0.824	0.971	0.917	0.776	0.798	0.815	0.836	0.736	0.806
	Fit indices	AIC	0.827		0.882		0.805		0.839		0.817	
		BIC	0.698		0.999		0.702		0.739		0.599	
REML	One-tailed LRT	Bottom-up	0.787	0.820	0.982	0.902	0.802	0.824	0.836	0.857	0.707	0.780
		Top-down	0.797	0.826	0.961	0.880	0.787	0.799	0.827	0.836	0.746	0.821
	Two-tailed LRT	Bottom-up	0.768	0.810	0.991	0.953	0.784	0.821	0.818	0.854	0.680	0.747
		Top-down	0.779	0.818	0.971	0.931	0.772	0.802	0.812	0.840	0.717	0.787
	Mixture LRT	Bottom-up	0.781	0.820	0.986	0.927	0.796	0.825	0.830	0.857	0.698	0.767
		Top-down	0.791	0.824	0.965	0.904	0.783	0.803	0.822	0.840	0.736	0.807
	Fit indices	AIC	0.807		0.866		0.806		0.839		0.821	
		BIC	0.701		0.999		0.707		0.746		0.600	

Note. α = Significance level for likelihood ratio tests in bottom-up and top-down model selection. LRT = Likelihood ratio test. Minimal = MEM with random intercepts for subjects and items. Maximal = MEM with random intercepts and random slopes for subjects and items. Subject and item random slopes models include random intercepts for subjects and items. N_T = 540,000 replications.

TABLE 3

Table 3. Proportion of model selection agreements between bottom-up and top-down strategies using likelihood ratio tests (α=0.05) with fit indices (AIC and BIC) for both ML and REML estimators.

	ML						**REML**					
	One-tailed LRT		**Two-tailed LRT**		**Mixture LRT**		**One-tailed LRT**		**Two-tailed LRT**		**Mixture LRT**	
	Bottom-up	**Top-down**	**Bottom-up**	**Top-down**	**Bottom-up**	**Top-down**	**Bottom-up**	**Top-down**	**Bottom-up**	**Top-down**	**Bottom-up**	**Top-down**
AIC	0.7963	0.8065	0.7728	0.7834	0.7868	0.7972	0.7980	0.8068	0.7746	0.7834	0.7886	0.7975
BIC	0.6708	0.6673	0.6821	0.6806	0.6763	0.6738	0.6709	0.6677	0.6839	0.6824	0.6773	0.6749

Note: Only likelihood ratio tests are reported using α=0.05 are reported here. LRT = Likelihood ratio test. N_T = 540,000 replications.

TABLE 4

Table 4. Relative proportion of incorrect model selections of each strategy divided by simulated scenario (true model).

True model	ML								REML							
	One-tailed LRT		Two-tailed LRT		Mixture LRT		AIC	BIC	One-tailed LRT		Two-tailed LRT		Mixture LRT		AIC	BIC
	Bottom-up	Top-down	Bottom-up	Top-down	Bottom-up	Top-down			Bottom-up	Top-down	Bottom-up	Top-down	Bottom-up	Top-down		
Minimal	0.4128	0.3634	0.4828	0.4396	0.4443	0.3975	0.3507	0.6221	0.3909	0.3380	0.4662	0.4170	0.4243	0.3731	0.3236	0.6105
Subject slopes	0.2027	0.1860	0.1957	0.1843	0.2006	0.1857	0.2009	0.1630	0.2031	0.1917	0.1979	0.1905	0.2023	0.1924	0.2069	0.1672
Item slopes	0.3033	0.2596	0.2882	0.2599	0.2978	0.2613	0.2648	0.2097	0.3115	0.2715	0.2965	0.2730	0.3060	0.2731	0.2712	0.2167
Maximal	0.0812	0.1910	0.0333	0.1162	0.0574	0.1555	0.1836	0.0051	0.0945	0.1988	0.0392	0.1194	0.0673	0.1614	0.1983	0.0056

