Is subjective perceptual similarity metacognitive?

Moharramipour Ali[1]*, Zhou William[1], Rahnev Dobromir[2], Lau Hakwan[1]*

[1] Center for Brain Science (CBS), RIKEN, Wako, Japan
[2] School of Psychology, Georgia Institute of Technology, Atlanta, United States

* Corresponding authors
Correspondence: alimoharrami1371@gmail.com, hakwan@gmail.com

Abstract

Perceptual similarity is a cornerstone for human learning and generalization. However, in assessing the similarity between two stimuli differing in multiple dimensions, it is not well-defined which feature(s) one should focus on. The problem has accordingly been considered ill-posed. We hypothesize that similarity judgments may be, in a sense, metacognitive: The stimuli rated as subjectively similar are those that are in fact more challenging for oneself to discern in practice, in near-threshold settings (e.g., psychophysics experiments). This self-knowledge about one's own perceptual capacities provides a quasi-objective ground truth as to whether two stimuli 'should' be judged as similar. To test this idea, we measure perceptual discrimination capacity between face pairs, and ask subjects to rank the similarity between them. Based on pilot data, we hypothesize a positive association between perceptual discrimination capacity and subjective dissimilarity, with this association being importantly specific to each individual.

Keywords: similarity judgment, subjective perceptual similarity, perceptual discrimination capacity, metacognition, subjective perception

Introduction

Subjective perceptual similarity between stimulus pairs has long been studied in human behavior. These studies explored various factors modulating similarity judgments, such as the effects of knowledge and expertise, contextual cues, and the order of presenting the stimuli (Shepard, 1964; Tversky, 1977; Smith, 1989; Smith & Heise, 1992; Medin et al., 1993). Different theories and quantitative models of similarity have also been proposed (Nosofsky, 1984; Shepard, 1987; Smith, 1989). For example, Roger Shepard famously formulated the universal law of generalization, according to which humans respond in the same way to stimuli of high similarity, and the probability of this generalization decays exponentially as the distance increases within a putative metric psychological space (Shepard, 1987). Later, Shepard's law was expanded to encompass general non-metrically structured spaces (Tenenbaum & Griffiths, 2001) and different accounts; notably, the rate-distortion theory was proposed to explain its nature (Sims, 2018). Intriguingly, recent research has demonstrated that the exponential similarity decay, coupled with a signal detection theory, can also effectively capture observations in visual working memory (Schurgin et al., 2020). There is also a rich history of studies utilizing similarity judgments, in combination with multidimensional scaling, to uncover the underlying perceptual dimensions of stimuli (Borg & Groenen, 2005; Hebart et al., 2020).

Similarity judgments are subjective, in that it is up to the subject to report how they feel about the stimuli. Accordingly, some researchers have argued that similarity judgments may reflect key aspects of conscious perception (Clark, 2000; Rosenthal, 2010; Malach, 2021; Lau et al., 2022; Tallon-Baudry, 2022; Zeleznikow-Johnston et al., 2023; Moharramipour & Lau, 2024). However, the essentially subjective nature of these judgments also led to the well-known critique that similarity is perhaps an ill-posed problem: there is, in a sense, no objective ground-truth as to how similar two things really are (Goodman, 1972; Medin et al., 1993). For example, Joe Biden may look more similar to Hillary Clinton than to Barack Obama, with respect to skin color. However, if we focus on gender-related facial features, Joe Biden may look more similar to Barack Obama. From the outset, it is unclear which visual features one should focus on. This presents a challenging obstacle to understanding the processes underlying similarity judgments, as mechanistic explanations of perception often rely on characterizing the observer as performing optimal inference, given existing constraints (Rao, 1999; Shen & Ma, 2016).

Following previous theoretical work (Lau et al., 2022), we hypothesize that subjective similarity judgments may be normative and rational, in the sense that they are made systematically based on the metacognitive access of our own perceptual abilities. Stimuli pairs judged to be more similar are, in fact, more challenging for oneself to discern in practice. If one judges two perceptual stimuli to be highly dissimilar, and yet fails to distinguish them in psychophysical tasks, the said similarity judgment can be regarded as 'incorrect' in a meaningful sense.

Revisiting the above example of how subjectively similar two faces are, 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.

Metacognition is commonly defined as the monitoring (and control) of one's own cognitive abilities (Morales et al., 2018). In the present study, we hypothesize that similarity judgments involve a type of implicit metacognition (Lau et al., 2022). When we make a similarity judgment, it reflects our own perceptual capacities.

The above is a non-trivial prediction, because an alternative hypothesis is 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. This alternative hypothesis is not implausible given that 'error' feedbacks are generally never given to subjects, to 'correct' them or train them, as they make these similarity ratings somewhat freely.

To test our hypothesis, we quantify the degree of subjective perceptual similarity between stimulus pairs by having participants freely rank similarity, without being given specific criteria, making it subjective, across a stimulus set (Figure 1A & 1B; subjective similarity judgment task). We also assess participants' perceptual discrimination capacity between the pairs. The stimulus pairs may be so obviously dissimilar that discriminating between them is just too easy (i.e. performance under normal conditions would be at the ceiling). To address this problem, we propose to use a psychophysical method to measure such discrimination capacity near perceptual threshold. We measure the participants' discrimination performance within the morph set that spans between the two stimuli (Figure 1C; near-threshold discrimination task). With this, we quantify the number of just-noticeable-differences (#JNDs; see legends of Figure 1C for explanation) between a pair. The #JNDs reflects the perceptual discrimination capacity, with its higher value indicating a higher capacity. We use faces as stimuli in our study due to their high-dimensional (i.e. multi-featural) nature, and the fact that these are naturalistic stimuli commonly encountered in everyday life. In subsequent sections, we use the notion "dissimilarity" instead of "similarity", so the hypothesized association with discrimination capacity is positive.

In summary, we hypothesize that there is a correlation between perceptual discrimination capacity (in near-threshold tasks) and subjective perceptual dissimilarity (as reflected by self-ratings of supra-threshold stimulus pairs) within each individual (Hypothesis 1, first row in Table 1). Specifically, perceptual discrimination capacities are higher in face pairs that are subjectively judged to be more dissimilar. Further – and critically– we hypothesize that this association is specific to each individual (Hypothesis 2, second row in Table 1), meaning that one's subjective perceptual dissimilarity is better explained by one's own perceptual discrimination capacity than other participants' (average) discrimination capacity. This would support the notion that 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. A complete overview of the hypotheses and their corresponding tests is provided in Table 1.

