Reviewer: Haiyang Jin

Review of “The importance of consolidating perceptual experience and contextual knowledge in face recognition” (PCI-RR#313_Stage1).

The authors proposed an interesting study to explore whether the video sequence (original vs. scrambled) will influence the performance of recognizing faces with varied contexts (“in show” vs. “out show”) both immediately after watching the videos and after 4 weeks. The results will potentially bring new insights into how the context, manipulated via video sequences, affects recognition performance/familiarization.

Thank you for those positive comments.

The authors have included many experimental design and analysis details in the report. It would be better if the authors could clarify some parts further.

For the power analysis reported in P6, the authors need to clarify the employed power analysis is intended for which of the four hypotheses. Specifically, the employed effect size seemed to come from one simple effect (P14), while no simple effects were employed for testing any of Hypothesis 2-4. Also, when GPower was used, whether the employed effect size was submitted to the (one-sided or two-sided) t-test or ANOVA?

We have now revised this aspect of the study. A sensitivity analysis has been computed for one-sided independent t-tests with a power of 0.9 and alpha level of 0.02. Our proposed sample size of 100 participants per condition (Original and Scrambled) is sufficient to detect the effect size for all the hypotheses based on our pilot study. We have added a new figure (Figure 1), amended Table 1 and revised the text (pg. 6).

For the stimuli, there are obvious differences in luminance between “In Show” and “Out of show” images. I’m not sure if the authors could do anything about this, but these luminance differences could be the main source of the differences in performance (thus a potential confounding to the effects of within-person variability). Moreover, it is not clear how the foil faces differ from the target stimuli (e.g., in luminance), it would be helpful to additionally display some examples of foil faces in Figure 2.

Unfortunately, there is nothing that we can do about the appearance of the In Show and Out of Show images. These differences are to be expected given the way that different shows are filmed and reflect natural variation in viewing conditions. However, this should not influence hypotheses 3 and 4. This is because the measure of recognition is based on a comparison between targets and foils within each condition. The key thing is that the targets and foils are similarly matched. We have now added some examples of the foils in Figure 3 and revised the text (pgs. 7-8).

For the planned analysis for Hypothesis 2, the proposed analysis is essentially the interaction between Test time (0hours vs. 4weeks) and Image condition (original vs. scrambled) with a particular direction. But the interaction itself is unlikely sufficient. For example, there is a possible case where the deduction in the scrambled condition is 0 while

the deduction in the original condition is -0.5 (i.e., an increase due to certain reasons, e.g., measurement errors). Then the deduction will be larger in the scrambled relative to the original condition, which matches the proposed supporting results (Table 1). However, probably the authors should not claim the hypothesis was supported by the case above. Instead, in addition to the proposed analysis (the interaction with a particular direction), the authors may also need to clarify at least one of the simple effects (for more please see the example in Jin, 2022). A similar comment applies to Hypothesis 4 as well.

Thank you – that is an interesting point. Erroneous support for Hypotheses 2 and 4 (outlined above) would be possible if there is an increase in recognition in the Scrambled condition. However, our pilot data (suppl. Fig. 2 & 3) shows that this scenario is unlikely to be the case. Moreover, if there are increases then this will be a matter for further consideration in the Discussion part of the article, and which may temper the conclusions reached. Nonetheless, we have amended the text (pgs. 4-5, 10-11) and Table 1 to make the directionality of the effect clear.

In addition to the list of exclusion criteria, the authors may also need to consider some other criteria. For example, since it is an online study, authors may also like to consider including the criteria to include/exclude participants with behavioral performance (e.g., exclude participants with too fast or too slow responses). Also, authors may also need to consider excluding anyone who is already familiar with the actors (both the actors for the 10 characters and potentially also the foil faces).

In our online pilot study, we found that the participants responded between 0.6 and 5 sec per trial. So, we are not intending to exclude participants on the basis of response time. We have, however, added some additional exclusion criteria to exclude participants who have seen other shows that include these actors and have changed the manuscript (pg. 11).

Also, I'm not sure whether the "consolidation" in the introduction and in Hypothesis 3 and 4 have the same meaning. Specifically, "consolidation" in the introduction seems to refer to the enhancement of the memory of the identity/face potentially with some manipulations (e.g., watch the videos/images again) while "consolidation" in the hypothesis only refers to what will happen during the delay time in general and no explicit enhancement manipulation will be applied. Therefore, it remains elusive whether there is "consolidation" (or enhancement in this study). Probably it is more appropriate to argue the effect as "how much participants will forget after the delay of 4 weeks".

