Haiyang Jin and I have some comments that require your consideration, and you should indicate how you have responded to these if and when you submit a revised version. As a matter of editorial guidance, please note that, because you have followed your preregistered analysis plan, you are not obliged to follow reviewer suggestions for additional exploratory analyses at this stage (though you can do so if you agree that they may be important/informative). For instance, having specified a NHST approach, you are not required to follow-up non-significant outcomes with equivalence tests, which would ask different question from the one that you preregistered (however, if you do decide to include them, then these tests should be clearly demarcated from the preregistered portions, as post-hoc).

Thank you for taking the time to review the manuscript Rob. We have outlined how we have responded to all the comments in red and indicated the corresponding changes to the manuscript.

My own main comment on reading this Stage 2 manuscript is that there may be too much emphasis given to the exploratory analyses, which have been allowed to play an undue role in driving your main conclusions from the study. It is perfectly acceptable to add exploratory analyses, but these have a subordinate status to the pre-registered analyses, which were agreed in advance to represent the most appropriate tests of your hypotheses. The pre-registered analyses and outcomes must therefore remain clearly in focus and should drive your main conclusions. If you believe, in retrospect, that these were not adequate tests of your hypotheses (not just because they did not produce expected outcomes), then you should explain why they may have been compromised, and make recommendations for how the hypotheses could be reassessed in future. You can use your exploratory analyses to support these discussion points, but you cannot implictly or explicitly swap them in for your pre-registered analyses as representing the preferred tests of your original hypotheses.

Thank you for clarifying that. We do understand that the exploratory analyses have a subordinate status to the pre-registered analyses. We have removed some of the exploratory analysis from the manuscript. We have also changed the way the exploratory analyses are described throughout the manuscript so that they do not read as an alternate test of our original hypotheses.

This comment relates to the framing of your results, and the relative balance between pre-planned and post-hoc parts in driving theoretical conclusions in Discussion. Although the exploratory analyses may be sensible, and suggestive, they are unconstrained in the researcher degrees of freedom available (meaning that p values lose their formal meaning), and their post-hoc nature means that any conclusions drawn from them must be highly tentative, and considered as suggesting ways to configure future hypothesis tests, rather than supporting any clear conclusions in their own right. You need to be scrupulous to avoid implying that your exploratory analyses represent the same severity of test (or even a better test) of your hypotheses than your pre-planned analyses do.

This framing is sometimes fairly overt. For instance, in Discussion, you state:

"… an individual differences approach could be a better way to explore the role of conceptual information on face recognition. Accordingly, we performed an exploratory analysis in which we compared conceptual knowledge and face recognition across all participants…. The difference between the immediate and delayed timepoints for Out of Show faces suggests that a period of consolidation may be necessary for the development of a more flexible representation that underpins face recognition."

We have now removed this paragraph and this part of the exploratory analysis from the manuscript (see response to Haiyang Jin).

"Interestingly, the strength of the relationship between the conceptual knowledge and face recognition was
significantly greater at the delayed timepoint, which again supports an important role for consolidation in memory. Thus, while our pre-registered analyses failed to show support for a greater effect after consolidation, our exploratory analyses show that conceptual knowledge is both quantitatively and qualitatively important in generating stable representations of people."

We have changed this paragraph to differentiate the exploratory analysis (qualitative differences between participants) from the pre-registered hypotheses (quantitative differences across groups). The conclusions from this exploratory analysis have been changed so that they are not taken as a direct test of the pre-registered hypothesis. Moreover, these have been described as being suggestive of an effect and that they may provide a useful approach for future research.

The tilt towards an emphasis on exploratory outcomes can also be more subtle and stylistic. For instance, in the final conclusion (repeated in your Abstract) you state:

"While planned analyses did not reveal a greater effect of conceptual knowledge after consolidation... Exploratory analyses showed that the level of conceptual knowledge was significantly correlated with face recognition, before and after consolidation. These findings highlight the importance of non-visual, conceptual information in face recognition during natural viewing."