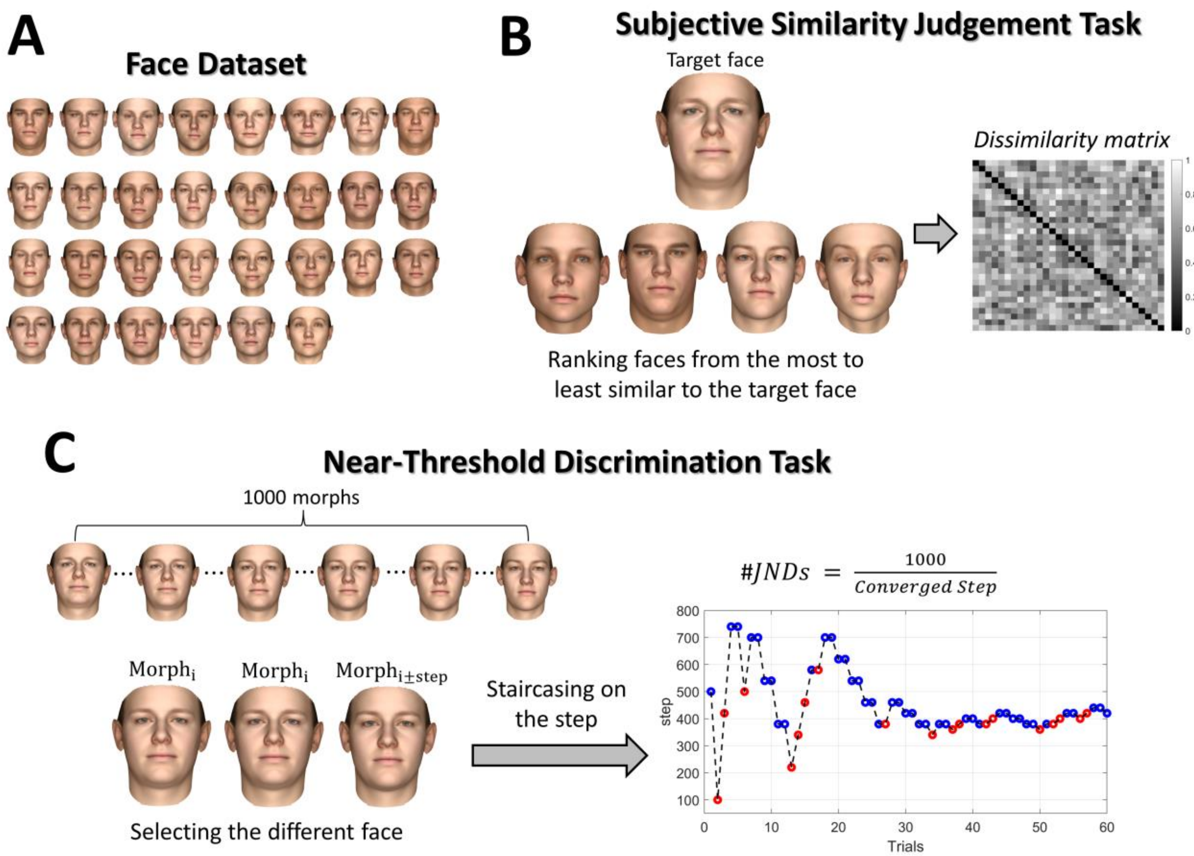


Figure 1. Experimental tasks for estimating subjective perceptual dissimilarity and perceptual discrimination capacity. (A) Illustration of 30 faces to be used in the present study. (B) The subjective similarity judgment task for estimating the level of subjective perceptual dissimilarity between face pairs. A target face on top and four other faces (candidates) on the bottom are shown to the participant in each trial. Participants are instructed to rank the candidate faces from the most to least similar with respect to the target face by clicking on them in order. Then, a 30x30 dissimilarity matrix is computed from the participant's responses across trials, with the value in each cell of the matrix indicating the level of subjective dissimilarity between a face pair. (C) The near-threshold perceptual discrimination task for measuring the discrimination capacity between two faces. One thousand morphs are created as intermediate transitions between two faces (based on a computational face model; see Methods for details). In each trial, three faces are shown simultaneously to the participants. Two are identical, and one is different from the other two by a certain degree (number of morph steps within the 1000-morph series). Participants are instructed to click on the face that is different from the other two. Task difficulty is maintained by titrating the number of morph steps needed for the different face to be barely detectable, using a standard 1-up-2-down staircase method. The converged (i.e., stabilized) value of the staircase indicates the number of morph steps required to maintain near-threshold performance (71% correct); thus, this value reflects the just-noticeable-difference (JND). Because these morph steps come from a series of 1000 morphs between a face pair (e.g., any two faces in 1A), if e.g. JND = 250 morph steps, we can also describe the two faces concerned as being 4 JNDs apart from each other. This general notion of the number of JNDs (#JNDs) between face pairs, which is just the total number of morph steps (1000) divided by the measured JND, allows us to describe the psychophysical discriminability between two faces, free from the non-standard physical unit of 'morph steps' (which depends on the

arbitrary specifics of the morphing procedure such as total number of steps used). Essentially, #JNDs indicates the perceptual distance between a face pair, in other words, how many JNDs are in between the face pair; thus, its higher value corresponds to a higher discrimination capacity.

Methods

Please note that the method section is written in the present tense as the experiment has not been done yet. We will change the tenses to past tense in the second phase submission.

Ethics information

The study received approval from the Ethics Review Committee at RIKEN, complying with all their ethical guidelines. Informed consent is obtained from participants before the experiment, and in appreciation of their participation, they are compensated with 3000 yen (approximately 20 US dollars) for each day of participation (roughly 90 minutes each day).

Design

Twelve participants are recruited for the study. They initially perform the subjective similarity judgment task, twice over the course of two days. After all participants complete this task twice, a set of 24 face pairs are selected for examination in the near-threshold discrimination task. The criteria for selecting the face pairs are described in the subsequent sections. Then, all participants are invited back to perform the near-threshold discrimination task over two days on these 24 face pairs, randomly splitting the pairs between the days.

Note that there are 48 sessions in total, across 12 participants. 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. The subjective similarity judgment task consists of 300 efficiently crafted trials to estimate the level of subjective perceptual dissimilarity between all face pairs. Participants perform this task twice, and the achieved dissimilarity values are averaged to further enhance the robustness. The near-threshold discrimination task comprises a total of 1440 trials, 60 trials per face pair, to effectively estimate the perceptual discrimination capacity between a systematically selected set of 24 face pairs. Furthermore, we recruit more participants if these initial 12 participants don't satisfy our data collection stopping rule described in the Sampling plan section.

Face data set

The basal face model (BFM) (Paysan et al., 2009) is used to select our face dataset and generate morphs between the faces for the near-threshold discrimination task. BFM is a widely used morphable model for generating graphical faces with two embedded vectors describing the shape and texture of the faces independently. We arbitrarily selected 30 faces from the BFM space, while ensuring a diverse set that also includes faces positioned close to each other in the BFM space. The top three shape dimensions were assigned systematically from a cylindrical coordinate with a 2.5 SD radius, and the subsequent top 47 shape dimensions and top 50 texture dimensions were assigned randomly from a uniform distribution ranging between -1.5 and 1.5 SD. The remaining less important shape and texture dimensions were set to zero. Figure 1A shows the selected 30 faces for the study.

Subjective similarity judgment task

In each trial, participants are presented with a visual arrangement consisting of one face positioned at the top (target face) and four other faces positioned at the bottom (candidate faces) (Figure 1B). Participants are instructed to rank the four bottom candidate faces based on their perceived similarity to the top target face by mouse-clicking on the faces in the order of most to least similar. Each clicked face immediately disappears from the screen, and the trial ends after all candidate faces are clicked one by one. If participants fail to complete the trial within 30 seconds, the trial is skipped, and any ranking assigned is disregarded. The aim of this task is to estimate the level of subjective dissimilarity between each face pair and to construct a dissimilarity matrix (Figure 1B) for each participant by analyzing their assigned rankings across trials.

