- We have responded to the reviewers' comments in this document.
- In the revised version of the manuscript, changes are highlighted in red with track changes.

Recommender:

Dear authors

Your Stage 1 RR manuscript has now been reviewed by two experts in the field. While they are general enthusiastic about your proposed research, they raised several points I'd like you to consider. I also included a few comments of my own:

1) Reviewer Haiyang Jin raises questions about the use of the term 'metacognitive' in the context of this study. While I agree with you that you are investigating metacognition here, he is right that it would be worth spelling out clearly why this is about metacognition rather than simply comparing objective thresholds with subjective ratings.

We further elaborated on why the study is about metacognition in the "introduction" section in lines 71-73: "Metacognition is commonly defined as the monitoring (and control) of one's own cognitive abilities. In the present study, we hypothesize that similarity judgments involve a type of implicit metacognition. When we make a similarity judgment, it reflects our own perceptual capacities."

Some descriptions can also be found in lines 59-70 and 102-104: " the idea is that such judgment would be made on a dimension in which all relevant features are optimally combined, such that along this dimension, the two faces are maximally distinguishable. Specifically, for this combination to be optimal, the choice of this dimension should be based on how perceptible each feature is to oneself. In other words, this process is not only about the physical stimulus itself, but rather, it reflects (implicit) metacognitive knowledge of one's own perceptual abilities."

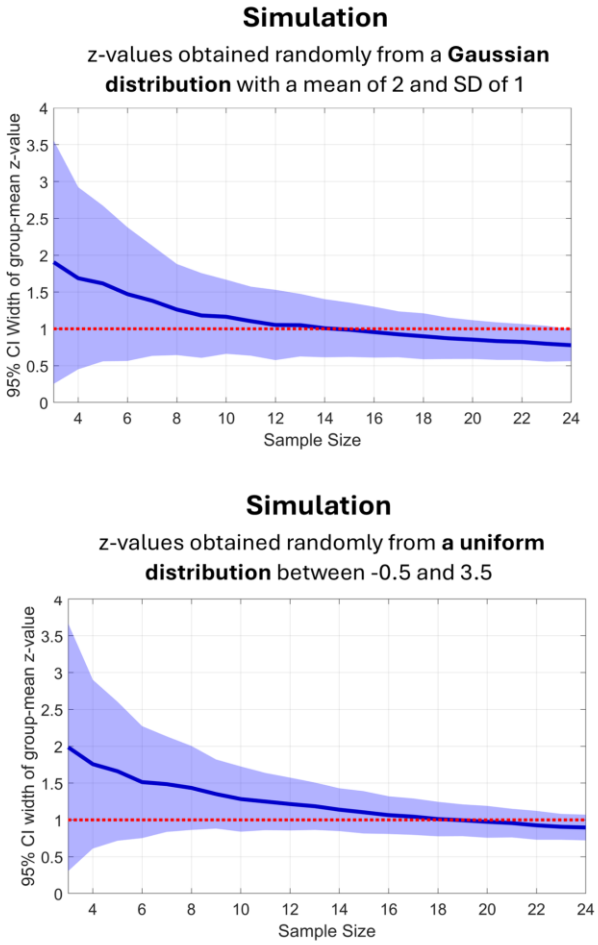
"subjective perceptual similarity may be metacognitive in nature, meaning that it concerns one's own perceptual capacities, not just the general physical similarity between stimuli."

2) It is not clear whether excluded participants will be replaced. For your stopping criterion based on confidence intervals, this doesn't technically matter. But for your minimum sample size of 12 it is important to clarify. I assume that's what you mean but please state this explicitly.

The excluded participants will be replaced. We clarified this in the "sampling plan" section in lines 347-349: "The data collection continues until we have 12 participants who successfully complete the experiment without meeting any of the exclusion criteria. If a participant meets the exclusion criteria, a new participant is recruited to replace the excluded participant".

3) Note that both reviewers raise the issue that you did not specify a maximum sample size. Please either do so or at least show (e.g. through simulations) that you can feasibly obtain your stopping criterion with a realistic sample size. It is imortant to minimise the risk of inconclusive results.

Judging from our pilot data, it is likely that the stopping criterion will be satisfied with our initial 12 participants. The width of the 95% CI was around 1.5 in our pilot data with only four participants (Figure 2B & 2D), and the stopping criterion will be met if we reach a 95% CI with a width less than 1. However, we agree with the reviewers that we need to define a maximum sample size to ensure the experiment ends. After our initial 12 participants, we recruit three more participants, each time failing to meet the stopping criterion. We will stop the recruitment when reaching a maximum of 24 participants, even if the stopping criterion is not met. A maximum sample size of 24 seems reasonable, considering our resources and the high likelihood of meeting the stopping criterion by that sample size that is justified by a simulation shown below. We clarified this in the "sampling plan" section in lines 370-374.



