UNIVERSITY OF AMSTERDAM

Department of Psychology
Nieuwe Achtergracht 129b,
1001 NK Amsterdam,
The Netherlands
j.b.vandoorn@uva.nl

Amsterdam, May 31, 2021

Dear Dr. Brown,

My co-authors and I are happy to submit the revision of our manuscript "Bayes Factors for Mixed Models" (COBB-D-21-00007) for publication in *Computational Brain & Behavior*.

We would like to thank you and the reviewers for the positive comments and the constructive feedback. Below we have provided a point-by-point response in bold. The adjusted passages in the manuscript are indicated in blue (additions) and red (deletions).

We are looking forward to your comments on this revision.

Sincerely,

Johnny van Doorn
(also on behalf of the co-authors)

Comments by the Editor

The review reports have many suggestions for improving the manuscript. I would particularly like to encourage you to think carefully about these three points:

1. It would be much better to avoid the ad hominem labels for the comparison methods. One of the reviewers suggests alternative labels. You may have your own ideas, of course. It would be helpful if labels could be attached which help remind the reader about a key characteristic of each method.
We have changed the labels for two comparisons to make them more informative and less personal. Specifically, we have changed the "Oberauer comparison" to the "Balanced null comparison", and the "Rouder comparison" to the "Strict null comparison" to specifically indicate which type of null model the comparison uses. We have decided to keep the individual model names as they are, since we felt the suggested names still require a fair bit of decoding. Additionally, they grow particularly complicated when writing the names in subscripts of the Bayes factor, or when using these to define the model comparisons. We also hope that the inclusion of Figure 1 serves to aid readers' memory in linking the numbers to the models' interpretation.

2. For readers less familiar with the problems of inference here, a figure will be helpful. R2 suggests something like this, w.r.t. the nested models. I do not know what the right kind of figure is, but maybe you have ideas.
We have added such a figure in the section introducing the models (see below). The figure includes the relevant parameter settings that distinguish between the models, and helps readers see the nestedness.

3. Perhaps this is best left to the commentators, but I wonder if there are some more general points to be made in the Discussion, about whether or not the very hypothesis testing that is discussed is a good way to progress theory development. One reviewer raises related issues (their "meta question"). This may well be beyond the scope of your work, but it could be interesting and important.
Our initial goal was to present some interesting cases that would spur discussion, but we did not want to restrict the content of the reactions to only the points that we raised. We hope that respondents to the special issue did not feel too restricted and have added a sentence to the discussion to clarify this (i.e., "We note that this is not an exhaustive list of questions worth discussing in the context of mixed model comparison and we welcome contributors to go beyond"). The applicability of Bayes factors in hypothesis testing is perhaps too general to discuss appropriately in the current article, although it would make for a great response article or even its own special issue.

