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

PsychOpen GOLD

Original Article

Check for updates

Recovering Crossed Random Effects in Mixed-Effects Models Using Model Averaging

José Ángel Martínez-Huertas[1], Ricardo Olmos[2]

[1] *Faculty of Psychology, Universidad Nacional de Educación a Distancia, Madrid, Spain.* **[2]** *Faculty of Psychology, Universidad Autónoma de Madrid, Madrid, Spain.*

Methodology, 2022, Vol. 18(4), 298–323, https://doi.org/10.5964/meth.9597

Received: 2022-05-31 • **Accepted:** 2022-12-08 • **Published (VoR):** 2022-12-22

Handling Editor: Francisco Abad, Universidad Autónoma de Madrid, Madrid, Spain

Corresponding Author: José Ángel Martínez-Huertas, Faculty of Psychology, Universidad Nacional de Educación a Distancia, C. de Juan del Rosal, 10, 28040 Madrid, Spain. E-mail: jamartinez@psi.uned.es

Abstract

Random effects contain crucial information to understand the variability of the processes under study in mixed-effects models with crossed random effects (MEMs-CR). Given that model selection makes all-or-nothing decisions regarding to the inclusion of model parameters, we evaluated if model averaging could deal with model uncertainty to recover random effects of MEMs-CR. Specifically, we analyzed the bias and the root mean squared error (RMSE) of the estimations of the variances of random effects using model averaging with Akaike weights and Bayesian model averaging with BIC posterior probabilities, comparing them with two alternative analytical strategies as benchmarks: AIC and BIC model selection, and fitting a full random structure. A simulation study was conducted manipulating sample sizes for subjects and items, and the variance of random effects. Results showed that model averaging, especially Akaike weights, can adequately recover random variances, given a minimum sample size in the modeled clusters. Thus, we endorse using model averaging to deal with model uncertainty in MEMs-CR. An empirical illustration is provided to ease the usability of model averaging.

Keywords

mixed-effects models, crossed random effects, random effects, model averaging, Akaike weights, Bayesian model averaging, AIC, BIC

Mixed-effects models with crossed random effects (MEMs-CR) are an affordable statistical model to analyze data with complex random structures where the levels of the random effects are crossed (e.g., Baayen et al., 2008; Hoffman & Rovine, 2007; Martínez-Huertas & Ferrer, 2022; Martínez-Huertas et al., 2022; Quené & van den Bergh, 2004;

This is an open access article distributed under the terms of the Creative Commons Attribution 4.0 International License, CC BY 4.0, which permits unrestricted use, distribution, and reproduction, provided the original work is properly cited.

Raudenbush, 1993). The usefulness of crossed random effects can be seen in different contexts like, for example, the sampling of participants and items in experimental research, or the sampling of students in neighborhoods and schools. In the first scenario, crossed random effects could capture variability within subjects and items in reaction times, as all the items are usually responded by all the participants (e.g., Baayen et al., 2008). In the second scenario, we could find more complex random structures where students would be nested within the crossed random effects of neighborhoods and schools, when schools can have students from different neighborhoods and students from the same neighborhood can go to different schools (e.g., Raudenbush, 1993). As we are going to explain later, these random effects also contain crucial substantive information to understand the variability of the processes under study. Please consider a hypothetical experimental design where response time (milliseconds -ms-) of multiple items by different participants were analyzed in two different conditions (control vs. experimental). Equation 1 presents a MEMs-CR that could be used to analyze such study:

$$Y_{tsi} = \gamma_{000} + \gamma_{100}(W_{tsi}) + U_{0s0} + U_{00i} + U_{1s0}(W_{tsi}) + U_{10i}(W_{tsi}) + e_{tsi} \quad (1)$$

where Y_{tsi} is the response time of participant s for item i in condition t, and W_{tsi} (control vs. experimental conditions) is a within factor where the items were presented in both control ($W_{tsi} = 0$) and experimental ($W_{tsi} = 1$) conditions. Fixed effects are represented with gamma (γ) letters and random effects are represented with U letters. γ_{000} is the mean of the control condition (intercept), γ_{100} is the experimental effect, U_{0s0} and U_{00i} are the random intercepts for subjects and items, U_{1s0} and U_{10i} are the random slopes for subjects and items, and e_{tsi} is the error term. Equation 1 assumes that there is a complex random structure with both random slopes for items and subjects, but there are many other intermediate parametrizations that could be considered for the same data by imposing null random slopes in each cluster. For example, there could be variability in the intercepts of both subjects and items, but we could find that all the items present the same experimental effect (i.e., null random slopes for items) or that the experimental effect tend to affect all the subjects in a similar way (i.e., null random slopes for subjects). The presence of not-null random slopes would mean that the estimated experimental effect is different for subjects and/or items.

It is common to know the fixed effects of the study when the experimental design is simple, like the one of the present simulation study. This allows to focus on the analysis of the variances of random effects and their implications for the target fixed effects (although, usually, the inclusion of random effects is used just as a way of statistical control). But more complex designs with different between and/or within effects, or crossed and/or nested effects, would require to, firstly, know the target fixed effects and, secondly, to test them after estimating the target random structure. Thus, even when researchers know the fixed effects of their study, some decisions must be made about the specification of the random structure. This could lead to important bias and errors of

PsychOpen GOLD

statistical inference in MEMs if inadequate random structures are fitted (e.g., Hoffman, 2015; Hox et al., 2018; Martínez-Huertas et al., 2022; Meyers & Beretvas, 2006; McNeish & Kelley, 2019; McNeish et al., 2017). Moreover, there is consensus about the necessity of including all relevant variability dimensions in conditions to potentiate ecological validity (Hoffman, 2015). Consequently, researchers have been studying the performance of different model selection strategies to establish random structures in MEMs and MEMs-CR (e.g., Barr et al., 2013; Martínez-Huertas et al., 2022; Matuschek et al., 2017). But model selection based on likelihood ratio tests, AIC and BIC of MEMs-CR seems to reach, approximately, an 80% of true random structure selection (see Martínez-Huertas et al., 2022). This result is promising and shows that model selection is undoubtedly a good practice for researchers, but also means that researchers would select an incorrect model 20% of the times, and that more complex designs are expected to present lower rates of true random structure selection. In these contexts, model selection would mean to make an all-or-nothing decision regarding to the inclusion of parameters like random slopes.

Model averaging attempts to use all the available information in the competing models to increase the precision of the model estimations and to deal with model uncertainty (e.g., Burnham & Anderson, 2002, 2004; Claeskens & Hjort, 2008; Kaplan & Lee, 2018; Konishi & Kitagawa, 2008). Model averaging would fit all the possible models and generate an average estimation for the target parameters by weighting the information from all possible models. This weighting is based on the relative fit index of each model. It is known that model averaging can reduce overconfidence avoiding threshold-based all-or-nothing decision making and is expected to be relatively robust against model misspecification (Hinne et al., 2020), and have been found to be a less risky option than model selection in terms of the bias of standard errors of fixed effects (Martínez-Huertas et al., 2022). Thus, these methodological tools are supposed to deal with the uncertainty of random effects and avoid errors of statistical inference, comparing to model selection that makes an all-or-nothing decision.

In the following lines, we present two model averaging perspectives: model averaging with Akaike weights and Bayesian model averaging (BMA) with BIC posterior probabilities.

