Response to Reviewers

The role of semantic encoding in production-enhanced memory: A registered report

Editor:

I have now obtained three very helpful reviews of your Stage 1 submission. The reviewers agree that this is a rigorous and promising proposal, while also offering a range of constructive suggestions to consider. Headline issues to address include the inclusion of literature to provide additional context and strengthen the study rationale, greater prospective interpretation of null results (which is addressed in your design template but could perhaps be strengthened in the main text), inclusion of additional methodological details (including graphic presentations), and consideration of additional or alternative analyses.

One of the reviewers suggests the addition of Bayesian analyses, which can be useful for providing positive evidence of no effect. If including these, please be sure to define all priors and other parameters precisely, and if also including frequentist tests, be sure to explain which outcomes (Bayesian or frequentist) will determine whether a hypothesis is supported or not.

I look forward to receiving your revised manuscript in due course.

Thank you for the opportunity revise and resubmit. We are sorry for the delay in getting these revisions to you. As a summary, here is a list of our most important responses:

1) We made edits to the introduction to strengthen the study rationale.

2) We expanded in the manuscript how null results may be interpreted.

3) We completed stimulus selection and now provide a list in the appendix.

4) We clarify parts of the methods (e.g., we now include a figure that illustrates the different learning procedures) and conducted additional extensive sample size calculations.

5) We carefully considered the alternative statistical analyses and clarify why we think the approach that we had originally suggested may nevertheless be most appropriate to address the specific goals of this proposal, but we also emphasize how alternative analyses could add to follow-up work (which we will help enable by making the raw data freely available).

Thank you for your consideration.

Reviews

Reviewed by anonymous reviewer, 12 May 2023 09:21

After carefully reviewing Stage 1, here are some information organized as a function of each key criterion. Overall, the stage 1 manuscript was original, and I look forward to seeing the next version.

1A. The scientific validity of the research question(s).

The authors provide a clear theoretically motivated research question. The research question is embedded in rich literature on the production effect and spreading activation to better understand the underlying processes driving the production effect with an original investigation with bilingual (English/German) speakers. In addition, the writing is clear and easy to follow.

P1 (reference): This is a minor point, but the production effect was well established before the work of MacLeod et al. (2010). For instance, Murray (1965) in Nature published an article called "Vocalization-at-presentation, Auditory Presentation and Immediate Recall.

We added the reference to Murray (1965) as well as some other early ones (Conway & Gathercole, 1987, Gathercole & Conway, 1988, Hopkins & Edwards, 1972) when the production effect is first introduced (p. 3).

P2 (alternative account): The RFM (revised Feature Model; Saint-Aubin et al., 2021 JML) which has been applied to account for the production effect in immediate recall, free recall, and reconstruction of order, might strengthen your argument here (or an alternative account). According to this computational account, production block rehearsal processed but add modality-dependent features (information related to the presentation of the information such as color, sound, pitch) and had no effect to my understanding on modality-independent features (information related to categorization and internal processes). It might help if the results are null as a likely alternative.

Thank you for pointing our attention to this very interesting account. We now mention this paper in the introduction to highlight that the production effect may not only be observed in long-term but also in short-term memory (p. 3 lines 19-20).

Additionally, we now discuss a family of computational models, including the RFM, which propose that speaking integrates additional features into memory traces that in turn improve retention of those memory traces. We think the key premises of these models are helpful for explaining the production effect in general, and we now use them as an initial framework for our research question. In fact, from what we can see in the literature, computational models seem to differ in their assumptions about whether speaking adds only sensorimotor features (i.e., modality-dependent features) or whether speaking can also add other features such as semantics (modality-independent features). We fully agree that our proposed study should help address this question. If speaking does add modality-independent features of items (i.e., semantics), then we should see an intact production effect in the semantic condition: in other words, if speaking helps participants activate and retain the meaning of words when they first encounter them, then recognition should improve when participants are later presented with pictures (not the original written words) that match this retained meaning. If on the other hand the production effect would greatly diminish or disappear (fail to be detected) in the semantic condition (when presented with pictures at test), this would be strongly supportive of the assumption that speaking only, or primarily, adds modality-dependent features to memory traces. We predict that the production effect will be present but relatively smaller in the semantic condition compared to the veridical condition. Our interpretation here would be that while speaking can add modality-independent features (i.e., speaking could engage semantic associations via spreading activation), the production

effect is probably not only based on semantic processing—otherwise we would not observe production effects for nonwords to the extent that we do. Rather, we hypothesize that both modality-independent and dependent features can contribute to the production effect, and that the production effect may improve when both types of features can be engaged (which should be possible in the veridical condition, where the original items are presented at test). We lay out these arguments in the manuscript on p. 4 lines 1-22 and p. 5 lines 1-2.