The level of subjective dissimilarity (dissimilarity value) between two faces is estimated by calculating the probability of one face being ranked lower than the rest of the faces when the other face is the target, as outlined in the following. The rankings given in all trials are segmented into sets of three, consisting of the target face and the combination of two of the four candidate faces (i.e., six sets per trial). Within each set, the face that ranked lower is marked as the odd face. Subsequently, the dissimilarity value between a face pair is determined by calculating the ratio of instances where one of the faces is marked as the odd face across all sets that include the face pair with either of them as the target face. It is noteworthy that when calculating this ratio, we account for a non-tested set with an obvious outcome, where one of the faces repeats, by adding 0.5 to both the numerator and the denominator. This fundamentally prevents getting a dissimilarity value of zero, as only the diagonal value of the dissimilarity matrix should be zero.

The tuple of five faces displayed in each trial is strategically selected using the InfoTuple method (Canal et al., 2020). This method guarantees that each trial offers informative data, thereby enhancing the estimation of the dissimilarity matrix. This essentially enables achieving a robust estimation of the dissimilarity matrix over a smaller number of trials. The trial selection procedure is similar to the one used by Canal et al. (Canal et al., 2020) and comprises the following steps:

1. The tuple set in the first 30 trials is selected at random while ensuring that each face is selected once as the target face.
2. The dissimilarity matrix is calculated as described above, and a 5-dimensional metric multidimensional scaling (MDS) (Borg & Groenen, 2005) is applied to the dissimilarity matrix to find its embeddings.
3. A cycle of 30 trials, showcasing each face once as the target face, is selected by the InfoTuple method using the embeddings. 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). Please refer to the original paper for a detailed explanation and the mathematics of the InfoTuple (Canal et al., 2020).

4. The dissimilarity matrix is calculated given all the data collected, and the embeddings are updated by applying a 5-dimensional metric MDS to the dissimilarity matrix and using the previous embeddings as the seed in the MDS algorithm.
5. Steps 3 and 4 are repeated for 9 iterations. We stop after 9 iterations as the dissimilarity matrix and embeddings reach a relatively stable state at this point. As completing 9 iterations is lengthy and can be exhausting, participants are given a short break after every 3 iterations.

The final obtained embeddings are used to recalculate the dissimilarity matrix by computing the Euclidean distances between the faces. This process results in a more accurate version of the dissimilarity matrix by refining inaccuracies in some cells of the original dissimilarity matrix due to insufficient data attributed to them or due to noise in responses (i.e., response inconsistencies). Therefore, this dissimilarity matrix, derived from the embeddings, is utilized in all the subsequent stages instead of the original dissimilarity matrix. Please see the supplementary Figure 1 for a schematic overview of the task design described above.

We also plan to test a 2-dimensional metric MDS for recalculating the dissimilarity matrix. Even though there might be significant information loss in using such a low dimensional MDS, it can further refine the matrix and also make the dissimilarity relations sharper (i.e., more distinct). A sharp dissimilarity relation can potentially make its association with perceptual discrimination capacity more salient. This was true in our pilot data. Both Hypothesis 1 and 2 (Figure 2) reached a $p < 0.05$ at the individual-level in all four participants.

We note that in our algorithm, prior to applying a metric MDS, a nonmetric MDS is employed to fill in the missing cells of the dissimilarity matrix (i.e., pairs with no ranking data). The missing cells are filled in by the Euclidean distances computed from the embeddings derived by the nonmetric MDS. This procedure is important in initial iterations in which there are a considerable number of missing cells. Then, the metric MDS is applied to this filled-in dissimilarity matrix. We don’t use the nonmetric MDS directly because the nonmetric MDS, unlike the metric MDS, doesn’t preserve the magnitude of the dissimilarity between pairs.

Last, it is important to mention that each participant performs the above task twice, each time on a different day. The average of the dissimilarity matrices obtained from each day forms the final dissimilarity matrix. Further, to evaluate the reliability of the obtained dissimilarity matrix from a session, we report the within-participant correlation between the dissimilarity matrices derived from each day. 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.

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.

Near-threshold discrimination task

The objective of this task is to estimate perceptual discrimination capacity in a face pair. A series of 1000 equally spaced morphs are generated along the line connecting a face pair in the BFM space. In each trial, three faces are shown to the participants: two identical faces, randomly selected from the morph set, and a third different face spaced by a certain number of morphs (step value) from the identical faces (e.g., morph 200, morph 200, and morph 300: here the step value is 100). The faces are displayed simultaneously at the center of the screen next to each other, and their arrangement is randomized in each trial (Figure 1C). Participants are instructed to identify and click on the different face.

A staircase (Cornsweet, 1962) with a 1-up and 2-down protocol is applied to the step value (i.e., the number of morphs between the different and identical faces), initiating from a step value of 500. After each incorrect response trial, the step value is increased, and it is decreased after two consecutive correct trials. The magnitude of the change in the step value gradually decreases over trials, reaching a minimum change of 20 steps. The task is terminated after 60 trials, allowing precise convergence of the step value. The converged step value indicates the just-noticeable-difference (JND), the minimum degree of differences between the faces required to achieve near-threshold discrimination performance (71% correct response). The average of the steps achieved within the last 5 changes is defined as the converged step. Essentially, a small JND, for example, 100, indicates that the two questioned faces are quite distinct, involving 10 JNDs (i.e., 1000 divided by 100) between them. We use the notion of the number of JNDs (#JNDs) to quantify perceptual discrimination capacity. The #JNDs indicates the perceptual distance between a face pair, in other words, how many JNDs are in between the face pair in a participant. Therefore, its higher value corresponds to a higher discrimination capacity. The #JNDs is simply calculated as 1000 (i.e., total number of morphs) divided by the JND.

In a session, there are 12 face pairs to undergo the staircase procedure. The staircases for each of these face pairs are interleaved, progressing concurrently. There is a cycle of 12 trials, featuring each staircase once in a random order. The trials are time-constrained, requiring participants to respond within 8 seconds. If participants fail to respond within this time window, the trial is skipped and reintroduced at the end of the cycle. To encourage participants to perform to the best of their abilities, they are provided with feedback on their responses. A green circle is displayed on the different face (i.e., correct answer) and a red cross on the identical faces (i.e., wrong answers) after they provide their response. Since the session is lengthy, with a total of 720 trials, participants are given a short break after every 180 trials.

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.

Selection of the pairs for the near-threshold discrimination task

Following the completion of the subjective similarity judgment task twice by all 12 participants, 24 face pairs are systematically selected to be examined in the near-threshold discrimination task. Practical constraints (time limitations; it takes 4-5 minutes to complete the near-threshold discrimination task for a face pair) limit us to examine only a small subset of the pairs. 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. Participants are re-invited for two sessions to perform the near-threshold discrimination task on these specific 24 pairs, completing 12 of the pairs in each session.

Measuring perceptual discrimination capacity, expressed as #JNDs, in a face pair, involves running a near-threshold discrimination task dedicated to that specific pair. Thus, considering practical constraints as described earlier, we have no choice but to examine only a limited subset of pairs (24 out of 435 possible pairs). However, this subset is carefully chosen to provide the most informative data for testing our hypotheses while fairly covering different ranges in the group-averaged dissimilarity matrix. The pairs with a controversial subjective dissimilarity degree across participants are particularly promising candidates. If the hypothesis holds true, these pairs should also exhibit controversial discrimination capacity across participants. Considering the inherent noise in our methods in estimating the dissimilarity values and the #JNDs, any effect should be better detectable on pairs with larger standard deviations, those that are more distinct, across participants. So, we select 18 pairs with controversial dissimilarity values across participants, and for the sake of comparison, we select 6 pairs with less (non) controversial values.

