Face familiarity and similarity: Within- and between-identity representations are altered by learning

Robin S. S. Kramer[1], Alex L. Jones[2], and Daniel Fitousi[3]

[1] School of Psychology, University of Lincoln

[2] School of Psychology, Swansea University

[3] Department of Psychology, Ariel University

Correspondence concerning this article should be addressed to Robin Kramer, School of Psychology, University of Lincoln, Lincoln LN6 7TS, UK. E-mail: remarknibor@gmail.com

Word count: 10,960

Abstract

Face familiarity is thought to alter distances between representations in psychological 'face space', resulting in substantial improvements in recognition. However, the underlying changes are not well understood. In Experiment 1 (n = 192), we investigated the effect of familiarity based on everyday exposure to celebrities. Participants judged the similarity of pairs of face photographs, and we found that greater familiarity increased perceived similarity for two images of the same person, while decreasing similarity for two images depicting different people. In Experiment 2 (n = 157), familiarity was manipulated through the learning of new identities by watching 5-minute video clips. Again, when judging the similarity of image pairs, familiarity increased the perceived similarity of images of the same person, while having the opposite effect on images depicting different people. In Experiment 3, we trained a computational model with images of 333 different identities (totalling 3,949 photographs) and manipulated its familiarity with two new identities. The changes in distances between novel images of these identities (a proxy for similarity) replicated our behavioural findings. Overall, we build upon recent evidence by demonstrating two transformations through which familiarity alters representational space to likely benefit face perception.

Keywords

face space, face familiarity, face similarity, face learning, Bayesian inference, computational modelling

Public significance statement

By combining behavioural evidence with computer simulations, we show that increasing face familiarity results in 1) an increase in the perceived similarity between different images of the same person, and 2) a decrease in the perceived similarity between images of different people.

Human faces represent an especially homogenous category, with every instance sharing both a set of features and their configuration. These constraints result in a broad similarity in appearance *between* faces, while alongside this, we see considerable variability present *within* each face across encounters. This variation is produced by many factors, including changes in facial expression, pose, aging, cosmetics, and so on. Because these two sources of variability may be comparable in size (Jenkins et al., 2011), viewers are often error-prone when discriminating between similar-looking people, as well as realising that different images depict the same person (e.g., Burton et al., 2010; Fysh & Bindemann, 2018; Jenkins et al., 2011). Crucially, these difficulties are limited to unfamiliar faces, and tasks involving familiar face discrimination and identification can be considered trivial (e.g., Bruce et al., 2001; Burton et al., 1999). Clearly, familiarity with a face has significant consequences for how it is processed (e.g., Megreya & Burton, 2006; for reviews, see Johnston & Edmonds, 2009; Ramon & Gobbini, 2018). However, little is known regarding how face familiarity affects our internal representations and, as a result, our performance.

Since faces vary along a variety of continuous dimensions (e.g., mouth width, nose length, etc.), one way to represent these internally is a theoretical 'face space', proposed to explain psychological similarity (e.g., Valentine, 1991; Valentine et al., 2016). Each face occupies a specific location, with each dimension describing the variation in some feature or characteristic. This could be a particular measure (e.g., face width, nose length) or a more global property (e.g., masculinity). However, its key principle is that faces located nearer to each other in face space are perceived to be more similar. So far, this framework has provided an explanation for numerous behavioural results including the effects of distinctiveness, inversion, caricaturing, adaptation, and the other-race effect (for a review, see Valentine et al., 2016).

Early instantiations of face space typically focussed on discriminating between different faces while tending to ignore the within-person variability present across instances of the same face (e.g., Valentine et al., 2016). Related, these models failed to address the important distinction between familiar and unfamiliar faces, and therefore to incorporate a mechanism of face learning. Recent

research has shown that these two concepts are intertwined (Burton, 2013). With each face varying idiosyncratically (Burton et al., 2016), face learning is improved through experiencing this face-specific variability (e.g., Corpuz & Oriet, 2022; Ritchie & Burton, 2017). In simple terms, familiarity with a particular face can be thought of as exposure to more (varied) instances of that face (e.g., Blauch et al., 2021; Kramer et al., 2018; Kramer, Young, et al., 2017).