Supplementary Figure 2. Simulations to determine a feasible sample size for our stopping criterion.
We ran two simulations: one generated z-values from a Gaussian distribution with a mean of 2 and SD of 1 and another from a uniform distribution ranging between -0.5 and 3.5. These example distributions seem reasonable given our expectations based on our pilot data and seem conservative enough. For example, the 95% CI width in our pilot data with four participants was around 1.5, however, in the presented simulations, the 95% CI width is, on average, around 1.7 for the same sample size of four. The shaded area indicates the 2.5th and 97.5th percentile of the 95% CI width obtained over 1000 simulations. Assuming the used distributions are realistic, there is a high likelihood of hitting the stopping criterion by reaching a maximum sample size of 24.

4) You are determining the 95% CI via bootstrapping. I assume that this is done by using the 2.5th and 97.5th percentile. Please clarify, because there are various adjustment methods available. As the exact limits here will affect your stopping decision and statistical inference it is critical to ensure there is no flexibility.

We thank the recommender for pointing this out. We define the 95% CI as the 2.5th to 97.5th percentile of the distribution derived by bootstrapping. We clarified this in the "Analysis plan" section in lines 384-385.

5) Perhaps I'm misunderstanding something, but in several places you state than higher #JND corresponds to "higher capacity," i.e., better performance. This sounds incorrect. Your staircase converges on ~71% correct threshold. Wouldn't fewer morphing steps therefore correspond to better performance (i.e. higher capacity)?

The converged morphing step (i.e., 71% correct threshold) indicates the just-noticeable difference (JND); thus, as the recommender stated, a smaller JND corresponds to a better performance and a higher capacity. However, we use "the number of JNDs (#JNDs)" in our analysis, which is defined as the number of all morphs (1000) divided by the obtained JND. For example, if the JND is 200, there would be five JNDs (i.e., 1000/200) between the two faces. Essentially, #JNDs indicates the perceptual distance between a face pair in a subject, thus, its higher value corresponds to a higher discrimination capacity. We use the notion #JNDs because it is free from the non-standard arbitrary unit of the total morph steps (e.g., if we use a different large number of total morph steps instead of 1000, the value of #JNDs would remain unchanged). An explanation of this can be found in the legend of Figure 1 (lines 123-135). We further elaborated on this in our "method section" under the "Near-threshold discrimination task" subsection in lines 287-291.

6) Design Table, Hypothesis 1, Sampling plan: When you write "execution" criteria, I assume you mean "exclusion"?

We appreciate the recommender's careful reading for noticing this typo error. The exclusion is correct. We corrected this error.

7) A *Discussion* section isn't required in a Stage 1 manuscript. It is fine to include one and it will be acceptable if this is changed completely at Stage 2 if the results suggest different interpretations at that stage. However, for simplicity and to avoid reviewer time being spent on a part of text that isn't needed and I'd suggest removing it.

We agree with the recommender that the discussion section is not necessary at this stage. However, since the reviewers did not leave any comment on it, we prefer keeping it. The manuscript is publicly available in BioRxiv, so the discussion section could be beneficial for readers.

Reviewer 1

We appreciate the reviewer's careful reading of our manuscript and are thankful for the valuable comments. Below, we address each of the reviewer's points.

1) Although the main research question is about "metacognitive", the manuscript surprisingly does not seem to explain what "metacognitive" means in the manuscript or the measures used in this study do not seem to tap into the popular understanding of "metacognitive". To my knowledge (I'm not an expert in metacognition), "metacognition of face ability" refers to whether the participant know how good is his/her face recognition ability. For example, both a person with bad face recognition ability knowing his/her recognition ability being bad and a person with good face recognition ability knowing his/her recognition ability being good have high metacognition in face recognition ability. But this does not seem to be measured by the tests in this manuscript. As such, the meaning of "metacognitive" in the manuscript (and its relationship with the potential understanding above) should be explained and clarified further.

The reviewer's understanding of metacognition is correct. Metacognition is commonly defined as the monitoring (and control) of one's own cognitive abilities. In this study, we have the same definition of metacognition, but the metacognitive access is implicit. We hypothesize that when we make a similarity judgment, it reflects our own perceptual capacities. So, our similarity judgment does not just concern the physical properties of the stimulus but an awareness of our capabilities in perceiving those properties which makes it metacognitive. Some explanations can be found in the "introduction" section in lines 59-70 and 102-104, and we further elaborated on it in lines 71-73.

2) Throughout the manuscript, "perceptual similarity" is emphasized as subjective (e.g., the term "subjective perceptual similarity"), whereas "discriminability ability" seems to be treated as objective/"quasi-objective". But both "perceptual similarity" and "discriminability ability" were reported/responded by participants subjectively. Thus, it remains unclear why there is such differences (subjective vs. objective) between "perceptual similarity" and "discriminability ability".

We refer to perceptual similarity as subjective since participants are not given specific criteria in ranking face similarity and they can base their judgments on any salient combination of facial features that stands out to them. There is no correct or wrong answer. Participants freely rank the face similarity. Generally, similarity judgments are considered subjective. For example, a child may look more like their father or mother in different people's eyes. There is no correct answer. However, in the discrimination task, there is a correct answer. The task aims to measure near-threshold discrimination capacity. There is no subjectivity in the answer; it is about whether the participant is capable of distinguishing the different face in a trial. We emphasized this in the "introduction" section in lines 46-47 and 80-82.