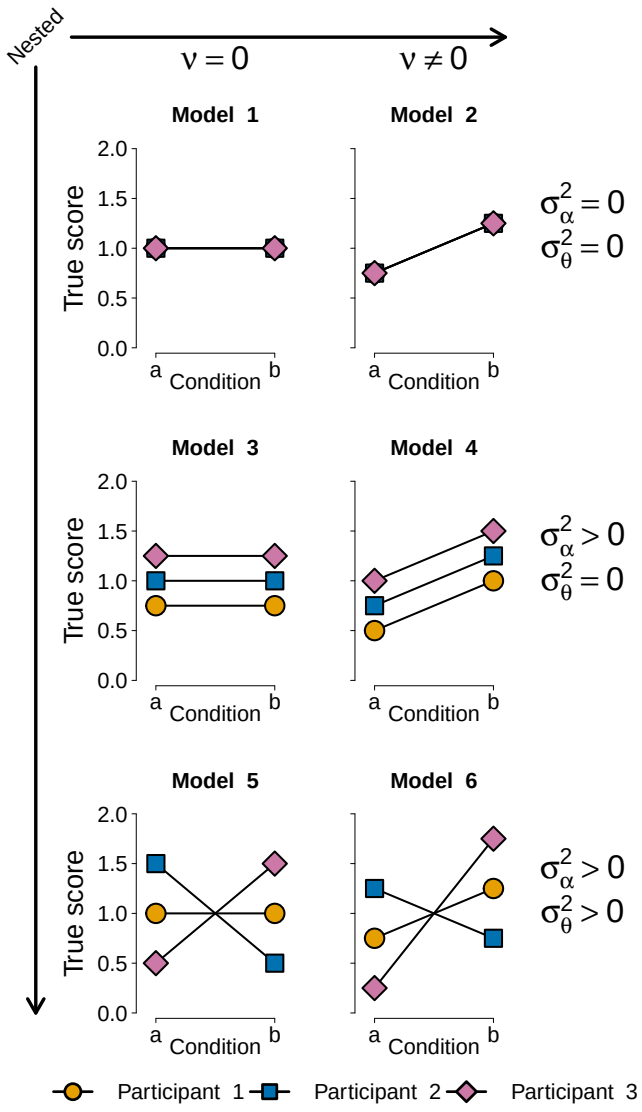

Figure 1: This figure was added as Figure 1 to the manuscript.

Comments by Reviewer 1

I found this to be a very clearly written and engaging paper which I think is well suited to raising important issues and encouraging debate about them. I particularly like the point about the amount of measurement error and the effect size prior, something that had not occurred to me before. I have one substantive but relatively minor point, and a few comments on details.

Re. Footnote 1: Why not assume log(RT) is normal, seems easy to do and closer to the truth. Your simulations all run through in the same form and your real data analysis will be more valid in the non-aggregated case. It is fairly standard practice to do this in the mixed model literature and I think a paper like this should encourage best practice. Also, probably best not mentioned where you do now in discussing aggregated analysis (where distributional assumptions focus on the mean and central limit makes that approximately normal) this is something to mention the first time you talk about including trial data.

We have adjusted the footnote accordingly, and have moved it to appear earlier (when we first discuss the RTs as the dependent variable). The footnote now reads:

> **"It is standard practice to log-transform the RTs, in order to satisfy the normality assumption of linear models. In the remainder of this article, we draw our synthetic observations from normal distributions where relevant, thus generating what can be considered log-transformed RTs."**

Minor points

"there are observations for each item" rather this should say that there CAN be (e.g., not in a recognition memory design there would typically not be as some items fill the new role and some the old randomly, but not both).

This is in the context of crossed effects, where all participants complete all items. We added some text to clarify this. The text now reads:

> **"For example, if both "participant" and "item" are used as grouping factors, the structure is crossed when all participants complete all items, because for each participant, there are observations for each item."**

"inflation of Type 1 and Type 2 error" should that be OR? Seems unlikely it can do both at once.

We have changed this to "or".

"The Oberauer comparison" much as I like Klaus, you don't reference any of his work to justify this nomenclature!

Please see our response to the editor's comment about the naming conventions.

"Moreover, by definition, adding a random effect inherently increases a models robustness to the added variance induced by outliers." This comes out of the blue, "outlier' suggests a contaminated process to me, but there is nothing like that in the simulation, perhaps this term needs to be clarified.
We have changed "outliers" to "extreme values in the tails of the distribution" to be more precise.

"Since there is overlap of the predictive space of Model 4 (fixed effect, but no random effect) and Model 5 (random effect, but no fixed effect)," is confusing, the overlap (random intercepts) is not mentioned, and "random effect" should be "random slope effects". I also don't understand "see also Figure 2".
We have elaborated more on what we mean here and point to the specific panel of Figure 2 that illustrates the principle. The question now reads:

> **"Since there is overlap of the predictive space of Model 4 (fixed effect, but no random slopes) and Model 5 (random slopes, but no fixed effect), there is a certain degree of model mimicry: random slopes in a model account for variations due to a fixed effect (e.g., see also the middle panel of Figure 3, where Model 6 receives far more support in the Strict null comparison than in the Balanced null comparison due to the random slopes in Model 5). Can we therefore meaningfully disentangle a fixed effect and random slopes, both statistically and theoretically?"**

Comments by Reviewer 2

I should start by saying that I (like I assume other potential reviewers) are intending to submit a comment on this article as per the call. Given that, my review focuses on urgent issues and communication suggestions in order to improve the readability of this manuscript. The authors present three examples that they use to demonstrate differences between model comparison techniques for mixed effects models. The first example demonstrates the effect of aggregation, the second the impact of aggregation with decreasing numbers of trials aggregated (which they use to represent measurement error) and the third is an example taken from the literature with a 2x2 design. They finish with a list of questions that arise when considering model comparison with multilevel models.

