Is subjective perceptual similarity metacognitive?

Ali Moharramipour[1]*, William Zhou[1], Dobromir Rahnev[2], Hakwan Lau[1, 3, 4]*

[1] Center for Brain Science (CBS), RIKEN, Wako, Japan
[2] School of Psychology, Georgia Institute of Technology, Atlanta, United States
[3] Center for Neuroscience Imaging Research, Institute for Basic Science, Suwon, South Korea
[4] Department of Biomedical Engineering, SungKyunKwan University, Suwon, South Korea

* 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 measured perceptual discrimination capacity between face pairs, and asked subjects to rank the similarity between them. We found a positive association between perceptual discrimination capacity and subjective perceptual dissimilarity, with this association being importantly specific to each individual. The results indicate that perceptual similarity judgment reflects and predicts one's own perceptual capacities, supporting our hypothesis that perceptual similarity judgment is metacognitive.

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 [1–5]. Different theories and quantitative models of similarity have also been proposed [2,3,6,7]. 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 [7]. Later, Shepard's law was expanded to encompass general non-metrically structured spaces [8] and different accounts; notably, the rate-distortion theory was proposed to explain its nature [9]. 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 [10]. There is also a rich history of studies utilizing similarity judgments, in combination with multidimensional scaling, to uncover the underlying perceptual dimensions of stimuli [11,12].

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 [13–19]. 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 [5,20]. 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 [21,22].

Following previous theoretical work [16], 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 [23]. In the present study, we hypothesize that similarity judgments involve a type of implicit metacognition [16]. 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 quantified 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 assessed 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 proposed to use a psychophysical method to measure such discrimination capacity near perceptual threshold. We measured the participants' discrimination performance within the morph set that spans between the two stimuli (Figure 1C; near-threshold discrimination task). With this, we quantified 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 used 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 Supplementary 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 Supplementary 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 Supplementary Table 1.

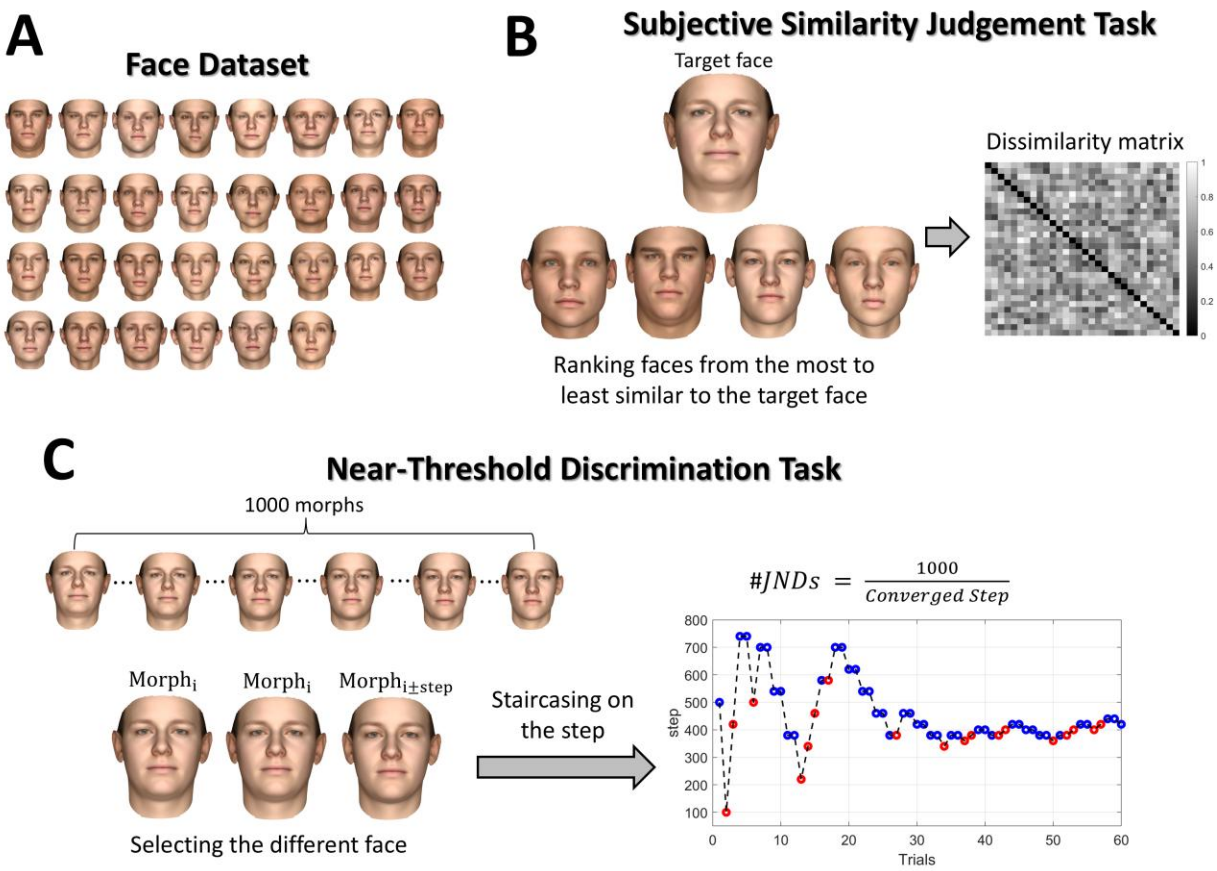


Figure 1. Experimental tasks for estimating subjective perceptual dissimilarity and perceptual discrimination capacity. (A) Illustration of 30 faces used in the present study. (B) The subjective similarity judgment task for estimating the level of subjective perceptual dissimilarity between face pairs. A target face on top and four other faces (candidates) on the bottom were shown to the participant in each trial. Participants were 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 was 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 were created as intermediate transitions between two faces (based on a computational face model; see Methods for details). In each trial, three faces were shown simultaneously to the participants. Two were identical, and one was different from the other two by a certain degree (number of morph steps within the 1000-morph series). Participants were instructed to click on the face that was different from the other two. Task difficulty was 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

Ethics information

The study received approval from the Ethics Review Committee at RIKEN, complying with all their ethical guidelines. Informed consent was obtained from participants before the experiment, and in appreciation of their participation, they were compensated with 3500 yen (approximately 23 US dollars) for each day of participation (roughly 90 minutes each day).

Design

Twelve participants were recruited for the study. They initially performed the subjective similarity judgment task, twice over the course of two days. After all participants completed this task twice, a set of 24 face pairs were 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 were 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 were 48 sessions in total, across 12 participants. Each participant performed four sessions on different days with each session taking more than 60 minutes. This provided us with enough data to perform our statistical analysis at the individual-level. The subjective similarity judgment task consisted of 330 efficiently crafted trials to estimate the level of subjective perceptual dissimilarity between all face pairs. Participants performed this task twice, and the achieved dissimilarity values were averaged to further enhance the robustness. The near-threshold discrimination task comprised 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 planned to recruit more participants if these initial 12 participants didn't satisfy our data collection stopping rule described in the Sampling section. However, since they met the stopping criterion, no further participants were recruited. As discussed in the Sampling section, this criterion ensured that the obtained statistics had sufficient precision to draw robust conclusions.

Note that this study was pre-registered and peer-reviewed at *Peer Community In (PCI)* as a Registered Report (Stage 1), with the preprint available at [24]. The study was conducted in accordance with the recommended Stage 1 protocol.

Face data set

The basal face model (BFM) [25] was 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 includes faces positioned both close to and far from each other within 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 were 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 were 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 disappeared from the screen, and the trial ended after all candidate faces were clicked one by one. If participants failed to complete the trial within 30 seconds, the trial was skipped, and any ranking assigned was disregarded. The aim of this task was 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 was estimated by calculating the probability of one face being ranked lower than the rest of the faces when the other face was the target, as outlined in the following. The rankings given in all trials were 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 was marked as the odd face. Subsequently, the dissimilarity value between a face pair was determined by calculating the ratio of instances where one of the faces was marked as the odd face across all sets that included the face pair with either of them as the target face. It is noteworthy that when calculating this ratio, we accounted 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 prevented 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 was strategically selected using the InfoTuple method [26]. This method guaranteed that each trial offered informative data, thereby enhancing the estimation of the dissimilarity matrix. This essentially enabled achieving a robust estimation of the dissimilarity matrix over a smaller number of trials. The trial selection procedure was similar to the one used by Canal et al. [26] and comprised the following steps:

1. The tuple set in the first 30 trials was selected at random while ensuring that each face was selected once as the target face.
2. The dissimilarity matrix was calculated as described above, and a 5-dimensional metric multidimensional scaling (MDS) [11] was applied to the dissimilarity matrix to find its embeddings.

3. A cycle of 30 trials, showcasing each face once as the target face, was 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 [26].

4. The dissimilarity matrix was calculated using all the data collected thus far, and the embeddings were updated by applying a 5Dl metric MDS to the dissimilarity matrix and using the previous embeddings as the seed in the MDS algorithm.

5. Steps 3 and 4 were repeated for 10 iterations. We stopped after 10 iterations as the dissimilarity matrix and embeddings were expected to reach a relatively stable state at this point. As completing 10 iterations plus the initial burn-in iteration is lengthy and can be exhausting, participants were given a short break after every three iterations.

Lastly, the embeddings obtained from applying a 5D metric MDS to the dissimilarity matrix were then used to recalculate the dissimilarity matrix by computing the Euclidean distances between the embedded face points. This process of reducing dimensionality to 5D 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, was utilized in all the subsequent stages instead of the original dissimilarity matrix. Please see Supplementary Figure 1 for a schematic overview of the task design described above.

We also tested a 2D metric MDS for recalculating the dissimilarity matrix. Even though there might be significant information loss in using such a low dimensional MDS, it may further refine the matrix and also make the dissimilarity relations more apparent (i.e., more distinct across pairs). Thus, we thought that focusing on only two major components of the dissimilarity relations may potentially make their association with perceptual discrimination capacity more salient. This prediction was based on our pilot data, in which both Hypothesis 1 and 2 reached a higher significance level (an individual-level $p<0.05$ in all four participants) when a 2D instead of a 5D MDS was used.

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

It is important to mention that each participant performed the above task twice, each time on a different day. The average of the dissimilarity matrices obtained from each day formed the final dissimilarity matrix. Further, to evaluate the reliability of the obtained dissimilarity matrix from a session, we assessed the within-participant correlation between the dissimilarity matrices derived from each day. As a reference, we compared it with the distribution of between-participant correlation computed by randomly correlating a dissimilarity matrix from a session in one participant with that of another participant. We expected 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 used a machine learning approach as an exploratory analysis (i.e., we planned not to 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:

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

$$L = -\frac{1}{N}\sum_{i=1}^{N} log(p_i)\ +\ \lambda_1 \sum_{m=1}^{M} \|x_m\|_1 +\ \lambda_2 \sum_{m=1}^{M} \|x_m\|_2{}^2$$

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 all 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 (i.e., suppressing unnecessary dimensions) and the scale of the embeddings, respectively. We used the Keras library in Python, with Adam optimizer, to minimize the loss function.