3) It is highly appreciated that the introduction discusses the potential alternative hypotheses, which to some extent addresses my concerns what other hypotheses may account for the correlations between perceptual similarity and ability judgements. However, (1) it remains elusive

whether these alternative hypotheses were mutual exclusive to the main hypotheses; (2) if not, it is unclear why these alternative hypotheses were not tested in the manuscript; (3) a potential relating issue is that the alternative hypotheses are too vague to test in practice. Since this is a registered report, there is possibility that the main hypothesis would not be supported (see later for potential issues in employing statistical evidence disconfirming the main hypotheses). In this case, it should be clarified what we can conclude from the findings (and the potential specific results).

We outlined a possible interpretation when our hypotheses are not supported by the data in Table 1, under the column "Interpretation given to different outcomes," as well as in the "introduction" section in lines 74-79. For example, if Hypothesis I does not reach the level of significance, we reject that perceptual similarity judgment is aligned with underlying psychophysical capacities. So, a plausible alternative interpretation (hypothesis) could be that similarity judgment is made based on an arbitrary subjective preference or fluctuations of focus on different visual features. It may not be within the scope of this study to formally test the truth of the alternative hypothesis as we may be required to design a new set of tasks for it. Basically, in the present study, we can merely suggest alternative interpretations for the data when our hypotheses are rejected, without formally testing them.

Further, we note that our primary hypothesis testing follows a frequentist approach. We confirm the hypothesis by rejecting the null hypothesis. However, our complementary tests also include a Bayesian measure that can shed light on how conclusively the evidence is in favor of the null hypothesis. Please see our response to the reviewer's sixth comment.

4) The analysis of Hypothesis 1 seems to be related to Simpson paradox, and therefore, it is necessary to explain how the obtained group-level correlation results are different from the correlations between two tasks. For instance, when we talked about correlations between two tasks, the popular understanding is that participants completed two tasks and their performance in two tasks were correlated among participants. But this is different from the group-level correlation calculated in the manuscript, which should be clarified.

It seems there is a misunderstanding, and we apologize if our explanation of the method was unclear. As far as we know, the Simpson paradox may possibly occur if we concatenate all participants' data and compute one group-level correlation. However, we don't do this in our analysis. In both hypotheses, we transform the individual-level statistic (e.g., in Hypothesis I, the correlation between the dissimilarity value and #JNDs in a participant) to a standard z-score. Then we test whether the group mean z-value is significantly above zero (lines 381-385 and 414-418). This is a common approach in the literature that tests the existence of a trend across participants. We also note that besides the group-level stats, we report the individual-level significance, as shown in Figure 2.

5) The test for the second hypothesis seems biased. To test the second hypothesis, only selected face pairs with large standard deviations (SDs) were included. Indeed, since in these trials different participants more likely have different responses for the same face pairs, using face pairs

with large SDs have higher probability to provide evidence support the second hypotheses. However, on the other hand, face pairs with smaller SDs are more likely to provide evidence that different participants made the same or similar responses for the same face pairs, which does not seem to provide support for the second hypotheses. Instead, it may suggest that the perceptual similarity responses are not unique to individuals. But the method considers more trials with large SDs, which is not appropriate. Maybe trials with smaller SDs can also be used to test the potential alternative hypotheses.

We appreciate the reviewer's comment, and we would like to clarify our approach. The pairs with high SDs could be considered more reliable samples for examining the hypothesis. Our measurements (i.e., the measured dissimilarity value and #JNDs) inherently have some level of noise. Consequently, the statistic for Hypothesis II obtained from pairs with smaller SDs is more affected by this noise and is thus less reliable than those from pairs with high SDs. Therefore, using face pairs with large SDs should not introduce a bias supporting Hypothesis II; rather, it makes the results more reliable. We note we include a few pairs with smaller SD in our study just as a form of control. If the hypothesis is true, we would like to see to what extent it is observable in these pairs with small SDs. We discussed these points in the "Selection of the pairs for the near-threshold discrimination task" section in lines 320-328 and in the "pilot data" section in lines 463-467.

To further clarify the above point, let's assume that the estimated dissimilarity values have an associated noise level of 0.1. When a difference between two dissimilarity values is 0.05, the direction of the difference can be trusted much less than when a difference is 0.4. Thus, a greater difference score (higher SD) gives us higher confidence that the observed difference in dissimilarity value is real.

6) It is great that the manuscript included the sample size planning section. But some key information is missing, and some procedures do not make sense in practice. First, as no statistical power analysis is conducted, the proposed sample size (i.e., 12) and the procedure of adding more participants does not guarantee sufficient statistical power. Although precision of CI is used as the stopping rule, it remains unclear why CI size of 1 is used. By assuming CI of 1 is sufficient somehow, there is not an upper limit of the sample size in the procedure, which brings the risk that the study would never end.

