Review of Ali et al – PCI RR

Ali and colleagues plan to test the hypothesis that judgments of perceptual similarity reflect a metacognitive awareness of one's own perceptual capacities. They propose to test this by comparing perceptual judgments of similarity for a set of faces with threshold measurements of perceptual discriminability between pairs of faces (using a morphed continuum).

There are two key hypotheses: 1) clear association between similarity judgments and perceptual discriminability, and 2) individual differences such that the association is stronger within than between participants.

This is a well-motivated proposal and the rationale for the work is clearly and comprehensively laid out. The experiments have been thoughtfully designed and the pilot data demonstrates feasibility and provides preliminary results that are supportive of the hypotheses.

Overall, this is a very solid proposal, but I have a few comments/suggestions/questions that the investigators might want to consider:

1) The nature of the subjective similarity task with multiple target faces presented below the sample face, means that participants are not just comparing the sample with one target, but considering all faces simultaneously. Might this introduce strong context effects and would it be better to present triplets to minimize such effects? I assume that part of the rationale is to speed up data collection, but perhaps this also leads to less stable data?
2) The investigators propose running 12 participants with four sessions per participant (2 similarity judgment, 2 threshold discrimination) with, for example, 24 pairs of faces for the perceptual discrimination tasks. The rationale for all these numbers is partly based on the pilot data, but the numbers seem arbitrary.

   Lines 145-147 – "Each participant performs four sessions on different days with each session taking more than 60 minutes. This provides us with enough data to perform our statistical analysis at the individual-level."

   Lines 262-264 – "Our decision to select 24 pairs is supported by our pilot study, as we achieved reasonably robust results by examining only 13 pairs, almost half of our planned 24 pairs."

   The basis for these statements is not clear. I was wondering if the authors could use the pilot data to run simulations to estimate how much data they actually need, both for each participant to reliably estimate their performance on each task and at the group level to estimate the relationship between performance on the two tasks. This would help increase confidence in the proposed plan and potentially avoid

collecting too little or too much data. I'm a little bit concerned about the latter and the burden currently placed on each participant – overly taxing the participants could actually lead to less reliable data.

3) Do the investigators have any sense of how stable/reliable the similarity judgments and perceptual discrimination judgments are? What is the test/retest reliability across days? In the context of the similarity ratings, they suggest that "subjective similarity ratings may be made based on whatever visual features that happen to be more salient, depending on one's fluctuating attentional states, or arbitrary preferences that aren't necessarily related to one's own performance in near-threshold psychophysical tasks." (Lines 71-74). To the extent that performance on the different tasks fluctuates, combining sessions across days may be worth reconsidering.
4) Lines 197-198 – can the investigators give some intuitive sense of how the trials are selected based on the embeddings.
5) I like the idea of using precision as the basis for the stopping criterion, but what is the rationale for choosing <1 as the desired 95% confidence interval? Might it be worth setting an upper limit for the number of participants that will potentially be recruited in case the precision does not converge as the investigators anticipate?