While familiarity improves face identification (e.g., Clutterbuck & Johnston, 2002, 2004, 2005), we have yet to confirm the underlying change(s) that may produce this shift in performance. However, two transformations within face space have been proposed as candidates. The first is an increase in the representational distance between different people, which would decrease the likelihood of their being confused with one another. Theoretical models have tended to emphasise this change (e.g., Valentine et al., 2016), which has also been the focus of empirical studies (e.g., Collins & Behrmann, 2020; Faerber et al., 2016). The second transformation is a decrease in the representational distance between instances of the same person. This facilitates the perception of these different images/instances as the same face while again emphasising the differences (distances) between people.

This second change is suggested by research on social trait impressions. For instance, the ratings attributed to different images of the same identity were more tightly clustered when that identity was familiar rather than unfamiliar (Mileva et al., 2019). Further, attractiveness ratings became more consistent over time, for different images of the same identity, as that individual was being learned (Koca & Oriet, 2023). Finally, higher ratings of likeness were given to *all* images of a person who was more familiar (Balas et al., 2023; Ritchie et al., 2018; see also Jenkins et al., 2011). Taken together, these results may be explained by a decrease in the representational distance between instances for a given identity with increased familiarity.

More direct evidence for this representational change has come from recent computational models (Blauch et al., 2021; Kramer et al., 2018; see also Fitousi, 2023), while only one study has considered this transformation behaviourally. White and colleagues (2022) created ‘identity

averages' by combining multiple images of a single person, with these considered to represent each person's centroid in face space. In addition, 'gender averages' were created by combining these identity-specific averages for a given gender, with these representing the (gender-specific) centroid of the overall face space. Analogous to the partitioning of between- and within-group variation during analysis of variance, this approach allowed for the direct comparison of these two sources of variation. The researchers found that the perceived similarity between an identity's images and its average was higher for familiar faces in comparison with unfamiliar ones, suggesting that images of the same person are more tightly clustered as a result of familiarity. However, the perceived similarity between identity averages and the gender average showed only weak evidence of a familiarity effect, and so failed to provide support for the idea that familiar people are represented at a greater distance from each other in face space. Therefore, the researchers argued that between-person representational distances might play a far smaller role in recognition than was previously thought.

However, as White and colleagues (2022) note, one issue with their study may be that within- and between-person representational distances were determined relative to average images. Although there is evidence to support the idea that averages are encoded internally when viewing face images (e.g., de Fockert & Wolfenstein, 2009; Kramer et al., 2015; Neumann et al., 2013), it remains to be seen as to whether we use these representations when determining perceived similarity rather than making a comparison between exemplars/instances (see Balas et al., 2023). In addition, we cannot be sure that our internal average representations sufficiently resemble these computer-derived stimuli for the purposes of investigating these hypothesised changes. Of course, the finding that instances were perceived as more similar to identity averages necessarily implies that they were also represented more closely to each other.

In the present set of experiments, we wished to test for the influence of familiarity on the two transformations described above without the involvement of averages. Instead, we employed a more direct approach by asking the following questions. First, does increased familiarity with people

produce a decrease in the perceived similarity between images of these different people? Second, does this increase in familiarity also increase the perceived similarity between different images of the same person? By utilising participants' pre-existing familiarity with faces (Experiment 1), manipulated familiarity via participants learning new identities (Experiment 2), and a simple computational model (Experiment 3), we demonstrate the presence of both representational transformations as a result of increased face familiarity.

Experiment 1: Prior familiarity

As discussed above, the study by White and colleagues (2022) utilised averages (both identity- and gender-specific) as the framework for investigating perceptual similarity and representational transformations in face space. Here, we took a more direct approach by considering the perceived similarity of pairs of images that were unaltered/unconstrained. We made no attempt to control for low-level image statistics or to remove colour information (cf. White et al., 2022) since recent studies have begun to emphasise the importance of incorporating 'ambient', natural variability when investigating face perception (see Burton, 2013; Jenkins et al., 2011). In addition, we used a fully crossed design, incorporating participants and stimuli from the UK and US. As a result, all images served as both familiar and unfamiliar stimuli across our participants, avoiding the possibility of confounding familiarity with image set (cf. White et al., 2022). Using this design, we asked participants to rate the similarity of pairs of images, along with their familiarity with the identities depicted, to investigate the relationship between these two factors.