Model Averaging With Akaike Weights

Model averaging with Akaike weights uses the AIC index to compute an average estimate of the target parameters (e.g., Akaike, 1979; Burnham & Anderson, 2002; Claeskens & Hjort, 2008; Kishino et al., 1991; Steele et al., 2014). This procedure involves fitting the competing models, M_1, M_2, ..., M_r, and computing the so-called Akaike weights using the AIC index (e.g., Akaike, 1974) for each competing model as relative evidence in favor of each one among all the competing models (Burnham & Anderson, 2002). As explained, the AIC indices of all R models are ranked as follows:

PsychOpen GOLD

$$\Delta_r = AIC_r - AIC_{min} \quad (2)$$

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

$$\omega_r = \frac{\exp(-\Delta_r/2)}{\sum_{r=1}^{R} exp(-\Delta_R/2)} \quad (3)$$

where ω_r is the resulting Akaike weight for each competing model based on Δ_r of all *R* competing models (Burnham & Anderson, 2002). Next, a weighted estimation of the target parameters is obtained by weighting the estimations of each model *r* using its ω_r. These averaged estimations have been found to generate estimations of SEs of fixed effects of MEMs-CR with small bias, being a less risky approach than model selection strategies (Martínez-Huertas et al., 2022). In the present study, model averaging was applied on the estimation of the random effects of MEMs-CR.

Bayesian Model Averaging With BIC Posterior Probabilities

BMA shares the same rationale than model averaging with Akaike weights about the computation of averaged estimates for target parameters, but using a Bayesian framework (e.g., Chatfield, 1995; Draper, 1995; Fragoso et al., 2018; Kaplan & Lee, 2018; Hinne et al., 2020; Steel, 2020). It is worth mentioning that this approach works with posterior probabilities and that Bayes factor is one of the most accurate approximations to them, but in this study we are going to use BIC as an approximation to the posterior probability of the competing models due to its computational simplicity and effective performance (see Schwarz, 1978, and Neath & Cavanaugh, 2012 for a demonstration of the derivation of prior distributions based on BIC). Once different competing models M_1, M_2, ..., M_r, have been fitted to the data set *D*, we can extract their BIC indices (BIC_1, BIC_2, ..., BIC_r). Again, it is necessary to rank all the *R* competing models using:

$$\Delta_r = BIC_r - BIC_{min} \quad (4)$$

where Δ_r represents the difference between the BIC of each *r* competing model and that of the best fitting model (BIC_{min}). Then, it is possible to approximate the posterior probability of each *r* model given data *D* as:

$$Pr\left(M_j \middle| D\right) = \frac{\exp(-\Delta_r/2)}{\sum_{r=1}^{R} exp(-\Delta_R/2)} \quad (5)$$

where $Pr\left(M_j \middle| D\right)$ is an approximation to the posterior probability of each competing model, and shares an important equivalence with Akaike weights as they are based on

PsychOpen GOLD

Δ_r of all R competing models. Similarly, a weighted estimation of the target parameters is obtained weighting the estimations of each model r using its posterior probability.

The Present Study

Random effects should be understood as informative parameters of the target processes under study, making them a relevant substantive part of the results report (Barr, 2013 presented a similar rationale on the use of random effects from a confirmatory perspective). Large intercept variances for subjects or items would indicate that there are important differences in the mean (e.g., some participants could be slower than others, or some items could be more difficult than others). Similarly, large slope variances for subjects or items would indicate that there are important differences in how the fixed effect and the dependent variable are related (e.g., some participants or items could be more sensitive to the experimental conditions than others). Thus, random effects should not be understood as of secondary interest, but as relevant and substantive information about the modeled processes. That is one of the main reasons to study the performance of model averaging, which aims to deal with model uncertainty, and to compare it with model selection all-or-nothing decision making about the inclusion of such random effects. In this line, it is known that the estimation of fixed effects does not present bias regardless of whether the random effects of the fitted model are correct or not, and that they are very similar across different random structures (Hoffman, 2015; Hox et al., 2018; Meyers & Beretvas, 2006). Previous research found that the estimation of the standard errors of fixed effects was more efficient for model averaging than for model selection (Martínez-Huertas et al., 2022). We think that such efficiency is related with the consideration of random effects.

In this line, there is a handicap when researchers want to recover the random structure of MEMs-CR using model averaging. This procedure uses all the available information from the competing models but, unfortunately, some models do not include all the parameters of interest. This is the case of random effects. Researchers could try to average random intercepts or random slopes when some of the competing models do not include such parameters by weighting null information (a zero) in model averaging. Thus, the usefulness of model averaging could be compromised recovering different random structures. The present study aims to evaluate the bias of averaged crossed random effects in simulation scenarios where the random structure of various competing models is incomplete. For this purpose, we are going to compare the average estimates of the two model averaging strategies introduced previously (Akaike weights and BMA with BIC posterior probabilities), with model selection of AIC and BIC indices as benchmarks. We think that there will be relevant differences between both perspectives because model averaging is supposed to deal with model uncertainty while model selection always does an all-or-nothing decision making. Given that all the MEMs-CR of the study are nested, the maximal MEM-CR (Barr et al., 2013; which is a MEM-CR with random

PsychOpen GOLD

intercepts and random slopes for both subjects and items in this study) was used as an additional benchmark to compare with the averaged estimations. All these conditions were studied in two different simulation scenarios under the presence of symmetric and asymmetric variances of random intercepts and slopes of subjects and items, that is, when both random effects present the same or different variability. Although the simulation scenario of symmetric variances of random effects is more artificial, it is useful to analyze if the quality of the estimations of variances of random intercepts and random slopes tends to be similar. The simulation scenario of asymmetric variances of random effects is more ecologic because of usually the random effects present different variabilities in real data sets.

Method

Simulation Study

An experimental design in psycholinguistics was simulated using Equation 1 as the generating model. Specifically, response time (ms) of multiple items by different participants were analyzed in two different conditions (control vs. experimental). Thus, a within-subject effect was simulated here. Two different sources of variability were simulated: participants and items. These random effects were fully crossed due to all the participants answered all the items. In our simulation study, we used similar fixed and random effects population parameters as those in previous related work (Baayen et al., 2008; Barr et al., 2013; Matuschek et al., 2017; Martínez-Huertas et al., 2022). Table 1 presents the simulation parameters. The within-subject effect was set to 0 in all simulation conditions (fixed effects do not present bias as has been evidenced previously: Hoffman, 2015; Hox et al., 2018; Meyers & Beretvas, 2006). The mean of the control condition (intercept) was set to 2,000 ms. The residual level-1 variance σ^2_e was 90,000, so a 300 standard deviation should be understood as the residual level-1 error. Given that the focus of this study was the study of random effects, almost all the random structure of this model was manipulated in the simulation. The cluster-specific random effects for subjects and items intercepts are represented by U_{0s0} and U_{00i}, respectively, and their variances σ^2_{0s0} and σ^2_{00i}, respectively) were set to 0, 10,000 and 40,000 ms (standard deviations equal to 0, 100 and 200 ms). The subject and item cluster-specific random slopes are represented by U_{1s0} and U_{10i}. Two different variances σ^2_{1s0} and σ^2_{10i}, respectively) were simulated for subjects and items random slopes (0 and 10,000 ms; standard deviations equal to 0 and 100 ms). An intercept-slope covariance for subjects ($\sigma_{U_{0s0},U_{1s0}}$) and items ($\sigma_{U_{00i},U_{10i}}$) was simulated equivalent to an intercept-slope correlation equal to 0 and .60. The random effects and the error term were normally distributed. Different sample sizes were simulated for subjects and items (sample size for subjects: 15, 30, 60, and 90 subjects; and sample size for items: 5, 10, 30, and 60 items). A total of 1,152 simulation conditions were considered

PsychOpen GOLD

in the present study, and 1,000 replications were generated for each simulation condition using R software. This makes a total of 1,152,000 replications.