First, the dissimilarity matrix is z-normalized within each participant to ensure that its scale is consistent across participants. Subsequently, the mean and SD of the dissimilarity matrix are computed across participants, and the quantiles of the mean values are derived. Within the first and the last quantiles, 3 pairs with the highest and 1 with the lowest SD are selected. Additionally, 6 pairs with the highest and 2 pairs with the lowest SD are chosen within the second and the third quantiles. This systematic selection ensures choosing 18 controversial and 6 non-controversial pairs that cover a diverse range in the group-averaged dissimilarity matrix.

Sampling plan

Participants who meet the following criteria are excluded from the analysis: Those who don't complete all four experimental sessions and those who show a lack of attentiveness to the task in any of the sessions. The lack of attentiveness in the near-threshold discrimination task is identified by non-converging staircases, indicated by a non-fluctuating increment in the step value over trials. Specifically, 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. In the subjective similarity judgment task, the lack of attentiveness is judged by comparing the consistency of responses between the first and second half of the session. Specifically, if the correlation between the dissimilarity matrices estimated from each half falls below 0.2, the session is considered bad with inadequate attentiveness. This correlation was 0.56 ± 0.086 (mean$\pm$SD) in our pilot data. 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. Note that the second phase of the study, involving the selection of the pairs and the near-threshold discrimination task, does not start until the quality of the data from all 12 participants in the subjective similarity judgment task is confirmed as not meeting the above mentioned exclusion criteria. Following this, any subsequent exclusion and recruitment of new participants do not modify the initially selected pairs for the near-threshold discrimination task.

After analyzing the data from these 12 participants, if the statistics fail to meet the following stopping criterion, we recruit more participants until the criterion is satisfied. The individual-level statistic is converted to z-values, and the 95% confidence interval of the group-mean z-value is derived (See the Analysis Plan section). We stop the experiment, if, in both Hypothesis 1 and 2, the width of this 95% confidence interval is less than 1. Moreover, we consider a hypothesis to be confirmed, if the group-mean z-value is significantly above zero, specifically, if the 95% confidence interval is above zero. Note that we set our stopping criterion independent of the significance testing and solely based on the precision of the effect (i.e., the confidence interval). We do not stop our experiment until achieving a high precision, so that we are confident that the effect is not being confirmed or rejected because of some extreme observations (Cumming, 2008; Lakens, 2014) . 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 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). 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.

Analysis plan

Hypothesis 1

Spearman correlation coefficient and its p-value are computed between the dissimilarity values and #JNDs of the examined 24 pairs in each individual. The Spearman correlations are converted to z-values using the Fisher z-transformation (Fieller et al., 1957) to conduct group-level statistical tests. The distribution of the group-mean z-value is computed by bootstrapping, iterated 100,000 times, and then its 95% confidence interval is derived by obtaining the 2.5th and 97.5th percentile of the distribution. The hypothesis is confirmed if this confidence interval is above zero. The following statistics are reported as complementary information: the p-value and the Bayes factor of a t-test applied to z-values (BayesFactor Matlab package is used: https://zenodo.org/badge/latestdoi/162604707), the p-value of a Fisher's combined probability test, combining individual-level p-values (Brown, 1975), and a Bayesian posterior distribution of population prevalence (Ince et al., 2021) and its 95% highest posterior density interval, considering the p-value of 0.05 as the individual-level significance threshold. The Bayesian posterior distribution quantitatively summarizes how prevalent a particular effect would be in the population, based on the number of participants tested in a study and their proportion showing the effect significantly. Furthermore, we note that if the 95% confidence interval of the group-mean z-value includes zero (i.e., a non-significant finding), we incorporate an equivalence test (Lakens et al., 2018) to further assess the significance of the null result (i.e., how confidently the effect can be considered non-significant). We assume a group-mean z-value range of -0.5 to 0.5 as the smallest effect size of interest; thus, if the 95% confidence interval falls within this range, we regard the effect to be confidently null.

Hypothesis 2

Each participant's #JNDs is z-normalized to ensure that the #JNDs range is consistent across participants. This normalization is crucial, given that some participants may exhibit generally higher #JNDs than others. Subsequently, a nonparametric permutation test is applied to each individual to assess the specificity of the relationship between their #JNDs and dissimilarity values, as follows:

1. A permutation set of #JNDs is constructed by randomly permuting the #JNDs across participants (i.e., for each pair, selecting the value in one of the participants at random), excluding the participant in question. Essentially, the permutation set simulates a new participant by mixing the existing participants.
2. The Spearman correlation coefficient is calculated between the permuted #JNDs and the dissimilarity values of the participant in question. It is noteworthy that 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.
3. Steps 1 and 2 are repeated 100,000 times to derive the distribution of the Spearman coefficients. This distribution represents the null hypothesis distribution in which there is no individual specificity.
4. The actual Spearman coefficient between the #JNDs and dissimilarity values of the participant in question is tested against the null distribution, and the p-value, indicating the significance level, is derived. The z-value is also calculated by subtracting the actual Spearman correlation from the null distribution's mean and then dividing it by the null distribution's SD.

Then, similar to Hypothesis 1, the 95% confidence interval of the group-mean z-value is derived through bootstrapping, and if it is above zero, the hypothesis is confirmed. The complementary statistics, outlined in Hypothesis 1, are also reported for Hypothesis 2.