Method

Transparency and openness

We used Bayesian approaches for our analyses, which do not rely on specifying the nature of data collection in advance (Dienes, 2011; Etz et al., 2018; Rouder, 2014). As such, the experiments reported in this article were not preregistered. Similarly, we did not carry out power analyses prior to data collection.

We report all rules for data exclusion, all manipulations, and all measures for these experiments. Data collection for all experiments reported in this article took place in 2023. (To address the question of demand characteristics, we collected additional data in 2024.) The data that support the findings of these experiments are openly available on the Open Science Framework at: https://osf.io/wafvq/?view_only=497c43191e5a4ac597e43750314658ce.

Participants

A sample of 192 participants (123 women, 67 men, 2 nonbinary; age M = 42.2 years, SD = 14.6 years) living in either the UK (n = 97) or US (n = 95) gave informed, onscreen consent before taking part in the experiment and were provided with an onscreen debriefing upon completion. Participants were recruited through the Prolific online platform, where eligibility was restricted to these two countries, and were paid £1.20 for their time. The data from seven additional participants were excluded due to responding incorrectly to at least one of the attention checks (see below). Both Experiments 1 and 2 received ethical approval from Ariel University's research ethics committee (ID: AU-SOC-DF-20230904) and were carried out in accordance with the provisions of the World Medical Association Declaration of Helsinki.

The sample sizes for this experiment and Experiment 2 were initially set as a compromise between available resources and our estimates of what was required to measure sufficiently precise effects with these experimental paradigms. Since we planned to use Bayesian analytical methods, we anticipated increasing the sample size where estimates of theoretically important predictors were imprecise/ambiguous. It is worth noting that Bayesian methods do not suffer from many of the

issues affecting frequentist analyses regarding optional stopping if the aim is simply sufficient precision rather than hypothesis confirmation (Rouder, 2014). In the end, our initial samples provided sufficiently unambiguous evidence and we did not choose to collect additional data.

Stimuli

We compiled a list of 166 individuals of national (rather than international) fame, with half of these selected for their celebrity within the UK and the other half within the US. These individuals were commonly known through appearances in visual media rather than, for instance, radio alone. The list of names was generated by contacting colleagues who lived in these countries, as well as through searching online. For each individual, we collected two different photographs using Google Images searches, with each photograph depicting the individual facing roughly front-on and with their face free from occlusions.

From this initial set, we formed 25 pairs of different people using the UK celebrities, and another 25 pairs from the US celebrities, equating visual descriptors within each pairing. One of the authors (unfamiliar with the majority of these identities) was responsible for creating suitable pairings that demonstrated some similarity in appearance (through inspection of the collected images). From the remaining list of unused identities, we selected a further 25 UK and 25 US celebrities who we believed would be the least known outside of the countries in which they were based. In total, this resulted in the inclusion of 75 UK and 75 US celebrities.

For these 150 identities, all collected images were cropped to contain only the head and neck, and in some cases, the top of the shoulders, and were resized to 380 x 570 pixels. (Backgrounds were not removed.) For the paired identities, we selected one image of each identity so that the two people appeared most similar to each other. For the remaining identities, both images were used in the study (see Figure 1).