We note that we didn't use this machine learning approach during the task (i.e., in our online application) because it was slower and also required cross-validations to select appropriate hyperparameters. This approach was more sophisticated than our main approach, which simply involved estimating ranking probabilities and applying MDS to reduce dimensionality and thereby reduce noise, and it could yield a better estimation of the embeddings and, consequently, the dissimilarity matrix. In our pilot study, using this machine learning approach, we got similar results to those shown in Supplementary Figures 3 and 4 obtained with the main pre-registered approach.

The λ_1 and λ_2 values that effectively reduce embeddings dimensionally and scale, while maintaining accuracy on a validation set, are considered optimal as they result in more generalizable, thus more robust embeddings. The λ_1 and λ_2 were fixed at 0.00025, as they fell within the optimal range for all session data, identified through grid search sampling of λ_1 and λ_2

values. The embeddings were initialized with 10 dimensions. However, the optimal λ_1 and λ_2 resulted in embeddings with 3 to 5 effective dimensions across all sessions, indicating that no more than five dimensions were required to represent how participants perceived the similarity among the 30 faces used in the study.

Near-threshold discrimination task

The objective of this task was to estimate perceptual discrimination capacity in a face pair. A series of 1000 equally spaced morphs were generated along the line connecting a face pair in the BFM space. In each trial, three faces were 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 were displayed simultaneously at the center of the screen next to each other, and their arrangement was randomized in each trial (Figure 1C). Participants were instructed to identify and click on the different face.

A staircase [27] with a 1-up and 2-down protocol was 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 was increased, and it was decreased after two consecutive correct trials. The magnitude of the change in the step value gradually decreased over trials, reaching a minimum change of 20 steps. The task was 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 was defined as the converged step. Essentially, a small JND, for example, 125, indicates that the two questioned faces were quite distinct, involving 8 JNDs (i.e., 1000 divided by 125) between them. We used 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 was simply calculated as 1000 (i.e., total number of morphs) divided by the JND.

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

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 were 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) limited us to examine only a small subset of the pairs. We expected a sample size of 24 pairs to be fairly adequate for detecting an effect, and it was further justified by our pilot study (Supplementary Figure 3), where we achieved reasonably robust results by examining only 13 pairs. Participants were re-invited for two more 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, involved running a near-threshold discrimination task dedicated to that specific pair. Thus, considering practical constraints as described earlier, we had no choice but to examine only a limited subset of pairs (24 out of 435 possible pairs). However, this subset was 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 were 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 measuring the dissimilarity values and the #JNDs, any effect should be better detectable on pairs with larger deviations, those that are more distinct, across participants. So, we selected 18 pairs with controversial dissimilarity values across participants, and for the sake of comparison, we selected 6 pairs with less controversial values.

First, the dissimilarity matrix was z-normalized within each participant to ensure that its scale was consistent across participants. Subsequently, the median and the median absolute deviation (MAD) of the dissimilarity matrix were computed across participants, and the quantiles of the median values were derived. Within the first and the last quantiles, 3 pairs with the highest and 1 with the lowest MAD were selected. Additionally, 6 pairs with the highest and 2 pairs with the lowest MAD were chosen within the second and the third quantiles. This systematic selection ensured choosing 18 controversial and 6 less-controversial pairs that covered a diverse range in the group-averaged dissimilarity matrix.

Sampling

A set of criteria were pre-registered to exclude participants from the analysis. However, none of the participants met any of the exclusion criteria. The criteria were as follows: Those who didn't

complete all four experimental sessions and those who showed a lack of attentiveness to the task in any of the sessions. The lack of attentiveness in the near-threshold discrimination task was identified by non-converging staircases, indicated by a non-fluctuating increment in the step value over trials. Specifically, a session in which there were more than 4 (out of 12) staircases with less than three downs in their last 20 trials was considered bad, lacking sufficient attentiveness. None of the sessions from any participants met this criterion. In the subjective similarity judgment task, the lack of attentiveness was 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 fell below 0.2, the session was considered bad with inadequate attentiveness. This correlation ranged between 0.50 to 0.78 (mean$\pm$SD$=0.65\pm0.08$) in all sessions, indicating good data quality. Before performing the main analysis (see the Analysis section), we planned to continue the data collection until we had 12 participants who successfully completed the experiment without meeting any of the exclusion criteria. As none of our initial 12 participants met the exclusion criteria, we proceeded with the main analysis.

Furthermore, a data collection stopping rule was pre-registered to ensure that the analyzed data from our initial 12 participants would be sufficient to draw robust conclusions. We planned to continue recruiting new participants until the criterion was met, with a maximum cap of 24 participants; no more than 24 participants would be recruited (Supplementary Figure 2). However, as the criterion was satisfied with our initial 12 participants, no additional participants were recruited, and the data collection was concluded. The criterion was as follows: the individual-level statistic was converted to z-values, and the 95% confidence interval of the group-mean z-value was derived (See the Analysis section). Data collection would be stopped if, in both Hypothesis 1 and 2, the width of this 95% confidence interval was less than 1. Moreover, we considered a hypothesis confirmed, if the group-mean z-value was significantly above zero, specifically, if the 95% confidence interval was 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 width). So that we are confident the effect is not being confirmed or rejected because of some extreme observations [28,29]. Given our sample size scale, we expected a considerable effect to have a group-mean z-value of at least above 0.5. So, a minimally significant scenario would involve 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 to safely reject or accept a hypothesis.

Analysis

Hypothesis 1

Spearman correlation coefficient and its p-value were computed between the dissimilarity values and #JNDs of the examined 24 pairs in each individual. The Spearman correlations were converted to z-values using the Fisher z-transformation [30] to conduct group-level statistical tests. The distribution of the group-mean z-value was computed by bootstrapping, iterated 100,000

times, and then its 95% confidence interval was derived by obtaining the 2.5th and 97.5th percentile of the distribution. The hypothesis was considered confirmed if this confidence interval was above zero. The following statistics were also computed 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 [31], and a Bayesian posterior distribution of population prevalence [32] 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 the likely prevalence of observing a particular effect in the population, based on the limited number of participants tested in the study and their proportion showing the effect significantly. Furthermore, we planned to apply an equivalence test [33], in case of a null finding, considering a z-value range of -0.5 to 0.5 as the smallest effect size of interest, to assess the significance of the null result (i.e., how confidently the effect can be considered non-significant). However, given that no null results were found, this test was not applied.

Hypothesis 2

Each participant's #JNDs was z-normalized to ensure that the #JNDs range was consistent across participants. This normalization was crucial, given that some participants exhibited generally higher #JNDs than others. Subsequently, a nonparametric permutation test was 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 was 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 simulated a new participant by mixing the existing participants.
2. The Spearman correlation coefficient was 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 were an enormous number of possible permutations (i.e., 11^24 unique permutations), which ensured the construction of a reliable null distribution.
3. Steps 1 and 2 were 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 was tested against the null distribution, and the p-value, indicating the significance level, was derived. The z-value was 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 was derived through bootstrapping, and if it was above zero, the hypothesis was considered confirmed. The complementary statistics, outlined in Hypothesis 1, were also computed for Hypothesis 2.

Pilot data

We conducted a pilot version of the study with four participants prior to pre-registration (i.e., publishing the Stage 1 registered report). The experimental design was similar to the main 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 main 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 (Supplementary Figures 3A and 3B), 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 (Supplementary Figures 3C and 3D), suggesting that the association between subjective perceptual dissimilarity and perceptual discrimination capacity is specific to each individual. 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.

We further explored the correlation between subjective perceptual dissimilarity and perceptual discrimination capacity in each face pair across participants in our pilot study (Supplementary Figure 4). 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, this also indicated that one's perceptual discrimination capacity can

explain one’s subjective perceptual dissimilarity. Similar inter-individual differences observable in subjective perceptual dissimilarity could also be found in perceptual discrimination capacity.

Results

Quality of the measures

Perceptual dissimilarity

Participants performed the subjective similarity judgment task twice on two different days. Similarity judgments are expected to be relatively consistent within individuals. If, however, similarity judgments are not made consistently, changing from one session to the next, then this would automatically reject our hypothesis that similarity judgments are made normatively based on one's perceptual capacities. Therefore, it was important to examine the consistency between participants' first and second sessions of the subjective similarity judgment task. To this end, within-participant correlations—the Spearman correlation between the dissimilarity matrices derived from the first and second sessions of the same participants—were computed. These were then evaluated against the distribution of between-participant correlation obtained by randomly correlating a dissimilarity matrix from a session in one participant with that of another participant. As shown in Figure 2, the within-participant correlations were significantly higher than the between-participant correlations. The z-value of the within-participant correlation relative to the distribution of the between-participant correlation was highly above zero in all participants, with most (ten out of twelve participants) surpassing 2.3. This indicates that similarity judgments were consistent within each individual and also unique to them.

Furthermore, given the self-consistency of similarity judgments, the within-participant correlation could reflect the performance of the main and alternatively pre-registered approaches in robustly estimating the dissimilarity matrix from the collected similarity ranking data. The within-participant correlation was 0.79 ± 0.071 (mean $\pm$ SD) when the alternatively pre-registered machine learning approach was used, whereas it was 0.72 ± 0.064 (mean $\pm$ SD) when the main pre-registered approach was used (see Supplementary Figure 5). The within-participant correlation was significantly higher (signed-rank test p-value<0.001) when the machine learning approach was used, indicating its better performance. It is noteworthy that the machine learning approach, even when using non-optimal hyperparameters, still outperformed the main pre-registered approach, further highlighting its superiority.

#JNDs

The #JNDs between face pairs were measured once for a participant, so their robustness cannot be evaluated in the same way as the perceptual dissimilarity above. However, given that the #JNDs, unlike the dissimilarity, is considered an objective measure (meaning it is not expected to change with factors like the participant's state of mind), its robustness can be assessed by the stability of its corresponding staircase, even with a single run. The ratio of reversals within the last 20 trials was about 0.3 ± 0.08 (mean $\pm$ SD) across participants in the examined face pairs staircases (see Supplementary Figure 6). Given the 1-up and 2-down protocol used, the maximum possible

reversal ratio is 0.6, meaning that each “up” is followed by a “down” and vice versa. A ratio of 0.3 (i.e., 30% of the trials involved a reversal within the last 20 trials) indicates fluctuations about half as much as the maximum. As a reference, the example staircase shown in Figure 1C has a reversal ratio of 0.35. These suggest that the staircases converged fairly well across participants. Therefore, the measured #JNDs were reasonably robust.