This statement de-emphasises and skips over the null result of the planned analyses by placing it within a subordinate clause ("While planned analyses did not…"), and then uses strong inferential language to interpret the exploratory results ("Exploratory analyses showed that…").

This may seem pedantic, but it is critical to the integrity of the RR format that proper weight be given to the pre-registered hypothesis tests, and that further exploratory analyses, which are not subject to the same level of bias control, are interpreted appropriately in this context.

We have restructured this section of the Abstract so that the description of our key planned comparison is described in a single sentence rather than within a subordinate clause. We have also avoided using strong inferential language to interpret the exploratory results.

I also mention three other issues below:

1) In my view, you need to give a more detailed explanation both of the rationale and of the procedure for the exploratory analysis in which you "correlated the similarity of the free-recall text and the similarity in the recognition of faces across all pairs of participants". A clear argument needs to be made that this is both a theoretically relevant and a statistically sound thing to do. Here (if you keep these analyses) and elsewhere, there should be less focus upon the significance of correlations, and more on their size. It is not surprising is a correlation with nearly 20k pairs achieves significance, and not particularly surprising if one with 200 pairs does, so it is more informative to discuss the estimated strength of correlation. This is especially so in an exploratory context, where the meaning of 'significance' is moot.

We have removed one of the exploratory analyses (see response to Haiyang Jin). We have also provided a more complete description and rationale for the use of the similarity analyses in the Discussion. This is a method that we were not aware of at pre-registration. Nonetheless, it provides an alternative qualitative approach to understanding the relationship between conceptual knowledge and face recognition.

We feel that the strength of the correlations (.18 - .24) are good within a psychological context of two very different dependent measures. Although we agree that significance is moot for correlations with very large dof, this is not how we measured significance. We used permutation testing to randomly assign values in the similarity matrix. We have changed the text in the Results to make this clear and we have removed the large dof, so that it avoids confusing the reader.

2) Plot style

You use different plotting styles for different outcomes. Figure 4 is a violin plot, and Figure 5 is a bar plot. If you are deciding to use different plotting styles across plots, then it should be clear to the reader why you have made this choice.

It looks like you might have used a violin plot for Figure 4 to more fully represent a non-normal distribution, but this creates concerns that the parametric t-test could be inappropriate to the data. As an aside, violin plots can be a nice way to illustrate a distribution, but as they are effectively a symmetrically-reflected density plot, you should check that they are doing a good job of representing your data (sometimes the density smoothing kernel can lead to misrepresentations). It can often be useful to overlay a jitter plot of individual observations, so that all data points are ultimately displayed.

We have changed Figure 4 with the two narrative tasks plotted separately. We are happy that these are doing a good job of representing the data.

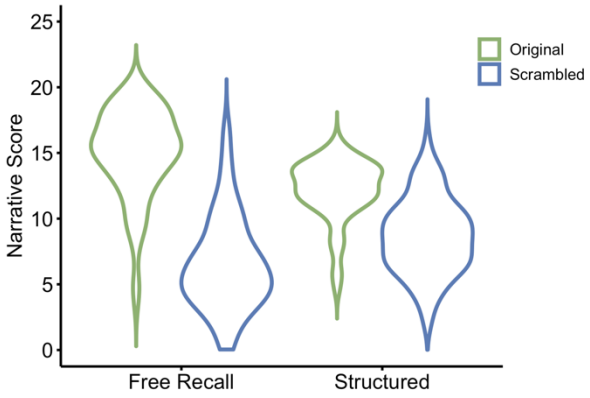


Figure 5 is a bar plot, with what I presume are +/- 1SE error bars (but this is not stated), showing the interaction of the within-subject factor of image, and the between-subject factor of condition, with different plots for each level of the within-subject factor of time. This is fine to grasp the overall pattern (although 95% CIs might be preferable to SEs, unless you have a strong reason to prefer the latter). I also think it would be more intuitive (for me) to split the panel by the between-subjects factor, so each plot represents the experiment for a different subject group. (You could also then use within-subject error bars within each panel). But this is really a stylistic preference.