Table 1

Simulation Parameters

Simulation conditions	Parameter		Values
Fixed conditions	Intercept	γ_{000}	2,000 ms
	Within-subject effect	γ_{100}	0 ms
	Residual level 1 variance	σ_e^2	300 ms
Manipulated conditions	Random intercepts for subjects	σ_{0s0}	0; 100; 200 ms
	Random intercepts for items	σ_{00i}	0; 100; 200 ms
	Random slopes for subjects	σ_{1s0}	0; 100 ms
	Random slopes for items	σ_{10i}	0; 100 ms
	Intercept-slope correlation for subjects and items	r_{01}	0; 0.60
	Sample size for subjects	N_1	15; 30; 60; 90 subjects
	Sample size for items	N_2	5; 10; 30; 60 items

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

Data Analysis

We assume that MEMs-CR naturally have random intercepts for both subjects and items. Also, we consider that the most complex MEM-CR will be the one with random intercepts and random slopes for both subjects and items. The former is called here *minimal*, whilst the latter is called here *maximal* MEMs-CR. Two intermediate random structures were also considered here: a MEM-CR with random intercepts for both subjects and items and random slopes for subjects (here, subject random slopes), and a MEM-CR with random intercepts for both subjects and items and random slopes for items (here, item random slopes). Thus, four MEMs-CR were fitted to the data using the restricted maximum likelihood (REML) approach. This makes a total of 4,608,000 analyses (1,152,000 replications x 4 MEMs-CR). All MEMs were fitted with the *lme4* package (Bates et al., 2015) in R software. *lme4*'s AIC and BIC fit indices were used to compute model selection and model averaging. Approximately, 99.79% of the estimated models converged. Please note that the simulated data was analyzed with random intercepts even when some simulation conditions set those parameters to zero to emulate a natural analytical strategy, and to compare the recovery of random intercepts and random slopes under the same simulation conditions.

First, model selection performance of AIC and BIC fit indices was evaluated. It was computed as the proportion of true model selection: if the model used to simulate the

PsychOpen GOLD

data was selected, it was considered a correct model selection. Otherwise, it was considered as incorrect (please note that selecting the *minimal* MEMs-CR was also considered correct in those simulation conditions that set intercepts to zero to ease the simulation design). In the present study, these results are useful to contextualize the mean true model selection of each fit index as benchmarks.

Second, the estimations of model averaging with Akaike weights and BMA with BIC posterior probabilities were computed. For this purpose, we applied the above-mentioned procedures using the four competing models of this simulation study, namely: minimal, subject random slopes, item random slopes, and maximal MEMs-CR. Specifically, we computed the average estimations of the random effects.

Third, the bias of the estimations of the variances of the random effects was computed using different measures. Bias was computed for all the random effects using the following formula:

$$\frac{1}{R}\sum_{r=1}^{R}\left(\widehat{\theta}_r - \theta\right), \tag{6}$$

where θ is the population parameter and $\widehat{\theta}_r$ is the sample estimate of each R replicate. The interpretability of bias was done considering the simulated random variance in not-null variance conditions, that is, calculating the percentage of bias regarding the 100 or 200 simulated variances. Root mean squared error (RMSE) was also computed for all the random effects using the following formula:

$$\sqrt{\sum_{r=1}^{R}\frac{\left(\widehat{\theta}_r - \theta\right)^2}{R}}, \tag{7}$$

where θ is the population parameter and $\widehat{\theta}_r$ is the sample estimate of each R replicate. To ease the interpretability of the results, the analyses were conducted in different simulation scenarios depending on (1) symmetric vs. asymmetric simulated population variances of random intercepts and slopes; and (2) zero vs. not-zero simulated population random effects.

Results

Model selection with AIC and BIC and fitting a full random structure (*maximal model*) were used as benchmarks to compare with model averaging performance. As we will see in the results, the worse performance of model selection could be attributed to incorrect model selections that do not include the target random effects. Given that all the MEMs-CR of this simulation study were nested (they just have different random structures), fitting a full random structure (here, the maximal MEM-CR) is used as

PsychOpen GOLD

another benchmark to compare with model averaging because it contains all the random effects under study.

First Simulation Scenario: Symmetric Variances of Random Intercepts and Slopes

AIC and BIC Model Selection

Table 2 presents the proportion (and standard deviation) of true model selection in the conditions of the simulation study. True model selection was higher for AIC in small sample sizes of subjects and items, comparing to BIC, when there were random slopes. These differences decrease as the number of subjects and items increase, and BIC obtains slightly higher true model performances in larger sample sizes. In general, BIC infra-parametrized the random structure of the model in almost all the simulation conditions (which favored the selection of MEMs-CR without random slopes), while AIC showed more accurate true model selections in demanding conditions with smaller sample sizes being more sensible to the presence of variability in random slopes. In this line, medium to large sample sizes were required in both clusters to obtain appropriate true model selection. These results show that researchers would select an important proportion of incorrect models in many of these simulation conditions (which means that true variances of random effects would be incorrectly set to zero). A similar pattern of results was found in the second simulation scenario (simulation conditions with asymmetric variances of random intercepts and slopes), but true model selection was even worse than in the first simulation scenario (simulation conditions with symmetric variances of random intercepts and slopes). As we will see, this could explain the advantages of model averaging in front of model selection.

PsychOpen GOLD

Table 2

Proportion (and Standard Deviation) of True Model Selection in the Simulated Conditions

Fit index	Number of subjects	Number of items	Minimal	Subject random slopes	Item random slopes	Maximal
AIC	**15**	**5**	.91 (.28)	.25 (.43)	.20 (.40)	.07 (.25)
	15	**10**	.90 (.29)	.45 (.50)	.42 (.49)	.22 (.41)
	15	**30**	.90 (.30)	.81 (.39)	.85 (.37)	.78 (.41)
	15	**60**	.90 (.30)	.92 (.28)	.92 (.27)	.96 (.18)
	30	**5**	.91 (.28)	.44 (.50)	.31 (.46)	.18 (.39)
	30	**10**	.91 (.29)	.72 (.45)	.63 (.48)	.51 (.50)
	30	**30**	.90 (.31)	.92 (.26)	.92 (.26)	.98 (.14)
	30	**60**	.89 (.31)	.94 (.23)	.93 (.26)	.99 (.02)
	60	**5**	.91 (.28)	.67 (.47)	.51 (.50)	.41 (.49)
	60	**10**	.91 (.29)	.88 (.32)	.82 (.38)	.84 (.36)
	60	**30**	.89 (.31)	.93 (.25)	.94 (.24)	.99 (.01)
	60	**60**	.89 (.31)	.93 (.26)	.93 (.26)	1.00 (.00)
	90	**5**	.91 (.29)	.77 (.42)	.60 (.49)	.57 (.49)
	90	**10**	.90 (.30)	.92 (.27)	.88 (.33)	.92 (.27)
	90	**30**	.88 (.31)	.93 (.25)	.94 (.24)	1.00 (.00)
	90	**60**	.88 (.32)	.93 (.25)	.94 (.24)	1.00 (.00)
BIC	**15**	**5**	.99 (.06)	.05 (.22)	.03 (.16)	.00 (.04)
	15	**10**	.99 (.05)	.11 (.32)	.09 (.28)	.01 (.11)
	15	**30**	.99 (.02)	.42 (.49)	.52 (.49)	.24 (.42)
	15	**60**	.99 (.02)	.79 (.41)	.87 (.34)	.69 (.46)
	30	**5**	.99 (.04)	.15 (.36)	.05 (.21)	.00 (.08)
	30	**10**	.99 (.02)	.34 (.47)	.20 (.40)	.07 (.26)
	30	**30**	.99 (.02)	.86 (.34)	.87 (.34)	.75 (.43)
	30	**60**	.99 (.02)	.99 (10)	.99 (.08)	.99 (.12)
	60	**5**	.99 (.03)	.34 (.47)	.11 (.32)	.04 (.20)
	60	**10**	.99 (.03)	.68 (.47)	.48 (.50)	.34 (.47)
	60	**30**	.99 (.01)	.99 (.08)	.99 (.08)	.98 (.13)
	60	**60**	1.00 (.00)	.99 (.01)	1.00 (.00)	1.00 (.00)
	90	**5**	.99 (.02)	.49 (.50)	.19 (.39)	.10 (.30)
	90	**10**	.99 (.02)	.83 (.38)	.68 (.47)	.59 (.49)
	90	**30**	.99 (.01)	.99 (.03)	.99 (.01)	.99 (.03)
	90	**60**	1.00 (.00)	1.00 (.00)	1.00 (.00)	1.00 (.00)