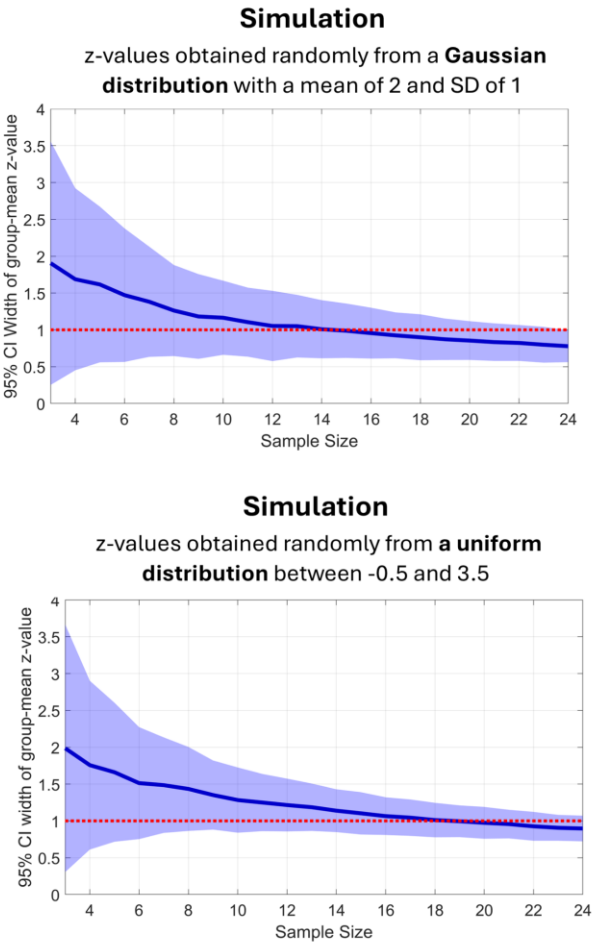
We agree with the reviewer that it would be beneficial to define a maximum sample size to ensure the experiment ends. Please see the "sampling plan" section lines 370-374: "We note that after our initial 12 participants, we recruit three more participants, each time the stopping criterion is not met. We repeat this until reaching a maximum of 24 participants. Given that our pilot data with only four participants show a 95% confidence interval with a width of around 1.5 (see Figure 2B & 2D), it is unlikely not meeting the stopping criterion before reaching our maximum sample size of 24 (see the supplementary Figure 2)."

We designed the experiment to enable enough power for statistical analysis at both the individual and group levels. At the individual level, for Hypothesis I, correlation is performed on

24 samples. This sample size seems fair for detecting the effect, especially considering that in our pilot study with 13 samples, the correlation was fairly reliable and was significant in two out of four participants (Figure 2A & 2B) (please see lines 311-314). Regarding Hypothesis II, having 12 participants and 24 pairs ensures constructing a reliable null distribution using the permutation test (i.e., 11^24 possible permutations). Please see the "Analysis plan" section, lines 406-408. 12 participants also seem sufficient for performing statistical analysis at the group level, considering that all the group-level tests were significant even in our pilot study with four participants (Figure 2B & 2D). Further, we note that given our stopping criterion, we may end up recruiting more than 12 participants.

Given our sample size scale, we expect a considerable effect to have a group-mean z-value of at least above 0.5. So, a minimally significant scenario involves a group-mean z-value of 0.5 with a 95% confidence interval width of less than 1. Considering this, we set our stopping criterion as the width of the 95% confidence interval being smaller than 1. Thus, this seems like enough precision to safely reject or accept a hypothesis. We added this explanation in the "sampling plan" section in lines 365-369.

We also ran the simulation below to demonstrate the feasibility of our proposed sample size in meeting the stopping criterion. The simulation properties were set given our expectations based on the pilot data.



Supplementary Figure 2. Simulations to determine a feasible sample size for our stopping criterion.

We ran two simulations: one generated z-values from a Gaussian distribution with a mean of 2 and SD of 1 and another from a uniform distribution ranging between -0.5 and 3.5. These example distributions seem reasonable given our expectations based on our pilot data and seem conservative enough. For example, the 95% CI width in our pilot data with four participants was around 1.5, however, in the presented simulations, the 95% CI width is, on average, around 1.7 for the same sample size of four. The shaded area indicates the 2.5th and 97.5th percentile of the 95% CI width obtained over 1000 simulations. Assuming the used distributions are realistic, there is a high likelihood of hitting the stopping criterion by reaching a maximum sample size of 24.

Second, with the stopping rule of CI size of 1, a non-significant result could be obtained; in this case, it remains unclear what conclusions could be drawn. As a registered report, tests for supporting null hypotheses also should be included.

Our hypothesis is confirmed if the 95% CI of the group-mean z-value is above 0. We are using a frequentist approach to hypothesis testing. Thus, if the 95% CI includes 0, given our high statistical power, such a result would strongly question the validity of the hypothesis.