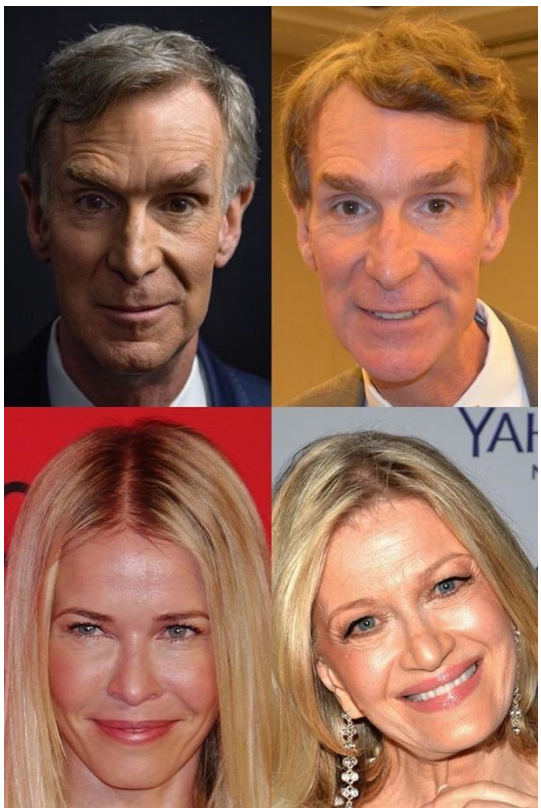

Figure 1. Examples illustrating two images of the same person (top row) and two images of different people (bottom row). Image attributions (top row left to right, bottom row left to right): Neil Grabowsky (cropped) [CC BY 2.0]; Raphael Perrino (cropped) [CC BY 2.0]; David Shankbone (cropped) [CC BY 2.0]; Andrew H. Walker (cropped) [CC BY 2.0].

Procedure

The main experiment was completed using the Gorilla online testing platform (Anwyl-Irvine et al., 2020). After consent was obtained, participants provided demographic information. Each participant was then presented with the 'same person' and 'different people' tasks, which were counterbalanced in order across participants.

In the 'same person' task, on each of the 50 trials, participants were presented with two different images of the same person. At the beginning of the task, instructions stated that these were "two photographs of the same person", and this information remained onscreen throughout the task. For each pair of images, participants were asked "how similar does the person look in these

photographs?" and responded using a 7-point scale with labelled anchors (1 = extremely different; 7 = extremely similar).

In the 'different people' task, on each of the 50 trials, participants were presented with two images of different people. At the beginning of the task, instructions stated that these were "photographs of two different people", and this information remained onscreen throughout the task. For each pair of images, participants were asked "how similar do the two people look in these photographs?" and responded using the same scale as above. In both tasks, responses were self-paced and trial orders were randomised for each participant.

In addition, we included an attention check within the randomly ordered presentation for each task, given that attentiveness is a common concern when collecting data online (Hauser & Schwarz, 2016). For the 'same person' task, we presented two identical images of an identity (not featured in the experiment itself). However, the internal features of the faces were replaced with the text "Instruction Manipulation Check: Select '7' as your response for this question". For the 'different people' task, we presented images of two identities (again, not featured in the experiment itself) who were very different in appearance (a Black man and a White woman). The internal features of the faces were replaced with the text "Instruction Manipulation Check: Select '1' as your response for this question".

Upon completion of both tasks, participants were presented with a 'familiarity' task. On each of the 150 trials, the name and photograph of an identity was displayed. This image was the one shown during the 'different people' task (for identities originally appearing in that task) or one of the two images shown during the 'same person' task (for identities appearing in that task). Participants were asked to rate their familiarity with the identity using a 7-point scale with labelled anchors (1 = extremely unfamiliar; 7 = extremely familiar). Onscreen instructions clarified that we were referring to their familiarity with each person *before* participating in the experiment.

Analytic strategy

For ‘same person’ trials, the familiarity rating given to that identity was used in our model. For ‘different people’ trials, we used the average familiarity rating, calculated from the values given to the two featured identities. Since this average failed to differentiate between particular situations (e.g., ‘moderate familiarity with both identities’ versus ‘high familiarity with one and low familiarity with the other’), we also estimated the model using a disaggregated familiarity rating for each trial, essentially doubling the number of observations per participant. However, the overall findings remained unchanged (with estimates being well within the posterior estimates for the model reported here) and so we report only the simpler model below. Figure 2 summarises participants’ responses prior to modelling.