PsychOpen GOLD

Bias and RMSE of the Estimations of the Variances of Random Effects of Subjects and Items

In the first simulation scenario, symmetric variances of random intercepts and slopes (100 and 100 standard deviations, respectively) were considered for the two clusters (subjects and items). Figure 1 presents the mean of bias and RMSE of the estimations of null symmetric variances of random effects of subjects and items in model averaging and model selection with AIC and BIC, and fitting a full random structure (maximal model). Given that a standard deviation = 100 was simulated in this section of the study, the bias can be easily interpreted as percentage of bias. Regarding to bias, it was found that fitting a full random structure (that is, fitting the *maximal* model) presented the largest bias in almost all the simulation conditions. The worse performance of fitting a full random structure was highlighted in the estimations of null random slopes, as the model would not be able to correctly distribute the variability of the random effects. Model selection and model averaging presented a similar pattern of results in these simulation conditions, being the main differences of bias related to the AIC and BIC indices. The variances of random slopes were found to present small bias in almost all the conditions for model selection and model averaging (that is, bias tend to present less than 10% of bias), while the variances of random intercepts showed larger bias in the most demanding conditions (reaching a 35% of bias in the most demanding conditions). Probably, this result is related to the lack of capacity to distribute the variability of not-null random slopes when the models are also estimating random intercepts whose true variance is null (note that this is an artifact of the simulation design where the less complex MEM-CR used to analyze the data—the minimal MEM-CR—included random intercepts of subjects and items). Given that BIC tend to infra-parametrize the random structure of the model, both Bayesian model averaging with BIC posterior probabilities and model selection with BIC presented relatively accurate results for the estimations of null symmetric variances of random effects of subjects and items. But, as we will see later, the same reason generates more bias in the estimates of not-null symmetric variances of random effects. Model averaging with Akaike weights and model selection with AIC showed a similar pattern of results, reaching appropriate levels of bias (less than 10%) in almost all the conditions of the random slopes. RMSE presented a similar pattern of results. In this sense, the sample variance decreases as both sample sizes increase. Similarly, sample variance was larger for the procedures using AIC than for the ones using BIC for the estimation of null symmetric variances of these random effects.

Figure 2 presents the mean of bias and RMSE of the estimations of not-null symmetric variances of random effects of subjects and items in model averaging and model selection with AIC and BIC, and fitting a full random structure (maximal model). As introduced above, a different pattern of results was found for the estimations of not-null symmetric variances in these simulation conditions. Regarding to bias, accurate estimates of variances of random intercepts of subjects and items were found for model averaging

PsychOpen GOLD

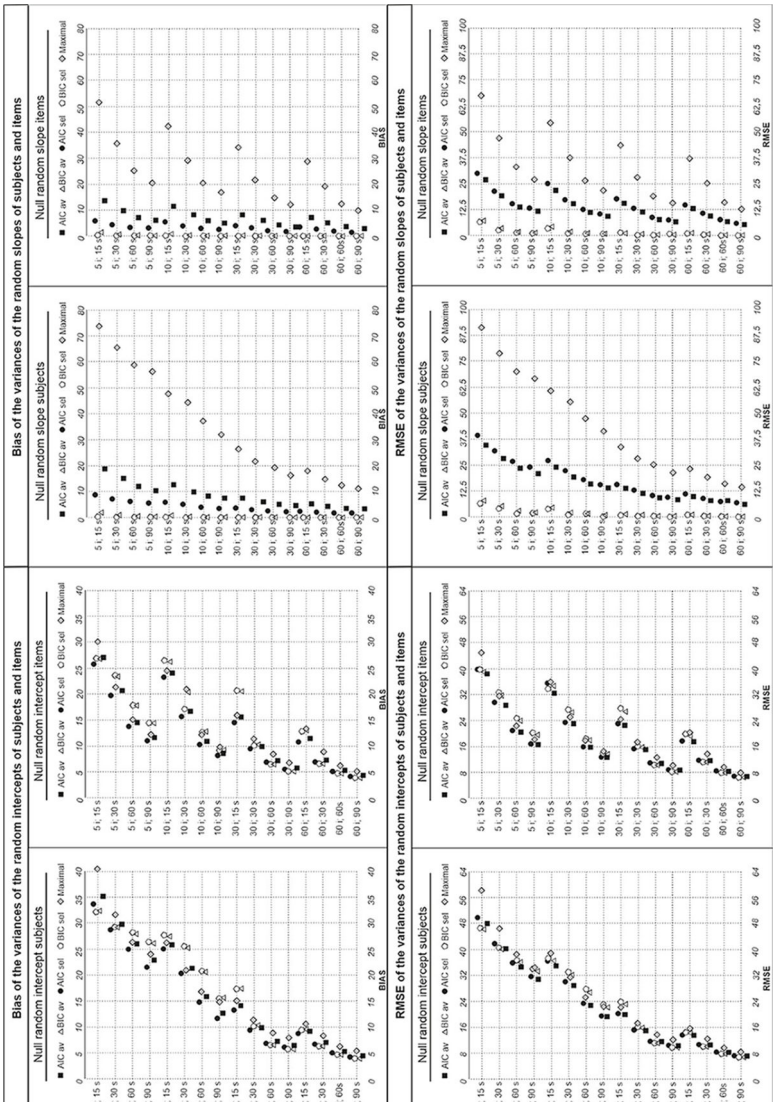
and model selection with AIC and BIC indices. That is, we found small bias (less than 10%) even in the most demanding conditions of the study. On the contrary, important bias was found for the estimations of the variances of random slopes when there are small sample sizes. In general, Bayesian model averaging with BIC posterior probabilities and model selection with BIC presented inappropriate bias (surrounding the 90% of bias in different conditions), requiring medium to large sample sizes in both clusters to obtain accurate results. Model averaging with Akaike weights and model selection with AIC obtained significantly better performances, but medium to large sample sizes of subjects (e.g., 60–90 subjects) were required to obtain accurate estimates of variances of random slopes when the number of items was equal or larger than ten. None of the conditions with five items reached an appropriate level of bias (being the bias of subjects random slopes larger than the bias of items random slopes). Moreover, fitting a full random structure (that is, fitting the *maximal* model) presented accurate estimations of the variances of random slopes and the random intercepts of items. But it showed a relevant underestimation of random intercepts of subjects (reaching a 30% of bias in some conditions) when there were only five items. In any case, the combination of the results of Figure 1 and Figure 2 present a negative balance for fitting a full random structure because of this strategy would generate significantly biased estimates when there are null random effects and requires medium to large sample sizes in both clusters to obtain accurate results. It is worth mentioning here that model averaging with Akaike weights presented less bias than model selection with AIC (approximately, a difference of 10–15% of bias was found in favor of model averaging) when there were small sample sizes. Again, it was found that RMSE decreases as both sample sizes increase.

