Thank you for submitting your Stage 1 manuscript to PCI-RR. As the recommender assigned to this manuscript, it is my role to perform an initial triage assessment to determine whether the submission is ready to be sent for external review, or requires some revision. This assessment is primarily with respect to the RR aspects of the proposal, rather than its specific topical content, on which I am not an expert. With RR still relatively new for many people, we most commonly find that some revisions are required at this initial stage. That is the case with your manuscript too.

The study that you propose seems to me a clever and elegant design to get at an interesting question regarding the effect of context on memory consolidation for faces. The sample size of your pilot data is also impressive, and puts you in a strong position to propose an empirically well-informed RR. However, the design is complex, and I believe that that logical links between your theoretical hypotheses (and proposed conclusions) and the actual analyses proposed are not sufficiently precise. Your approach to the critical issue of power also needs further consideration.

The design is complex, and the issues are quite complex, and I have limited time to draft this letter, so it is possible that there may be some misunderstandings in what I have written. I am happy to discuss or clarify any such points by email. I should also note that, as with any review comments, you are not required to make changes suggested, but you do always need to provide a suitable rebuttal or rationale for your chosen course of action.

I shall list first the major issues, as I see them, with respect to the requirements of RR. In addition, I will add a number of more minor points noted in passing. These also include comments from a fellow recommender who has also read your manuscript. I hope that these comments are helpful and I look forward to seeing a revised version of this proposal if you choose to address these comments.

Thank you for the careful review of our Stage 1 manuscript. It is been extremely helpful, because, as you guessed, we are new to RR. We have made significant changes to the manuscript. We have completely revised the text and Table to be more precise about the hypotheses and the critical statistical comparison to test each hypothesis.

Major (RR) points

1) Table 1 lists specific targeted hypotheses, which you divide into 3 main questions (Hypotheses 1-3), with further sub-questions. However, the dependence of your conclusions upon different possible combinations of outcomes is not clear. For example, you state in the paper that “The comparison between In Show and Out of Show face images will be important to determine whether the effect of context on face memory is specific to the visual context in which the images were originally shown”, but your analysis plan does not include any direct comparison between in show and out of show faces.

We are now explicit about the specific comparisons that test our hypotheses on the difference between In Show and Out of Show face images – see Hypothesis 3 & 4 (pgs. 4-5, 10-11, Table 1).

With specific regard to Table1, I will unpack just one example (Hypothesis 1), but the same general issues apply to all hypotheses.

Hypothesis 1 is that your video manipulations lead to differences in understanding the narrative or context. (It is not clear what role “or context” plays here, because your tests seem to be focused on testing the understanding of the narrative, but maybe these terms just require more precise definition). There are two sub-hypotheses, relating to free-recall and structured questions, and a significant effect of context would confirm an influence of context. The direction of this influence would tell you the direction of that effect. I note in passing that, given the theoretical background of the experiment and your pilot data, it would seem perfectly reasonably to propose one-tailed critical tests here, though you are of course free to choose two-tailed tests if you wish to be able to interpret an effect in either direction.

We have refined Hypothesis 1 and have linked this to the key statistical comparisons. We have also specified the direction of the effect and consequently propose 1-tailed tests. We have used this approach to analyse the pilot data and thus determine the power necessary for the current study (pgs. 4, 9, Table 1).

However, it is not clear how your conclusions depend upon the combination of possible outcomes for H1.1 and H1.2. Will you conclude that there is an effect of video manipulation if EITHER test is significant, or only if both are? This needs to be explicit, and it may have implications for alpha correction for multiple comparisons. Arguably, if a conjunction of significant outcomes is required (both tests significant), then alpha correction is not necessary, but if a disjunction is sufficient (either test significant), you need to protect against inflation of Type I error (see https://doi.org/10.1007/s11229-021-03276-4).

We have now amended the manuscript to indicate that the effect of video manipulation will need to be significant in the free recall and structured question analyses. This will show us that the manipulation has an effect on the understanding of the narrative or context. The pilot data suggests that both tests will be highly significant. (pg. 9, Table 1)

2) At a finer grain, a similar issue arises within each sub-hypothesis. You propose a one-way ANOVA to test for differences between conditions, but the actual inference depends upon the outcome of post-hoc tests, and so the overall ANOVA is in fact redundant. I believe that you would be better to pre-register the planned contrasts of interest directly as your critical test(s) (skip the ANOVA). You state that planned comparisons showing lower scores for the Scrambled and German condition compared to Original will confirm hypothesis 1.1. This suggests that you are proposing a conjunction test – you will only conclude in favour of the hypothesis if BOTH tests are significant, not if only one is. Is that correct?