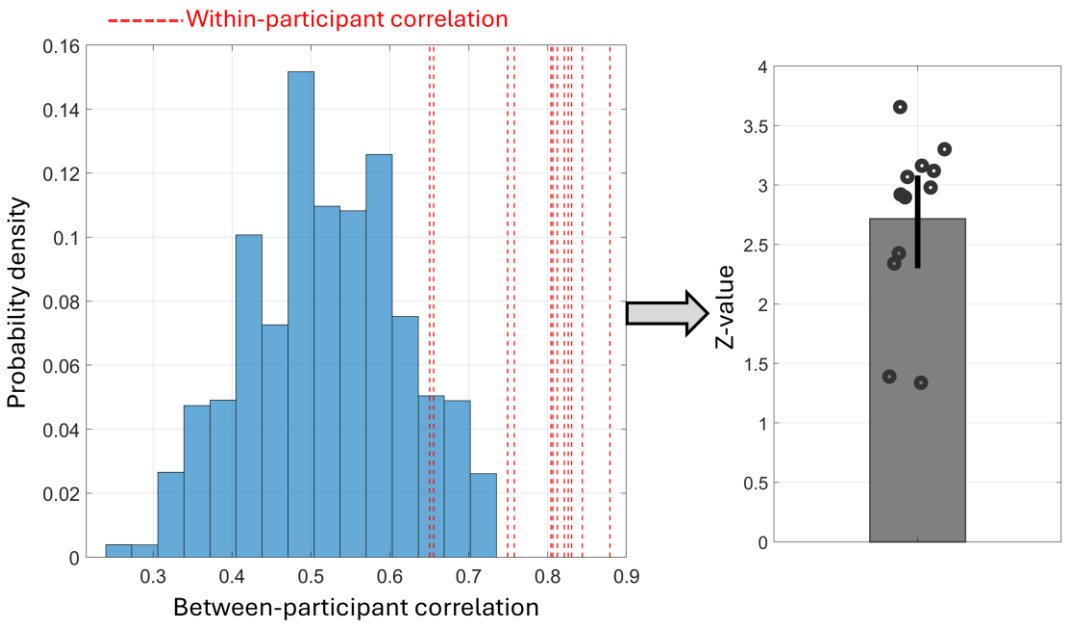


Figure 2. Self-consistency and uniqueness of the obtained 30x30 perceptual dissimilarity matrices (see Figure 1B) across participants. The left panel shows the histogram of the between-participant correlations computed by randomly correlating the dissimilarity matrix derived from one session of a participant with that of another participant. The dotted red lines indicate the within-participant correlation, representing the correlation between the dissimilarity matrices derived from the first and second sessions of the same participant. The right panel shows the z-values of the within-participant correlations computed relative to the between-participant correlation histogram in the left panel. The bar indicates the mean z-value, the error bar shows its 95% confidence interval, and each dot on the plot corresponds to a participant. The dissimilarity matrices were highly consistent within each individual and unique to them.

Perceptual discrimination capacity (#JNDs) and subjective perceptual dissimilarity

We found a strong positive association between individuals’ perceptual discrimination capacity and their subjective perceptual dissimilarity (Figure 3). Face pairs that were harder to discern in the near-threshold task were subjectively perceived as more similar in the (supra-threshold) similarity judgment task. This association was highly significant at the group level. The mean z-value was 2.63, with a 95% confidence interval of [2.18, 3.10]. Other complementary statistical tests also indicated high significance: a t-test on the z-values yielded a p-value<0.00001 and a BF>10000, and

a Fisher's combined probability test resulted in a p-value<0.00001. Furthermore, significance was reached at the individual-level in most (ten) participants. This equates to a Bayesian population prevalence of 0.82 with a 95% highest posterior density interval of [0.56, 0.97], indicating that the true likelihood of observing an individual-level significance, following our experimental design, lies between 56% and 97%.

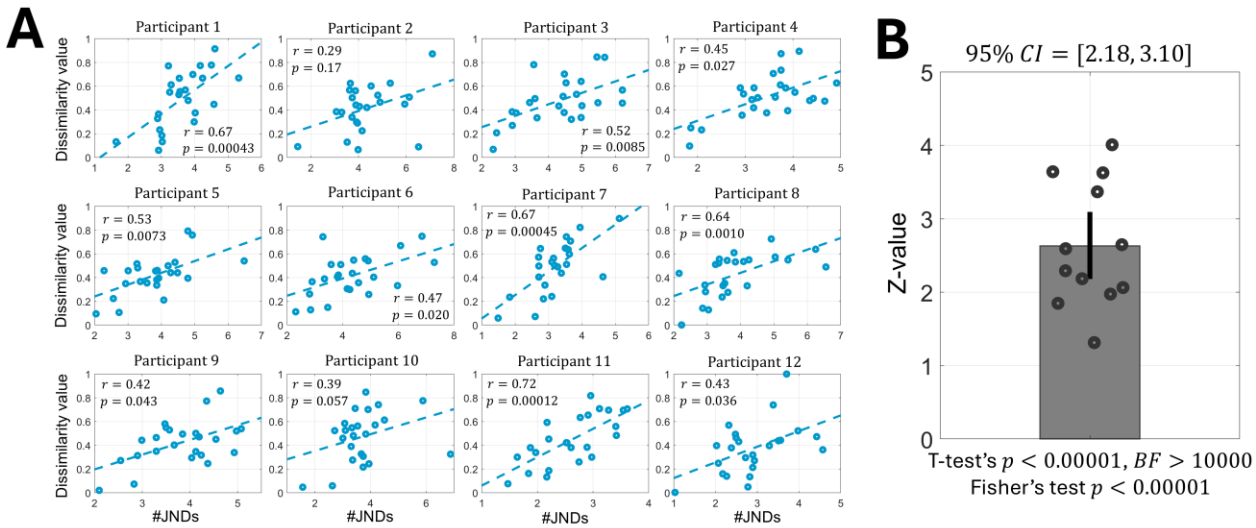


Figure 3. Relationship between perceptual discrimination capacity and subjective perceptual dissimilarity in each participant. (A) Each subplot illustrates the relation between perceptual discrimination capacity, quantified in #JNDs (x-axis), and subjective perceptual dissimilarity (y-axis) in each participant, 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 in A were transformed to z-values for group-level hypothesis testing. The bar shows the mean z-value across participants, the error bar represents the 95% confidence interval of this 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 smaller than 0.00001 and a Bayes factor (BF) above 10000. Additionally, employing a Fisher's combined probability test, combining the individual-level p-values, similarly resulted in a p-value smaller than 0.00001. The results suggest a strong positive association between perceptual discrimination capacity and subjective perceptual dissimilarity at the group-level, confirming Hypothesis 1, with significance reached in most (ten) participants at the individual-level.

Importantly, we discovered that this association between perceptual discrimination capacity and subjective perceptual dissimilarity was quite specific to each individual. As depicted in Figure 4, one's subjective perceptual dissimilarity was better correlated with one's own perceptual discrimination capacity than with that of others. This finding was significant at the group level, with a group-mean z-value of 1.19 and a 95 confidence interval of [0.77, 1.60]. Additional complementary statistics further highlighted the high significance of the result: a t-test on the z-values resulted in a p-value of 0.00021 and a BF of 151, and a Fisher's combined probability test

yielded a p-value of 0.000062. At the individual-level, this finding reached significance in five participants, equating to a Bayesian population prevalence highest posterior density interval of [0.14, 0.66]. This implies that given our experimental design and factors such as measurement noise, the true likelihood of observing an individual-level significance lies between 14% and 66%.

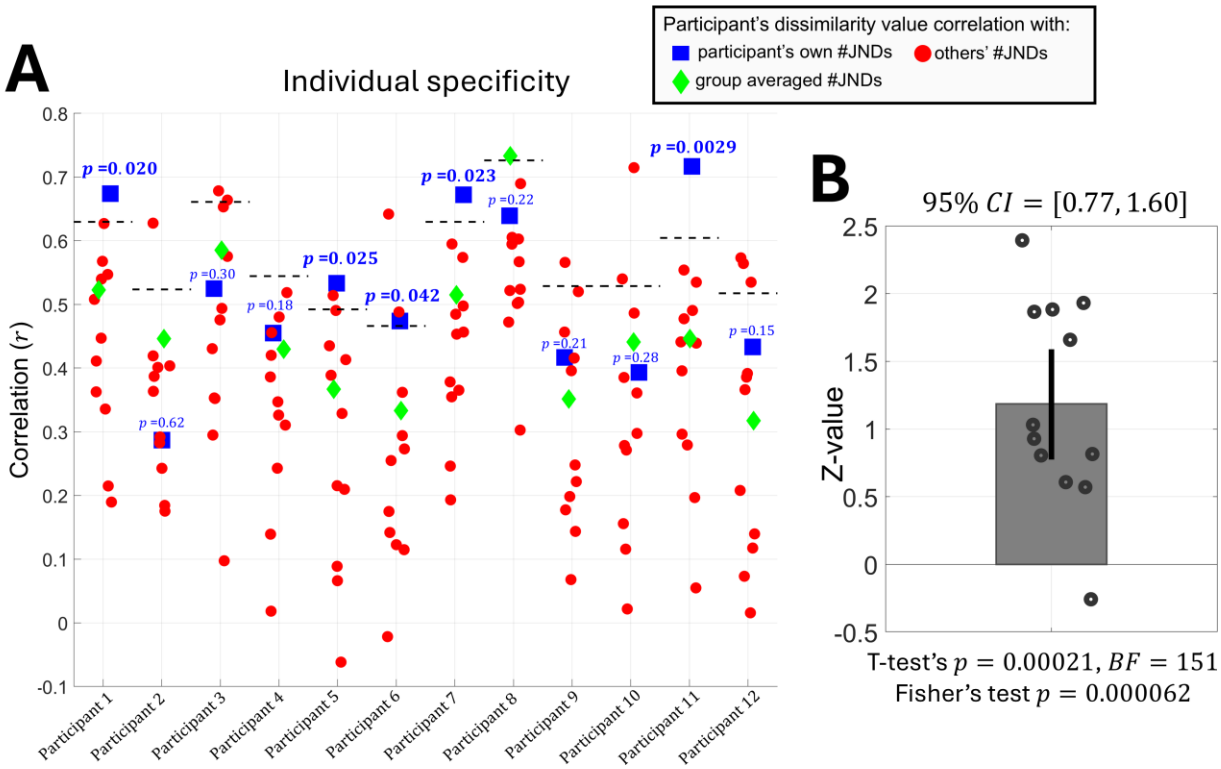


Figure 4. Individual-specificity of the relationship between perceptual discrimination capacity and subjective perceptual dissimilarity. (A) The blue square, red circle, and green diamond indicate the Spearman correlation coefficient between the dissimilarity value in each participant and their own #JNDs, other participants' #JNDs, and the group averaged #JNDs (excluding the participant in question), respectively. The dotted horizontal black line denotes the Spearman correlation value at which a permutation test yielded a p-value of 0.05, rejecting a null hypothesis of no individual-specificity. The null hypothesis distribution consisted of possible Spearman correlations between the dissimilarity value in the target participant and others' #JNDs. The p-value rejecting the null hypothesis in each participant is shown in blue at the top of the blue squares. (B) Individuals' specificity statistics in A were converted to z-values for group-level hypothesis testing. The bar shows the group-mean z-value, the error bar indicates its 95% confidence interval, and each dot corresponds to a participant. The confidence interval was above zero, indicating group-level significance. This demonstrates that the association between perceptual discrimination capacity and subjective perceptual dissimilarity is quite specific to each individual, confirming Hypothesis 2, with significance reached at the individual-level in five participants. The Other complementary statistics, similar to Figure 3B, are displayed below the bar plot, further highlighting the strong group-level significance.

