Thank you for your revised Stage 2 report. I think that you have handled many of the review comments well, but that there are still some outstanding issues to be addressed. I found it necessary to ask Haiyang Jin (Reviewer#1) for a further round of external review, in particular to consult on an issue that I find problematic (or at least insufficiently well explained). I am quoting directly from my request to HJ below, to provide the context for his review comments.

"My concern is about the exploratory analysis of narrative and face similarity, and whether the method used is a valid way to operationalise the question. The narrative similarity algorithm is not clearly explained, and so I think the reader is unsure exactly what this measurement represents. Second, the face similarity metric (which you queried at Stage 1) seems very odd to me, because it is just the number of faces that were both recognised by a pair of participants. If I understand this correctly, this means that if I were a participant who recognised all faces, then my 'similarity' score with each other participant would simply be equal to the number of faces that they recognised. Moreover, for any participant, the maximum 'similarity' that they can have with any other participant is fixed at the number of faces they recognised.

This implies, for instance, that:

- for a pair of participants in which one person recognised 30 faces, and another recognised 15, the 'similarity' of recognition is 15.

- but for a pair of participants in which both people recognised exactly the same sub-set of 15 faces, the 'similarity' is also 15

- and for a pair of participants who recognised exactly the same sub-set of 7 faces, the 'similarity' is 7.

This metric, if I have interpreted it correctly, does not seem to capture what I would intuitively think of as 'similarity', missing out on critical variation in the specificity of the pattern of faces remembered.

I would very much value your consultation on this particular issue, if you have time to give it."

Please see HJ's review for a more knowledgeable take on the same issue. Please bear in mind that your paper needs to be fully understandable both to people who are and are not already familiar with the 'narrative similarity' approach you have used.

Thanks Rob – we will reply to Haiyang Jin's comments to avoid duplicating the response.

In addition:

I do not think you have answered my comment on plot style fully. In particular: (a) why do you choose different plot styles (violin, and bar) for different plots; (b) why do you choose to use SE for error bars (rather than, say 95% CIs).

We chose violin plots for the narrative scores to show a more detailed representation of the data distribution. This was important for showing how much variation was evident across participants and how this could be used in the exploratory analysis using LSA. The difference between conditions is large in this analysis. So, these plots also show the main effects. We chose box plots and SE for error bars in the other data figures. In these analyses, the effect sizes are smaller, so these provide a simple way of evaluating the key comparisons from our registered hypotheses.

Having spent some time re-reading your Methods, I noted that it was rather difficult to find some key pieces of information. Specifically, I think it would be helpful if the 'Design' statement named the levels per factor (not just the number of levels), and if it were more clearly stated somewhere prominent in the Methods how many faces were viewed per participant per condition (and how many in total). Although the Stage 1 parts of the manuscript should not be substantially changed at Stage 2, I think that these small amendments would improve the readability.

We agree that it is difficult to easily find this information. We have changed the text on page 7 to give information on the levels per factor: Condition (Original, Scrambled), image type (In Show, Out of Show), and timepoint (Immediate, Delayed).

The number of faces is stated on pg. 8.

For each In Show or Out of Show face for each face memory test, two foils of different identities were selected that matched the targets in terms of age, expression, hairstyle, lighting, and general appearance (Colloff et al., 2021). 19 target images (Out of Show image not available for one actor) and 40 foils were used in each face recognition memory test."

However, we have added an additional sentence to provide a summary of the total number of faces

A total of 30 In Show images and 29 Out of Show images were shown at the immediate test, and a new set of 30 In Show images and 29 Out of Show images were shown at the delayed test.

Fig 5 legend typo: "Higher recognition ws evident"

Good spot - thanks

I therefore invite a revision of this Stage 2 manuscript to address/clarify the outstanding issues.

Best wishes,

Rob McIntosh

PCI RR recommender

*by **Robert McIntosh**, 22 Apr 2024 08:30*
Manuscript: **https://osf.io/nqez9**
version: Perceptual_Contextual_Reg_Report_PCIRR_S2_v2.docx

Review by Haiyang Jin, 22 Apr 2024 06:17

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

Review of "The importance of conceptual knowledge when becoming familiar with faces during naturalistic viewing" (PCI-RR#669_Stage2).