Table 1 Experimental Design Table

Question	Hypothesis	Outcome Measures	Sampling plan	Analysis Plan	Interpretation given to different outcomes
How does perceptual discrimination capacity relate to subjective perceptual dissimilarity?	There are higher perceptual discrimination capacities in pairs that are perceptually more dissimilar. In other words, we hypothesize that perceptual discrimination capacity is positively correlated with subjective perceptual dissimilarity.	Subjective perceptual dissimilarity and perceptual discrimination capacity between stimuli pairs are measured in each participant through two different psychophysical tasks called the subjective similarity judgment task (Figure 1B) and the near-threshold discrimination task (Figure 1C). The subjective similarity judgment task assesses the level of subjective dissimilarity (dissimilarity value) between stimulus pairs. The near-threshold discrimination task measures the number of just-noticeable-differences (#JNDs) between stimulus pairs, quantifying the perceptual discrimination capacity. A higher #JNDs indicates a higher capacity in distinguishing a stimulus pair.	Twelve participants are recruited, each completing four sessions over four days, spending two days on the subjective similarity judgment task and two days on the near-threshold discrimination task. Participants failing to complete all four sessions and those displaying a lack of attentiveness to the task in any of the sessions are excluded from the analysis. The criteria for a lack of attentiveness are described in the method section. After completing the data collection on 12 participants who don't meet any of the exclusion criteria, if the results don't satisfy our following experiment's stopping criterion, we recruit more participants until the criterion is satisfied. However, we end the experiment once we reach a maximum of 24 participants, regardless. The stopping criterion is met if the width of the 95% confidence interval of the group-mean z-value (see the method section) is less than 1, and the hypothesis is confirmed if this 95% confidence interval is above zero. * The hypothesis was confirmed in the pilot study (Figure 2A & 2B)	In each participant, the Spearman correlation between the #JNDs and dissimilarity values is computed. Then, the Spearman correlations are transformed into z-values using the Fisher z-transformation. Subsequently, the 95% confidence interval of the group-mean z-value is derived through bootstrapping with 100,000 iterations.	If the described group test doesn't reach the level of significance, yet at the individual level, the correlation reaches significance ($p<0.05$) within a certain few participants, we can interpret that the positive association between perceptual discrimination capacity and subjective perceptual dissimilarity holds true for certain individuals and doesn't generalize to the entire population. Failure of the group test may indicate that subjective perceptual dissimilarity is made rather arbitrarily, based on subjective preferences, and does not reflect underlying psychophysical capacities.
Is the association between perceptual discrimination capacity and subjective perceptual dissimilarity specific to each individual?	A participant's subjective perceptual dissimilarity is better explained by their own perceptual discrimination capacity than by a group-averaged perceptual discrimination capacity. To put it differently, we hypothesize that the positive association between perceptual discrimination capacity and subjective perceptual dissimilarity is specific to each individual. Essentially, subjective dissimilarity reflects a metacognitive assessment of one's own perceptual discrimination capacity, rather than general knowledge about the physical differences of the stimuli.	The level of individual-specificity of the relationship between perceptual discrimination capacity and subjective perceptual dissimilarity is measured by a nonparametric permutation test, which assesses whether one's own dissimilarity value is more strongly correlated with one's own #JNDs than with others' #JNDs. Essentially, the above test reflects how specific this relationship is in each individual in terms of z-values.	The participants' exclusion and the experiment's stopping strategies remain the same as above. Similar to the above hypothesis, the hypothesis is confirmed if the 95% confidence interval of the group-mean z-value is above zero, and the experiment stopping rule is met if the width of this confidence interval is smaller than 1. * The hypothesis was confirmed in the pilot study (Figure 2C & 2D)	First, the #JNDs is z-normalized within each individual to ensure that its scale is consistent across participants, and then a nonparametric permutation test is applied to each participant separately as follows: Briefly, a permutation #JNDs set is constructed by randomly permuting the #JNDs of all participants, excluding the participant in question. The Spearman coefficient is then computed between the permuted #JNDs set and the dissimilarity values of the participant in question. This process is iterated 100,000 times to establish the distribution of the Spearman coefficient, representing the null hypothesis distribution. Finally, the actual Spearman coefficient between the dissimilarity values and the #JNDs of the participant in question is compared against the null distribution, and its z-value is computed. Subsequently, the 95% confidence interval of the group-mean z-value is derived by bootstrapping, similar to the above hypothesis.	If the described group test fails to reach the significance level, but there are certain individuals with significant statistics ($p<0.05$), we can draw a similar interpretation as the above that the hypothesis holds true only for certain people. Failure of the group test, if the first hypothesis holds true, may suggest that subjective perceptual dissimilarity is made based on general stimulus properties that predict the psychophysical performance of human subjects and is not metacognitive in the sense of reflecting direct access to one's own perceptual capacities.

Pilot data

We conducted a pilot version of the study with 4 participants. The experimental design was similar to the currently proposed experiment (Figure 1), with a few differences. Participants performed the subjective similarity judgment task only once. After all of them completed the task, 13 pairs of faces were selected, and the participants were all invited back to perform, in a different session, the near-threshold discrimination task on these pairs. In this pilot study, we randomly selected the pairs while ensuring that most of them have high dissimilarity values SD across participants, indicating that the degree of subjective dissimilarity is quite 'controversial', i.e., individual-specific. The subjective similarity judgment task and the near-threshold discrimination task were conducted similarly to the proposed experiment, except that there was no time constraint on both tasks, and no trial-by-trial behavioral feedback was provided during the near-threshold discrimination task. Additionally, in one participant, the near-threshold discrimination task comprised 50 instead of 60 trials. We applied the same analysis approach described in the method section on the pilot data.

Hypothesis 1 was confirmed in the pilot study (Figure 2A & 2B), indicating that subjective perceptual dissimilarity and perceptual discrimination capacity are highly correlated. The correlation was significant ($p<0.05$) at the individual-level in 2 out of 4 participants. At the group level, the mean z-value across participants was 2.33, with a 95% confidence interval between 1.65 and 3.02. A t-test on the z-values yielded a p-value of 0.012, and a Bayes factor (BF) of 6.11. Moreover, a Fisher's test, combining the individual-level p-values, resulted in a p-value of 0.00011.

More importantly, Hypothesis 2 was also confirmed in the pilot study (Figure 2C & 2D), suggesting that the association between subjective perceptual dissimilarity and perceptual discrimination capacity is specific to each individual. To put it differently, others' perceptual discrimination capacity cannot account for one's subjective perceptual dissimilarity as well as their own perceptual discrimination capacity. The statistic was highly significant ($p<0.05$) at the individual-level in 3 out of 4 participants. At the group-level, the mean z-value across participants was 2.26, with a 95% confidence interval between 1.45 and 2.94. A t-test on the z-values resulted in a p-value of 0.016 and a BF of 5.07, and a Fisher's test yielded a p-value smaller than 0.00001.

In the main study, we anticipate observing even stronger statistics not only at the group-level but also at the individual-level. We expect that testing more stimulus pairs and having more participants lead to observing stronger results at the individual-level for Hypothesis 1 and Hypothesis 2, respectively.

We further explored the correlation between subjective perceptual dissimilarity and perceptual discrimination capacity in each face pair across participants in our pilot study (Figure 3). Given the small sample size (i.e., four participants), no meaningful statistical conclusions can be inferred. However, it is notable that the correlations were strongly positive, particularly in the controversial pairs: those with controversial degrees of subjective dissimilarity across participants. Essentially, Figure 3 also indicates that one's perceptual discrimination capacity can explain one's subjective perceptual dissimilarity. Similar inter-individual differences observable in the subjective perceptual dissimilarity could also be found in the perceptual discrimination capacity. However, this is not apparent in the less controversial pairs. Nonetheless, this may not necessarily suggest that the association doesn't exist in the less controversial pairs. The measures obtained from our psychophysical tasks inevitably contain some noise which may make them to be not precise enough to capture the subtle differences across participants in the less controversial pairs. In the main experiment, we expect to obtain a clearer picture by having more participants and experimental sessions.