While Figure 2 showed the central tendency of bias and RMSE of the estimations of not-null symmetric variances of random effects of subjects and items, it is worth mentioning that the distribution of bias was substantively different for model averaging and model selection. Figure 3 presents different histograms of bias of model averaging with Akaike weights and model selection with AIC for some simulation conditions of not-null symmetric variances of random slopes of subjects and items (1000 replications each). The mean performance of both procedures was very similar in the different simulation conditions (although we saw slight advantages around the 5–10% in favor of model averaging), but the distribution of the performances revealed that making incorrect all-or-nothing decisions in model selection would be very harmful for model estimations. These differences are highlighted in the most demanding conditions with small sample sizes (in fact, the median showed the maximum bias in such conditions, that is, a 100% of bias). In this line, model averaging weights all the available information of the model and, thus, its estimations present a relevant level of uncertainty, but it is a less risky option than model selection to estimate crossed random effects in mixed-effects models.

PsychOpen GOLD

Figure 1

Bias and RMSE of the Estimations of Null Symmetric Variances of Random Effects of Subjects and Items in Model Averaging and Model Selection With AIC And BIC, and Fitting a Full Random Structure (Maximal Model)

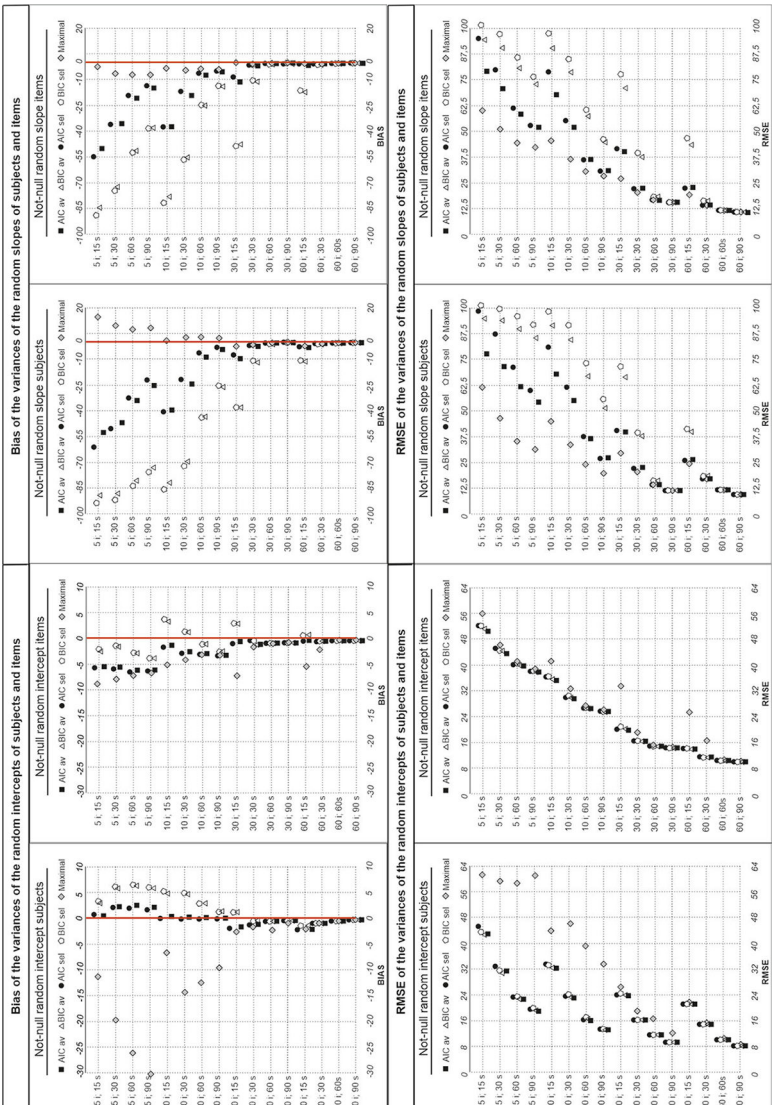


Note. AIC av. = Model averaging with Akaike weights. BIC av. = Bayesian model averaging with BIC posterior probabilities. AIC sel. = Model selection with AIC. BIC sel. = Model selection with BIC.

PsychOpen GOLD

Figure 2

Bias and RMSE of the Estimations of Not-Null Symmetric Variances of Random Effects of Subjects and Items in Model Averaging and Model Selection With AIC and BIC, and Fitting a Full Random Structure (Maximal Model)

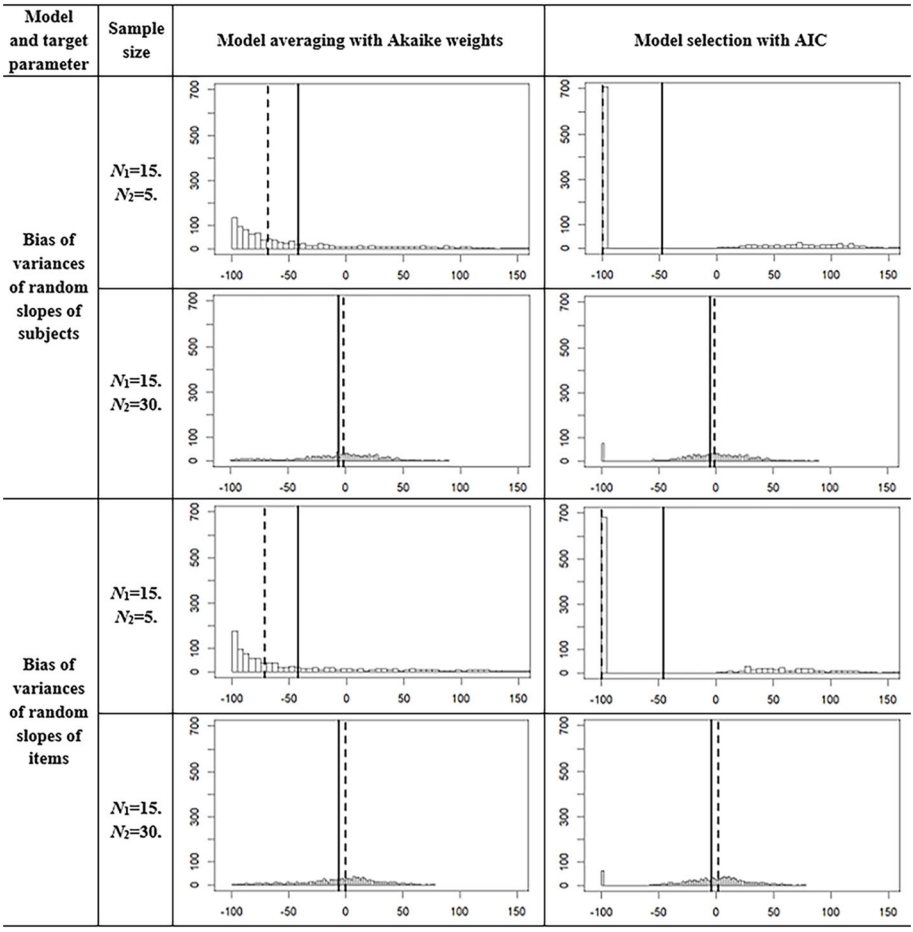


Note. AIC av. = Model averaging with Akaike weights. BIC av. = Bayesian model averaging with BIC posterior probabilities. AIC sel. = Model selection with AIC. BIC sel. = Model selection with BIC.