Good point. We are using the term consolidation to mean the generation of more stable and long-lasting memories. We are not suggesting that consolidation will lead to increased recognition over time, but that recognition of faces consolidated in a coherent context will be enhanced relative to those learnt in the absence of a coherent context. We have changed the introduction to make this point more clearly pg. 4.

Minor points:

1. It would be helpful if authors could summarize the whole experimental design, which seems to be 2*2*2?

We have changed to the text to provide an overview of the experimental design (pg. 7).

2. P.4. It is stated that "However, if recognition memory for faces is greater in the Original condition, then this would suggest the importance of contextual knowledge." Probably authors intended to argue that "if the contextual knowledge is important, the recognition memory for faces is greater in the Original condition.

   Yes, that is a much better way of phrasing what we intended (pg. 4).

3. P.5. Please clarify what is "a normal range" to be used for CFMT.

   Normal range is within 2 SD of the mean of CFMT. We have changed the text to make this clear (pg. 6)

4. P.9. How many raters will be?

   As in the pilot study, there will be 2 raters. We have modified the text to make this clear (pg. 9).

5. Table 1. "the null effect" does not seem to be applied appropriately. The authors do not seems to apply Bayesian methods or Equivalence tests to test the potential support for the null hypothesis. If the authors do not obtain significant results with the NHST they specified, it remain unclear whether the evidence is inconclusive or it supports the null hypothesis. Therefore, the authors cannot claim "A null result would suggest that…" or "A null result for In Show images…", etc., especially when they used a more stringent alpha (i.e., 0.02).

   We have based our sample size on the sensitivity analysis of the pilot data. Given that we have appropriate power, our assumption is that if the statistical tests used for each hypotheses are not significant then we will not be able to reject the null hypothesis, showing that the effect is not present or, if present, is smaller than the target effect size.

Reviewer: Lisa Debruine

My review will be focused on methodological and analysis issues; I won't be commenting on the appropriateness of the background or hypotheses.

Methods

- Please include an explicit cutoff (or strategy for calculating it after data collection if it relies on the sample distribution) for the Cambridge Face Memory Test for determining whether face perception is in the normal range, even if this is mentioned in the referenced paper.

We now state that the normal range will be based on CFMT scores that are within 2 SD of the mean (pg. 6).

- Participants will be screened for familiarity with the actors from Life on Mars. I've never seen that show, but checked iMDB for the actors and am familiar with John Simm from 24 Hour Party People, and Archie Panjabi was in 134 episodes of the very popular series The Good Wife. Will you screen for familiarity with the actors (perhaps after the main test is completed?)

Good point. We will now include an initial questionnaire in which we list other TV shows that include these actors (pg. 11).

- As the stimuli are completely essential for interpreting and perhaps replicating and extending your findings, do you plan to make them available? A managed archive, such as the UK Data Service might be appropriate if you are worried about copyright issues. At the very least, a clear description of the exact timestamps and scrambling for each video should be shared with no restrictions, plus links for finding the still face images for the recognition tests.

We are concerned about copyright issues. However, we will be happy to provide the images in a managed archive such as the UK Data Service.

- It should be possible to find an out-of-show image for all actors used in the study; it will make analysis more straightforward (but is possible to accommodate in a mixed design)

We were unable to find 'out of show' images for one actor

Analysis

- Thanks for including such a clear table of the hypotheses, analysis plan, and interpretation!

Thank you for that positive feedback.

- One method I use to see if an RR's analysis section is sufficiently detailed is to see if I can write the code to analyse a simulated data set. I probably didn't get everything right below, but that will give you clues where the RR should be clearer.

Thank you for doing this. It does make it clear where we could make the RR clearer. Our responses to the specific questions from the simulation are as follows:

How many raters will rate the free recall test? I've guessed 10

As in the pilot study, there will be two raters for the free recall test (pg. 9).

Are the foils for the in and out of show versions of the same actor the same foil identity (from 2 different shows) or two different identities?

The two foils for each target image have different identities. We have changed the text to make this clear (pgs. 7-8).

I'm not 100% sure if you're assessing the questions individually or after aggregation.

Inter-rater reliability will be assessed for both the free recall and structured question test aggregated across questions using intra-class correlation coefficients (ICC) in a two-way mixed model with agreement definition (pg. 9).

Are the analyses scores average sums across rater?

Yes, the scores on the free recall test will be the average across raters (pg. 9).

Is there any possibility that subjects or raters will skip a trial? If so, how will you handle that?

If there is any possibility of missing trials, a mixed model might be more appropriate here.

The code will be set up so that it is not possible for subjects or raters to miss a trial.