We have completely changed all the key comparisons to test each hypothesis directly. We now focus on the key between group comparisons for the Original compared to the Scrambled condition (see below).

I do wonder whether – in the economy of your experiment – the German condition is earning its keep. It is very thorough, but it seems expensive - in terms of participant numbers - to have a second control group, particularly given the very similar results between the scrambled and German groups in your pilot data. This is up to you, but if you do have two control groups, you have to specify explicitly whether your conclusions depend on a difference in the treatment group against either, against both (or perhaps even against both combined).

We take the point. From the pilot data, we found no clear differences between the Scrambled and German conditions. So, we now focus on the Scrambled condition.

3) The role of hypothesis 1 is not entirely clear to me. You interpret it as a test of the role of context on episodic memory, but the way that the hypothesis is actually stated is more like a manipulation check. That is, given that we know that context leads to changes in understanding, Hypothesis 1 will test whether our video manipulation has been successful in manipulating context. A null result would not challenge schema formation theory, but rather would suggest that your manipulation did

not do what you intend. That is, I think that Hypothesis 1 may be better understood and presented as your outcome neutral condition, or manipulation check, whereby a significant result confirms that your experiment is capable of testing its key experimental hypotheses (2&3). A null result would mean that you would not be able to interpret the outcome of the experimental hypotheses in terms of the effects of context. I may be wrong about this, but it seems more sensible to me.

We have modified the interpretation of a significant result for this hypothesis as being a manipulation check that is important for testing subsequent hypotheses (2&4). (pgs. 4-5, 9-11, Table 1)

4) The above point related to an issue of how theoretical interpretations will be informed by the combination of outcomes across your hypotheses. This also applied between hypotheses 2 and 3. The proposed theoretical interpretation of a significant advantage for the original condition would be in terms of the role of consolidation, but this would be legitimate only if the advantage for the Original condition at the delay time were greater than that for the Original condition immediately. That is, your ability to test for a role of consolidation depends on the change over time between immediate and delayed (i.e. comparison between effects of condition at times 1 and 2), not on a static snapshot of performance after a delay.

If so, this needs to be what your critical test will target. I suggest that it would be more simple and straightforward not to propose an ANOVA of condition by time that follows up the interaction term, but to isolate the interaction of interest directly by appropriate subtraction between conditions and then to run a between-groups t-test on that value. This would also allow you to specify your effect size of interest as Cohen’s d (or similar).

This is a good point. To test the effect of consolidation, we now focus on the specific interaction by subtracting across time points in the Original and Scrambled conditions and then comparing across these conditions using a between groups t-test. (pgs. 4, 9-10, Table 1)

5) You state that ICC greater than 0.75 will indicate good reliability between raters. Be clear on what role this value plays in the logic of your experiment. Is this an outcome-neutral criterion that must be passed in order for the experiment to be deemed capable of testing the key hypotheses. If so, this should be stated explicitly.

We have chosen this value as an indication of good reliability between raters. It is also lower than the reliability from the pilot study. So, we are confident that our reliability will be higher than this. However, if the reliability was lower than this, it would not rule out the ability to test the key hypotheses. We have amended the text to state this (pg. 9)

6) You propose an admixture of frequentist and Bayesian tests. Either is legitimate, but if you are using Bayesian tests, then these need to be fully specified and justified (in terms of the priors used), and supported by an appropriate sensitivity analysis that will differ from the power analysis for frequentist tests. (More guidance on this can be provided if necessary.) In general, you might be better to settle on a frequentist or Bayesian approach – if you wish to take a Bayesian approach to quantify evidence for H0, then why not also use it for H1? Or, conversely, if you wish to use a frequentist method to test H1, then you could consider a Two-One-Sided-Test approach to test H0 (where this is specified in terms of smallest effect size of interest).

We have decided to stick to frequentist tests throughout. We have also made the hypotheses more focussed and specified the critical statistical test that will be needed.

7) The effect size that you target should relate to the specific comparisons that you will conduct, which requires you to specify the smallest expected effect size, or the smallest effect size of interest for each of your critical comparisons, and show that each of your tests has sufficient power to meet your targeted thresholds.

Your sample size is predicated on an effect size of s η2 = .103 (Cohen's f = 0.338) from the critical delayed In Show face recognition ANOVA from the pilot data. There are several problems here:

(i) this is not the unique critical test of your hypothesis; and it is only one of the hypotheses under test.

(ii) Based on the fact that you are following these pilot data up because they showed an encouraging effect, the effect size estimate you are using is likely to be inflated. Should you be taking this into account?