PsychOpen GOLD

Figure 3

Histograms of Bias of Model Averaging With Akaike Weights and Model Selection With AIC for Some Simulation Conditions (1000 Replications Each)



Note. N_1 = Number of subjects. N_2 = Number of items. x axis ranges from -110 to 150. y axis ranges from 0 to 700 replications. Continuous lines represent the mean. Discontinuous lines represent the median.

Second Simulation Scenario: Asymmetric Variances of Random Intercepts and Slopes

In the second simulation scenario, asymmetric variances of random intercepts and slopes (200 and 100 standard deviations, respectively) were considered for the two clusters (subjects and items). Additionally, we only considered the estimations of not-null variances because of their results did not present large differences with those of null symmetric variances of random effects.

PsychOpen GOLD

Thus, we only present the estimations of not-null asymmetric variances of random effects of subjects and items in model averaging and model selection in this section for the sake of brevity. Figure 4 presents the mean of bias and RMSE of the estimations of not-null asymmetric variances of random effects of subjects and items in model averaging and model selection with AIC and BIC.

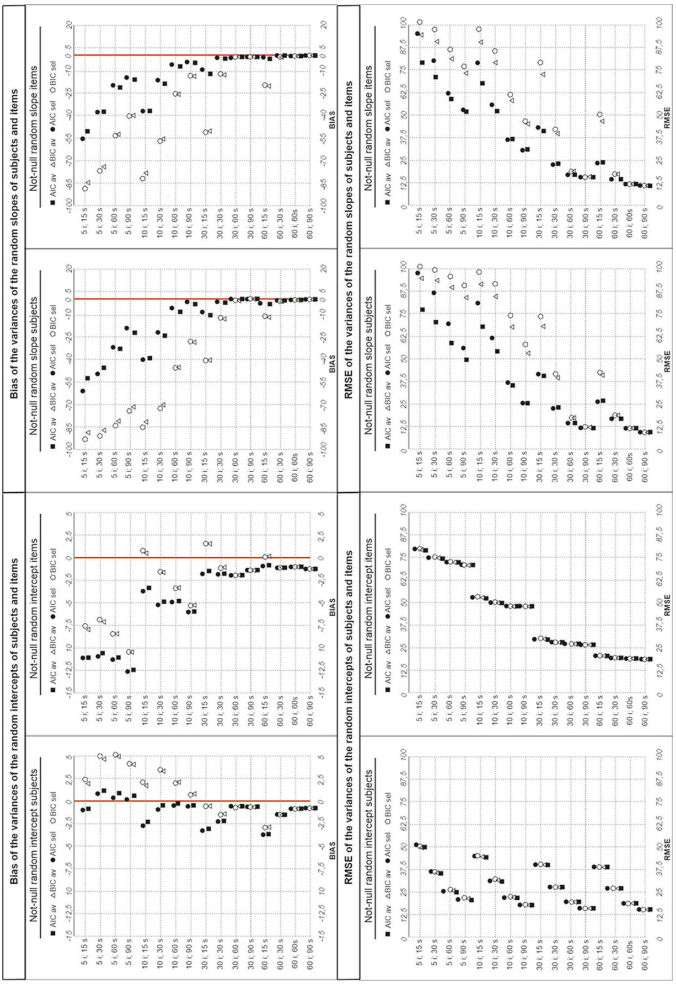
First, it was found an absence of bias of the estimations of the random variances of the intercepts of subjects and items. In the most demanding conditions, the most severe bias was found to be around the 6% (that is, approximately, -12.5 bias was for the estimation of the 200 standard deviations). Similarly, we found more bias of the estimations of random intercepts in items than in subjects. The comparison of bias and RMSE of the random intercepts of subjects and items show that there is a relevant sample variance of the estimates that decreases as the sample sizes of both clusters increase.

Second, important bias was found for the estimations of the variances of the random slopes in the more demanding conditions. Again, the model averaging and the model selection derived from the BIC index presented a significantly worse performance than the ones of AIC index (the results of both indices are only comparable when there are larger sample sizes in both clusters). Given that the simulated random slopes had a standard deviation of 100, we can easily interpret the results as the percentage of bias. Bias was large in all the conditions where the sample sizes of both clusters were small, that is, for 5–10 items and 15–30 subjects. When one of the clusters presented a medium/large sample size (e.g., 60–90 subjects or 30–60 items), the recovery of the variances of both random effects was accurate (bias was around 8%), and the results were near (but larger than) the 10% of bias even in the more extreme conditions of small sample sizes of items. On the contrary, none of the conditions with 5 items reached an appropriate level of bias (being the bias of subjects random slopes larger than the bias of items random slopes), and the 10 item conditions presented a similar pattern of results with lower bias. Again, RMSE of the random slopes of subjects and items decrease as the sample sizes of both clusters increase. In this line, it is worth mentioning that model averaging with Akaike weights presented a significantly lower bias and RMSE than model selection with AIC index in the most demanding conditions, that is, when the sample sizes were smaller in both clusters. This means that the bias of model averaging with Akaike weights, although relevant, was lower than the one of model selection with AIC, and that the sample variance of the estimates followed the same pattern of results.

PsychOpen GOLD

Figure 4

Bias and RMSE of the Estimations of Not-Null Asymmetric Variances of Random Effects of Subjects and Items in Model Averaging and Model Selection With AIC and BIC



Note. AIC av. = Model averaging with Akaike weights. BIC av. = Bayesian model averaging with BIC posterior probabilities. AIC sel. = Model selection with AIC. BIC sel. = Model selection with BIC.

PsychOpen GOLD

Empirical Illustration

In this section, we present an empirical illustration about the use of model averaging with Akaike weights in one of the simulated data sets of this study. In this data set, 90 subjects answered 60 items. Given that this is a simulated data set, we know that the population model has a standard deviation of 100 (a variance equal to 10,000 ms was simulated) for the random slopes of subjects, and a zero-variance for the random slopes of items. Then, the correct model that should be fitted to the data is a MEM-CR with not-null variances for the intercepts of both clusters, a not-null variance for random slopes of subjects and a null variance for random slopes of items. But the reality is that applied researchers typically do not know the population model of the random structure underlying their data. Imagine that a researcher aiming to analyze this data set is considering four possible MEMs-CR with different random variances (e.g., *minimal*, *subject random slopes*, *item random slopes*, and *maximal*). Such researcher would fit the four competing models to obtain their estimates and their AIC indices. The minimal MEM-CR presents an AIC = 77309.33, the model with subject random slopes has an AIC = 77207.52, the model with item random slopes presents an AIC = 77311.59, and the maximal MEM-CR has an AIC = 77209.74. Considering that the model with item random slopes presented the lower AIC, then it is possible to apply Equation 2 to obtain a measure of the relative fit of each model comparing to the best fitting model: 101.81, .00, 104.07, and 2.23, respectively. These measures are used to compute the Akaike weights for each of the competing models using Equation 3:

$$\omega_{minimal} \simeq .000,\ \omega_{subject\ randon\ slopes} \simeq .753,\ \omega_{item\ random\ slopes} \simeq .000,\ \text{and}\ \omega_{maximal} \simeq .247.$$

Then, we can compute the average estimate for all the parameters of the model. For example, consider that we want to compute the average estimate of the variances of the random slopes. To do so, we would multiply the estimation of the variance of the random slopes of each of the competing models by their respective Akaike weight, and then the resulting weighted estimates would be summed. Considering this example, the averaged variance of the subject random slopes would be:

$$\text{Averaged } \sigma_{1s0} = .000^*.000 + .753^*95.800 + .000^*.000 + .247^*95.814 = 95.804.$$

Similarly, the averaged variance of the item random slopes would be:

$$\text{Averaged } \sigma_{10i} = .000^*.000 + .753^*.000 + .000^*13.953 + .247^*11.382 = 2.817.$$

Table 3 presents a summary of the elements used to compute model averaging with Akaike weights, and the averaged estimates for each variance of the target random effects of this simulation study.