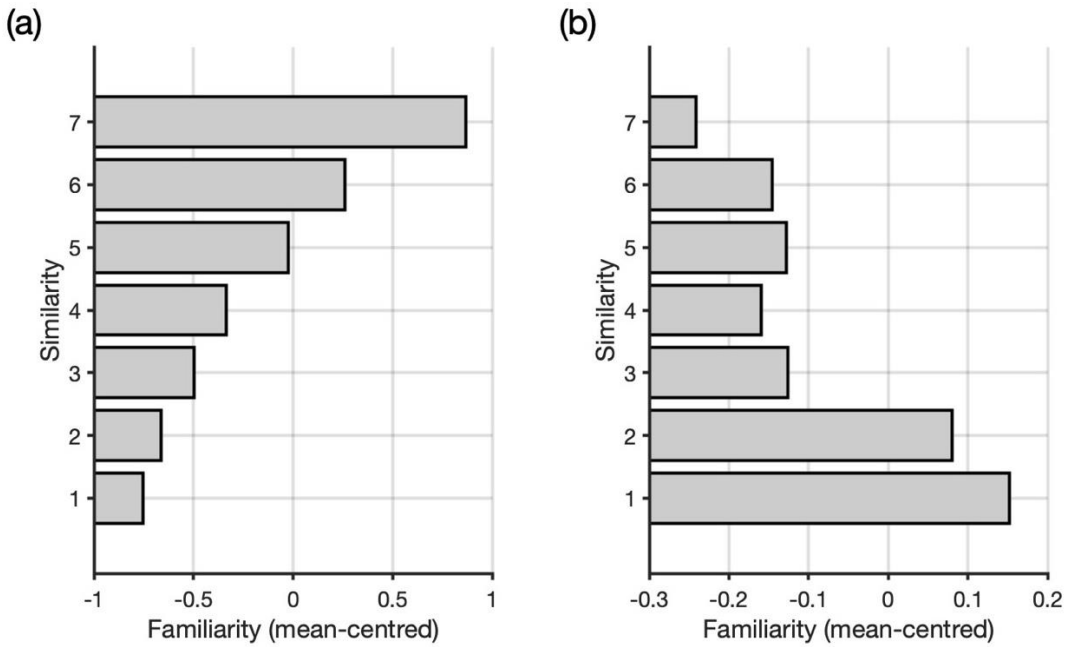

Figure 2. The mean familiarity rating for each level of rated similarity, presented separately for (a) ‘same person’ and (b) ‘different people’ trials. These means ignore the hierarchical structure of the data and therefore only illustrate the pre-modelled responses. (No error bars are presented since inferences should not be made from the data in this form.)

We used model-based Bayesian inference to interpret the data, specifically by fitting a hierarchical linear regression model. Similarity ratings were predicted from continuous (participant-mean centred) familiarity ratings, a dummy-coded variable indicating trial type (whether the trial represented the 'same person', code as one, or 'different people', coded as zero), and their interaction. The group-specific (or random) effects included an intercept for each participant, as well as a participant-specific slope for the effect of familiarity, allowing for variation in the effect familiarity had on each individual participant's similarity ratings.

We set weakly-informative priors on all model parameters (Gelman et al., 2017) that had little influence on the data, and used a Gaussian likelihood (i.e., we assumed similarity ratings were normally distributed, analogous to ordinary least squares). For the coefficients representing the intercept, familiarity ratings, trial type, and the interaction, a Gaussian distribution with a mean of zero and a standard deviation of ten were used, which entertain very large effects in either direction. A half-Gaussian distribution with a standard deviation of three was used for the error variance of the likelihood. For the group-specific effects (the participant intercept and participant familiarity slope), two Gaussian distributions with a mean of zero and a standard deviation of one were used. Models were estimated using the Bambi and PyMC packages (Capretto et al., 2020; Salvatier et al., 2016) in the Python programming language. Four Markov Monte Carlo chains were run, with each having 3,000 tuning steps and 4,000 samples drawn from the posterior. The model converged, and all parameters had an $\hat{R} < 1.01$.

Model interpretation

Our model specified the interaction between trial type and familiarity, which represented the difference between trial types for the association between familiarity and similarity. We recovered the estimates for the familiarity slope within each trial type by adding the familiarity coefficient to the interaction coefficient, and relied on the interaction coefficient itself as the difference between

trial type conditions for the familiarity slope. To make inferences about the hypothesis of differences between these slopes, we used the posterior probability of effects being in a specific direction (Makowski et al., 2019), calculated via the proportion of the posterior distribution being above or below zero, $p(\boldsymbol{\theta} > 0 \text{ or } \boldsymbol{\theta} < 0)$, given the observed data. This was similar in intention to classical null-hypothesis significance testing but provided the probability that the hypothesis was different to zero given the data, and not the converse (Welsch et al., 2020). We also estimated 94% highest density intervals (HDI) of all posterior estimates, which showed the credible range of effects given the observed data and model.