(iii) The effect size estimate could be inflated for a second reason, which is that your in show faces for the pilot experiment were actually seen in the viewed clips, which will not be the case in your proposed RR. You should consider whether this is liable to be consequential for the size of your effect, and – if it is – then try to adjust your targeted effect size accordingly.

Good points. We have completely revised the hypotheses and the specific statistical tests that will test them. The effect size is now based on the key comparison that tests the hypothesis with the lowest power (Hypothesis 2). (pgs. 4, 9-10, Table 1)

Major (more general) points

8) Your hypotheses are stated as questions (e.g. "Is there a difference in face recognition at a delay following stimulus encoding?"), which is a little unusual.

We have changed all the hypotheses to statements rather than questions.

9) In Figure 2 (assuming that these images are representative), it would appear that the in show and out of show stimuli differ quite dramatically in terms of lighting conditions and backgrounds. It may be that your design means this does not matter, but you should make a more specific argument for why this is the case (or adjust your in show and out of show stimuli to be better-matched).

This is an important aspect of the design. Our main comparison is based on the In Show images. These are images that the participants have not seen, but nonetheless are similar in appearance to those seen at encoding. Our hypothesis is that our recognition of these faces will be enhanced by context, but only after the effect of context has been consolidated in memory (i.e. at the delayed time point). However, we would also like to test Out of Show images that are very different in appearance to those that are shown at encoding. This will help us test secondary hypotheses about within-person variation (Hypotheses 3 & 4). This is based on evidence that shows recognition of faces depends on the range of within-person variation. In this paradigm, the within-person variation is low. So, we would expect that the effect would be more evident for the In Show images. We have changed the text to make this clearer (pgs. 3-5, 7).

10) You do not provide any description or examples of the foil stimuli, but these are critical to be sure that they are valid foils.

We have included a description of the foils stating that they will be matched on age, expression, hairstyle, lighting and general appearance. (pg. 7)

11) You do not seem to provide information about how long the clips are, how many there are etc… I can only find the statement that “Three 20-minute (1170s) movies constructed from audio-visual clips from the first episode of BBC TV series Life on Mars will be used as stimuli.” In general, your methods need to be described precisely and unambiguously (and, ideally, if you could make materials open to the reviewers, it might help in evaluation of the proposal).

We have now provided this information on pg. 6

12) As far as I can tell (maybe I missed it?) you do not even specify the number of trials that are completed in the face-recognition tests, nor the dimensions or timing or duration of stimuli.

We have added this information on pg. 7-8.

13) Related to the above, is the number of trials sufficient to support a valid calculation of d’ (or would a more simple metric eg. HR-FAR be more suitable?).

Our pilot data suggests that we have a sufficient number of trials to calculate d’ for each participant.

14) Your calculation of d’ builds in compensation for floor and ceiling effects, but how common do you expect floor and ceiling effects to be? If they are very rare this may be fine, but if they are more common it could be problematic, and you might need to consider adjusting your stimuli to avoid them and/or adding exclusion criteria for performance level.

We found that this was quite rare (~8%) in our pilot data, so we think that the adjustment should be fine.

15) The above would be very usefully informed by your pilot data. The pilot experiment is good, and very relevant, and I think it might work better to have a fuller report of the pilot experiment prior to the description of the planned study (rather than afterwards). This is particularly so, given that the original n (albeit not the final sample) of the pilot is in fact larger than that of the planned RR. Perhaps it could even be styled as an Exploratory Experiment 1, rather than as a mere pilot reported at the end. The present description of the pilot study is rather telegraphic about details (e.g. What do Test 1 and 3 refer to? And what happened to Test 2?). Your report of these data could usefully include some more insight into general levels of performance and distribution of scores on the tests. The bar charts with standard errors may demonstrate salient patterns, but they do not provide much insight into the underlying data.

We have changed the text to describe the experiments in the pilot data more clearly. We have also included the mean and standard deviation across groups (pgs. 12-15).

16) Make it clear that the sample size relates to complete valid datasets after exclusions.

We have made it clear that the sample size relates to data set after exclusions (pg. 6).

17) How likely is it subjects will do the 24h and long-term experiments in time? Responding to email at short notice seems potentially unreliable.

We have removed the 24h time point. We have also indicated that participants will have 48h window to complete the test at the 4 week time point (pgs. 8-9).

18) Will subject assignment be balanced across conditions in some way?

Yes.

19) Do you have any strategies to minimise the potential for participants to lie e.g. about German language skills, or knowledge of Life on Mars etc?

We now do not have a German condition. We also screen people to ask if they have seen Life on Mars. It was first broadcast in 2006 and it is no longer available on BBC iPlayer or Netflix. So, we think that it is very unlikely that the participants we recruit will have had the opportunity to see it.