Thank you for addressing the potential concerns. The manuscript is in better shape. The authors have definitely put a lot of effort into the revision.

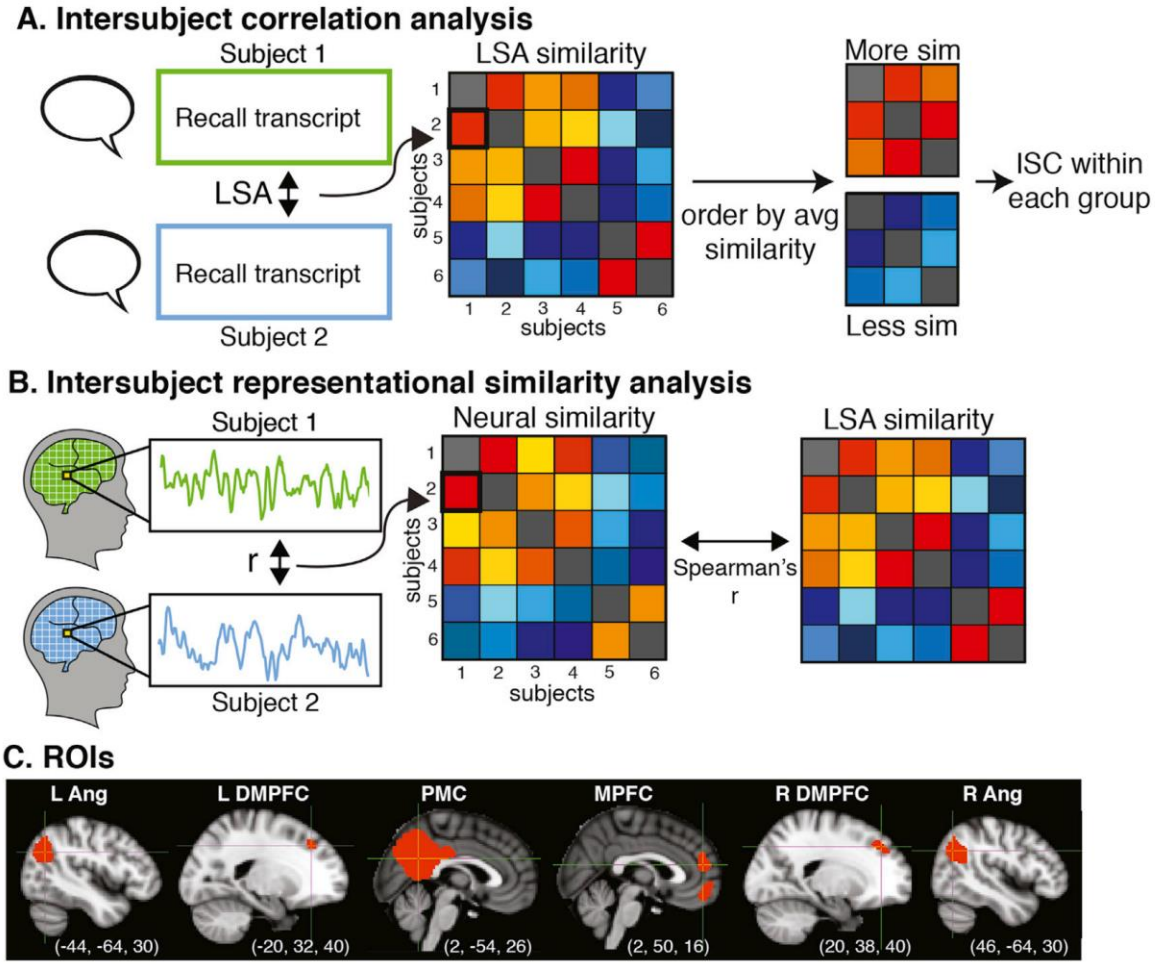
This review focuses only on the exploratory analysis of narrative and face similarity. If this exploratory analysis is to be included in the final version, I think more effort is needed.