Figure 5 illustrates the association between perceptual discrimination capacity and subjective perceptual dissimilarity in each face pair across participants. This association was positive in most pairs and particularly strong in eight pairs: Three showed a Spearman correlation above 0.58 (p-value<0.05), and five others showed a correlation above 0.50 (p-value<0.1, equivalent to a one-tailed p-value<0.05, pre-assuming only positive associations are meaningful). A t-test on the positivity of the association across pairs yielded a p-value of 0.00021 and a BF of 140. These observations further support our hypotheses, indicating that individuals with a lower capacity in discriminating a pair tended to judge that pair as more similar.

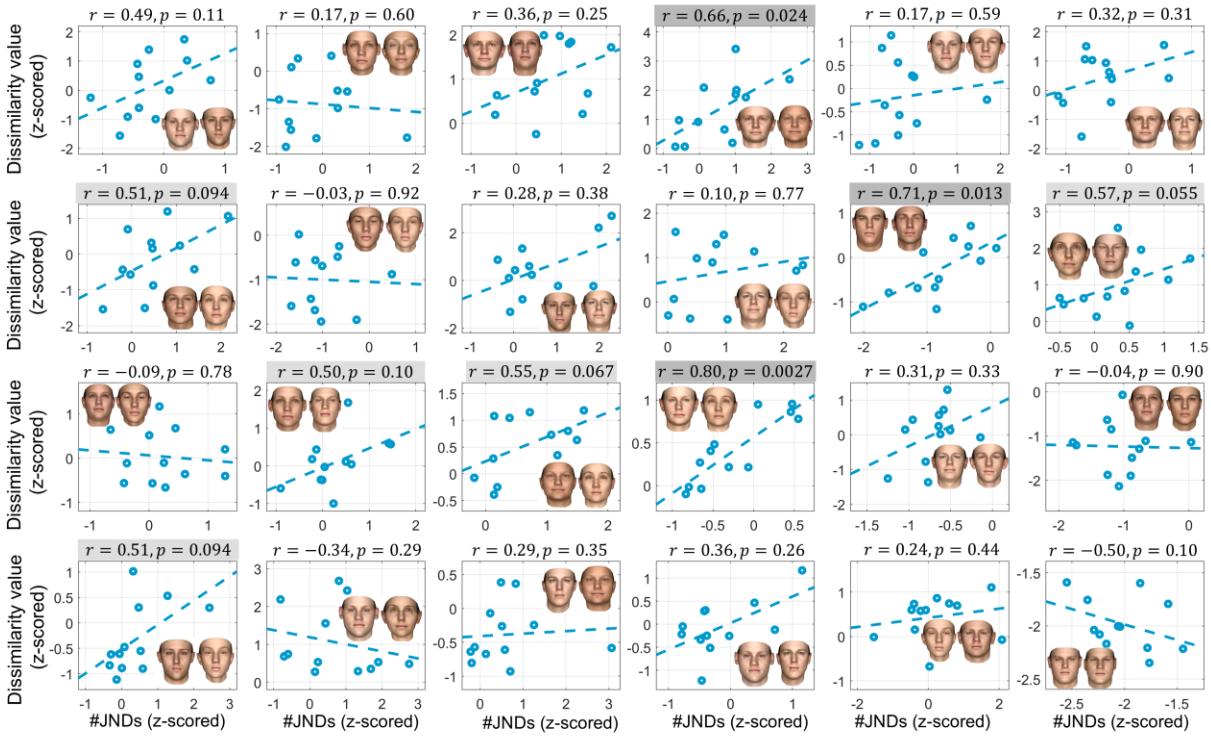


Figure 5. Relationship between perceptual discrimination capacity and subjective perceptual dissimilarity across participants in each face pair. Each subplot represents the association between the #JNDs and dissimilarity value across participants in a given face pair, with each dot corresponding to a participant. The dissimilarity value and #JNDs were z-scored within each participant to ensure a consistent scaling across participants. r and p displayed on top of each subplot indicate the Spearman correlation coefficient and its associated p-value, respectively. Pairs with a $p < 0.05$ are highlighted in gray, and those with a $p < 0.1$ (i.e., equivalent to a one-tailed $p < 0.05$) in a lighter gray. The subplots are organized such that the face pair with the most controversial dissimilarity value across participants (measured by the mean absolute deviation) is shown in the top left, and the pair with the least controversial value is shown in the bottom right. A positive association between perceptual discrimination capacity and subjective perceptual dissimilarity was observed in most pairs, indicating that participants with a lower capacity in discriminating a pair tended to judge that pair as more similar, which supports our hypotheses. A t-test on the r value being positive across pairs yielded a p-value of 0.00021 and a BF of 140. It is noteworthy that

this positive association was more salient in the more controversial pairs (e.g., compare the first three rows with the last row).

Moreover, the association was more apparent in the controversial pairs (i.e., pairs with a controversial degree of perceptual dissimilarity across participants) than in the less controversial pairs. As shown in Figure 5, the weakest associations were observed in the less controversial pairs (see the last row in this figure). This was expected as the psychophysiological measures (i.e., #JNDs and perceptual dissimilarity) inevitably contain noise, and such noise can obscure the detection of true relationships that may exist across participants, particularly when there is little variability across them (i.e., less signal). Despite this intuitive expectation, six less controversial pairs were intentionally examined in the study to assess whether the hypothesized relationship could still be observed in them, given the measurement noise. The results suggest that while our current design is precise enough to detect the individual-specific association between perceptual discrimination capacity and subjective perceptual dissimilarity in controversial pairs, it may not be precise enough when the pairs are less controversial, assuming our hypothesis holds true in all pairs. We measured #JNDs once, and perceptual dissimilarity twice in two sessions, with them then averaged. Therefore, conducting further repeated measures, especially for #JNDs, seems necessary to capture this association across all pairs.

As discussed earlier in the "Quality of the measures" section, the alternatively pre-registered machine learning approach outperformed the main pre-registered approach in estimating the dissimilarity matrices. Therefore, we decided to present the results using the dissimilarity matrices derived by the machine learning approach in the main figures (Figures 2 to 5). The results using the dissimilarity matrices from the main pre-registered approach are instead shown in the supplementary figures (See Supplementary Figures 7 to 10). However, we adhered to our pre-registration and exclusively used the main pre-registered approach at all stages, including selecting the face pairs, examining the data collection stopping criterion, and testing our hypotheses. Nevertheless, there were no discrepancies in the results; the stopping criterion was met, and the hypotheses were confirmed using both approaches. The only difference was that the results were stronger with the machine learning approach, which can be attributed to its more accurate estimation of the dissimilarity matrices.

We also tested the use of a 2D MDS instead of a 5D MDS in our main pre-registered approach. We hoped that a 2D MDS, by representing the dissimilarity relations in only two dimensions, would preserve the most robust aspects of these relations, thereby making their association with perceptual discrimination capacity more salient. However, as shown in Supplementary Figure 11, this was not the case. Although the association, as well as its individual specificity, were significant, they became weaker when a 2D MDS was used. Not only that, the within-participant correlation of the dissimilarity matrices, along with their individual uniqueness (evaluated by comparison to the between-participant correlation), also weakened. Therefore, these suggest that reducing dimensionality further to two dimensions caused more harm (information loss) than benefit (e.g., noise reduction).

Discussion

In the present study, we used a near-threshold psychophysical task to quantify perceptual discrimination capacity, which indicates one's capability to distinguish two stimuli (Figure 1C). We examined 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 explored 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 first conducted a small pilot study which provided initial support for both hypotheses (see Supplementary Figure 3). Then, the results from our main study robustly confirmed them (Hypothesis 1 in Figure 3 and Hypothesis 2 in Figure 4).

Our results may suggest that subjective similarity judgment is, in a specific sense, metacognitive: it reflects and predicts one's own perceptual capacity. In essence, perceptual capacities may 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 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 [23,34,35]. 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 [15]. 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.

These results could also 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 [36,37]. If the qualitative characters of conscious perception are determined by the relevant similarity relations, as some researchers assume [13–19], does it mean that these artificial agents are conscious [19]? Or, at least, does it mean that they contain the essential information that is encapsulated within human perceptual experiences? The answer is probably no, considering the metacognitive perspective described above. That is, for the similarity judgment to be relevant for subjective experiences, according to our findings, 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.

It is noteworthy that similarity relations are commonly represented in a metric psychological space that includes underlying perceptual dimensions [10–12]. Our results suggest that the scale and weight of these dimensions are specific to each individual and tied to their perceptual capacity. Individuals with a greater capacity to perceive particular dimensions (i.e., better at noticing fine details along those dimensions or their combinations) may have an expanded scale for those dimensions (i.e., further stretched) in their psychological space. This could make those dimensions more influential in the similarity judgment of those individuals. Furthermore, as famously formulated by Roger Shepard, generalization exponentially decays as the distance increases in the psychological space [7]. Thus, higher perceptual capacity, by increasing the psychological distance, could cause a stronger reduction in generalization—a steeper exponential decay. Alternatively, similarity judgment between two stimuli has been conceptualized by Amos Tversky as a function of their common features and the features that are unique to each (i.e., distinct from the other) [2]. More common features increase perceptual similarity, while more distinct features reduce it. In this framework, higher perceptual capacity could presumably enhance the identification of both shared and distinct features. However, our observation suggests that the impact of a higher perceptual capacity was more pronounced on distinct features, leading to a reduced perceived similarity.

Finally, we note that we used face stimuli in our study because they are high-dimensional and commonly encountered in daily life. However, it may be necessary to examine how our metacognitive account holds in other types of visual stimuli, such as colors or artificially constructed stimuli (e.g., Greebles; [38]), and even across different sensory domains [39]. Moreover, future studies could use multi-featural stimuli that, unlike faces, do not generally rely on holistic perception, such as stimuli with varying shapes, sizes, and colors, to assess how perceptual capacity within each individual feature influences similarity judgment.

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.

Acknowledgments

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.

Declaration of interests

The authors declare no competing interests.