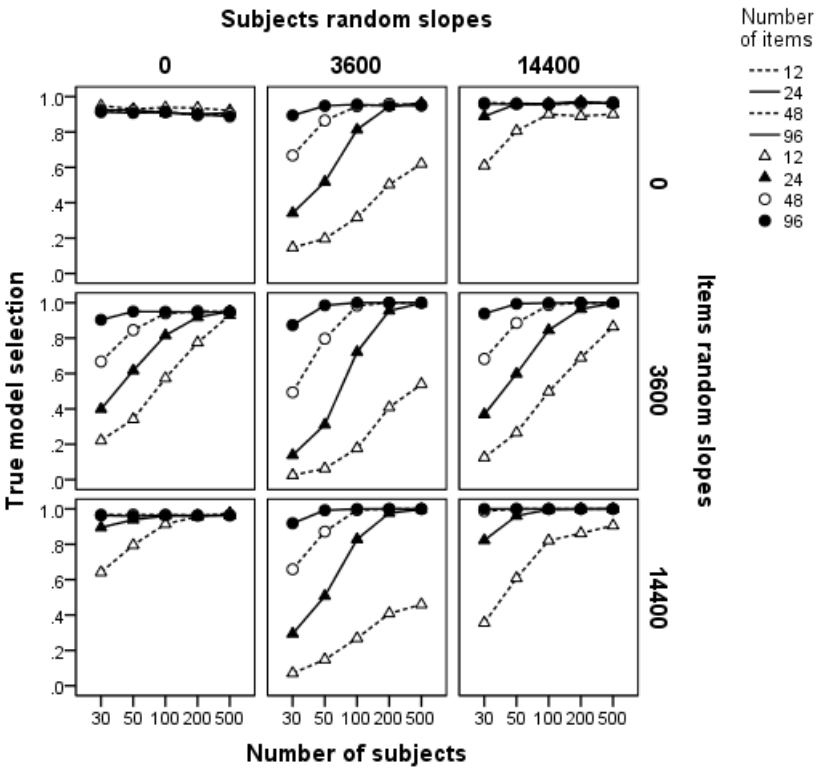
Note: LRT = Likelihood ratio test with α=0.05. Minimal = MEM with random intercepts for subjects and items. Maximal = MEM with random intercepts and random slopes for subjects and items. Subject and item random slopes models include random intercepts for subjects and items. N_T = 540,000 replications.

TABLE 5

Table 5. Mean (and standard deviation) percentage of bias of SEs of fixed effects in bottom-up model selection (α=0.05) and model averaging (Akaike weights).