PsychOpen GOLD

Table 3

Empirical Illustration of the Use of Model Averaging With Akaike Weights in a Simulated Data Set

	Model fit and Akaike weights (ω_r)				
Estimated models	**Minimal**	**Subject rand. slopes**	**Item rand. slopes**	**Maximal**	
AIC	77309.33	77207.52	77311.59	77209.74	
$AIC_r - AIC_{min}$	101.81	.00	104.07	2.23	
ω_r	.000	.753	.000	.247	
	Estimates of competing models				
Random effects	**Minimal**	**Subject rand. slopes**	**Item rand. slopes**	**Maximal**	**Averaged estimates**
σ_{0s0}	152.80	112.79	152.80	112.80	112.79
σ_{00i}	84.13	84.21	78.57	78.52	82.80
σ_{1s0}	.00	95.80	.00	95.81	95.80
σ_{10i}	.00	.00	13.95	11.38	2.82

Note. $N_1 = 90$. $N_2 = 60$.

The resulting estimates of random slopes of model averaging with Akaike weights are close to the simulated parameters. In this example, there was important evidence in favor of the model with subject random slopes, and the models with null subject random slopes were very unlikely given the estimated variability in the data set. In more demanding scenarios with less evidence in favor of a model, Akaike weights would be less extreme weighting the estimations of all the competing models. The same procedure would apply for the estimation of any of the parameters of the model, including the fixed effects and their standard errors.

Discussion

Whilst random effects are usually considered of secondary interest, they contain crucial substantive information to understand the processes that are being modeled in MEMs-CR. For example, random intercepts mean that there are individual differences in the mean process, and random slopes mean that there are individual differences in the target fixed effect of the researchers. Here, we endorse the use of random effects as a confirmatory hypothesis testing approach (see also Barr, 2013). Thus, finding substantive variability in the modeled processes is relevant from both theoretical and methodological point of views. Given that model selection makes all-or-nothing decisions regarding to

PsychOpen GOLD

the inclusion of model parameters, we evaluated if model averaging could deal with model uncertainty and to unbiasedly recover random effects of MEMs-CR. We found a relevant influence of the strategy on the bias of random effects, in favor of model averaging comparing to both model selection and fitting the maximal MEM-CR. Additionally, an empirical illustration was provided to ease the usability of model averaging for applied researchers.

Summary of Findings

We compared the bias of the average estimations of random effects of MEMs-CR using Akaike weights and BIC posterior probabilities, using AIC and BIC model selection and fitting a full random structure (here, *maximal* model) as benchmarks. Additionally, two simulation scenarios were considered where the random variances of intercepts and slopes of subjects and items were symmetrical or asymmetrical. The first simulation scenario was more artificial but allows to study the recovery of the variances of random intercepts and slopes with the same metric, while the second simulation scenario was more ecological as usually the variances of random intercepts tend to be larger than the ones of random slopes in experimental psychology. In general, AIC index obtained higher true model selection performances than BIC in different demanding conditions like having small sample sizes in any of the clusters (subjects and/or items), but incorrect random structures were selected in many cases (specially, under-parametrizing the random structure, that is, setting random effects equal to zero when they are different from zero). This was found to be crucial to comprehend the advantages of model averaging in front of model selection.

In scenarios with null random effects, no relevant differences were found between model averaging and model selection. Bias was larger in more demanding conditions (e.g., small sample sizes), but it was not alarming. This means that both model averaging and model selection provide virtually unbiased estimates of population random effects that do not differ from zero, once minimal sample sizes are raise. Presumably, this simulation scenario is not very common in empirical research.

In scenarios with not-null random effects, which are, presumably, the most common empirical scenario, interesting differences were found between model averaging and model selection. First, no relevant differences were found between them when random intercepts were estimated. This means that random intercepts were estimated unbiasedly in almost all conditions of the simulation study, that is, including the presence of symmetrical and asymmetrical variances of random intercepts and slopes. Second, some differences were found between model averaging and model selection when random slopes were simulated, especially in random slopes of subjects. That is, given a minimum sample size, model averaging could estimate random slopes more accurately than model selection (the differences between the two strategies mainly appeared in random slopes of subjects, but their differences were less important for random slopes of items). But the

PsychOpen GOLD

main advantage of model averaging was using all the available information while model selection made all-or-nothing decisions that were incorrect on many occasions (in some scenarios, we even found a median of the 100% percent of bias in random slopes when a null variance was settled to a cluster). These differences were dependent of the smaller sample sizes, as they probably limited the available information for the fit indices, being model averaging capable of overcoming, at least in part, this limitation.

Some relevant differences were also found between the versions of model averaging, that is, Akaike weights and BIC posterior probabilities. In general, it was found that Akaike weights obtained a better performance than BIC posterior probabilities when there are not-null random effects (which is, in fact, the most common scenario in applied research). Again, once minimum sample sizes were raised, Akaike weights could estimate unbiased random effects while BIC posterior probabilities obtained larger bias. In any case, the estimations of both versions of model averaging were affected by lower sample sizes in any of the simulated clusters.

Additionally, interesting differences were found between model averaging with Akaike weights and fitting a full random structure (here, *maximal* model). Model averaging was capable of recovering the random effects under conditions with large sample sizes in the target clusters. But fitting a full random structure (without using a model selection strategy; see for example Barr et al., 2013; Martínez-Huertas et al., 2022; Matuschek et al., 2017 for different model selection strategies) was found to present large bias in the random effects that were analyzed in this simulation study if there are null population variances for random effects. Thus, fitting a full random structure (without a previous model selection) would lead to large bias in the estimation of random effects. Probably, the variability of the different clusters would be incorrectly distributed among the estimated random effects (e.g., the model could not be able to determine if the variability of a cluster is related to the intercepts or the slopes, and also the model could mix the variability between clusters). In any case, the estimates of fitting a full random structure were accurate when all the estimated random effects exist at the population level, which is not always known by applied researchers. Thus, we would like to endorse the use of model selection and model averaging instead of just fitting a full random structure to the data. In this sense, model averaging would consider all the available information of the different competing models based on their model fits, being more robust against model misspecification (see also Hinne et al., 2020) than fitting a full random structure.