With our main frequentist approach, it is not possible to directly evaluate the significance of the null result as one can with the Bayesian approach. However, since our stopping criterion is the 95% CI width of smaller than 1, it should give us enough precision to safely reject or accept a hypothesis (please see our response to the reviewer's above comment). Therefore, in case of rejecting a hypothesis, the evidence is very weak and most likely is also in favor of the null hypothesis. We also note that within the complementary statistics that we plan to report (stated in lines 386-388), the Bayes factor (BF) of the t-test could be used to shed more light on how conclusively the evidence is in favor of the null hypothesis.

Third, the procedure of adding participants conflicted with the experiment procedures. It was introduced that all participants will complete task 2 only after all participants complete task 1, and Hypothesis 1 can only be tested when both task 1 and 2 were completed by all participants. But if the criteria were not met (e.g., CI for H1 is larger than 1), more participants would be added. 1) it remains unclear how many participants will be recruited additionally before the next round of analysis. Only 1 or more? 2) Since more participants would be recruited, it is possible that the face pairs for task 2 would change. Then will the first 12 participants re-complete the task 2? It remains elusive what the specific procedures would be.

We appreciate the reviewer's careful reading of our proposed plan. After our initial 12 participants, we recruit three more participants each time the stopping criterion is not met, up to a maximum of 24 participants. We clarified this point in lines 370-371.

We select the 24 pairs for task 2 after confirming the data quality of our 12 participants in task 1, as not meeting the exclusion criterion (please see lines 343-346). After this point, the 24 pairs are fixed and do not change with any recruitment of new participants. We explained this point in lines 353-354 and 374-377: “It is also worth noting that the recruitment of new participants does not alter the pairs used in the near-threshold discrimination task. The newly recruited participants perform the task on the same pairs selected based on our initial 12 participants.”

Fourth, it is unclear whether and how 95% HDI would be used for hypothesis testing. For instance, what conclusions could be drawn if the 95% HDI includes 0? Also, what is the prior for calculating Bayes factor?

These statistics are complementary to provide a more comprehensive view of the significance level. We don't use them to confirm or reject a hypothesis. 95% HDPI indicates the prevalence of observing the effect at the individual-level (please see lines 391-394). This would complement the evaluation of significance, as, for example, the effect can turn out to be significant at the group-level (i.e., the trend exists in the majority of participants), with only a few participants showing the effect significantly at the individual level.

To calculate the Bayes factor of the ttest, we use a publicly shared Matlab package that uses a Cauchy prior (reference was added in lines 387-388).

Minor-points:

1) Since the correlation between the measures is the main interest, and both tests were conducted on two separate days for each participant, the reliability of each task should be reported.

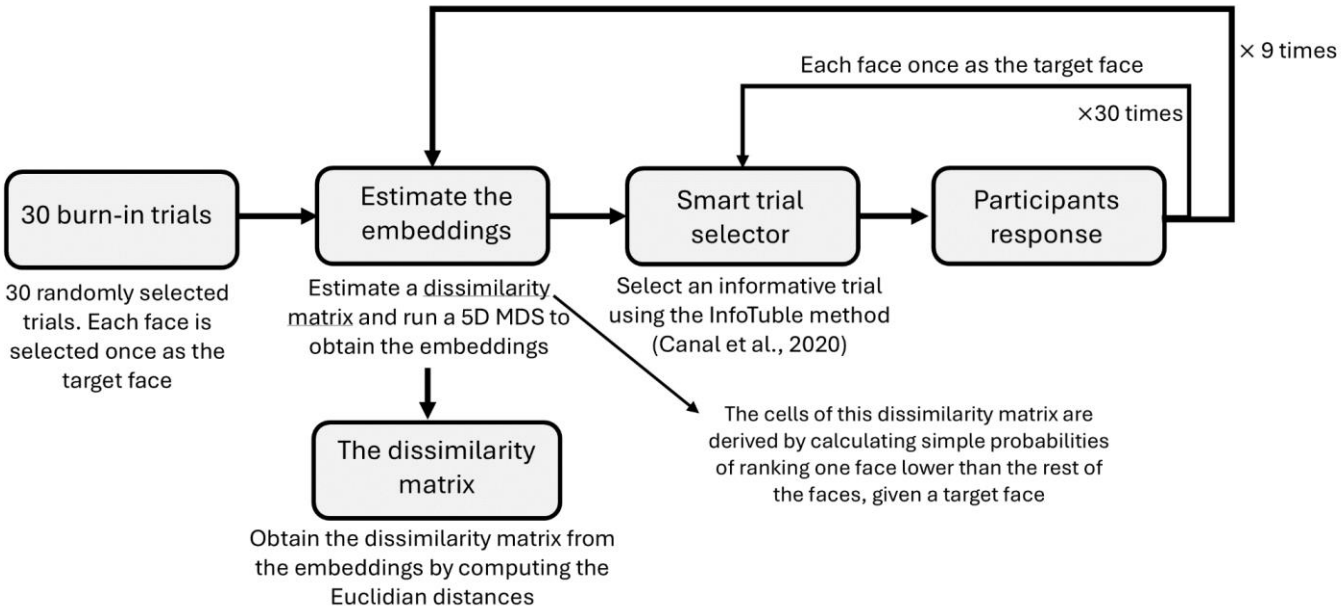
The reliability of the dissimilarity matrix can be evaluated by comparing the matrix obtained from the first and second sessions of task 1. We agree with the reviewer on the importance of reporting this information. So, we report the correlation between the dissimilarity matrices from the first and second sessions to indicate within-participant reliability. As a reference, we also provide the distribution of between-participant correlation by randomly correlating a matrix from a session in one participant with that of another participant. It is