1B. The logic, rationale, and plausibility of the proposed hypotheses, as applicable.

The logic, the rationale, and the plausibility of the proposed hypotheses are clear and well-founded based on theory. Overall, the introduction was well written.

P3 (null results or alternative results): I would encourage the authors to provide further clarification of the implications of the null results. What are the theoretical implications of the results if nothing works as expected? Potentially important to consider the implication of transfer appropriate processing in the rationale.

The implications of the null or contrary results were more clearly laid out in the table, but we had failed to include that information in the main text. We incorporated some of the information from the table into the main text to rectify this mistake (p. 13 lines 15 to 23, p. 14 lines 1-23, p. 15 1-23, and p. 16 lines 1-9), thus more clearly addressing the implications of unpredicted results.

In addition, we now discuss more clearly why we may observe a null effect based on reviewers' comments, including how the difficult of the semantic condition may increase or the decrease the probability of observing a production effect. In this context, we also discuss two theoretical accounts that could explain null effects (p. 16 lines 18 to 23, p. 17 lines 1-23, and p. 18 lines 1-12).

1C. The soundness and feasibility of the methodology and analysis pipeline (including statistical power analysis or alternative sampling plans where applicable).

Overall, the methodology, the analysis pipeline, and the statistical power analysis were okay. I have some minor suggestions or clarification questions.

P4 (recording): My understanding is participants will be tested remotely and they will have to record their "production". The authors mentioned an inclusion criterion of 95% for "production" could the authors clarify how this will be assessed (apologies if I miss that information)?

This is correct. Participants will be recorded when they produce the word (or not). These auditory files will be checked for accuracy offline by research assistants. Files in which participants followed the instructions (either read aloud or do not speak aloud, as indicated by task instructions) will be coded as 1 and files were the participants did not follow the instructions (speak aloud even though they were not supposed to speak, etc.) will be coded as 0. Files in which the correct response is heard but cut-off on a read aloud trial will be coded as 1 as long as it is clear what the spoken word matches the intended one. We clarified this in the manuscript (p. 29 lines 4-8).

P5 (design question): This is my misunderstanding. Will you present information at encoding in both languages or only one language? Are they any theoretical differences in having pure or mixed lists based on the spreading of activation?

At encoding, words will always be presented in L1 (participants' native language). We clarified this in the manuscript (p. 10 line 13, p. 27 line 9; also see the newly included Figure 1, p. 30). To our knowledge, there are currently no studies that test whether language switching during encoding impacts later memory in a production effect paradigm. In our earlier work, we have shown that the production effect may be enhanced in a second language versus a first language (in non-mixed lists at encoding, Brown & Roembke, 2024, *Memory & Cognition*). If participants had to switch during encoding, one could hypothesize that activation spread is inhibited, as activation of the non-target language has to be reduced on a current trial to minimize interference.

P6 (stimuli): I believe the authors' description seems like they will be very careful in the stimuli selection, but I would appreciate having the stimuli lists as it is a major point and has been shown to drastically affect the results in the past.

This is a fair point. We have now completed stimulus selection and provide these in the appendix of the manuscript (p. 46). We also include picture numbers (referring to the MultiPic database number), so that it is possible to match each word to the exact picture that will be used (as the database sometimes has more than one picture for the same word).

P7 (Bayesian analysis): I do not want to impose statistical preference, but could the authors add Bayesian analyses?

We carefully considered the possibility of including planned Bayesian analyses here because we agree that they would be informative in this line of work. We ultimately came to the conclusion that, because we predict differences between conditions rather than equivalence, the planned frequentist analyses should at least be sufficient to address our specific predictions. We agree that Bayesian statistics could be a valuable next step in follow-up experiments that aim to test equivalency (based on what we find here). We are also very open to the possibility of conducting exploratory Bayesian analyses (if that is deemed appropriate here). Unfortunately, we believe we do not have the necessary expertise to plan Bayesian analyses a priori in a principled way as would be required for a registered report. For example, we may not be able to foresee all possible issues. Of course, data will be provided in an open source format, so it will be possible for someone with the necessary expertise in Bayesian analyses to conduct these exploratory analyses on the data set from this proposal, and we hope that this could inspire further collaborative work with a Bayesian approach.