The table is much clearer than the text about the setup of this

You predict a main effect of show (in/out), but don't specify how you would interpret an interaction between show and condition. Which interaction patterns would be consistent or inconsistent with your hypothesis? Must this pattern be present in both original and scrambled conditions to support the hypothesis?

The prediction here is an interaction between show and type, whereby the target-foil difference is larger for in-show than out-of-show (for both original and scrambled conditions?

We do not have any predictions about interaction patterns. The key thing to show that recognition is related to the appearance at encoding (i.e. In Show > Out of Show). We have

now changed the proposed analysis to test hypothesis 3 to include a one-tailed t-test just for the Original condition. We have changed the text on pgs. 11, 15 and Table 1.

The interpretation part of the table for this hypothesis is confusing, as it refers to an interaction, but the proposed test is a t-test

Sorry that is a typo. It should say a non-significant result would imply

- A huge benefit of this method is that your analysis code will be done and super-easy for reviewers to check for consistency with your plan in the stage 2 RR.
- I've suggested a mixed effects analysis where appropriate instead of aggregated analyses.

We hope we have answered all questions raised from your helpful simulation.

- One general tip I have is to show plots of your predicted effects for each hypothesis. This makes it a lot easier for a reader to understand what is going on with this complex design.

Yes, we can definitely see how this could be helpful in the absence of data. Hopefully, the plots from the pilot data provide a schematic of the predicted effects.

Power Analysis

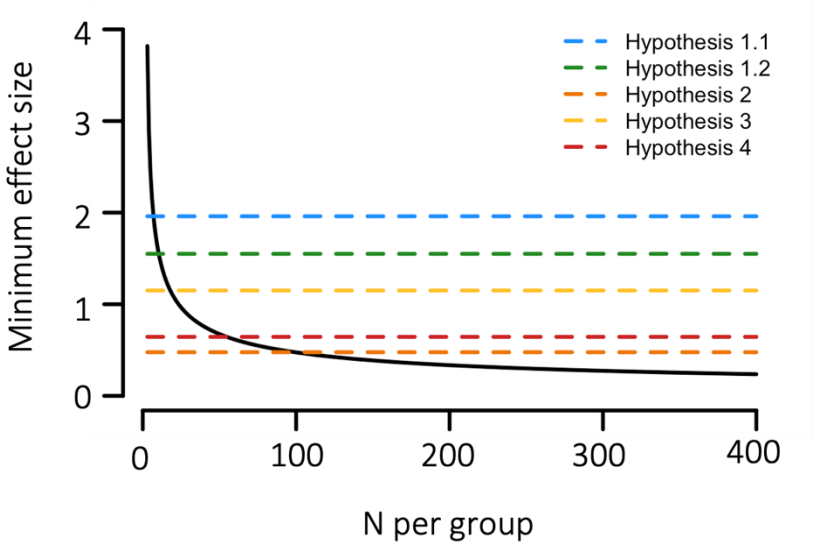
- There is only one power analysis, for Hypothesis 2, using d = 0.476, from the pilot data. In Table 1, you describe this as a "smallest effect size of interest". However, values from pilot data are likely to be poor estimates (or we wouldn't need to full study). I'd usually recommend powering for smaller effect than the pilot data, but see my further comments first on sensitivity analyses.
- I appreciate that you are explicit in your alpha criterion for power, but am wondering where the 0.02 figure came from.

A power of 0.9 and alpha level of 0.02 is for our target journal (Cortex)

- I'm happy to advise on conducting power analyses by simulation (using the pilot data) for more complex analyses that don't have an analytic solution.
- Instead of power, I'd encourage you to report sensitivity analyses for each hypothesis, reporting the smallest effect size you have 90% power to detect at your sample size of 100 in each condition (with 0.02 alpha for all tests?) and the size of this effect in the pilot data for comparison (with CIs if possible).

We conducted a sensitivity analysis (see Figure 1) for a one-tailed independent t-test with a power of 0.9 and alpha level of 0.02. This showed a rapid initial decrease in the minimum effect size that could be detected, with improvements being relatively marginal beyond around 100 participants per group for our smallest theoretically important effect size (d = 0.476, see orange dashed line in Figure 1). We chose this as our sample size, as it allowed us to detect effect sizes of a similar magnitude to that found in our pilot work and also kept the experiment feasible from a practical perspective. This is a 'medium' effect size (see Cohen,

1988), and we consider that effect sizes smaller than this are unlikely to have practical relevance for everyday face recognition performance, so it also constitutes the smallest effect size of interest for this work.



Sensitivity analysis showing the detectable effect size for a one-sided independent t-test with a power of 0.9 and alpha level of 0.02. The dashed lines represent the effect sizes found in the pilot data for each hypothesis.