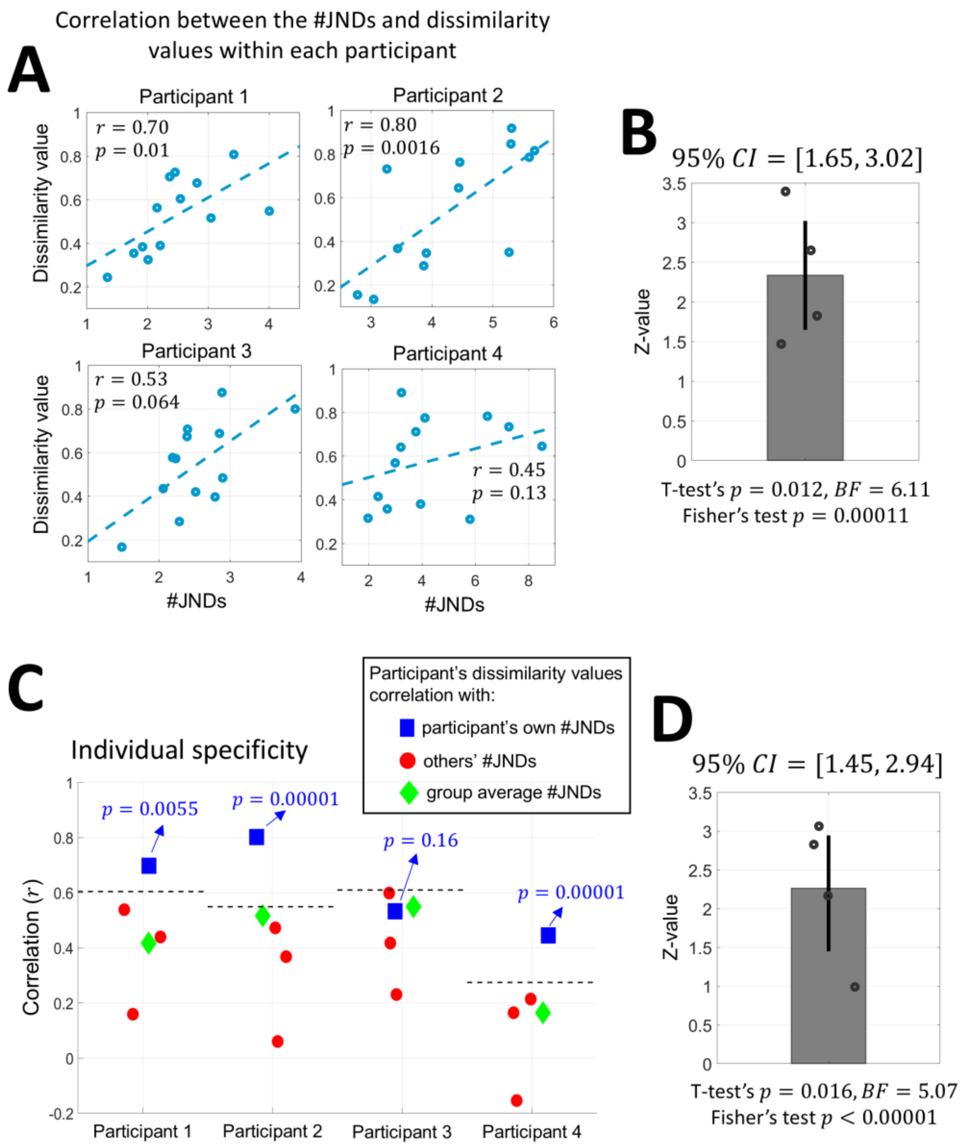


Figure 2. Relationship between perceptual discrimination capacity and subjective perceptual dissimilarity in the pilot study (N = 4 participants) at the individual participant level. (A) Correlation between the perceptual discrimination capacity expressed by #JNDs and the subjective perceptual dissimilarity within each participant. Each subplot illustrates the correlation in an individual participant, with each data point corresponding to a face pair. *r* indicates the Spearman correlation coefficient, and *p* denotes its associated p-value. (B) Individuals' Spearman correlations of A were transformed to z-values for group-

level hypothesis testing. The bar plot shows the mean z-value across participants, the vertical line represents the 95% confidence interval of the group-mean z-value, calculated through bootstrapping, and each dot on the plot corresponds to a participant. As displayed under the plot, analysis of the group-level effect by applying a t-test on the z-values yielded a p-value of 0.012, and a Bayes factor (BF) of 6.11. Alternatively, employing a Fisher's combined probability test, combining the individual-level p-values, resulted in a p-value of 0.00011. (C) Individual specificity of the relationship between perceptual discrimination capacity and subjective perceptual dissimilarity. Blue square, red circle, and green diamond indicate the Spearman correlation coefficient between each participant's dissimilarity values and the participant's own #JNDs, other participants' #JNDs, and the group averaged #JNDs, respectively. The dotted horizontal black line denotes the Spearman correlation value corresponding to a permutation test's p-value of 0.05, rejecting the null hypothesis and indicating that the correlation is specific to each individual. The correlation would not be specific to each individual (i.e., null hypothesis) if one participant's dissimilarity values are as equally correlated to the other participants' #JNDs as the own participant's #JNDs. The p-value rejecting the null hypothesis in each participant is shown in blue at the top of the blue squares. The result of the permutation test suggests that the relationship between perceptual discrimination capacity and subjective perceptual dissimilarity was highly specific to each individual in three out of four participants. (D) Individuals' specificity statistics of C were converted to z-values for group-level hypothesis testing. The remaining descriptions of the plot are similar to those in B.