1D. Whether the clarity and degree of methodological detail is sufficient to closely replicate the proposed study procedures and analysis pipeline and to prevent undisclosed flexibility in the procedures and analyses.

Overall, I believe the authors provide great clarity. The stimuli would help facilitate the clarity of the methodological details.

P7 (method figure): Given the novelty of the procedure, it might be very helpful for the reader to have an illustration of the procedure for encoding and the test.

Thank you for this suggestion. We now include Figure 1 (p. 15) that illustrates the trial procedures during all different learning and recognition tasks.

1E. Whether the authors have considered sufficient outcome-neutral conditions (e.g. absence of floor or ceiling effects; positive controls; other quality checks) for ensuring that the obtained results are able to test the stated hypotheses or answer the stated research question(s).

Overall, I believe the authors have considered many key aspects of the quality of the results. Additional information about outcome-neutral conditions could be beneficial (e.g., what happens if performance is too difficult or too easy).

We agree that we should decide a priori how floor and ceiling effects may be handled, should they arise despite our design choices to counteract them as much as possible. After conducting our hypothesis testing analyses, we will assess if overall performance across all conditions is above chance. Even if all conditions are above chance, we will have to look at the overall pattern of performance to determine whether for example an increased production effect in the semantic condition could be mostly the result of an artifact (due to for example overall worse performance in the semantic than the veridical condition) (p. 16 lines 18-23, p. 17 lines 1-23, and p. 18 lines 1-12). To make our interpretation process as a priori as possible, we walk through all the possible scenarios in the manuscript and their possible interpretations (p. 13 line 15 to p. 16 line 9).

Based on previous work and certain aspects of the task design, we think ceiling effects are unlikely. For example, in our previous work (Brown & Roembke, 2024, *Memory & Cognition*) with a very similar set-up as in the veridical conditions (online recruitment, RWTH Aachen student community, study and test in German/L1, same number of items at study/test), participants' mean hit rate for words read aloud was 83 percent and for words read silently it was 62 percent, while the false alarm rate was 25 percent. In addition, in each condition participants must study 80 words, only half of which are presented again at recognition, which we assume is challenging. At the same time, the veridical conditions in the current proposal use more imaginable words than in previous studies (out of necessity for the test conditions), which could make the task easier. Even so, other work has shown production effects in conditions that also improved overall accuracy (e.g., MacLeod et al., 2010, using a semantic judgement task to rule out "lazy reading" as an explanation of the production effect), and they did not seem to result in a ceiling effect. Nevertheless, we will consider the possibility of ceiling effects if we see near-perfect performance (e.g., above 95%) in the spoken veridical condition (presumably the easiest one).

Having said that, as also discussed in a response to a different reviewer, we think floor effects would be more likely than ceiling effects: It is possible that semantic conditions are too difficult, where transfer-appropriate processing may contribute to a drop in performance. In addition, there is the possibility that participants will simply struggle with mapping the

words they read at study to the pictures and translations they encounter at test, even though we make the word-meaning mapping as easy as possible by selecting only high-frequency words and pictures with a modal name agreement of 75% or above. In addition, we will conduct post-hoc tests for all conditions to clarify whether performance was above chance, thus clarifying whether conditions may have been affected by floor effects. Finally, we will also look at whether performance in the spoken semantic condition is at least as good as the performance in the veridical silent condition. Even if the semantic condition is overall more difficult, it should not be *too* difficult if participants can achieve at least the level of recognition accuracy as they do in the veridical silent condition.

Reviewed by Miguel Vadillo, 10 May 2023 13:55

Before I start my review of the manuscript, I must disclose that I am not an expert in psycholinguistics or semantic processes and therefore I cannot judge the extent to which the current study is innovative or relevant for the area. On the other hand, I can confirm that the text is accessible for non-expert readers and, at least in my humble opinion, it looks like an interesting project, definitely worth pursuing. Logically most of my comments will be focused on methodological aspects, although to be fair, I only have minimal comments that the authors/editor need not agree with.

The authors have done an excellent job with their simulation-based power analysis, taking as reference different combinations of descriptive statistics from a related previous experiment. They have manipulated orthogonally means, SDs… but, incidentally, they have rather ignored what could be a relevant parameter of the simulation analysis. In particular, if I am not mistaken, they always assume a correlation of 0.5 between dependent measures. This is reasonable, but it is not impossible that the true correlation turns out to be substantially lower than that. Their dependent variable is a d', which is computed as a difference score (essentially, hits minus false alarms). Difference scores are notoriously unreliable (see the famous Hedge et al. paper on BRM), which means that they do not correlate very well with anything. Perhaps it would make sense to run sensitivity analysis with lower correlation estimates and adjust sample size accordingly depending on the results.