Results

After estimating the model, we predicted the expected similarity score given to trials at five levels of mean-centred familiarity (-2, -1, 0, 1, and 2 units below/above the mean), for both trial types. These predictions are shown in Figure 3 and describe the pattern of responses in the data. While similarity ratings were unsurprisingly higher for ‘same person’ trials in comparison with ‘different people’ trials, the effect of familiarity varied in magnitude and direction for the two trial types. Overall, the model explained 43.8%, 94% HDI [43.1%, 44.5%], of the variance in similarity ratings.

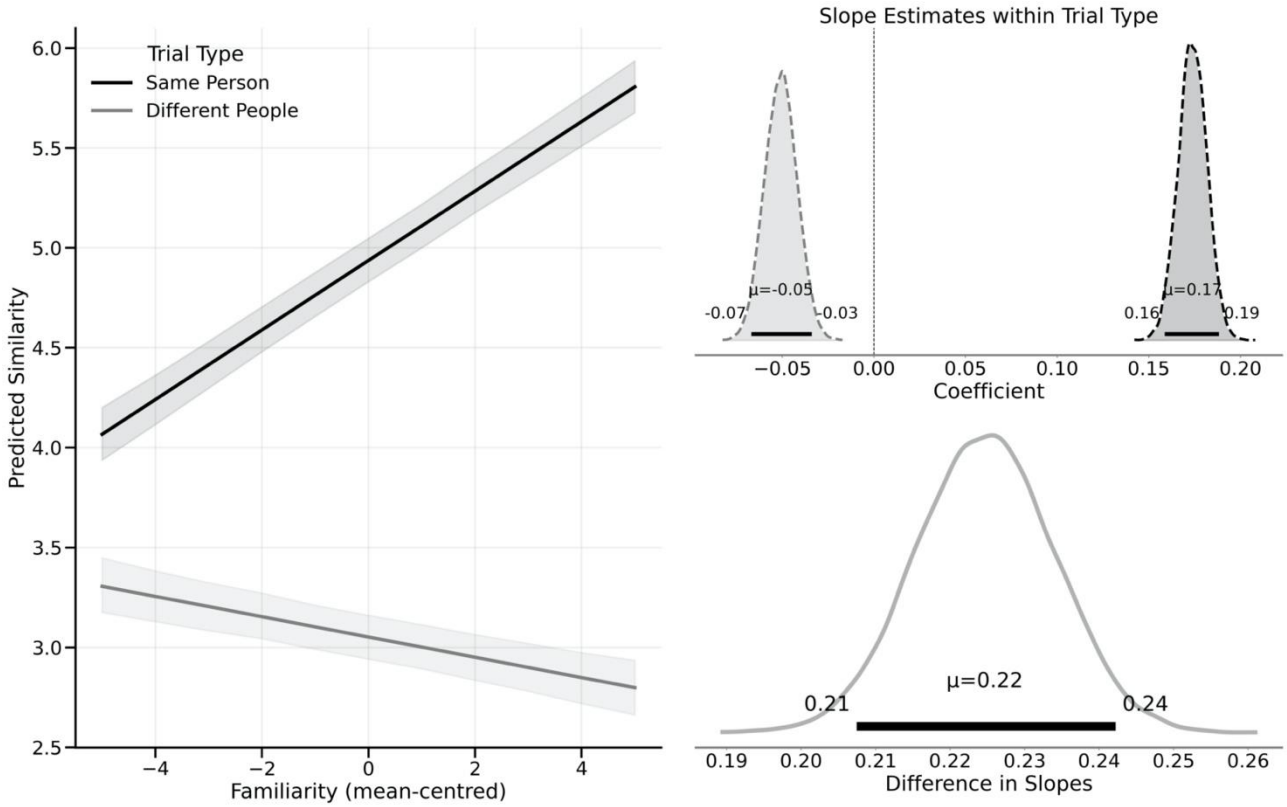

Figure 3. Left panel – model predictions representing the overall pattern of effects and 94% HDI across increasing familiarity and trial types. Top right – posterior distributions of the familiarity slope within trial type, with mean and 94% HDI. Each distribution excludes zero as a credible hypothesis. Bottom right – posterior distribution of the difference between slopes.