expected for the within-participant correlation to be higher than the between-participant correlation. We added this description in the "method" section in lines 240-245.

However, for the #JNDs, we only have one run of the staircase for each face pair. Therefore, we cannot conduct a reliability check like the above. Task 2 involves examining 24 pairs, with participants completing the task over the course of two sessions. The 24 pairs are randomly split between the sessions, with participants completing 12 of the pairs in a session (please see lines 148-150, and 293-294). However, we note that the fairly large number of trials (60 trials) included in each staircase helps obtain reliable #JNDs estimates. In addition, the trajectory convergence of a staircase could indicate the reliability of the estimated #JNDs. A staircase with a higher ratio of reversals in its later trials could be considered more reliable. Therefore, we report the ratio of reversals in the last 20 trials of each pair's staircase and its statistics across participants. In an absolute ideal case, given our 1-up and 2-down staircase protocol, the ratio of reversals in the last 20 trials would be 0.6. We added this description in the "method" section in lines 302-306. Further, note that we exclude participants who show a lack of attentiveness to the task in the first place, as we stated in lines 339-342: "a session in which there are more than 4 (out of 12) staircases with less than three downs in their last 20 trials is considered bad with lacking sufficient attentiveness." Achieving more than three downs in the last 20 trials by chance without attention in our task is around 13.6% (confirmed with both mathematical calculations and computer simulations).

2) The steps to get the dissimilarity matrix seem to be quite complicated, and maybe a figure together with the text explanation would help. (I made all my comments with the assumptions that all the steps to get the dissimilarity matrix is appropriate.)

We incorporated a similar approach to (Canal et al., 2020) as we described in lines 196-197. We also explained the step-by-step procedure in lines 182-224. However, to further help the readers understand the method, we made a supplementary figure as the reviewer suggested.



Supplementary Figure 1. Schematic of the subjective similarity judgment task design

For the reviewer's information, we note that in the revised version, we added an alternative method for obtaining the embeddings and the dissimilarity matrix, which is based on a machine learning algorithm that can potentially work better. We plan to test this method in our experiment as an additional exploratory analysis (i.e., we don't use this approach in our main hypotheses testing and the stopping criterion). We explained this approach in the "method" section in lines 246-267:

"In addition to the above approach for deriving the embeddings and the dissimilarity matrix, we plan to try a machine learning approach as an exploratory analysis (i.e., we don't use this method in our main hypotheses testing and the stopping criterion). This approach starts with random embeddings and iteratively updates them to minimize a loss function, which penalizes wrong similarity rankings derived from the embeddings. The loss is constructed using a sigmoid activation function in a binary cross-entropy as follows:

$$p = \frac{1}{1+e^{-k(d_{dissim}-d_{sim})}},\ d_{sim} = \left\|x_{target} - x_{sim}\right\|_2,\ d_{dissim} = \left\|x_{target} - x_{dissim}\right\|_2$$

$$L = -\frac{1}{N}\sum_{i=1}^{N} log(p_i) + \lambda_1 \sum_{m=1}^{M} \|x_m\|_1 + \lambda_2 \sum_{m=1}^{M} \|x_m\|_1{}^2$$

Where x_m represents the vector embedding of face m; d_{sim} and d_{dissim} are the Euclidean distances between a target face and a face ranked as more similar and a face ranked as less similar to the target face by the participant, respectively; k corresponds to the ranking difference (e.g., 2 for a face ranked first and a face ranked third), putting more emphasis on clearer similarity

comparisons; N indicates the number of segmented trio comparisons (with six trio segments in a trial); λ_1 and λ_2 are the hyperparameters of L1 and L2 regularizations which help to control the sparsity and scale of the embeddings. We use the Keras library in Python, with Adam optimizer, to minimize the loss function.

We note that we don't use this machine learning approach during the task (i.e., in our online application) because it is slower, requiring cross-validations and careful selection of the hyperparameters. This approach is more sophisticated than our main approach, which involves estimating probabilities and running MDS, but it has the potential to yield a better estimation of the embeddings and the dissimilarity matrix. In our pilot study, using this approach, we got similar results to those shown in Figures 2 and 3."

Reviewer 2

We are grateful for the reviewer's careful reading of our manuscript and for their valuable comments. Below, we address each of the reviewer's points.

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?