Also related to power planning, is the number of trials per condition similar to Fawcett et al.? Otherwise, the SD estimate entered into the power simulation might not be adequate (more trials should result in less variability in effect estimates across participants).

Thank you for pointing our attention to these issues. To answer your second question first, Fawcett et al. (2022) had the same number of study trials as we plan to (80) but more test trials (80, each with both a target and lure presented simultaneously). Even so, it seems to us that using the Fawcett et al. (2022) standard deviations for power estimations should be conservative because their semantic vs. non-semantic (in their case "unrelated" lures) manipulation was between-subjects, while ours will be within-subject.

As suggested, we looked more closely at how different correlation sizes would affect the power estimates for the Fawcett et al. data (the d-prime means and SDs reported in their paper). As suspected, the power does indeed decrease for lower correlations (see power analyses with the labels "_Fawcettetal2022" and "_corr01", "_corr02", etc., on OSF). The remaining problem, as we saw it, was how to determine which correlation size would be most

plausible. We decided to look to our own previously-reported data (Brown & Roembke, 2024, *Memory & Cognition*) derived from a production effect task that is the same as the one we are proposing here, including the same number of trials in different conditions, a 2 × 2 within-subjects design (including the spoken vs. silent production effect manipulation), and with d-prime scores as the dependent variable. In our previous data, we observed a mean dependent measure correlation of about 0.4. We first ran a power analysis with N=75 using the means and SDs in Fawcett et al., as well as the dependent measure correlation matrix we observed in our own data: this yielded 92.5 percent power for detecting an interaction between the semantic manipulation and the production effect (see power analysis labelled "Fawcettetal2022_owndatacormatrix" on OSF). We then ran a power analysis with N=75 and the exact d-prime means, SDs, and correlation matrix from our previous data: this yielded 98% power for detecting an interaction between the production effect and our second factor of interest (in our case this was a first or a second language; see power analysis labelled "owndata_cormatrix"). We additionally ran this power analysis with our own data with different correlation sizes (instead of the observed correlation matrix): even with a correlation size of 0.2, N = 75 yielded >80% power. Thus, we think a sample size of 75 per experiment seems to yield reasonable power, assuming that the results we will observe will be similar to actual results from a similar conceptual manipulation and research question (Fawcett et al., 2022) and actual results from a similar task (Brown & Roembke, 2024).

In the manuscript we now report the additional power analyses that incorporate aspects of our previously-reported data (p. 20 lines 17-20, p. 22 lines 8-23, and p. 23 lines 1-6). We do not mention the analyses varying the correlation sizes with the Fawcett et al. (2022) data, but these are viewable on OSF, and we are happy to report these in the manuscript as well if desired.

On different places, the authors argue that if they do not find a significant production effect this could be due to the difficulty of the task (p. 11 and elsewhere). But I wonder if the opposite prediction could be made: Given that the production effect is defined as an increase in accuracy relative to baseline, would it be easier to detect the effect in conditions where baseline performance is relatively low? In other words, without additional information, I don't think that the difficulty of the task on its own provides a sufficient explanation for any failure to observe the effect.

This is an interesting idea—it is generally easier to increase performance from 0.6 to 0.7 than from 0.9 to 1. If performance is generally lower in the semantic condition, it may be easier for the read aloud items to get a boost. Based on this, one may hypothesize that we may actually observe a higher production effect in the semantic conditions than in the veridical ones. We turned to look at the literature to see if any reported data show this pattern. From what we can see, there is at least some data showing an increased production effect in conditions that either improved or did not impair overall recognition/recall. For example, performance was compared for words (Experiment 1) and nonwords (Experiment 6; MacLeod et al., 2010). Memory performance was overall worse for nonwords compared to words, but the production effect was not increased for nonwords (MacLeod et al., 2010). In another study that used both high-frequency and low-frequency words (Experiment 2, Jones & Pyc, 2013),

lower recall was observed for low-frequency than high-frequency words, but this again did not interact with the impact of production. In our own work (Brown & Roembke, 2024), we observed lower recognition performance for L1 than L2 in addition to a higher production effect for L2. Based on this work, it should be at least possible that the production effect could be greater when recognition is overall greater.