Data and code availability

The data and the code used in the study are publicly accessible from the GitHub repository below:

https://github.com/AliMoharramipour/Subjective-Dissimilarity-and-Discrimination-Capacity-

https://doi.org/10.5281/zenodo.15166140

References

1. Shepard, R.N. (1964). Attention and the metric structure of the stimulus space. J. Math. Psychol. *1*, 54–87.
2. Tversky, A. (1977). Features of similarity. Psychol. Rev. *84*, 327.
3. Smith, L.B. (1989). A model of perceptual classification in children and adults. Psychol. Rev. *96*, 125.
4. Smith, L.B., and Heise, D. (1992). Perceptual similarity and conceptual structure. In Advances in psychology (Elsevier), pp. 233–272.
5. Medin, D.L., Goldstone, R.L., and Gentner, D. (1993). Respects for similarity. Psychol. Rev. *100*, 254.
6. Nosofsky, R.M. (1984). Choice, similarity, and the context theory of classification. J. Exp. Psychol. Learn. Mem. Cogn. *10*, 104.
7. Shepard, R.N. (1987). Toward a universal law of generalization for psychological science. Science *237*, 1317–1323.
8. Tenenbaum, J.B., and Griffiths, T.L. (2001). Generalization, similarity, and Bayesian inference. Behav. Brain Sci. *24*, 629–640.
9. Sims, C.R. (2018). Efficient coding explains the universal law of generalization in human perception. Science *360*, 652–656.
10. Schurgin, M.W., Wixted, J.T., and Brady, T.F. (2020). Psychophysical scaling reveals a unified theory of visual memory strength. Nat. Hum. Behav. *4*, 1156–1172.
11. Borg, I., and Groenen, P.J. (2005). Modern multidimensional scaling: Theory and applications (Springer Science & Business Media).
12. Hebart, M.N., Zheng, C.Y., Pereira, F., and Baker, C.I. (2020). Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. Nat. Hum. Behav. *4*, 1173–1185.
13. Clark, A. (2000). A theory of sentience (Clarendon press).
14. Rosenthal, D. (2010). How to think about mental qualities. Philos. Issues *20*, 368–393.
15. Malach, R. (2021). Local neuronal relational structures underlying the contents of human conscious experience. Neurosci. Conscious. *2021*, niab028.
16. Lau, H., Michel, M., LeDoux, J.E., and Fleming, S.M. (2022). The mnemonic basis of subjective experience. Nat. Rev. Psychol. *1*, 479–488.
17. Tallon-Baudry, C. (2022). The topological space of subjective experience. Trends Cogn. Sci. *26*, 1068–1069.

18. Zeleznikow-Johnston, A., Aizawa, Y., Yamada, M., and Tsuchiya, N. (2023). Are color experiences the same across the visual field? J. Cogn. Neurosci. *35*, 509–542.

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

20. Goodman, N. (1972). Seven strictures on similarity.

21. Rao, R.P. (1999). An optimal estimation approach to visual perception and learning. Vision Res. *39*, 1963–1989.

22. Shen, S., and Ma, W.J. (2016). A detailed comparison of optimality and simplicity in perceptual decision making. Psychol. Rev. *123*, 452.

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

24. Moharramipour, A., Zhou, W., Rahnev, D., and Lau, H. (2024). Is subjective perceptual similarity metacognitive? bioRxiv, 2024.06.13.598769. https://doi.org/10.1101/2024.06.13.598769.

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

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

27. Cornsweet, T.N. (1962). The staircase-method in psychophysics. Am. J. Psychol. *75*, 485–491.

28. Cumming, G. (2008). Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better. Perspect. Psychol. Sci. *3*, 286–300.

29. Lakens, D. (2014). Performing high-powered studies efficiently with sequential analyses. Eur. J. Soc. Psychol. *44*, 701–710.

30. Fieller, E.C., Hartley, H.O., and Pearson, E.S. (1957). Tests for rank correlation coefficients. I. Biometrika *44*, 470–481.

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

32. Ince, R.A., Paton, A.T., Kay, J.W., and Schyns, P.G. (2021). Bayesian inference of population prevalence. Elife *10*, e62461.

33. Lakens, D., Scheel, A.M., and Isager, P.M. (2018). Equivalence testing for psychological research: A tutorial. Adv. Methods Pract. Psychol. Sci. *1*, 259–269.

34. McCurdy, L.Y., Maniscalco, B., Metcalfe, J., Liu, K.Y., De Lange, F.P., and Lau, H. (2013). Anatomical coupling between distinct metacognitive systems for memory and visual perception. J. Neurosci. *33*, 1897–1906.

35. Fleming, S.M., Ryu, J., Golfinos, J.G., and Blackmon, K.E. (2014). Domain-specific impairment in metacognitive accuracy following anterior prefrontal lesions. Brain *137*, 2811–2822.

36. Kawakita, G., Zeleznikow-Johnston, A., Tsuchiya, N., and Oizumi, M. (2023). Comparing color similarity structures between humans and LLMs via unsupervised alignment. ArXiv Prepr. ArXiv230804381.

37. Marjieh, R., Sucholutsky, I., van Rijn, P., Jacoby, N., and Griffiths, T.L. (2023). Large language models predict human sensory judgments across six modalities. ArXiv Prepr. ArXiv230201308.

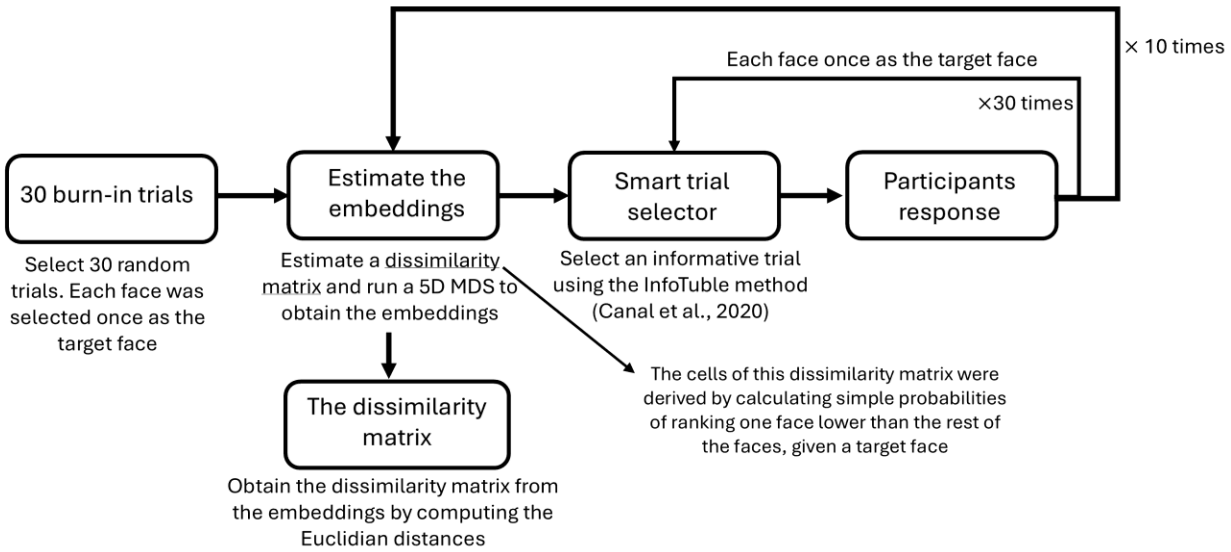
38. Gauthier, I., and Tarr, M.J. (1997). Becoming a “Greeble” expert: Exploring mechanisms for face recognition. Vision Res. *37*, 1673–1682.

39. Faivre, N., Filevich, E., Solovey, G., Kühn, S., and Blanke, O. (2018). Behavioral, modeling, and electrophysiological evidence for supramodality in human metacognition. J. Neurosci. *38*, 263–277.

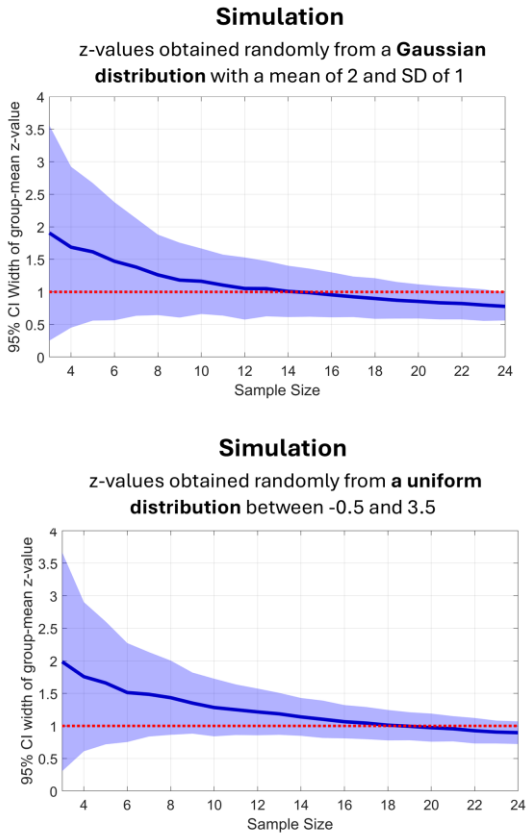
Supplementary Table 1 Experimental Design Table

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 were measured in each participant through two different psychophysical tasks: the subjective similarity judgment task (Figure 1B) and the near-threshold discrimination task (Figure 1C). The subjective similarity judgment task assessed the level of subjective dissimilarity (dissimilarity value) between stimulus pairs. The near-threshold discrimination task measured 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 were 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. We planned to exclude participants failing to complete all four sessions and those displaying a lack of attentiveness to the task in any of the sessions. The criteria for a lack of attentiveness are described in the method section. However, no participants were excluded as none met these exclusion criteria. We further planned to recruit more participants until a data collection stopping criterion was satisfied, with a maximum cap of 24 participants. However, as we met the stopping criterion with our initial 12 participants, no new participants were recruited. The stopping criterion involved achieving a 95% confidence interval of the group-mean z-value (see the method section) with a width of less than 1. Furthermore, the hypothesis was considered confirmed if this 95% confidence interval was above zero. * The hypothesis was confirmed (see Figure 3)	In each participant, the Spearman correlation between the #JNDs and dissimilarity values was computed. Then, the Spearman correlations were transformed into z-values using the Fisher z-transformation. Subsequently, the 95% confidence interval of the group-mean z-value was 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 others' 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 was measured by a nonparametric permutation test, which assessed whether one's own dissimilarity value was more strongly correlated with one's own #JNDs than with others' #JNDs. Essentially, the above test reflects how specific this relationship was in each individual in terms of z-values.	The participants' sampling and the data collection stopping strategies were the same as above. Similar to the above hypothesis, we considered the hypothesis confirmed if the 95% confidence interval of the group-mean z-value was above zero. We also planned to recruit participants until we achieved a 95% confidence interval with a width of less than 1, considering a maximum cap of 24 participants. * The hypothesis was confirmed, and the experiment stopping criterion was met (see Figure 4).	First, the #JNDs was z-normalized within each individual to ensure that its scale was consistent across participants, and then a nonparametric permutation test was applied to each participant separately as follows: Briefly, a permutation #JNDs set was constructed by randomly permuting the #JNDs across all participants, excluding the participant in question. The Spearman coefficient was then computed between the permuted #JNDs set and the dissimilarity values of the participant in question. This process was 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 was compared against the null distribution, and its z-value was computed. Subsequently, the 95% confidence interval of the group-mean z-value was 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.