Theoretical and Methodological Considerations

Applying a MEM-CR assumes decisions about their parametrization given that true population models are not known. Model averaging proposals use all the available information of different models to deal with uncertainty. Two general conclusions can be made from our analyses: model averaging and model selection show similar results for

PsychOpen GOLD

recovering null random effects, and model averaging shows less bias than model selection for recovering not-null random effects, especially for Akaike weights and random slopes of subjects. These conclusions are conditioned to minimum sample sizes in both clusters (that is, subjects and items). Thus, we would like to endorse model averaging with Akaike weights as a relevant tool to recover random effects in studies with medium to large sample sizes.

There is consensus about the necessity of including all relevant dimensions in experimental conditions to potentiate ecological validity (Hoffman, 2015). Thus, the importance of these methodological tools that allow to deal with the uncertainty of random effects is significantly emphasized. Therefore, model averaging approaches could have important theoretical implications for empirical research because they can avoid errors of statistical inference and to unbiasedly estimate random effects. Thus, model averaging allows to use the evidence in favor of each possible model without losing relevant information of their random effects by making questionable decisions about the random structure of MEMs-CR. The empirical illustration of this study aims to provide an example of the applicability of model averaging to the estimation of crossed random effects for applied researchers.

Limitations and Future Directions

The present study simulated different random structures for MEMs-CR, but only two population parameters (a null vs. not-null value) were considered to analyze the recovery of random effects. Also, a simple experimental design with two fixed effects (the intercept and the within-effect) was simulated due to our objective was to evaluate the recovery of random effects using model averaging. However, we think that model averaging would be an affordable tool to recover small random effects, which are more difficult to estimate, avoiding all-or-nothing decisions of model selection strategies. Also, given the importance of the design of the study to establish model random structures, it would be interesting to expand these results to other relevant designs like longitudinal studies (see Martínez-Huertas & Ferrer, 2022, for an illustration of the use of MEMs-CR in different longitudinal designs). Other complex designs, like models that use cross-level interaction effects or level-2 predictors, could be benefited of using model averaging approaches because these effects are largely influenced by the variability around the fixed effects. Moreover, these complex models would require considering more competing models that: (1) would generate a very complex model selection scenario, and (2) could bias all the results of fixed effects if incorrect random structures were selected (e.g., setting some random effects equal to zero). Also, future research should explore other average estimates like evidence ratios, which can be understood as the evidence in favor of a model as the ratio of the weights (Burnham & Anderson, 2002). Additionally, some readers might find some parallelism of model averaging with ensemble methods and other statistical procedures that combine the predictions of different models or

PsychOpen GOLD

estimators into a single estimate to improve the generalizability or robustness of the predictions of algorithms. Model averaging and other ensemble methods share a common underlying logic of computing a single estimate from different sources of information. In this study, we analyzed the usefulness of model averaging using AIC and BIC fit indices in MEMs-CR, but the research area of ensemble methods is more broaden. In this sense, the rationale of the present simulation study could be generalized to other interesting statistical procedures that use penalized estimation criteria to compute model estimations like penalized least squares for structural equation modeling (Huang, 2022), or regularized partial correlation networks (Epskamp & Fried, 2018).

Conclusions and Recommendations

Model averaging attempts to use all the available information of the competing models to deal with model uncertainty, beyond model selection based on AIC and BIC. Unbiased estimates of random effects were found in both model averaging with Akaike weights and BMA with BIC posterior probabilities under conditions with sufficient sample sizes in the target clusters. In general, 60 units per cluster were found to be an appropriate sample size to obtain very accurate estimations of the variances of random effects in the conditions of the present simulation study. But some differences were found in favor of model averaging with Akaike weights in demanding conditions like small sample sizes, which was also capable of estimating unbiased variances of random effects if one of the clusters presented large sample sizes. This means that larger number of subjects (e.g., 60 subjects) could lead to compensate the handicaps of smaller sample sizes of items (e.g., 10 items). Whilst the performance of model averaging is questionable under some simulation conditions, it supposes an alternative to deal with model uncertainty even in those scenarios where using model selection would require a risky all-or-nothing decision regarding to the inclusion of parameters like random slopes. Thus, we recommend using model averaging Akaike weights, and to use both model averaging approaches with small sample sizes to analyze their convergences and divergences.

Funding: The authors have no funding to report.

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

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

References

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

PsychOpen GOLD

Akaike, H. (1979). A Bayesian extension of the minimum AIC procedure of autoregressive model fitting. *Biometrika, 66*(2), 237–242. https://doi.org/10.1093/biomet/66.2.237

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://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*, Article e328. 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., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

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

Chatfield, C. (1995). Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society. Series A (Statistics in Society), 158*(3), 419–444. https://doi.org/10.2307/2983440

Claeskens, G., & Hjort, N. L. (2008). *Model selection and model averaging*. Cambridge Books.

Draper, D. (1995). Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society. Series B (Methodological), 57*(1), 45–70. https://doi.org/10.1111/j.2517-6161.1995.tb02015.x

Epskamp, S., & Fried, E. I. (2018). A tutorial on regularized partial correlation networks. *Psychological Methods, 23*(4), 617–634. https://doi.org/10.1037/met0000167

Fragoso, T. M., Bertoli, W., & Louzada, F. (2018). Bayesian model averaging: A systematic review and conceptual classification. *International Statistical Review, 86*(1), 1–28. https://doi.org/10.1111/insr.12243

Hinne, M., Gronau, Q. F., van den Bergh, D., & Wagenmakers, E. J. (2020). A conceptual introduction to Bayesian model averaging. *Advances in Methods and Practices in Psychological Science, 3*(2), 200–215. https://doi.org/10.1177/2515245919898657

Hoffman, L. (2015). *Longitudinal analysis: Modeling within-person fluctuation and change*. 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*. Routledge.

PsychOpen GOLD

Huang, P. H. (2022). Penalized least squares for structural equation modeling with ordinal responses. *Multivariate Behavioral Research, 57*(2–3), 279–297. https://doi.org/10.1080/00273171.2020.1820309

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

Kishino, H., Kato, H., Kasamatsu, F., & Fujise, Y. (1991). Detection of heterogeneity and estimation of population characteristics from field survey data: 1987/88 Japanese feasibility study of the Southern Hemisphere minke whales. *Annals of the Institute of Statistical Mathematics, 43*(3), 435–453. https://doi.org/10.1007/BF00053365

Konishi, S., & Kitagawa, G. (2008). *Information criteria and statistical modeling*. Springer Science & Business Media.

Martínez-Huertas, J. A., & Ferrer, E. (2022). Mixed-effects models with crossed random effects for multivariate longitudinal data. *Structural Equation Modeling*. Advance online publication. https://doi.org/10.1080/10705511.2022.2108430

Martínez-Huertas, J. A., Olmos, R., & Ferrer, E. (2022). Model selection and model averaging for mixed-effects models with crossed random effects for subjects and items. *Multivariate Behavioral Research, 57*(4), 603–619. https://doi.org/10.1080/00273171.2021.1889946

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., & 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://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://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

Neath, A. A., & Cavanaugh, J. E. (2012). The Bayesian information criterion: Background, derivation, and applications. *Wiley Interdisciplinary Reviews: Computational Statistics, 4*(2), 199–203. https://doi.org/10.1002/wics.199

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

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

PsychOpen GOLD

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

Steel, M. F. J. (2020). Model averaging and its use in economics. *Journal of Economic Literature, 58*(3), 644–719. https://doi.org/10.1257/jel.20191385

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

EUROPEAN ASSOCIATION OF METHODOLOGY

Methodology is the official journal of the European Association of Methodology (EAM).

leibniz-psychology.org

PsychOpen GOLD is a publishing service by Leibniz Institute for Psychology (ZPID), Germany.

PsychOpen GOLD