Even so, it is possible that any difficulty in the semantic condition could either artificially deflate or inflate the production effect. For example, if participants have problems recognizing the items at test as the ones they were presented earlier in the semantic condition. For example, a participant may remember that s/he earlier read the word "ostrich" but then not recognize the picture of a bird at test as an ostrich (but rather an emu). If this were the case, participants' performance in the semantic conditions may be very noisy, so that it would be hard to observe any effect (even if it existed). We now clarify this in the text (p. 16 line 18 to p. 18 line 12). To help interpretation as to whether any observed effects are due to an artifact (driven by lower performance in the semantic than veridical conditions), we will conduct post-hoc tests for all conditions to clarify whether performance was above chance. In any case, we must be clear that we probably cannot definitely rule out such an artifact, even if recognition is above-chance. We now clarify this in the text as well (p. 18 lines 2-5).

Of course, there are also other, more theoretically-based reasons why we may not observe an effect, such as—as pointed out by other reviewers, transfer-appropriate processing (performance in a memory test is thought to be best if conditions during study and test match; this is not the case in the semantic conditions) or the assumption that speaking only adds modality-dependent features to memory traces. We now make this point in the manuscript as well (p. 13 line 15 to p. 16 line 9).

Perhaps this is standard practice in this area of research, but I found surprising that there are twice as many "old" items in the recognition test (20 previously in blue + 20 in white) as "lures" (just 20). Wouldn't this bias participants towards responding "old"? I know that SDT disentangles sensitivity and criterion, but SDT comes with a number of assumptions that might not hold, in which case d' estimates might be affected by response criterion. Wouldn't it be better to include 40 lures?

We think this is a valuable point: this target/foil make-up may bias participants towards answering old (as there are more old than new items at recognition). This in turn may increase the overall false alarm rate. Such a bias should be equally observed in the read aloud/read silently conditions, and therefore it may not influence the production effect, but this is of course difficult to know without investigating this issue more directly.

For our purposes, we were motivated primarily by keeping our paradigm consistent with the multiple studies reported by MacLeod et al. (2010), and with our own previous and ongoing work (Brown & Roembke, 2024), all of which includes an equal number of lures (20), words read aloud (20) and words read silently (20). Sticking with the original design allows us to remain internally consistent with the production effect studies that we conducted before. This in turn gives us a clearer idea of the overall difficulty level of task (especially the veridical conditions); this has been especially helpful when estimating power for proposed experiments.

We also examined whether previous work has used the 1/2 target or 2/3 target setup at recognition. While the above design seems common in the production effect literature, they were also others, including some where there was an equal number of old and new items. A footnote reported by MacLeod et al. (2022) summarizes: "In many previous production experiments (e.g., MacLeod et al., 2010; Hourihan & MacLeod, 2008; Ozubko et al., 2012a), 2/3 of the items on the recognition test have been studied and 1/3 have been new. This makes the number of aloud, silent, and new items equal at test, but it also produces an unequal number of old and new items. We do know, though, from observing the benefit in recall experiments (Conway & Gathercole, 1987; Lin & MacLeod, 2012; Experiment 2 here) and in other recognition experiments in which the numbers of old and new items were equal (e.g., Forrin et al., Experiment 2, 2012) that this choice does not matter" (p. 1006). While we agree with the claim that the production effect is robust across different types of recognition tests, we still see this issue as worth investigating in its own right, with a direct comparison of different target/foil distributions. We have added this specific suggestion in the manuscript (p. 14 footnote 2) and we acknowledge the potential drawbacks of our chosen paradigm, even though it offers the advantage of consistency with our ongoing research line. We are also happy to consider this point in the discussion, if thought necessary.

On page 22 I found it odd that the authors explain that they will test spoken vs. silent conditions as indexed by Hedges' g. Whether or not they report effect sizes is independent from how they will test hypothesis. So, the sentence sounds a bit weird, because it seems to imply that the t-test will be run on Hedges' g. On a different note, I am also not sure there are good reasons to report Hedges' g instead of the more familiar Cohen's d. Essentially, both effect sizes estimates measure the same thing, with the only exception that g corrects for a small bias in small samples. But with N = 75 d and g will probably agree to the second decimal and the equations for d are far more familiar for the average reader. I see no good reason for reporting g (although of course this is not incorrect or invalid).