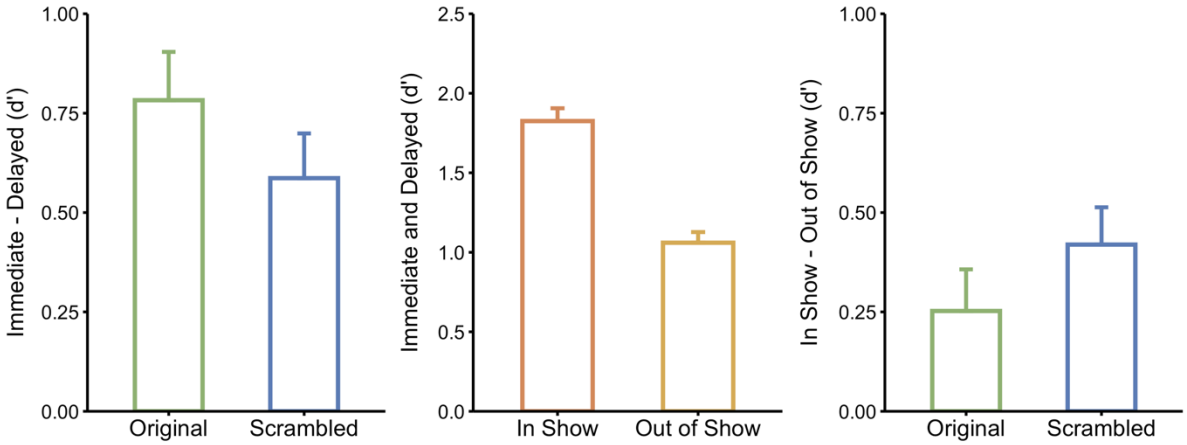
We have made it clear in the legend for this figure (now Figure 6) that the error bars are SE. Our preference is not to split the between-subjects factor (Original, Scrambled) as this is the key manipulation. So, we have left this Figure unchanged.

3) Plot specificity

You should be sure to include plots that specifically represent the data on which the inferential tests have actually been performed. Hypotheses 1.1 and 1.2 are tested and reported separately, but they seem to be conflated into a total score in Figure 4, for which there is no corresponding hypothesis test. For Hypothesis 2, the relevant data would be the immediate-delayed difference per group (preferably with 95% CIs); similarly you should directly represent the relevant collapsed data for H3 and H4.

We have modified Figure 4 (see above) so that it has the data showing the comparison used for Hypotheses 1.1 and 1.2.

We have also generated a new Figure 5 (see below) with 3 panels that provides the key comparisons for Hypotheses 2-4.



Review by Haiyang Jin, 03 Feb 2024 06:24

Review of "The importance of conceptual knowledge when becoming familiar with faces during naturalistic viewing" (PCIRR Stage 2).

I'm Haiyang Jin, and I always sign my review.

Haiyang, thank you for taking the time to review this manuscript and for your helpful comments.

The manuscript tested the role of conceptual information in learning new faces. Participants were instructed to watch a movie either in the original sequences or a scrambled sequence. Their performance in recognizing identities in the movie was tested the same day and after 4 weeks. The pre-registered analysis did provide support for some of the pre-registered hypotheses but not all of them. Some exploratory analyses were performed, and it was concluded that conceptual information plays a critical role in face familiarization.

The manuscript follows the pre-registered information in general and performs additional expletory analyses. However, some concerns described below should be addressed before the recommendation.

T-tests were used to test the pre-registered hypotheses following Stage 1 report. But some t-test results were not statistically significant. In other words, it remains unclear whether there were no differences between the tested two conditions, or it was inconclusive. Although it was not pre-registered, Bayesian or equivalence tests (Lakens et al., 2018) should be additionally performed.

Our pre-registered hypotheses specified a NHST approach, rather than equivalence tests. So, we would prefer not to deviate from this plan. We have, however, made new Figures that clearly show the data that provides the key comparisons that are used to test each pre-registered hypothesis.