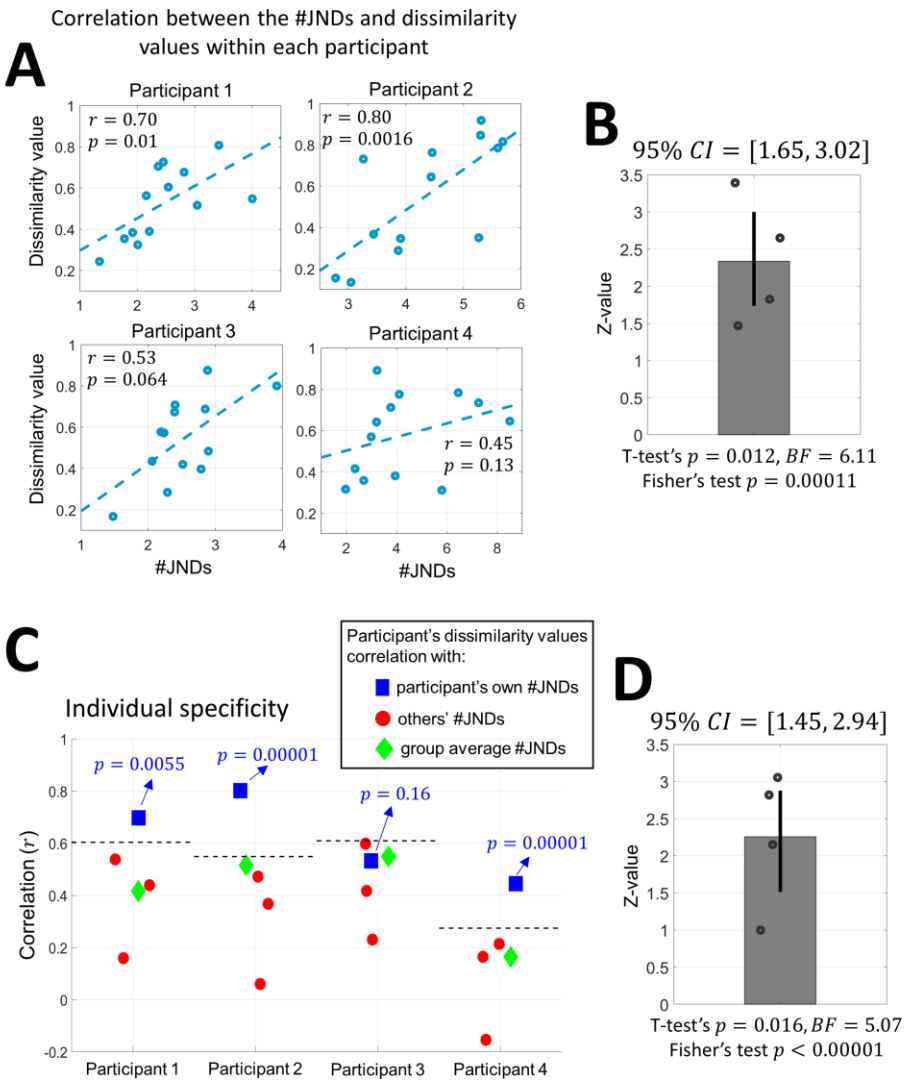
Supplementary Figures



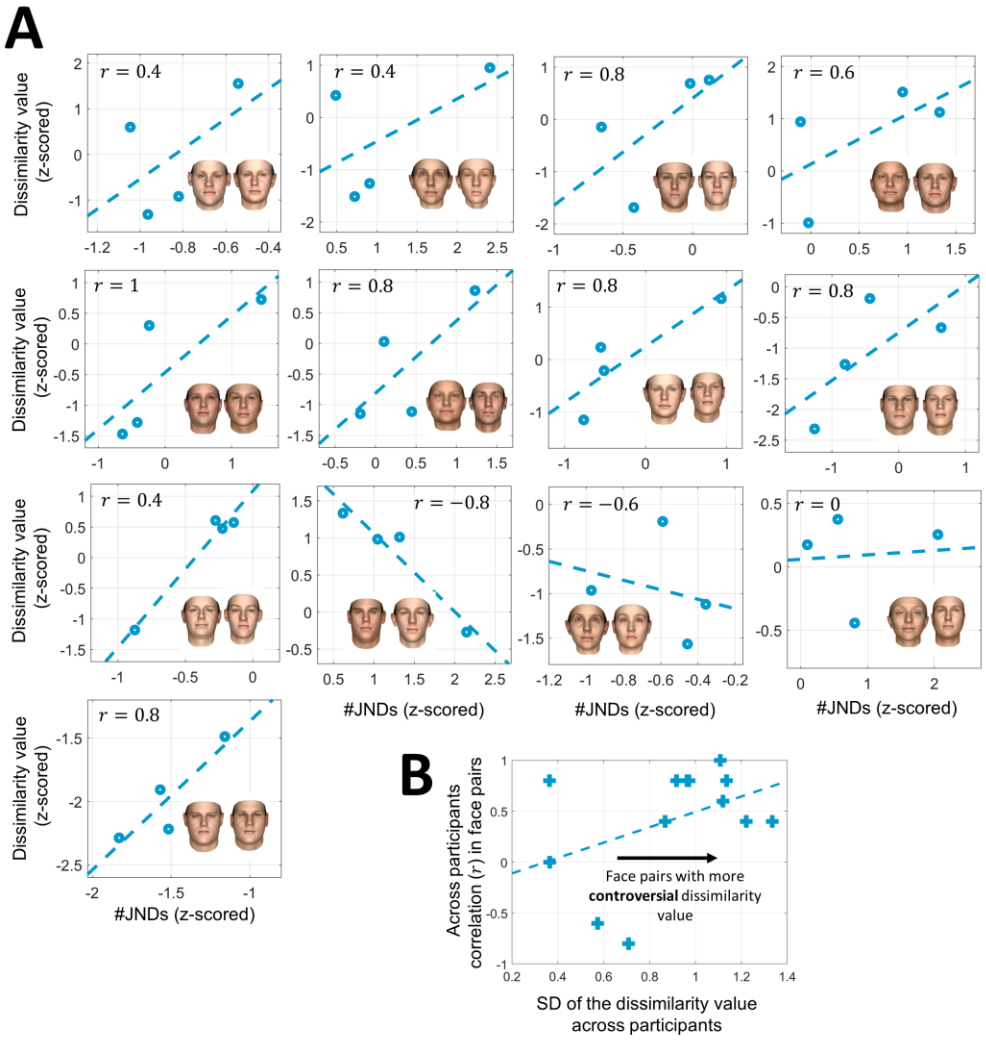
Supplementary Figure 1. Schematic of the subjective similarity judgment task design.



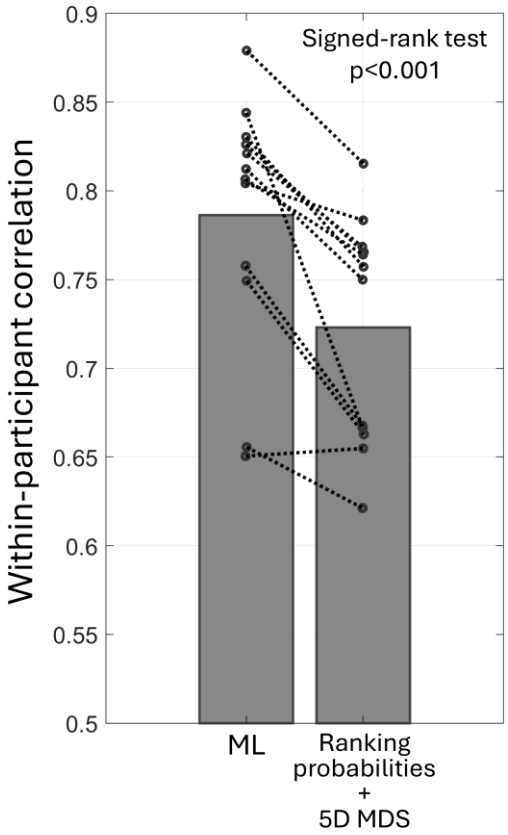
Supplementary Figure 2. Simulations to determine a feasible sample size for our data collection stopping criterion. We ran two simulations: one generated z-values from a Gaussian distribution with a mean of 2 and SD of 1 and another from a uniform distribution ranging between -0.5 and 3.5. These example distributions seem reasonable given our expectations based on our pilot data and seem conservative enough. For example, the 95% confidence interval (CI) width of the group-mean z-value 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 was a high likelihood of hitting the stopping criterion by reaching a maximum sample size of 24. We, in fact, reached the stopping criterion (i.e., a CI width smaller than 1) with our initial 12 participants in the main study.



Supplementary Figure 3. Results in the pilot study with N = 4 participants, who performed the subjective similarity judgment task once in a session and then the near-threshold discrimination task once in another session on 13 pre-selected pairs. (A, B) Relationship between perceptual discrimination capacity and subjective perceptual dissimilarity in each participant. The descriptions of the plots are similar to those in Figure 3. (C, D) Individual specificity of the relationship between perceptual discrimination capacity and subjective perceptual dissimilarity. The remaining descriptions of the plots are similar to those in Figure 4.