Thank you for pointing out these two issues. We clarified in the text that the tests will be run on d-prime scores and not Hedge's g (p. 32, line 2, p. 35, line 1). Regarding the use of Hedge's g rather than Cohen's d, if it is the case that the two values should be similar for the planned sample size, then our inclination would be to retain Hedge's g, but also make sure that the reader is aware of its likely equivalence to Cohen's d in this case. To help clarify this, we include a footnote that both explains why we initially chose Hedge's g (based on the assumption that it is less biased than Cohen's d, as recommended by Lakens, 2013), and that it is likely similar to Cohen's d for our selected sample size (p. 16 footnote 3). Because the data will be available in an open source format, readers will also be able to compute the Cohen's d value to verify this.

Just a suggestion for the authors, if the production effect is at least partly based on semantic processes, they would expect the production effect to influence other semantically-driven effects, like false memories in the RDM paradigm. This could be an idea for future research.

This is an interesting idea, thank you for sharing it with us. There is actually (at least) one study that did exactly that and found cross-language recognition of lures (Sahlin et al.,

2005). Given the relevance to our own work, we now introduce this research as part of the introduction (p. 6 lines 1-23, and p. 7 lines 1-2).

Reviewed by anonymous reviewer, 12 May 2023 01:57

This is a very interesting article, which will contribute not only to the current production effect literature, but also will be useful for overall linguistics theories. The hypothesis and predictions are well explained and appropriate for the study design. The power analysis is well justified and available on OSF. I recommend this manuscript to move to the next stage, with some minor revisions:

1. Additional information about spreading activation models and multi-step activation would be beneficial for better understanding of the hypothesis and predictions.

We expanded on spreading activation models when they are first introduced as well as reference more recent computational work that corroborates the older studies on mediated priming (p. 7 lines 4-10, p. 7 lines 14-17).

2. Specially for production and spreading activation (page 5, line 10), more details/information is needed here to then link it to the production effect.

When we wrote the first draft of this manuscript, we were not aware of any other work that has investigated the impact of production on spreading activation. Our goal with this study was to follow-up on Fawcett et al.'s (2022) findings that semantic processing contributes to the production effect. As a potential mechanism for these findings, we apply a spreading activation model.

Since then, we have become aware of a recent study by Tsuboi et al. (2021), where they investigated the impact of reading aloud vs. silently on priming (Experiment 2). They found that priming was increased for the read-aloud condition, consistent with the hypotheses that we put forward in this manuscript. We incorporated this novel study into the introduction (p. 8 lines 20-22, and p.9 lines 1-2).

Overall, it appears that the psycholinguistic differences between reading silently and aloud are under-researched. As such, we agree with you that our study stands the potential to contribute to overall (psycho)linguistic theories and not just the production effect literature. We make this point clearer in the manuscript (p. 8 lines 5-8). In addition, we expanded on the differences that have been described between reading aloud and reading silently (p. 8 lines 17-20).

3. I find the wording in page 6, line 13 "that semantic encoding plays a role in the production effect" misleading, sine, as the authors mention, the production effect is found even in the absence of semantic representations. Maybe replacing that by the wording used further down "contributes to the production effect".

We made this change (p. 5 line 11).

4. Page 7, line 7, has this been claimed before?

This is a novel prediction (to our knowledge). We now clarify this in the text (p. 10 line 3).

5. Page 7, line 16-17, do they do only the recognition task twice, or both learning and recognition twice?

Both will be done twice; we clarified this in the text (p. 10 line 14).

6. Page 19, line 11: "classical" should be "classic"

We made this change (p. 16 line 13).

7. For the data analysis section, I suggest also considering LMM instead of ANOVAs, to allow the authors to include item and participant effects and multiple comparisons.

We have considered this suggestion in detail and believe that it is not as advantageous to use LMMs with d-prime as the dependent measure because d-prime is calculated by averaging performance over a set of trials/stimuli. It would be possible to analyze hit rate with LMMs; however, by doing so, we would give up the advantages that are associated with analyzing d-prime.

As a compromise, we considered to also include a by-item analysis ANOVA as well. We understand that this is not the same as including random effects of subject and item in the same analysis but it may still be informative. However, a by-item analysis of d-prime is complicated by the fact that subjects see different stimuli, resulting in lots of missing data across subjects. Thus, our planned by-subject ANOVA seems to fit our planned design and dependent variable choices.

Nevertheless, we are open to the possibility of conducting LMM analyses of hit rate as part of additional exploratory analyses. In addition, we will of course provide the data in a format that is appropriate for LMMs. We predict that LMM analyses should show the same fixed effects as ANOVAs. We agree that the variances and correlations between by-item and by-subject variances are potentially very interesting, but they are not our specific goals and may be more suitable for exploratory analyses, motivating future studies.