Some of the exploratory analyses should be re-considered. For example, the approach in

calculating similarity in face recognition may not be appropriate. The proposed approach only considered the identities remembered by both participants as “similar” but ignore the fact that the identities that were not remembered by both participants could also be considered as “similar”. Thus, researchers may consider using the correlation for binary variables instead to calculate the similarity.

Thanks for this suggestion. This is an interesting point and one that we struggled with when setting up the analysis. One way to think about this is with respect to the narrative similarity analysis. For example, two participants who report similar content will be given a high narrative similarity score. On the other hand, two participants who report limited incoherent content will receive a low similarity score. For face recognition, participants who recognise the same faces will be given a high similarity score. However, (to align with the narrative similarity) participants who do not recognise the same faces are given a low similarity score. Otherwise, we would have participants who both have a lot of errors being rated as having a high similarity. We believe a correlation for binary variables would have this problem. So, we would prefer to keep our current analysis.

In the exploratory correlation analyses, “there was some overlap in conceptual understanding between the groups” does not seem to be a good justification for combining data together. The potential issue has been addressed by previous literature (Figure 2, Makin & Orban de Xivry, 2019). If the analysis were kept in the manuscript, the individual points from each group probably should be displayed in different colors in correlation figures. (Figure 6)

Thank you. Yes, we understand this point. We agree that there is a potential confound when combining different groups that differ in the two dimensions into a correlation. We have therefore decided to remove this analysis from the manuscript.

Also, researchers may need to clarify whether the exploratory analyses were testing the same pre-registered hypotheses or new hypotheses. Please also clarify whether there were any other exploratory analyses performed but not reported. A related potential issue is to clarify whether multiple comparison correction was applied on exploratory analyses (or why it is not needed here).

The exploratory analysis that we now report uses a different approach to measure the link between conceptual knowledge and face recognition. Rather than looking at overall differences between groups, this analysis looks at qualitative similarity between participants. Nonetheless, it does allow us to make a link between these measures and to determine if this relationship changes over time.

No other exploratory analyses were performed. The exploratory analyses were significant following multiple comparisons (Bonferroni-Holm).

Throughout the manuscript, “consolidation” was used. However, “consolidation” means “the action or processing of making something stronger or more solid”. But the reported effects all seem to be “fewer decreases” rather than “stronger” (for a 4-week test relative to

the same-day test). The authors may need to use other words to describe the observed effects.

Yes, we do use consolidation to show a smaller decrease in recognition. We interpret a smaller decrease in recognition as being consistent with a stronger or more solid memory. From a memory perspective, our ability to recall things from the past is always prone to forgetting. Consolidation of these memories makes them less resistant to forgetting. We have checked through the manuscript to make it clearer that our prediction is not for an increase in recognition following recognition.

Minor points:

1. The alpha level should be consistent across the manuscript. For example, the alpha of 0.02 was used for the first half of the manuscript (mainly pre-registered analyses). It is unclear whether the alpha of 0.05 was used for the other half (e.g., P. 17. "The correlation at the delayed timepoint was significantly greater than the correlation at the immediate timepoint ($z = -1.69$, $p = .046$).")

We have now removed this analysis.

2. (P.12) It is probably better to avoid using "comparing" when the analysis is correlation. (First paragraph in exploratory analysis)

We have now removed this analysis.

3. (P.19) The abbreviation (the TV series Life on Mars (LoM)) probably should have been defined at an earlier time.

We refer to the TV series in full in the Methods for clarity. So, this represents the first time we use the abbreviation.

Reference

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. Advances in Methods and Practices in Psychological Science, 1(2), 259–269. https://doi.org/10.1177/2515245918770963

Makin, T. R., & Orban de Xivry, J.-J. (2019). Ten common statistical mistakes to watch out for when writing or reviewing a manuscript. eLife, 8, 1–13. https://doi.org/10.7554/eLife.48175