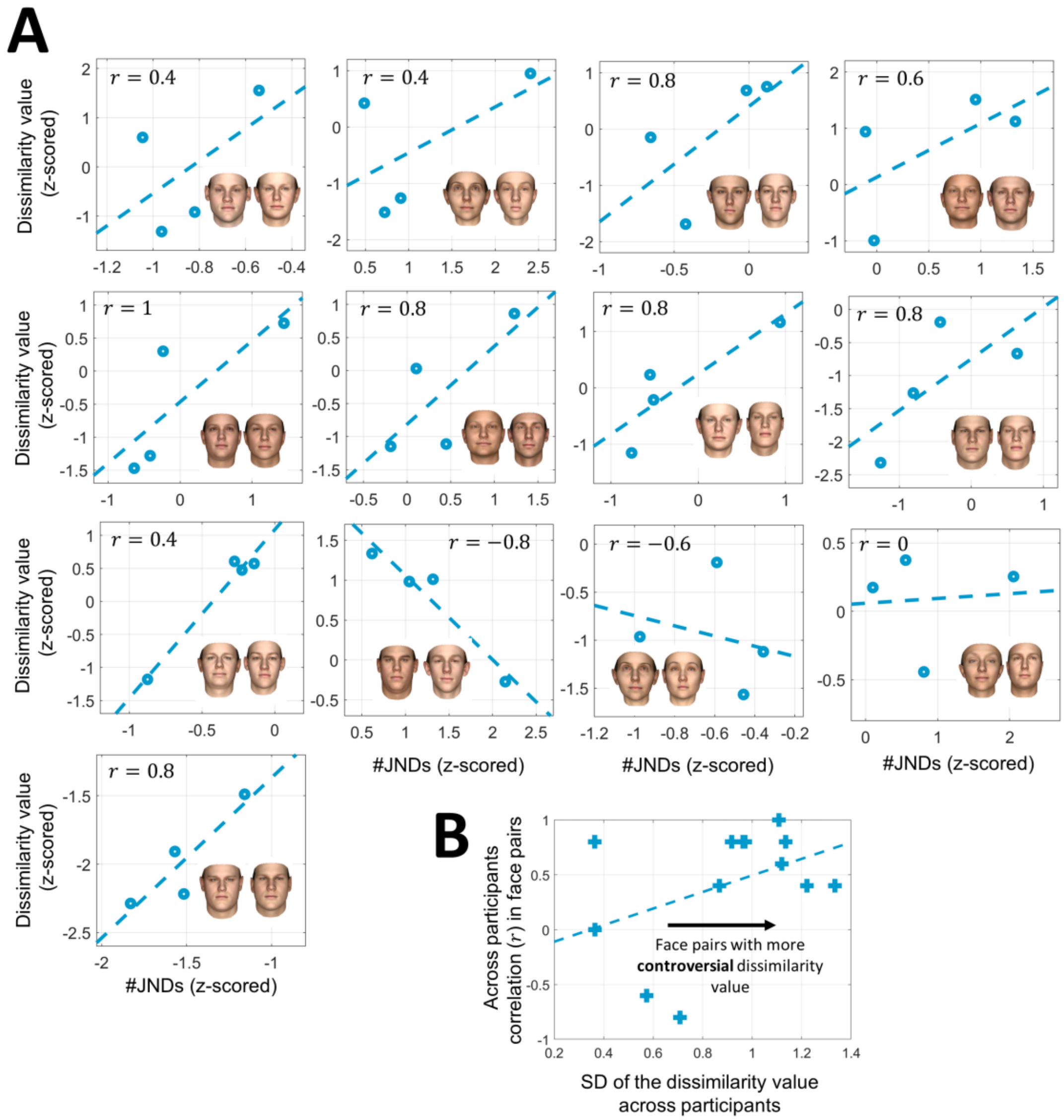


Figure 3. Relationship between perceptual discrimination capacity and subjective perceptual dissimilarity across participants (pilot study N = 4 participants). (A) Across participants' correlation between the #JNDs and the dissimilarity values in different face pairs. Each panel corresponds to a face pair, sorted based on their controversy in the level of subjective dissimilarity across participants. The top left panel shows the most controversial pair (i.e., one with the highest dissimilarity value SD across participants), and the bottom right panel showcases the least controversial pair. Both the #JNDs and the dissimilarity values were z-normalized within each participant. *r* indicates the Spearman correlation coefficient, and each dot on the plots corresponds to a participant. (B) Relationship between the face pairs Spearman correlation coefficient shown in A and their controversy in the level of subjective dissimilarity across participants (i.e., SD of the dissimilarity values across participants). The relationship between perceptual discrimination capacity and subjective perceptual dissimilarity across participants was more salient in highly controversial pairs.

Discussion

In the present study, we use a near-threshold psychophysical task to quantify perceptual discrimination capacity, which indicates one's capability to distinguish two stimuli (Figure 1C). We aim to examine whether this perceptual discrimination capacity measured at near-threshold is associated with subjective perceptual similarity rankings (Figure 1B) given at suprathreshold (Hypothesis 1). More critically, we seek to explore whether this association is specific to each individual, meaning that one's perceptual discrimination capacity can best explain one's own subjective perceptual similarity compared to that of others' (Hypothesis 2).

We conducted a pilot version of the study and confirmed both Hypotheses on our pilot data (Figure 2). However, to further and more precisely investigate our hypotheses, we intend to conduct a larger-scale study with more participants and experimental sessions. Given the high significance level observed in our pilot data, we expect a high likelihood of confirming the hypotheses in the main experiment.

If our hypotheses hold true, it may suggest that subjective similarity judgment is, in a specific sense, metacognitive: The self-knowledge of one's perceptual capacity guides one's subjective similarity judgment, and this may occur automatically and implicitly. In essence, perceptual discrimination capacities serve as a ground truth basis for making similarity judgments. A more accurate perceptual similarity judgment could be defined as the one with a more precise metacognitive read-out of one's own perceptual discrimination capacities. Similarly, the instability in perceptual similarity judgments could be considered as the result of inaccurate metacognitive assessment of one's own perceptual capacities.

Consequently, higher cortical brain areas, particularly the prefrontal cortex, may play a critical role in perceptual similarity judgments, given that its activity has been demonstrated to be associated with perceptual metacognition (McCurdy et al., 2013; Fleming et al., 2014; Morales et al., 2018). Of course, the current study does not directly test this hypothesis about neural mechanisms. Others have suggested that perceptual similarity information resides within the sensory cortices (Malach, 2021). In light of this, we are currently investigating whether perceptual similarity representations can be found beyond the visual areas, such as the lateral prefrontal cortex, using fMRI.

Finally, if our hypotheses are correct, perhaps it could shed light on one conundrum regarding large language models and consciousness. Recently, it has been reported that these models built with current technology in artificial intelligence can give human-like similarity ratings (Kawakita et al., 2023; Marjieh et al., 2023). If the qualitative characters of conscious perception are determined by the relevant similarity relations, as some researchers assume (Clark, 2000; Rosenthal, 2010; Malach, 2021; Lau et al., 2022; Tallon-Baudry, 2022; Zeleznikow-Johnston et al., 2023; Moharramipour & Lau, 2024), does it mean that these artificial agents are conscious (Moharramipour & Lau, 2024)? Or, at least, does it mean that they contain the essential information that is encapsulated within human perceptual experiences? The answer is probably no, if the metacognitive perspective described above is correct. That is, for the similarity judgment to be relevant for subjective experiences, according to our hypothesis, they need to reflect one's own perceptual capacities. What these models do is simply to mimic what humans say in general, and as such, their similarity judgments at best reflect common world knowledge about the physical characteristics of the stimuli, but they are not about one's own perceptual capacities (of which these models have none). There is, thus, a critical difference between humans and those models, in terms of what the similarity judgments mean for them.

Author contributions

A.M. contributed to conceptualization, project planning and design, methodology application, data collection, data analysis, visualization, writing, review, and editing.

W.Z. contributed to methodology application, data analysis, review, and editing.

D.R. contributed to conceptualization, project design, review, and editing.

H.L. contributed to conceptualization, supervision, project planning, design, analysis, review, and editing.

Funding

This study is supported by internal funding at the RIKEN Center for Brain Science (to H.L.).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Conflict of interest disclosure

The authors declare no competing interests.

Data, scripts, code, and supplementary information availability

The pilot data, the code used in the pilot study and the code that will be used in the main experiment are publicly accessible from the GitHub repository below: https://github.com/AliMoharramipour/Subjective-Dissimilarity-and-Discrimination-Capacity-

References

Borg I, Groenen PJ (2005) *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media.

Brown MB (1975) 400: A method for combining non-independent, one-sided tests of significance. *Biometrics*, 987–992.

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

Clark A (2000) *A theory of sentience*. Clarendon press.

Cornsweet TN (1962) The staircase-method in psychophysics. *The American journal of psychology*, **75**, 485–491.

Cumming G (2008) Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better. *Perspectives on psychological science*, **3**, 286–300.

Fieller EC, Hartley HO, Pearson ES (1957) Tests for rank correlation coefficients. I. *Biometrika*, **44**, 470–481.

Fleming SM, Ryu J, Golfinos JG, Blackmon KE (2014) Domain-specific impairment in metacognitive accuracy following anterior prefrontal lesions. *Brain*, **137**, 2811–2822.

Goodman N (1972) Seven strictures on similarity.

Hebart MN, Zheng CY, Pereira F, Baker CI (2020) Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature human behaviour*, **4**, 1173–1185.

Ince RA, Paton AT, Kay JW, Schyns PG (2021) Bayesian inference of population prevalence. *Elife*, **10**, e62461.

Kawakita G, Zeleznikow-Johnston A, Tsuchiya N, Oizumi M (2023) Comparing color similarity structures between humans and LLMs via unsupervised alignment. *arXiv preprint arXiv:2308.04381*.

Lakens D (2014) Performing high-powered studies efficiently with sequential analyses. *European Journal of Social Psychology*, **44**, 701–710.