Overall, I think the paper touches on a number of important issues. I struggled somewhat with the naming of the different methods after the proposers, and would suggest the reviewers change this for increased clarity. One limitation was that the investigators only consider treatments as 2 levels. I think it would be worth noting how moving to more than two condition levels would impact the findings and research questions.
We have expanded the footnote on effect coding on page 6 to be more informative about

including > 2 factor levels, and to clarify that we only consider designs with 2 levels in the manuscript. In the cases that we consider, we think that our conclusions will not systematically differ for designs with more levels. The footnote now reads:

> **"For designs with > 2 factor levels, multiple coding vectors are used. In the remainder of the manuscript we consider only cases with two levels. The questions we pose here do not fundamentally change when the number of levels is increased."**

Detailed comments

While I respect the use of the method proposers throughout the manuscript as a method of attribution, it does raise two issues for me. The first is that it does unfortunately make the paper very difficult to read. I find myself moving between the pages constantly to understand later figures and comparisons. As a compromise, I would suggest attributing the methods to informal email exchange and then naming them in a meaningful way. I've included an example/suggestion that I'm happy for the authors to use, but I also think they might think of something better, which is also fine. Suggested model names:
Model 1 = Intercept Only (IO)
Model 2 = Fixed Effect (FE)
Model 3 = Random Intercepts (RI)
Model 4 = Fixed condition, Random Intercept (FC-RI)
Model 5 = Random intercepts and slopes (RS-RI)
Model 6 = Full model (Full)
Suggested model comparison names:
RM ANOVA (what does RM stand for in this instance? The acronym was used but not clarified)
Oberauer comparison: Full-RSRI comparison
Rouder comparison: Full-RI comparison
Please see our response to the editor's comment about the naming conventions.

The second is that I couldn't find a footnote stating that that Klaus Oberauer and Jeff Rouder have consented to have their informal emails used in this matter. I assume that they have (and if they haven't it would be impossible to reverse this decision), but I think a footnote stating this would be useful. Neither are authors of this manuscript so I think this is important to make clear.
We have changed our naming conventions, so this is no longer relevant. To clarify, we did obtain consent for the previous naming convention.

I also think it would be useful to have a greater discussion of the existing literature discussing model comparison with mixed effects models. As it stands, there is a citation to the Barr (2013) paper and Rouder (2016), but I would be surprised if there isn't more discussion on the topic given the popularity

of both model comparison and mixed effects models - I've discussions of the null model come up in R^2 calculations for mixed effects models, for example.

The focus of our manuscript is to have model comparison through Bayes factors, so we intend to limit discussion of articles outside of that scope. We include the Barr references primarily to motivate our own research questions, and since these references are especially seminal in the field of mixed model comparison.

Example 1: It is difficult to differentiate the individual colours in Figure 1. I would suggest the following to increase readability. Firstly, increase the jitter in all of the figures, and introduce it in the second row, and make the point size in the first row smaller to reduce overplotting. Secondly, add lines for each individual to represent the impact for condition, but add these lines as the bottom layer on the plot with high transparency. Then you can add the mean estimate in a thicker, darker line on top so that it pops out from the individual data. This will allow the reader to see variability due to condition as well as overall data variability.

We have increased the jitter in the plots and decreased the point size in the top row in order to avoid overplotting. While creating Figure 1, we debated how to display the data and ran into the issue raised here. First, connecting the dots in the upper left panel (full data for all participants) creates the false impression that trial 1 in condition A is linked to trial 1 in condition B, trial 2 in condition A is linked to trial 2 in condition B, and so on. Second, taking the participant averages and connecting those averages will lead to lines that do not connect to data points in the upper left panel. This is the reason why we instead chose to include the second row of plots. While it shows the data for only 5 participants, it does highlight individual differences, the participant averages, and how the data is structured. We include both versions of the plot below, to show how it looks. We do not feel strongly about this choice, so in case you prefer the second option (Figure 3 here), we can change it.