The familiarity slope estimates for both trial types are also shown in Figure 3. For 'same person' trials, the slope was positive, b = 0.17, [0.16, 0.19], $p(\boldsymbol{\theta} > 0)$ = 100%. In contrast, the slope for 'different people' trials was negative and smaller in magnitude, b = -0.05, [-0.07, -0.03], $p(\boldsymbol{\theta} < 0)$ = 100%. Figure 3 also reveals little overlap between these distributions, indicating they are credibly different. The interaction coefficient of the model supported this, b = 0.22, [0.21, 0.24], $p(\boldsymbol{\theta} > 0)$ = 100%. That is, the familiarity slope for 'same person' trials was on average 0.22 units greater than that for 'different people' trials.

Ruling out demand characteristics

Our results demonstrated that greater familiarity was associated with an increase in similarity between different images of the same person, along with a decrease in similarity between images of different people. However, we relied solely on subjective ratings of similarity (following on from White et al., 2022). The potential concern here is that some version of demand characteristics could have played a role. For instance, participants might have chosen to assign similarity more liberally to familiar identities, perhaps because they 'knew' those images depicted the same person.

We pre-empted this issue by explicitly informing participants (via onscreen instructions throughout) that image pairs did, or did not, depict the same person. As such, we aimed to remove the potential for participants to rely on their own judgements of 'same person versus different people' when rating similarity. It is also important to note that we found two contrasting effects of familiarity for the 'same person' and 'different people' tasks. That is, familiarity resulted in higher ratings of similarity for two images of the same person but lower ratings for images of two different people. Therefore, it seemed unlikely that any demand characteristics would have led to these two opposing patterns.

Even so, we chose to investigate this concern experimentally. To this end, we collected data from an additional 40 participants (21 women, 18 men, 1 nonbinary; age M = 35.8 years, SD = 13.4 years) living in the UK or US and recruited/paid through Prolific. These participants completed a shortened version of the experiment – only 20 'same person' trials and 20 'different people' trials (half UK celebrities and half US celebrities in each task), rather than 50 of each. The aim was for participants to have a representative experience of the experiment, but we were not interested in their ratings here. All other design aspects mirrored the original version of the experiment (counterbalancing task order, a subsequent familiarity task, etc.). Crucially, upon completion, participants were asked three open-ended questions, one after the other and in a set order, to explore their strategies and beliefs.

The first question asked, "When completing the similarity ratings, did you use a particular strategy? (If you didn't, just write "no strategy" or something similar.)" In response, 22 participants

reported no strategy. For the remaining participants, none mentioned familiarity or how it might have influenced their ratings. Responses tended to describe, for instance, focussing on particular parts of the face or a consideration of how likely they might mistake two different identities for each other.

The second question asked, "Did you feel like you should respond in a particular way when rating the 'same person' and/or 'different people' image pairs? (If you didn't feel like you should respond in a particular way, just write "no" or something similar.)" In response, 34 participants reported that they had not felt like they should have responded in a particular way. For the remaining participants, none mentioned familiarity or how it might have influenced their ratings. Three responses suggested participants felt that they should provide higher similarity ratings for the 'same person' trials, with no other comments of relevance.

The third question asked, "We are interested in whether familiarity influenced your ratings of similarity. Can you guess what we predicted we would find for the 'same person' and/or 'different people' similarity ratings? (If you can't think of a guess, just write "not sure" or something similar.)" In response, 16 participants reported that they were unsure. For the remaining participants, there were 13 responses that included some mention of familiarity. Of these, 10 provided a directional hypothesis – that increased familiarity would either increase or decrease similarity judgements (presumably in relation to 'same person' and 'different people' trials respectively, although often the specific trial type was not specified). Importantly, no responses proposed two different familiarity-similarity relationships for the two trial types.

Taken together, we found no evidence that demand characteristics could account for our earlier results. No participants spontaneously mentioned familiarity when asked about their strategies or whether they felt the need to respond in a particular way. Indeed, even when told that we were interested in how familiarity influenced similarity, only a small number of participants guessed correctly as to one of the relationships we had predicted, but importantly, none included any mention of the idea that we predicted two different patterns for our two trial types.