Ranking multiple faces in a trial drastically speeds up the data collection, allowing us to obtain the dissimilarity matrix much faster. However, we are uncertain whether the impact of such context effect is significant on the stability of the data in our case. The stimuli shown in a trial are all faces, and they are diverse enough not to cause a significant context effect. Thus, we believe the pros and cons of using multiple stimuli outweigh the potential advantages of using triplets.

We also note that Canal et al., 2020, have previously demonstrated the significant benefit of a design involving ranking multiple stimuli using simulation and real experiments.

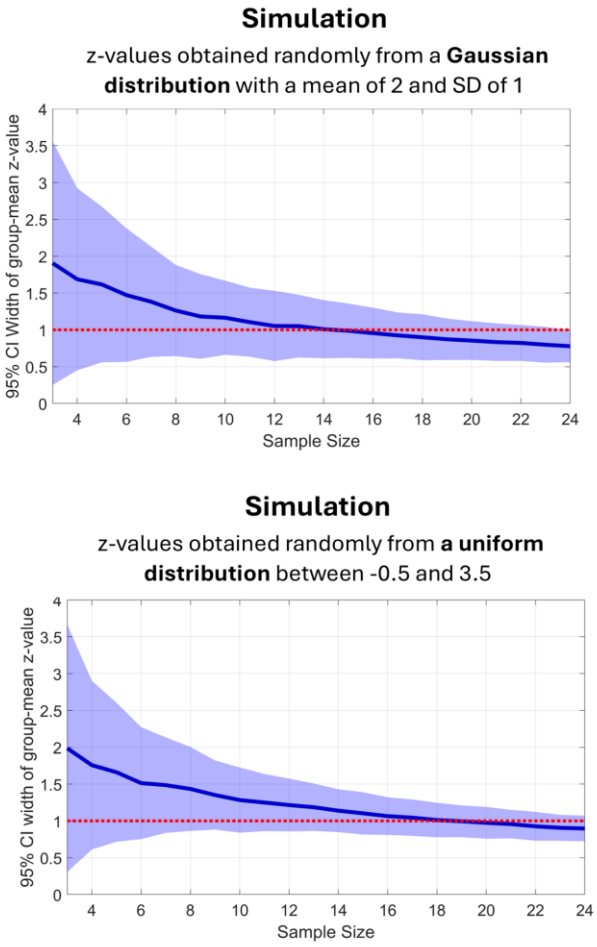
Canal G, Fenu S, Rozell C (2020) Active ordinal querying for tuplewise similarity learning. In:, pp. 3332–3340.

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.

Our pilot study, involving only four participants (each attending two sessions) and testing only 13 pairs, yielded highly significant results (Figure 2). In our main study, we plan to have at least three times more participants (i.e., 12), with each attending double the number of sessions (i.e., four sessions) with testing almost twice as many pairs (i.e., 24 pairs). Further, we have a stopping criterion, so we may end up recruiting more than 12 participants. We added a further

explanation on why we have chosen these numbers in the manuscript in lines 311-314 and 406-408: “A sample size of 24 pairs should be fairly adequate for detecting an effect, and it is further justified by our pilot study, as we achieved reasonably robust results by examining only 13 pairs, almost half of our planned 24 pairs.” “with 12 participants and 24 pairs, there are an enormous number of possible permutations (i.e., 11^24 unique permutations), which ensure constructing a reliable null distribution”

Given the significance of our pilot study, it feels intuitive that the chosen sample sizes should be more than sufficient. However, we also conducted the simulation below to demonstrate the feasibility of our proposed sample size in meeting the stopping criterion. The simulation properties were set given our expectations based on the pilot data.



Supplementary Figure 2. Simulations to determine a feasible sample size for our stopping criterion.
We ran two simulations: one generated z-values from a Gaussian distribution with a mean of 2 and SD of 1 and another from a uniform distribution ranging between -0.5 and 3.5. These example distributions seem reasonable given our expectations based on our pilot data and seem conservative enough. For example, the 95% CI width in our pilot data with four participants was around 1.5, however, in the presented simulations, the 95% CI width is, on average, around 1.7 for the same sample size of four. The shaded area indicates the 2.5th and 97.5th percentile of the 95% CI width obtained over 1000 simulations. Assuming the used distributions are realistic, there is a high likelihood of hitting the stopping criterion by reaching a maximum sample size of 24.

Last, It is worth noting that we include multiple rest periods during a session to prevent participants from becoming too fatigued and, thereby, performing less reliably (lines 215-217 and 30-301).

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.

We agree with the reviewer that it is important to evaluate the reliability of the measures. Participants perform the similarity judgment task twice in two different sessions, which allows us to evaluate its reliability. We report the within-participant correlation between the dissimilarity matrices derived from each session. As a reference, we also report the distribution of between-participant correlation by randomly correlating a matrix from a session in one participant with that of another participant. We expect that the within-participant correlation to be higher than the between-participant correlation. We added this description in the "method" section in lines 240-245. We point out that our statement quoted by the reviewer represents a possible interpretation in case our hypothesis is not supported by the data. The reliability check above could help to evaluate this interpretation as the reviewer noted.