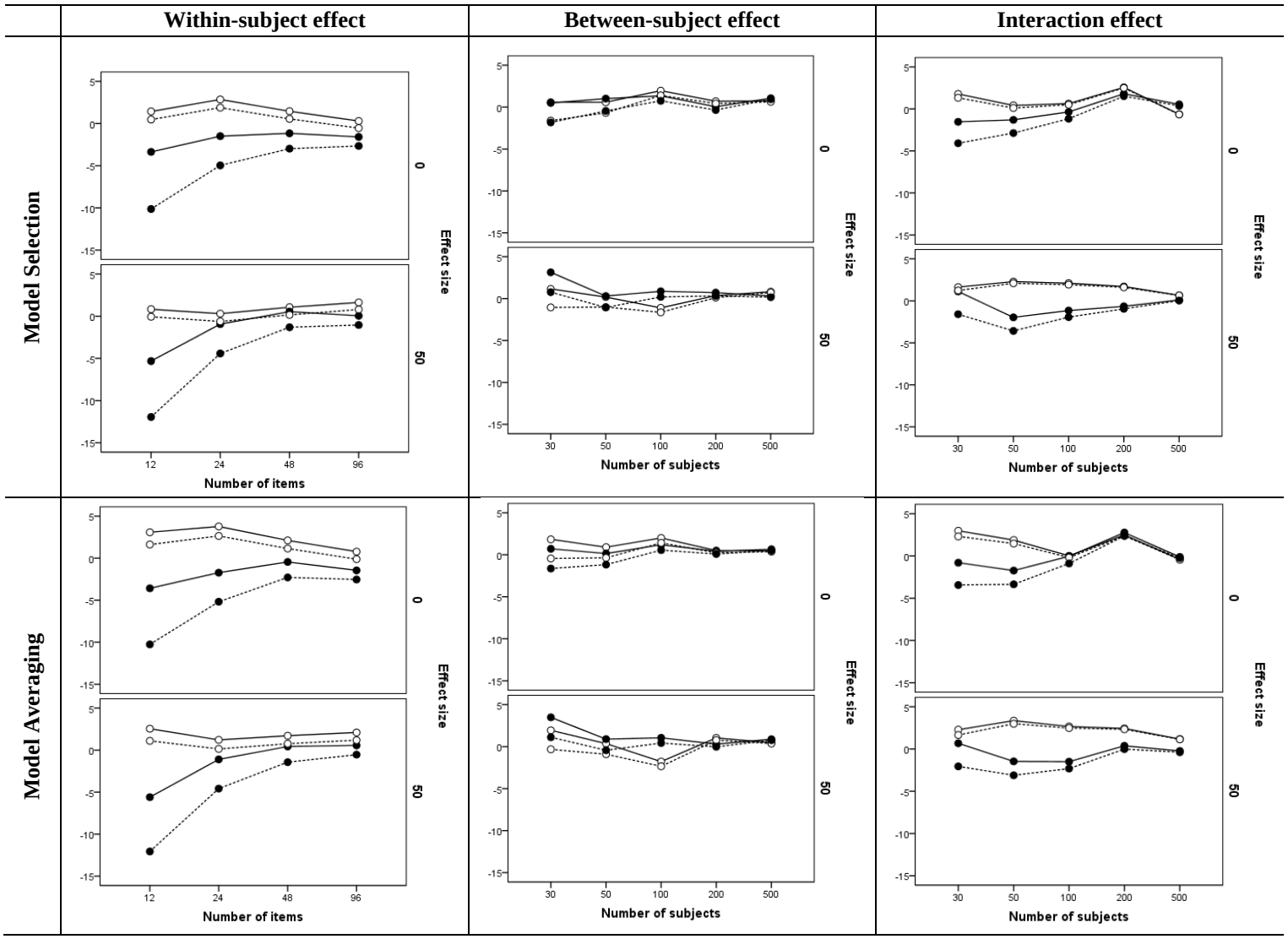
			Within-subject effect				Between-subject effect				Interaction effect			
			ML		REML		ML		REML		ML		REML	
True Model	**Items**	**Subjects**	**Model Selection**	**Model Averaging**	**Model Selection**	**Model Averaging**	**Model Selection**	**Model Averaging**	**Model Selection**	**Model Averaging**	**Model Selection**	**Model Averaging**	**Model Selection**	**Model Averaging**
All			-3.04 (13.93)	-2.65 (13.92)	-0.75 (14.70)	-0.29 (14.62)	0.10 (8.80)	0.03 (8.69)	0.99 (8.96)	0.93 (8.88)	-0.71 (7.56)	-0.38 (7.47)	0.01 (7.83)	0.36 (7.70)
Minimal [0; 0]	12	50	-0.19 (6.91)	1.26 (6.24)	0.57 (8.35)	2.57 (7.36)	-1.29 (8.10)	-1.08 (8.03)	-0.30 (8.31)	-0.04 (8.25)	-0.81 (5.30)	-0.38 (5.05)	-0.49 (5.61)	0.09 (5.35)
		200	1.49 (5.85)	2.73 (5.44)	2.01 (7.47)	3.67 (6.79)	-0.40 (4.33)	-0.36 (4.29)	-0.21 (4.42)	-0.17 (4.35)	1.55 (3.68)	1.64 (3.67)	1.66 (3.79)	1.76 (3.78)
	96	50	2.16 (4.27)	3.16 (4.08)	2.56 (5.23)	3.67 (4.61)	-2.49 (9.61)	-2.50 (9.60)	-1.04 (9.83)	-1.07 (9.82)	1.16 (4.25)	1.81 (4.11)	1.49 (5.11)	2.21 (4.60)
		200	4.91 (2.89)	5.53 (2.52)	5.04 (3.17)	5.75 (2.74)	-0.35 (5.86)	-0.39 (5.87)	-0.06 (5.88)	-0.09 (5.88)	3.26 (4.21)	3.54 (4.12)	3.30 (4.26)	3.65 (4.18)
Subjects [14,400; 0]	12	50	-1.21 (12.29)	-.96 (10.44)	0.23 (12.63)	1.20 (11.17)	2.64 (11.95)	2.78 (10.57)	3.66 (11.83)	3.76 (10.78)	-4.38 (11.23)	-5.26 (9.25)	-3.39 (11.25)	-3.75 (9.65)
		200	-0.24 (9.98)	0.97 (8.40)	0.16 (10.57)	2.05 (9.04)	-0.35 (8.66)	-0.97 (7.29)	0.33 (8.72)	-0.63 (7.27)	1.34 (9.47)	1.69 (7.76)	1.58 (9.54)	2.09 (7.74)
	96	50	-2.60 (10.24)	-2.54 (10.23)	-0.60 (10.46)	-0.53 (10.44)	-1.83 (10.32)	-1.87 (10.31)	-0.24 (10.53)	-0.28 (10.52)	-4.13 (10.08)	-4.17 (10.07)	-2.16 (10.29)	-2.20 (10.28)
		200	-0.46 (4.95)	-0.40 (4.94)	0.04 (4.97)	0.11 (4.96)	0.77 (6.07)	0.72 (6.06)	1.11 (6.09)	1.05 (6.08)	1.25 (5.13)	1.19 (5.12)	1.76 (5.16)	1.70 (5.15)
Items [0; 14,400]	12	50	-9.59 (23.54)	-9.16 (23.36)	-3.18 (26.64)	-2.88 (26.40)	-1.51 (7.49)	-0.91 (7.62)	-0.52 (7.71)	0.17 (7.85)	1.25 (4.00)	1.52 (4.11)	1.54 (4.49)	1.89 (4.45)