First, I agree that the narrative similarity algorithm is not clearly explained, which presents extra difficulties for readers. The key component of the narrative similarity is 'embedding', but it is not well explained in the manuscript (and probably most readers, especially face processing researchers, are not familiar with this terminology). Also, it is unclear how the narrative contents were submitted to the embedding (e.g., are all words embeddings were averaged or using other ways?). Without sufficient information, it remains elusive how the averaged embedding connects to content understanding or narrative. One sanity check (or clarifying what the average embedding means) might be testing the narrative similarity among participants with varied free recall scores graded by the two raters. For example, does participants with higher free recall scores also have a higher embedding similarity (relative to participants with different free recall scores)? Another important aspect to be considered is whether the potential relationship between narrative and face similarity is specific to the embedding currently used. In other words, if embeddings generated from other corpus (rather than the one currently used) were used, do we still find the same or similar relationships between narrative and face similarity?

We have changed the text in the methods to introduce the technique to people who may not be familiar with Latent Semantic Analysis (LSA).

LSA is a technique in natural language processing and information retrieval that helps to uncover the underlying structure in a collection of text by analyzing the relationships between the words. It helps to uncover relationships between text datasets by mapping words and documents into a continuous semantic space. In this space, similar words and documents are positioned closer together, reflecting their underlying semantic relationships. In this study, we have compared the free-recall text summary of the narrative between different pairs of participants. The similarity between texts that is measured using LSA is taken as the overlap in semantic (or conceptual) understanding about the movie they have watched. The logic behind this analysis is that participants may have picked up on different pieces of conceptual information from the movie. This analysis will provide a measure of this overlap.

We agree that this is a technique that that researchers in face perception my not be familiar with. Nevertheless, it is an established method for analysing text (cf Jafarpour, A., Piai, V., Lin, J. J., & Knight, R. T. (2017). Human hippocampal pre-activation predicts behavior. Scientific reports, 7(1), 5959; Nguyen, Vanderwal, & Hasson (2019) Shared understanding of narratives is correlated with shared neural responses. NeuroImage, 184, 161-170). For example, the same analysis protocol was used by Nguyen and colleagues (2019) to compare narrative understanding similarity between participants in a previous study (see figure from this study below). In this study, they then correlated narrative similarity (using LSA) with neural similarity (using fMRI) between participants.



We have changed the methods to make this clear. We hope this provides enough information for the reader.

Second, the validity of the face similarity metric is not sound. Although it is explained in the reply that its main motivation is to align the analysis of narrative similarity, using the number of recognized faces by both participants to index the face similarity does not seem to match our intuitive understanding. Please consider the examples provided by Prof. Robert McIntosh:

- 1) for a pair of participants in which one person (participant A) recognised 30 faces, and another (participant B) recognised 15, the 'similarity' of recognition is 15.
- 2) but for a pair of participants in which both people (participant B and C) recognised exactly the same sub-set of 15 faces, the 'similarity' is also 15.
- 3) and for a pair of participants (participant D and E) who recognised exactly the same sub-set of 7 faces, the 'similarity' is 7.

For participants A, B, and C, it is likely that readers intuitively think B and C are more similar but they are not that similar to A (without considering other conditions, e.g., number of faces both participants failed to correctly report). However, the index currently used would suggest A is similar to B and B is similar to C. Mismatch between intuition and what the index suggests also can be found between 2) and 3).

A reasonable index of face similarity here would be the correlations between binary variables (i.e., correct and incorrect responses by both participants). At least, correlations are closer to our understanding of similarity (e.g., two participants with good performance are similar; two participants with worse performance are also similar). If a different word or understanding of "similarity" is used in this analysis, authors may need to use a different terminology and develop a new index. But the currently used index does not seem to capture our intuitive understanding of similarity or the content authors would like to capture as described earlier.

Thank you both for explaining the issue. The problem is clearly the use of the word similarity. Rather than calculating similarity, our analysis calculates the overlap in correct face recognition. A higher score between participants represents the shared *accurate* recognition. This maps onto the narrative measure, which also measures the overlap in conceptual knowledge. In simple terms, greater overlap in semantic understanding (higher LSA score) predicts a greater overlap in the faces that are recognised.

Thanks for pointing this out and sorry for being slow on the uptake. We have changed the text and Figure 7 to make this clearer.

A related minor point is that it is needed to explain what the points denote in Figure 7. And sample sizes or degrees of freedom should be included in the correlation results.

Points denote an individual pairing of participants. We had chosen to removed degrees of freedom from the correlation results, as we used a permutation test to determine the significance. We felt that indicating degrees of freedom would be misleading as it would suggest that we had performed a standard correlation analysis.