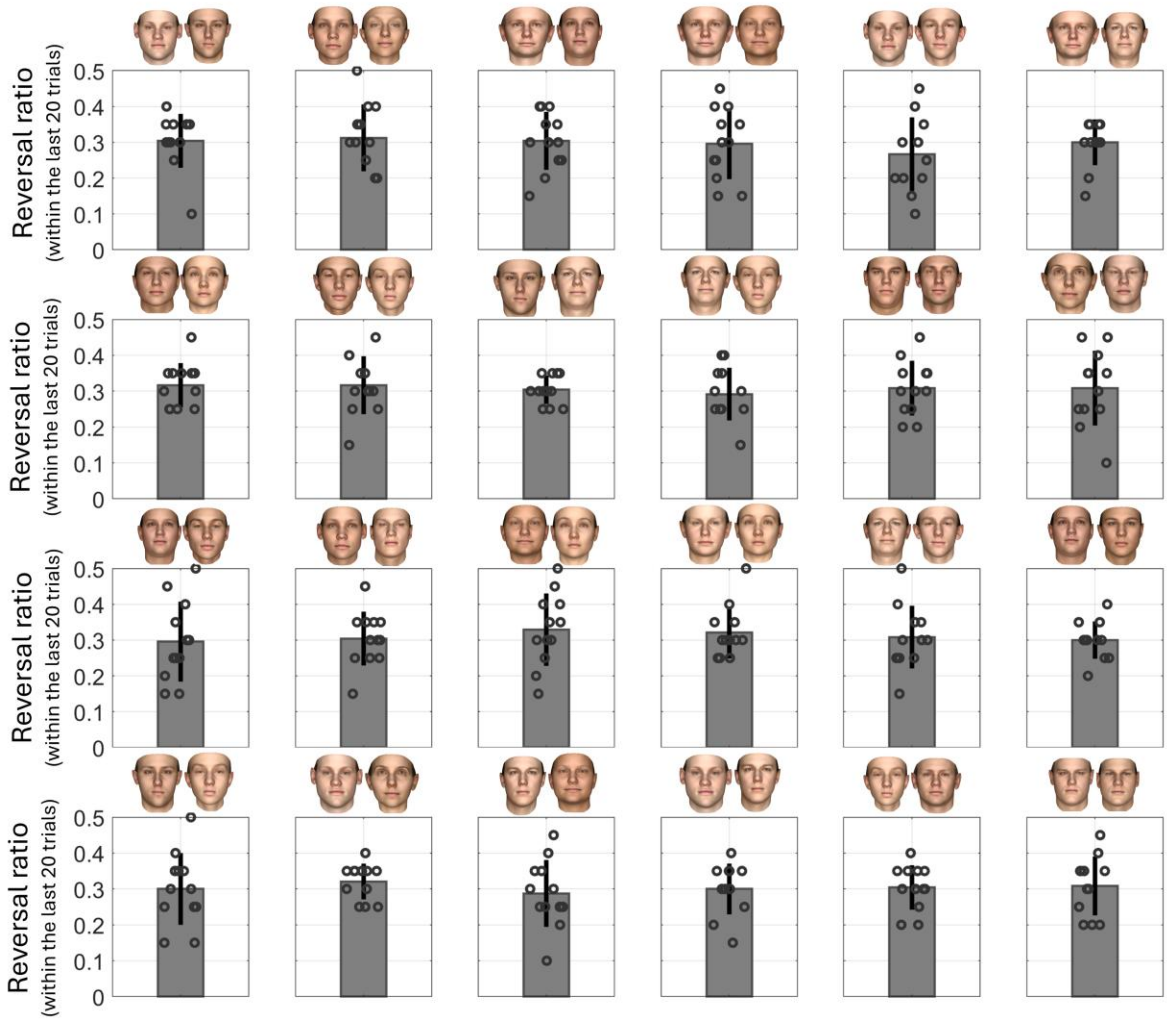


Supplementary Figure 4. Relationship between perceptual discrimination capacity and subjective perceptual dissimilarity across participants in each face pair (pilot study N = 4 participants). (A) The descriptions of the plots are similar to those in Figure 5. (B) Relationship between the face pairs Spearman 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 association between perceptual discrimination capacity and subjective perceptual dissimilarity across participants was more salient in highly controversial pairs.

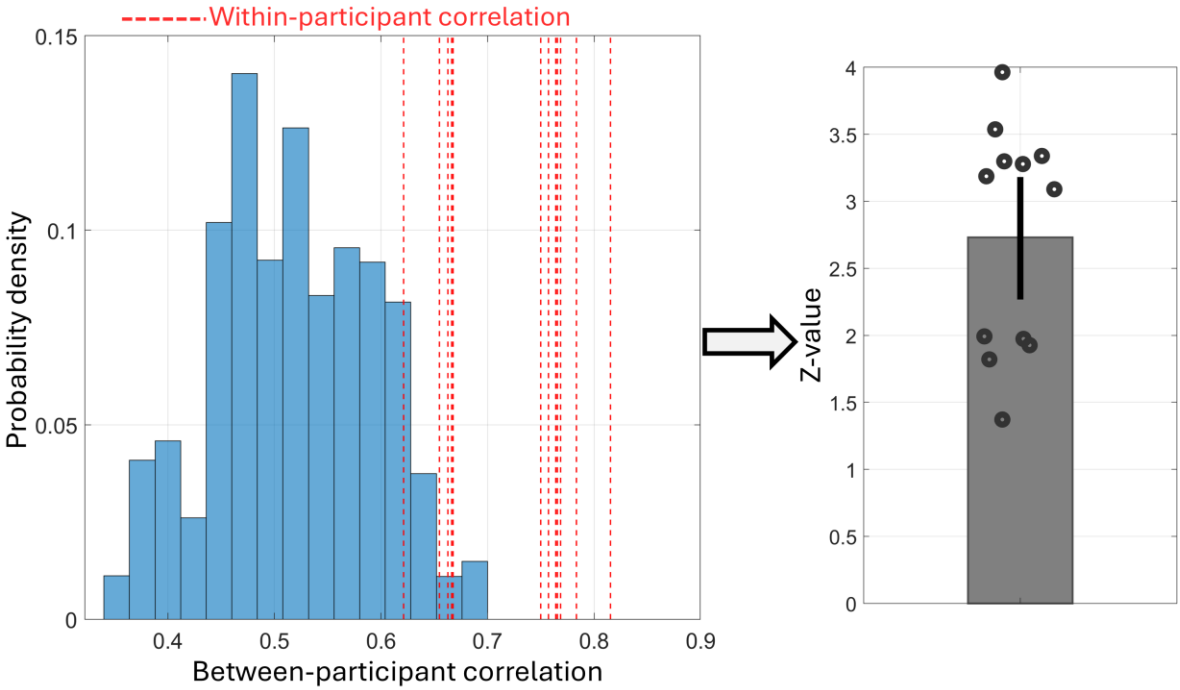


Supplementary Figure 5. Comparison of the alternatively pre-registered machine learning approach and the main pre-registered approach in estimating the dissimilarity matrix. The correlation between the dissimilarity matrices from the first and second sessions of the same participants (i.e., within-participant correlation) was higher when derived using the machine learning approach (signed-rank test p-value < 0.001). Thus, the machine learning approach could more robustly estimate the dissimilarity matrix from the collected subjective similarity ranking data.

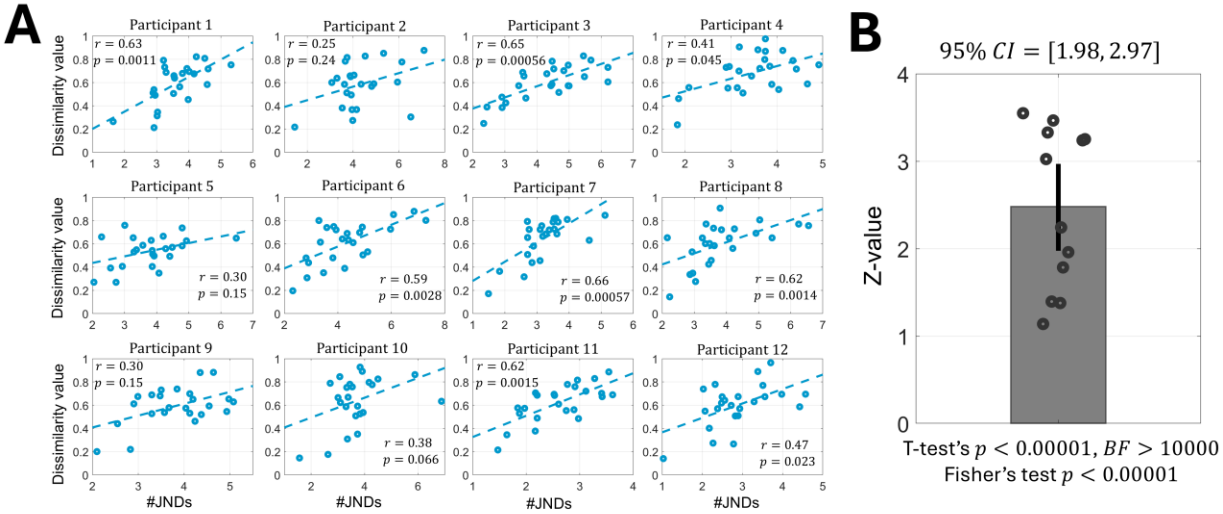
To recap, the main pre-registered approach involved estimating ranking probabilities and then applying a 5D MDS to reduce dimensionality and thereby reduce noise. In contrast, the alternatively pre-registered machine learning approach was designed to directly find the underlying embeddings by minimizing a loss function that penalizes mismatches between the similarity rankings inferred from the embeddings and those made by the participant.



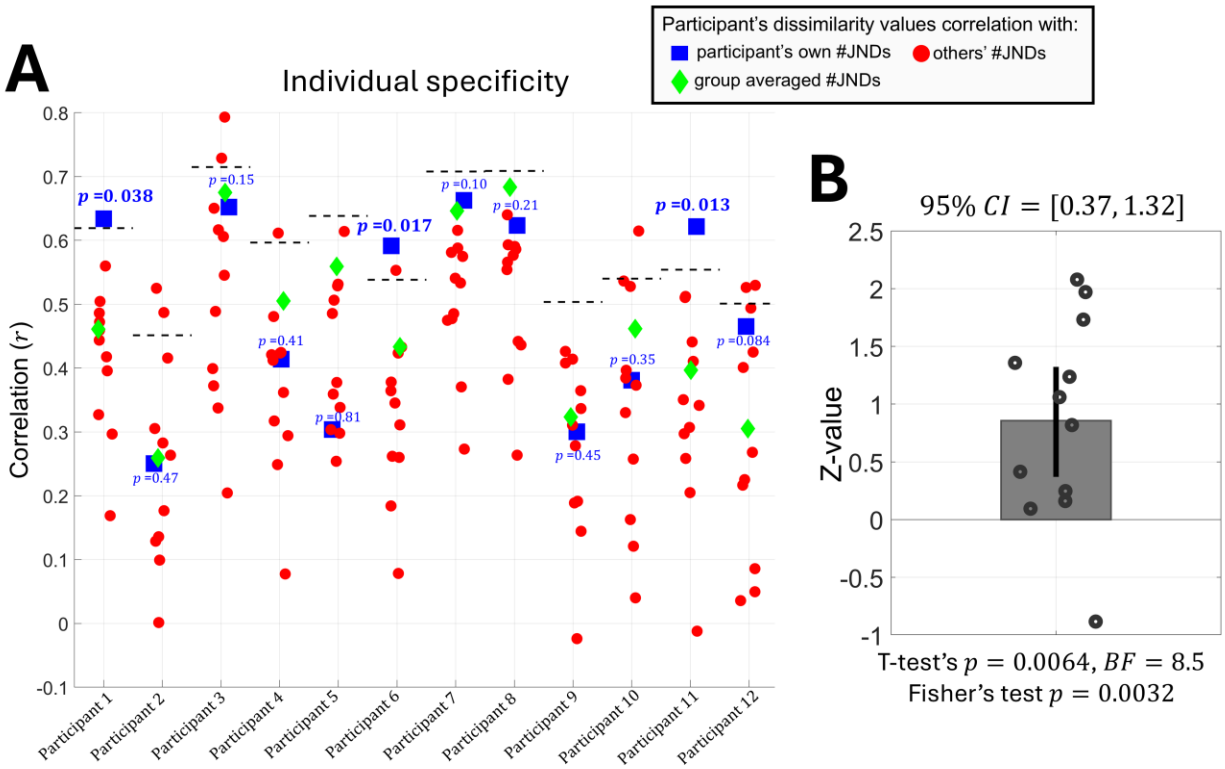
Supplementary Figure 6. Quality of the measured #JNDs (stability of the staircases). Each panel displays the ratio of reversals within the last 20 trials of each face pair's staircase. The bar shows the mean value, and the error bar displays its standard deviations (mean ± SD) across participants, with each dot representing a participant. The maximum possible ratio is 0.6, given the 1-up and 2-down protocol used. Therefore, on average, the staircases had half as much fluctuation as the maximum, which involves each "up" followed by a "down" and vice versa. As a reference, the example staircase shown in Figure 1C has a reversal ratio of 0.35 within its last 20 trials. Considering these, the staircases were fairly stable across participants, indicating that the measured #JNDs were reasonably robust.