		200	-13.77 (25.62)	-13.52 (25.89)	-6.15 (28.33)	-5.93 (28.59)	1.94 (5.44)	0.63 (4.67)	2.19 (5.54)	0.88 (4.68)	0.16 (2.87)	0.39 (3.39)	0.30 (2.94)	0.49 (3.48)
	96	50	-1.44 (8.73)	-1.29 (8.66)	-0.53 (8.86)	-0.32 (8.77)	-1.88 (9.43)	-1.15 (9.67)	-0.43 (9.65)	0.36 (9.89)	1.61 (3.73)	4.54 (4.03)	1.74 (4.12)	4.84 (4.43)
		200	-3.54 (9.22)	-3.03 (9.26)	-2.57 (9.32)	-2.05 (9.36)	0.65 (5.00)	1.70 (5.55)	0.96 (5.02)	2.02 (5.58)	1.37 (1.51)	1.28 (2.50)	1.42 (1.66)	1.33 (2.57)
Maximal [14,400; 14,400]	12	50	-11.24 (20.57)	-11.48 (19.82)	-5.59 (22.79)	-6.13 (22.02)	2.74 (10.88)	1.98 (10.63)	3.65 (11.10)	3.01 (10.93)	-2.83 (11.19)	-2.22 (9.91)	-1.36 (11.74)	-0.89 (10.36)
		200	-11.90 (24.64)	-12.68 (24.92)	-4.73 (27.39)	-5.60 (27.52)	0.99 (8.86)	0.98 (7.25)	1.53 (9.01)	1.37 (7.37)	-2.39 (9.16)	-1.14 (8.08)	-2.78 (9.26)	-0.90 (8.18)
	96	50	-.28 (8.12)	2.09 (8.57)	1.10 (8.23)	3.50 (8.69)	-2.47 (10.10)	-2.41 (10.18)	-0.91 (10.31)	-0.84 (10.38)	-1.57 (9.36)	-2.37 (10.06)	0.02 (10.15)	-0.79 (10.27)
		200	-1.39 (7.13)	0.27 (7.38)	-0.71 (7.21)	0.96 (7.45)	-2.79 (5.11)	-2.40 (5.91)	-2.49 (5.13)	-2.09 (5.94)	0.74 (5.10)	1.63 (5.12)	1.06 (5.12)	1.95 (5.14)

Figure 1. Proportion of true model selection for relevant interactions between simulation conditions in bottom-up model selection with one-tailed likelihood ratio test (α=0.05, ML estimator).



Note: y-axis = Proportion of true model selection. x-axis = Number of subjects. Different lines and symbols are used to define number of items. Columns and rows define the simulation scenario by the presence of random slopes [0; 3,600; 14,400].

Figure 2. Percentage of SE bias of fixed effects for relevant interactions between simulation conditions in bottom-up model selection with one-tailed likelihood ratio test (α=0.05) and model averaging (Akaike weights).



Note: y-axis = SE bias [range = -15–5%]. Discontinuous lines = ML estimator. Continuous lines = REML estimator. Random slopes [○ = 0; ● = 14,400] for subjects or items (depends on x axis factor).