Peer Community In Registered Reports

A registered test of the role of contextual information in perceptual learning of faces

A recommendation by ***Robert McIntosh*** *based on peer reviews by* ***Haiyang Jin*** *of the STAGE 2 REPORT:*

Kira N. Noad and Timothy J. Andrews (2024) The importance of conceptual knowledge when becoming familiar with faces during naturalistic viewing. OSF, ver. 4, peer-reviewed and recommended by Peer Community in Registered Reports. https://osf.io/thgrz

Submitted: 18 January 2024, Recommended: 21 May 2024

Cite this recommendation as:
McIntosh, R. (2024) A registered test of the role of contextual information in perceptual learning of faces. *Peer Community in Registered Reports*, 100669. 10.24072/pci.rr.100669

Published: 21 May 2024

When we familiarise with new faces over repeated exposures, it is generally in situations that have meaning for us. Seeing a face more often tends to go along with learning more about the person, and their likely contexts and actions. In this Registered Report, Noad and Andrews (2024) tested whether meaningful context during exposure improves the consolidation of faces into long-term memory. Participants were shown video clips from the TV series Life on Mars, either in their original chronological sequence, which provides meaningful context, or in a scrambled sequence. It was expected that the original sequence would provide a better conceptual understanding, and this was confirmed by free recall and structured question tests. Face recognition memory was tested with images of the actor from the original clips ('in show') and the same actor from another show ('out-of-show'), to test whether memory was modulated by the similarity of appearance to that at encoding. Face recognition was tested immediately after exposure and after four weeks, to allow time for consolidation. As expected, recognition memory was better for participants in the meaningful context condition, and for in-show faces. However, meaningful context did not lead to less forgetting of the faces at the follow up test, even for in-show faces, which did not support the original predictions. An exploratory analysis found that a metric of overlap between pairs of participants' conceptual understanding was related to overlap in the set of faces they recognised. This relationship was stronger after four weeks, which suggests increased interaction of conceptual knowledge and face recognition after consolidation. The Stage 2 manuscript was assessed over two rounds of review, and the recommender judged that the manuscript met the Stage 2 criteria for recommendation. **URL to the preregistered Stage 1 protocol:** https://osf.io/8wp6f

Level of bias control achieved: Level 6. *No part of the data or evidence that was used to answer the research question was generated until after IPA.* **List of eligible PCI RR-friendly journals:**

- Advances in Cognitive Psychology
- Collabra: Psychology
- Cortex
- Experimental Psychology
- F1000Research
- Journal of Cognition
- Peer Community Journal
- PeerJ
- Royal Society Open Science
- Studia Psychologica
- Swiss Psychology Open

References:

1. Noad, K. & Andrews, T. J. (2024). The importance of conceptual knowledge when becoming familiar with faces during naturalistic viewing [Stage 2]. Acceptance of Version 4 by Peer Community in Registered Reports. https://osf.io/thgrz

Reviews

Evaluation round #3

DOI or URL of the preprint: https://osf.io/pa8qc
Version of the preprint: Perceptual_Contextual_Reg_Report_PCIRR_S2_v3.docx

Authors' reply, 19 May 2024

We have replaced

Our prediction was that conceptual knowledge would affect face recognition after the information has been consolidated into memory.

with

Our prediction was that this contextual manipulation would affect face recognition after the information has been consolidated into memory.

Decision by Robert McIntosh, posted 18 May 2024, validated 18 May 2024

Very minor revision

Thank you for these thorough responses to the second round of review. The LSA is now much more clearly explained, and the shift in language from 'similarity' to 'overlap' is helpful, and liable to avoid misunderstandings. The exploratory analysis might be further questioned conceptually, but the important thing is it is now more clear exactly what you have done, and so readers will be able to make up their own minds about this aspect.

Before recommending this Stage 2 ms, I have one small change to request in the abstract. You say, "To determine if the context in which the movie was viewed had an effect on face recognition, we compared recognition in the original and scrambled condition. We found an overall effect of conceptual knowledge on face recognition." I think you should change 'conceptual knowledge' to 'context' here, to avoid the stronger implication of a causal pathway via conceptual knowledge (which is plausible, but not tested by your study).

I'd also like to take this chance to let you see the Stage 2 recommendation that I have drafted for your study. If you see anything that you think is incorrect and needs to be amended then please let me know.

Best wishes,

Rob

Draft of recommendation text

A registered test of the role of contextual information in aiding perceptual learning of faces

When we familiarise with new faces over repeated exposures, it is generally in situations that have meaning for us. Seeing a face more often tends to go along with learning more about the person, and their likely contexts and actions. In this Registered Report, Noad and Andrews (2023) test whether the provision of meaningful context during exposure improves the consolidation of faces into long-term memory. Participants were shown TV show video clips, ordered either in their original chronological sequence, which provides meaningful context, or in a scrambled sequence, which disrupts the meaning. It was expected that the original sequence would provide a better conceptual understanding of the narrative, and this was confirmed by free recall and structured question tests. Face recognition memory was tested with images of the actor from the original clips ('in show') and the same actor from another show ('out-of-show'), to test whether memory was modulated by the similarity of appearance to that at initial encoding. Face recognition memory was tested shortly after exposure and following a four-week delay, to allow time for consolidation. Recognition memory was better for participants in the meaningful context condition, and for in-show faces, as expected. However, meaningful context did not lead to less forgetting of the faces at the follow up test, even for in-show faces, which did not support the original predictions. An exploratory analysis suggested that a metric of overlap between pairs of participants' conceptual understanding of the clips was related to overlap in the set of faces they recognised, and this was relationship was stronger at the follow up test, which may suggest that the interaction of conceptual knowledge and face recognition changes with consolidation.

The Stage 2 manuscript was assessed over two rounds of review, and the recommender judged that the manuscript met the Stage 2 criteria for recommendation.

Evaluation round #2

DOI or URL of the preprint: https://osf.io/nqez9

Version of the preprint: Perceptual_Contextual_Reg_Report_PCIRR_S2_v2.docx

Authors' reply, 15 May 2024

Download author's reply

Download tracked changes file

Decision by Robert McIntosh, posted 22 April 2024, validated 22 April 2024

Invitation to revise Stage 2 RR

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.

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).

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.

Fig 5 legend typo: "Higher recognition ws evident"

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

Best wishes,

Rob McIntosh

PCI RR recommender

Reviewed by Haiyang Jin, 22 April 2024

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?

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.

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.

Evaluation round #1

DOI or URL of the preprint: https://osf.io/r7s8a?view_only=a59c340099d44a8db190a1b382b3b4d8
Version of the preprint: 1

Authors' reply, 27 March 2024

Download author's reply
Download tracked changes file

Decision by Robert McIntosh, posted 19 March 2024, validated 19 March 2024

Invitation to revise Stage 2 RR

Thank you for submitting your Stage 2 report for this study (and congratulations on completing the experiment).

I apologise that it has taken a while to get this decision letter to you. I have had one review available, from Haiyang Jin, for some time, but have been unable to elicit a response from the other Stage 1 reviewer. Given the time that has elapsed, I have decided to act as the second reviewer myself, and have now read the Stage 2 manuscript closely.

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).

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.

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."

"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."

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.

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.

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.

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.

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.

Reviewed by Haiyang Jin, 03 February 2024

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.

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.

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.

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)

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).

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.

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).")

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

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

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