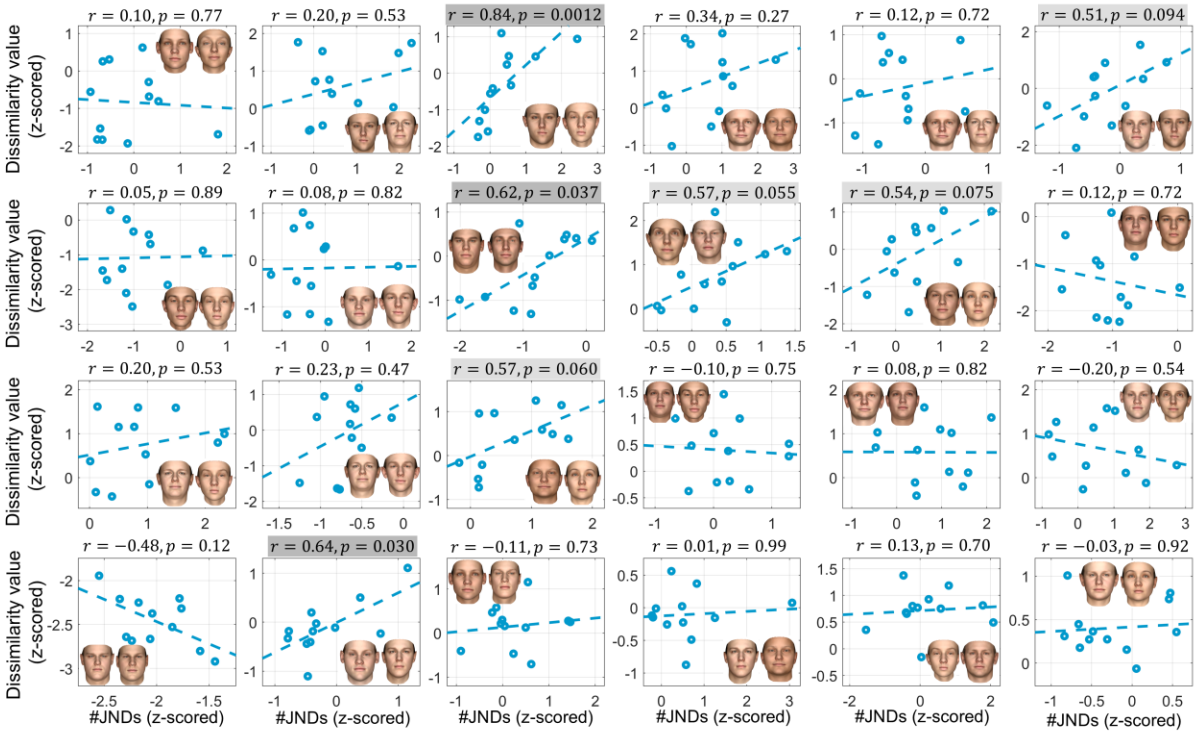
Supplementary Figure 7. Same as Figure 2, but with the dissimilarity matrices derived using the main pre-registered approach (i.e., estimating ranking probabilities and then applying a 5D MDS to reduce dimensionality and thereby reduce noise). The dissimilarity matrices derived using the main pre-registered approach could reliably capture each participant's specific subjective perceptual dissimilarity. This was comparable, though slightly weaker, compared to the specificity of the dissimilarity matrices derived with the alternatively pre-registered machine learning approach (see Figure 2, where the z-value was above 2.3 in most (ten) participants).



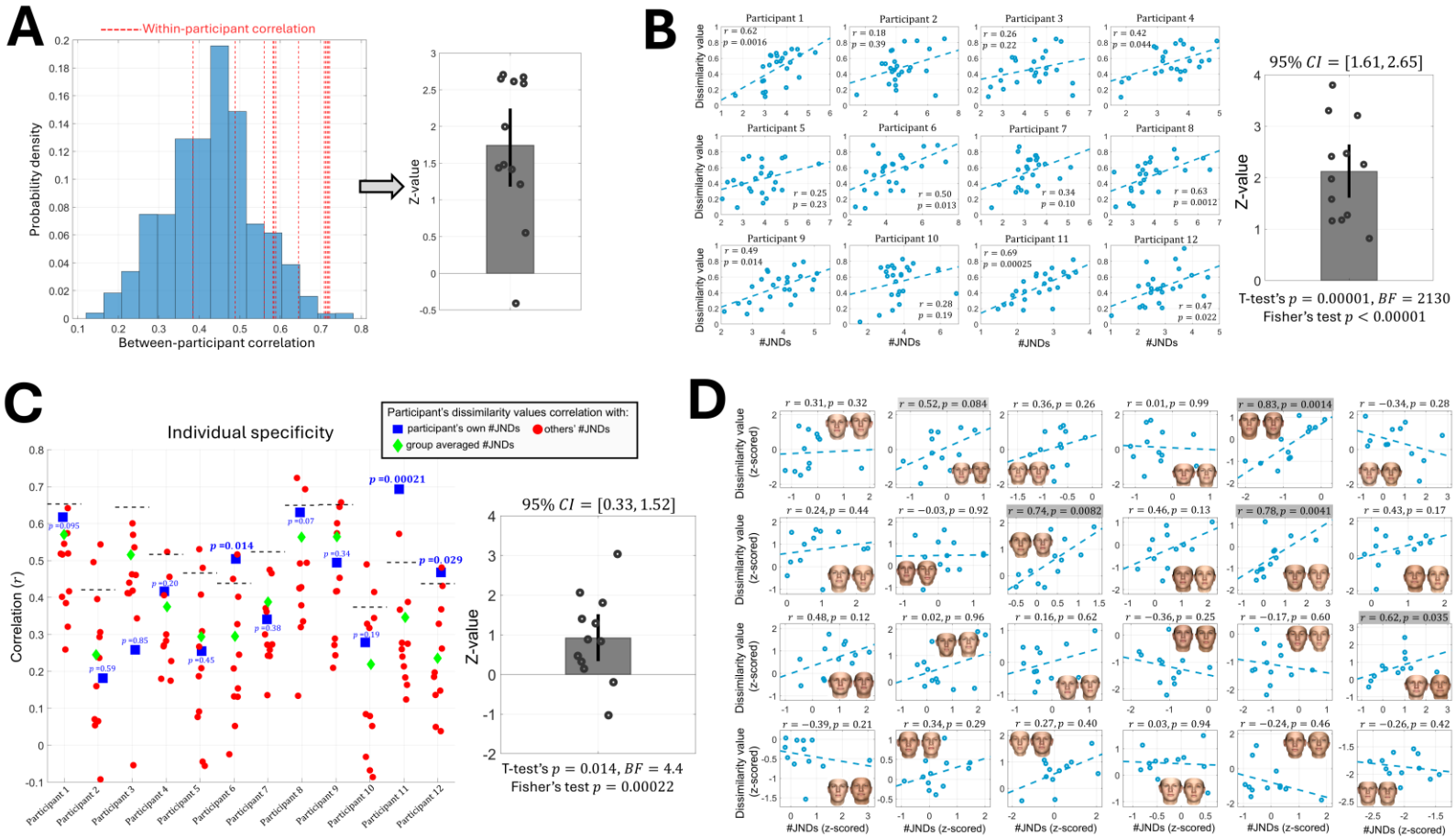
Supplementary Figure 8. Same as Figure 3, but with the dissimilarity matrices derived using the main pre-registered approach (i.e., estimating ranking probabilities and then applying a 5D MDS to reduce dimensionality and thereby reduce noise). Similar to Figure 3, a significant positive relationship was observed between perceptual discrimination capacity and subjective perceptual dissimilarity at the group-level, with significance reached at the individual-level in most (eight) participants. Note that as shown in panel B, the 95% confidence interval was above zero, confirming Hypothesis 1, and its width was less than one, indicating that the data collection stopping criterion was met.



Supplementary Figure 9. Same as Figure 4, but with the dissimilarity matrices derived using the main pre-registered approach (i.e., estimating ranking probabilities and then applying a 5D MDS to reduce dimensionality and thereby reduce noise). Similar to Figure 4, the association between perceptual discrimination capacity and subjective perceptual dissimilarity was quite specific to each individual (see the group-level statistic in panel B), with significance reached at the individual-level in three participants. Note that as shown in panel B, the 95% confidence interval was above zero, confirming Hypothesis 2, and its width was less than one, indicating that the data collection stopping criterion was met.



Supplementary Figure 10. Same as Figure 5, but with the dissimilarity matrices derived using the main pre-registered approach (i.e., estimating ranking probabilities and then applying a 5D MDS to reduce dimensionality and thereby reduce noise). Similar to Figure 5, the association between perceptual discrimination capacity and subjective perceptual dissimilarity was positive in most pairs across participants, supporting our hypotheses. A t-test on the r value being positive across pairs yielded a p-value of 0.0081 and a BF of 5.8. Also, this association was more salient in more controversial pairs (compare the last row, which shows less controversial pairs, with the other rows).



Supplementary Figure 11. (A) same as Figure 2, (B) same as Figure 3, (C) same as Figure 4, (D) same as Figure 5, but with the dissimilarity matrices derived using the main pre-registered approach incorporating a 2D instead of a 5D MDS. Hypothesis 1 was confirmed, reaching individual-level significance in 7 participants (see panel B). Hypothesis 2 was also confirmed, reaching individual-level significance in three participants (see panel C). As shown in Panel D, perceptual discrimination capacity was positively associated with subjective perceptual dissimilarity across participants in most face pairs (A t-test on the *r* value being positive across pairs yielded a p-value of 0.015 and a BF of 3.4). Panel A indicates that using a 2D MDS could also capture individuals' unique subjective perceptual dissimilarity, though it performed noticeably worse than using a 5D MDS (see Supplementary Figure 7) and the alternative machine learning approach (see Figure 2). Furthermore, the use of a 2D MDS made the relationship between perceptual discrimination capacity and subjective perceptual dissimilarity (panel B), as well as its individual-specificity (panel C), appear weaker than those observed in supplementary Figures 8 and 9 (use of a 5D MDS), and Figures 3 and 4 (use of the alternative machine learning approach). Therefore, forcing the dissimilarity relations into two dimensions caused more harm (information loss) than benefit (e.g., noise reduction).