The authors note that it is difficult to differentiate what causes the extreme favour in the Rouder model in example 1. Why wasn't the 6,4 model comparison used to help investigate this?

This is a great suggestion, and something we also noted when working on this article. However, this approach could already be part of a more substantive reaction to this article (i.e., proposing a multi-step procedure where Models 4, 5, and 6 are all compared against each other). We felt that including such a conclusion would already answer some question instead of sparking a discussion.

Example 2: I thought the use of collapsing increasingly large groups of trials was a nice methodology to use! Figure 2 could benefit to labelling the rows according to the data simulation method. That, in addition to different naming convention of the model comparison techniques could make it easier to quickly understand what the Bayes Factor should be given each data generating scenario.

We have updated the naming for the model comparisons to be more informative.

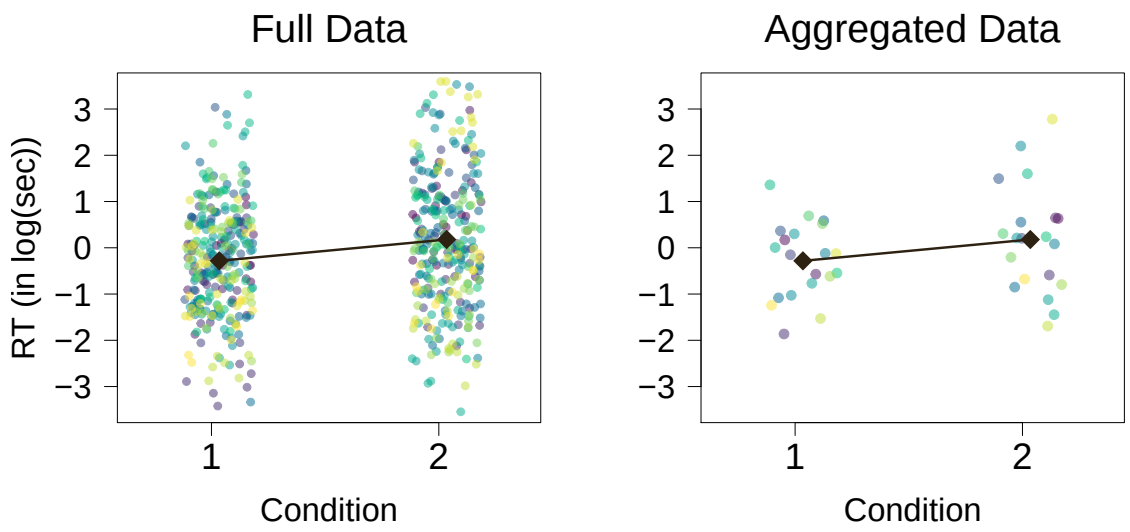

Figure 2: The jittered version that is in the revision.