Lakens D, Scheel AM, Isager PM (2018) Equivalence testing for psychological research: A tutorial. *Advances in methods and practices in psychological science*, **1**, 259–269.

Lau H, Michel M, LeDoux JE, Fleming SM (2022) The mnemonic basis of subjective experience. *Nature Reviews Psychology*, **1**, 479–488.

Malach R (2021) Local neuronal relational structures underlying the contents of human conscious experience. *Neuroscience of consciousness*, **2021**, niab028.

Marjieh R, Sucholutsky I, van Rijn P, Jacoby N, Griffiths TL (2023) Large language models predict human sensory judgments across six modalities. *arXiv preprint arXiv:2302.01308*.

McCurdy LY, Maniscalco B, Metcalfe J, Liu KY, De Lange FP, Lau H (2013) Anatomical coupling between distinct metacognitive systems for memory and visual perception. *Journal of Neuroscience*, **33**, 1897–1906.

Medin DL, Goldstone RL, Gentner D (1993) Respects for similarity. *Psychological review*, **100**, 254.

Moharramipour A, Lau H (2024) Open Review of Kawakita et al's "Is my 'red'your 'red'?"(2023) PsyArVix.

Morales J, Lau H, Fleming SM (2018) Domain-general and domain-specific patterns of activity supporting metacognition in human prefrontal cortex. *Journal of Neuroscience*, **38**, 3534–3546.

Nosofsky RM (1984) Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, memory, and cognition*, **10**, 104.

Paysan P, Knothe R, Amberg B, Romdhani S, Vetter T (2009) A 3D face model for pose and illumination invariant face recognition. In:, pp. 296–301. Ieee.

Rao RP (1999) An optimal estimation approach to visual perception and learning. *Vision research*, **39**, 1963–1989.

Rosenthal D (2010) How to think about mental qualities. *Philosophical Issues*, **20**, 368–393.

Schurgin MW, Wixted JT, Brady TF (2020) Psychophysical scaling reveals a unified theory of visual memory strength. *Nature human behaviour*, **4**, 1156–1172.

Shen S, Ma WJ (2016) A detailed comparison of optimality and simplicity in perceptual decision making. *Psychological review*, **123**, 452.

Shepard RN (1964) Attention and the metric structure of the stimulus space. *Journal of mathematical psychology*, **1**, 54–87.

Shepard RN (1987) Toward a universal law of generalization for psychological science. *Science*, **237**, 1317–1323.

Sims CR (2018) Efficient coding explains the universal law of generalization in human perception. *Science*, **360**, 652–656.

Smith LB (1989) A model of perceptual classification in children and adults. *Psychological review*, **96**, 125.

Smith LB, Heise D (1992) Perceptual similarity and conceptual structure. In: *Advances in psychology* , pp. 233–272. Elsevier.

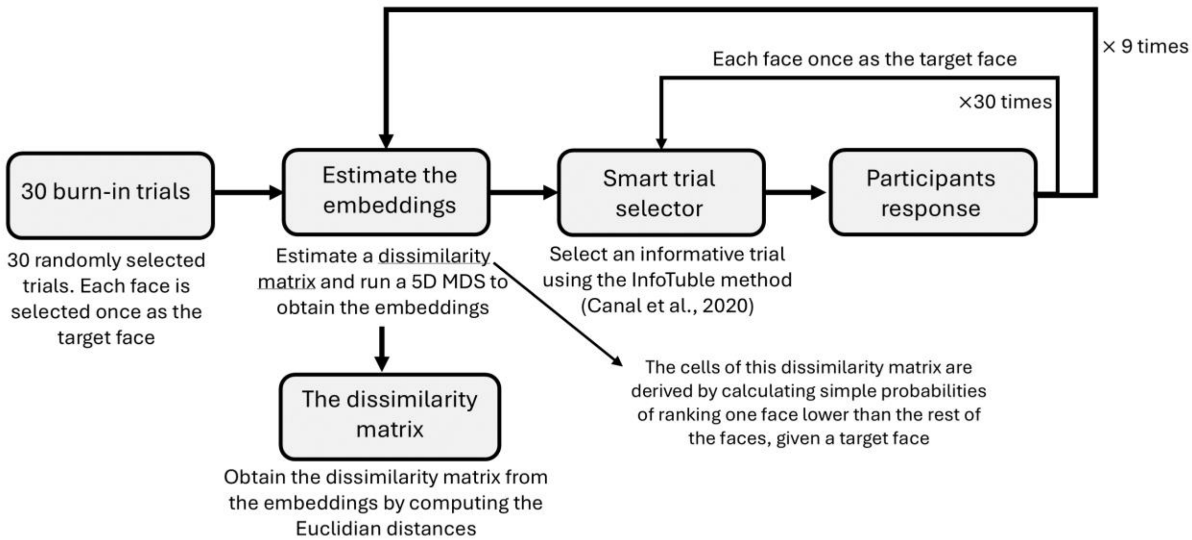
Tallon-Baudry C (2022) The topological space of subjective experience. *Trends in Cognitive Sciences*, **26**, 1068–1069.

Tenenbaum JB, Griffiths TL (2001) Generalization, similarity, and Bayesian inference. *Behavioral and brain sciences*, **24**, 629–640.

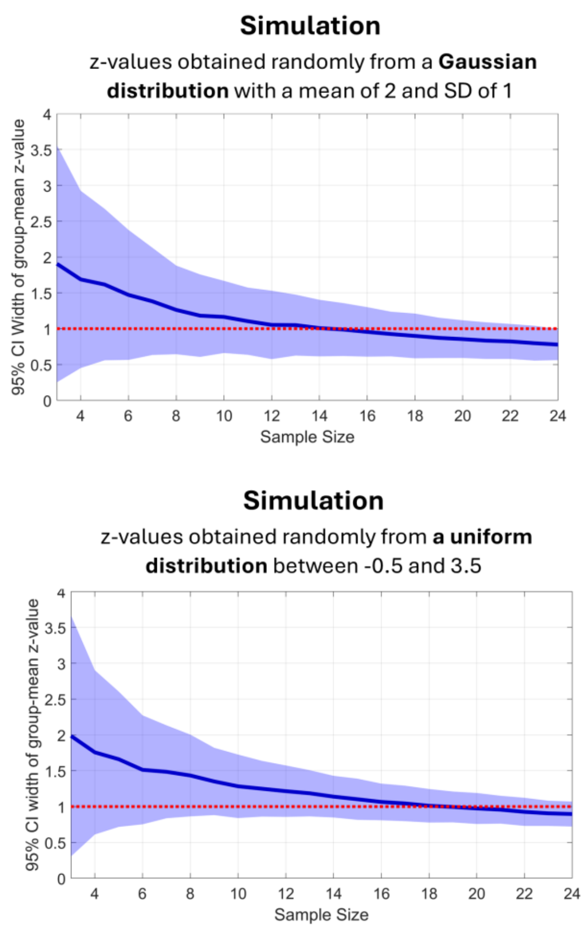
Tversky A (1977) Features of similarity. *Psychological review*, **84**, 327.

Zeleznikow-Johnston A, Aizawa Y, Yamada M, Tsuchiya N (2023) Are color experiences the same across the visual field? *Journal of Cognitive Neuroscience*, **35**, 509–542.

Supplementary Figures



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



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.