As for the reliability of the estimated #JNDs, since participants perform the near-threshold discrimination task (staircase) only once per face pair, we cannot obtain a test-retest measure like the above. However, we note that the fairly large number of trials (60 trials) included in each staircase helps obtain reliable #JNDs estimates. In addition, the trajectory convergence of a staircase could indicate the reliability of the estimated #JNDs. A staircase with a higher ratio of reversals in its later trials could be considered more reliable. Therefore, we report the ratio of reversals in the last 20 trials of each pair's staircase and its statistics across participants. In an absolute ideal case, given our 1-up and 2-down staircase protocol, the ratio of reversals in the last 20 trials would be 0.6. We added this description in the "method" section in lines 302-306.

4) Lines 197-198 – can the investigators give some intuitive sense of how the trials are selected based on the embeddings.

We utilized the InfoTuple method code, implemented and shared by Canal et al., 2020. The InfoTuple method selects the tuple that maximizes a mutual information estimate which involves two entropy terms: intuitively, one term favors tuples whose rankings are uncertain given the current embeddings, while the other discourages inherently ambiguate tuples that are expected to remain uncertain even if the embeddings are revealed. So, it aims to select an informative tuple whose rankings are unknown but yet can be answered reliably (consistently). We added this description in the "method" section in lines 204-210. The calculation of the mutual information involves complex mathematics, which is detailed in the original study.

Canal G, Fenu S, Rozell C (2020) Active ordinal querying for tuplewise similarity learning. In:, pp. 3332–3340.

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?

We agree with the reviewer that it is worth setting an upper limit. A maximum sample size of 24 seems reasonable, considering our resources and the high likelihood of meeting the stopping criterion by that sample size (please see the simulation provided in our response to the second comment of the reviewer). We added this information in the "Sampling plan" section in lines 370-374: "We note that after our initial 12 participants, we recruit three more participants, each time the stopping criterion is not met. We repeat this until reaching a maximum of 24 participants. Given that our pilot data with only four participants show a 95% confidence interval with a width of around 1.5 (see Figure 2B & 2D), it is unlikely not meeting the stopping criterion before reaching our maximum sample size of 24 (see the supplementary Figure 2)."

Given our sample size scale, we expect a considerable effect to have a group-mean z-value of at least above 0.5. So, a minimally significant scenario involves a group-mean z-value of 0.5 with a 95% confidence interval width of less than 1. Considering this, we set our stopping criterion as the width of the 95% confidence interval being smaller than 1. Thus, this seems like enough precision to safely reject or accept a hypothesis. We added this explanation in the "Sampling plan" section in lines 365-369.

For the reviewer's information, we note that in the revised version, we added an alternative method for obtaining the embeddings and the dissimilarity matrix, which is based on a machine learning algorithm that can potentially work better. We plan to test this method in our experiment as an additional exploratory analysis (i.e., we don't use this approach in our main hypotheses testing and the stopping criterion). We explained this approach in the "method" section in lines 246-267:

"In addition to the above approach for deriving the embeddings and the dissimilarity matrix, we plan to try a machine learning approach as an exploratory analysis (i.e., we don't use this method in our main hypotheses testing and the stopping criterion). This approach starts with random embeddings and iteratively updates them to minimize a loss function, which penalizes wrong similarity rankings derived from the embeddings. The loss is constructed using a sigmoid activation function in a binary cross-entropy as follows:

$$p = \frac{1}{1+e^{-k(d_{dissim}-d_{sim})}},\ d_{sim} = \left\|x_{target} - x_{sim}\right\|_2,\ d_{dissim} = \left\|x_{target} - x_{dissim}\right\|_2$$

$$L = -\frac{1}{N}\sum_{i=1}^{N} log(p_i) + \lambda_1 \sum_{m=1}^{M} \|x_m\|_1 + \lambda_2 \sum_{m=1}^{M} \|x_m\|_1^{\ 2}$$

Where x_m represents the vector embedding of face m; d_{sim} and d_{dissim} are the Euclidean distances between a target face and a face ranked as more similar and a face ranked as less

similar to the target face by the participant, respectively; k corresponds to the ranking difference (e.g., 2 for a face ranked first and a face ranked third), putting more emphasis on clearer similarity comparisons; N indicates the number of segmented trio comparisons (with six trio segments in a trial); λ_1 and λ_2 are the hyperparameters of L1 and L2 regularizations which help to control the sparsity and scale of the embeddings. We use the Keras library in Python, with Adam optimizer, to minimize the loss function.

We note that we don't use this machine learning approach during the task (i.e., in our online application) because it is slower, requiring cross-validations and careful selection of the hyperparameters. This approach is more sophisticated than our main approach, which involves estimating probabilities and running MDS, but it has the potential to yield a better estimation of the embeddings and the dissimilarity matrix. In our pilot study, using this approach, we got similar results to those shown in Figures 2 and 3."