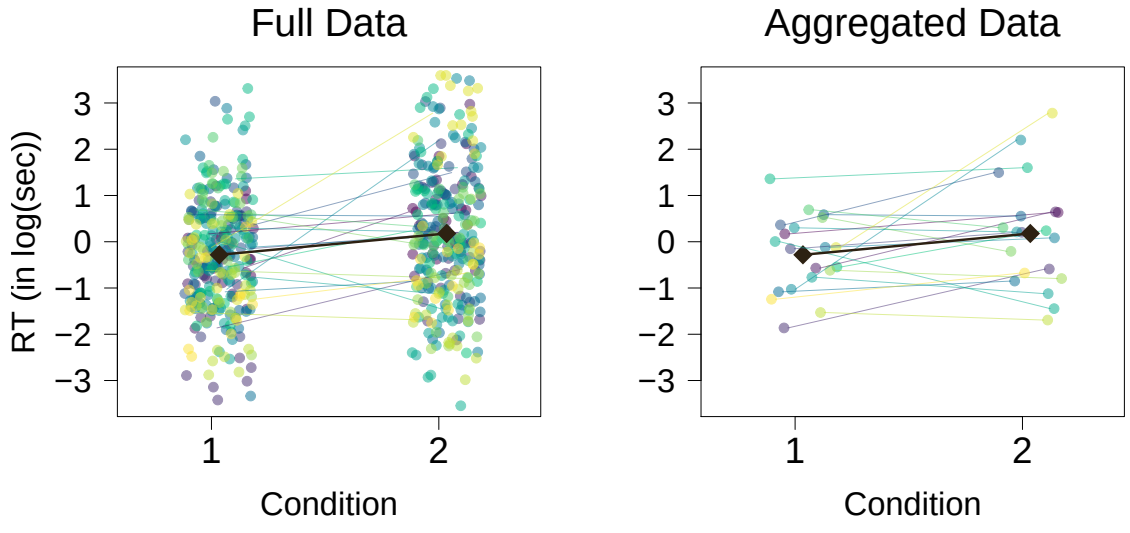

Figure 3: The jittered version that also connects each participant's mean score.

With regards to the comment on priors, it would be useful to see a prior predictive distribution compared to the actually observed distribution to understand the implications of the priors.
This would be a great addition. However, the BayesFactor package uses improper priors on the intercept and sample noise, and we therefore cannot sample from these priors. Van Heide and Grünwald (2020, PB&R) attempted to sample from these default priors, and ran into similar issues.

Example 3: The points on figure 3 could again be jittered, maybe with some degree of transparency so that the distribution of points can be viewed.
We have added the jittering and made the data points more transparent for the full data.

Final conclusions: I really like the final table! I wonder if it could be moved to the beginning of the manuscript to provide some scaffolding for the overall direction?
We have moved the table to the front, after the introduction.

Comments by Reviewer 3

The authors discuss the important topic of how to test experimental effects in within-subject designs using linear mixed models. The manuscript is well written and very accessible. The following issues focus on improving the presentation of the content as I assume that substantive issues should be discussed in the special issue itself.

(1) Given the importance of priors for Bayes factors and for the specific topics discussed in the paper, I think it would be helpful to provide a more detailed specification of the prior specification in an appendix. This would facilitate the discussion since commentators do not have to look up details in the original papers by Morey, Rouder et al. Moreover, some model assumptions should be clarified, for instance, does the model assume a zero correlation of random intercepts and random slopes (p. 7)? This may not always be plausible, for instance, when there are ceiling effects leading to a negative correlation of random intercept and random slope.
The prior specification is clarified at the start of the Examples section, where we state which prior distributions are used for the random and fixed effects. We have added the specification for the grand mean and error variance, such that it now reads:

> **"The BayesFactor package specifies Jeffreys's prior on the grand mean and error variance (i.e., $f(\mu, \sigma^2) \propto \frac{1}{\sigma^2}$) and does not explicitly model correlations between random slopes and intercepts."**

Furthermore, we have added a note on page 7 about the correlation between random

intercepts and slopes to this paragraph:

> **"Additionally, in the remainder of this article we do not discuss explicitly modeling the correlation between the random slopes and random intercepts. Failure to account for correlated random effects can lead to misleading results, for instance in the context of ceiling effects, where participants with high intercepts will be inherently limited in their effect."**

(2) To facilitate the presentation of the different model comparisons, it might be helpful to add a diagram that illustrates the relations of the models, that is, how the models are nested (e.g., M1 nested in M3; and M3 in M4 etc.). Such a figure could include a few color arrows indicating which are the relevant pairwise comparisons that are discussed.
This is a great suggestion. Please see our response to the editor's comment for a more elaborate response.

(3) I am not sure whether personalized labels such as "Rouder comparison" and "Oberauer comparison" are the best choices for such a special issue. I think it would be preferable to use more neutral and intuitive labels (e.g., "FE+RE comparison" and "FE-only comparison"; "M5 vs. M6 comparison" and "M3 vs. M6 comparison"; and there are probably better terms). This is of course a matter of taste and merely a suggestion.
Please see our response to the editor's comment about the naming conventions.

(4) The abstract could state the scope of the manuscript more clearly, namely, that the paper focuses on factorial experimental designs with within-subject factors which are analyzed using linear mixed effects models.
We have updated the abstract accordingly, such that it now reads:

> **"Although Bayesian linear mixed effects models are increasingly popular for analysis of within-subject designs in psychology and other fields, [...]"**

(5) Concerning Section 2.1 and 2.2, I wondered whether the author deem it to be desirable that the aggregation over some batches of trials results in similar Bayes factor / evidence for the hypothesis test? Or is the goal merely to compare the effect of aggregation? As a remedy, it could be stated more explicitly what seems to be desirable in this scenario.
We did not want to be too prescriptive here, and focus mainly on stating/comparing the behaviors of the respective comparisons. We have added a sentence on page 21 that combines the observations of the two sections:

> **"Combined with the results from Section 2.1, this suggests that the process**

of aggregation mainly affects those model comparisons where only one of the models under consideration includes random slopes."

(6) On p. 18, the authors state that "The Oberauer comparison is relatively stable in the top and bottom panel, which confirms the balance between the number of trials and their accuracy." This may merely be due to the fact that with such a small number of trials and participants, models 5 and 6 are overly complex, implying that the Bayes factor does necessarily indicate ambiguous evidence because the data are just not informative. Put differently, it may be the case that this model can only provide clear evidence when analyzing large data sets. Shortly mentioning these issues would highlight a core benefit of the BF, namely, the tradeoff between model fit and complexity.
In the bottom panel, the Oberauer comparison is relatively stable around a log(BF) of 5, which is quite strong evidence in favor of the alternative model (BF ≈ 150). Our point in the text is concerned with the stability of the Bayes factor, which is a separate behavior from its magnitude. Furthermore, observing this stability for both an indecisive and a decisive Bayes factor underscores the balance we mention.

(7) Using the log BF for comparison makes a lot of sense. However, when interpreting the results and comparisons, readers may overlook the log scale. Hence, it could be helpful to remind the reader every once in a while that logBF=3 actually indicates "strong" evidence (BF=20).
We have added this reference point when we first discuss the logarithmic scale of the Bayes factors (p. 14):

"Note that $\log(\mathbf{BF}_{A,N}) = 3$ **corresponds to** $(\mathbf{BF}_{A,N}) \approx 20$**."**

(8) Somewhere in the paper, it should be discussed more explicitly that all examples in the paper assume the presence of random slopes. It depends on the specific context whether this assumption is plausible. There may be effects that are in fact rather stable and similar in effect size across individuals. In this case, models 3 and 4 might not be so bad after all.
We have included a note on this on p. 12. We implicitly hope that readers will realize that a single model comparison is not enough in such a research setting and look forward to the responses. If more than two models are compared, than Models 3 and 4 are certainly relevant, but when forced to pick a single comparison, we would recommend a comparison that includes random slopes. The added note reads:

"The comparison of Model 3 to Model 4 on the full data might be applicable under the strict assumption that there are no random slopes. In the remainder of this manuscript, we focus on scenario's where random slopes cannot be excluded a priori."

(9) Concerning the Discussion: A very general question that could be added would be whether the questions asked by the authors are the right questions at all. Put differently, are there issues not yet addressed by the examples and the proposed questions? Making this "meta-question" explicit could be beneficial and lead to more different perspectives and comments. On the other hand, it could make the special issue too diverse - limiting the target paper to a small set of relevant questions would clearly delineate the scope of the discussion. I do not have a clear opinion, but I am sure that the editor can weigh in on this.

We feel that both readers and respondents are already free to criticize our proposed scenario's and the corresponding questions. To clarify this point, we have added a sentence to the discussion:

> **"We note that this is not an exhaustive list of questions worth discussing in the context of mixed model comparison and we welcome contributors to go beyond."**

Minor comments

p. 3: I think one would usually not use left- vs. right handed as a grouping factor with random effects. Instead, such a factor would be modelled using a fixed effect. Maybe, the authors can find a more intuitive example for a random-effect clustering factor.

We have changed the example from handedness to geographical region.

p. 19: "the instability of the Oberauer comparison" → this is not clear: it appears that the Oberauer comparison is relatively stable whereas the Rouder comparison shows more variability?

Here we refer to the relatively minor instability of the Oberaur (Balanced null comparison) that is still present. We now clarify this in the text: "We suspect that the (relatively minor) instability of the Balanced null comparison [...]"

Figure 2: It would be helpful to add gray horizontal lines at zero.

We have added the lines.