Experiment 2: Learned familiarity

The first experiment investigated pre-existing face familiarity as a continuum (see Kramer et al., 2018), with the findings demonstrating both hypothesised representational changes. Simply, greater familiarity was associated with an increase in similarity between different images of the same person, along with a decrease in similarity between images of different people. Interestingly, the effect of familiarity was smaller for between-person judgements (a shallower slope in Figure 3), providing some support for the results of previous work, where this transformation also played a smaller role in representational changes (White et al., 2022).

Next, we directly manipulated familiarity (now treated as a dichotomy) through having participants learn three previously unfamiliar faces during the experiment. In this way, we could control the amount of familiarity and provide a stronger argument for a causal relationship between familiarity and perceived similarity. In addition, we were able to investigate whether relatively minimal exposure (only 5 minutes) was sufficient to produce an effect of familiarity. As in Experiment 1, participants rated the similarity between pairs of images, although here, they first learned three of six previously unfamiliar identities.

Method

Participants

A sample of 157 participants (73 women, 82 men, 2 nonbinary; age M = 43.3 years, SD = 13.2 years) gave informed, onscreen consent before taking part in the experiment and were provided with an onscreen debriefing upon completion. Participants were recruited through the Prolific online platform, where eligibility was restricted to people living in the US, and were paid £2 for

their time. The data from 34 additional participants were excluded due to responding incorrectly to at least one of the attention checks (see below). There was no overlap between this sample and those who participated in Experiment 1.

Stimuli

We selected two pairs and two individual celebrities from the UK identities that were used in Experiment 1. For each of these six identities, we collected several high-quality video clips from YouTube where the person appeared approximately front-on ± 45° and was alone in the frame (with minimal or no speaking by off-camera individuals). The majority of these were interviews in which the video frame either included the person from the waist upwards or only showed their head and shoulders. In all cases, the face was clearly visible and large enough to facilitate learning. We then combined multiple short clips (ranging from four to seven individual segments) taken from different videos/interviews to produce a final 5-minute video for each identity (the minimum length of time needed to form a robust face representation; Popova & Wiese, 2023). These six videos were 1280 x 720 pixels in size, shown in colour, and with the audio included.

In addition to these videos, we reused the trials featuring these six identities from Experiment 1. Therefore, for the two pairs, we used their two 'different people' trials, and for the two individuals, we used their two 'same person' trials.

Finally, we divided these six identities into two sets. Set A comprised one pair of identities and one individual identity. Set B comprised the remaining pair and individual.

Procedure

The main experiment was completed using the Gorilla online testing platform (Anwyl-Irvine et al., 2020). After consent was obtained, participants provided demographic information. Each

participant was then presented with a 'learning' task, where they were shown three 5-minute videos. These videos depicted the identities from either Set A or Set B, with this assignment counterbalanced across participants. The onscreen instructions at the start of this phase were as follows: "You will be shown 3 videos of people being interviewed. Each video is 5 minutes long, made up of a few different interviews. Please watch the videos carefully and learn to recognise each person's face." Participants were also instructed to have their device's sound turned on while viewing these videos. The order of the three videos was randomised for each participant, with each video only playing once (i.e., participants were unable to replay the videos).

We also inserted two attention checks during the 'learning' task, which appeared after the first and second videos finished playing. For each of these, participants were provided with two buttons next to each other onscreen, labelled "LEFT" and "RIGHT". The first attention check instructed participants as follows: "Attention check. Please click the LEFT button now (in less than 10 seconds) to show you're paying attention." The second attention check instructed participants to click the right button.

After the 'learning' task, all participants completed the same 'similarity' task. This comprised two 'different person' trials and two 'same person' trials, featuring the six identities from Sets A and B. These four trials were presented following the same procedure as in Experiment 1, i.e., using the same onscreen instructions and labelled response scale. In addition, we presented the two attention checks used in Experiment 1. Responses for this task were self-paced and the order of these six trials was randomised for each participant.

Finally, participants were presented with a 'familiarity' task. This was identical to the one used in Experiment 1 except that participants were only presented with the six identities from Sets A and B. As in the previous experiment, onscreen instructions clarified that we were referring to their familiarity with each person *before* participating in the experiment.

Analytic strategy

While the aim for this experiment was to manipulate participants' familiarity with novel (UK) identities, we collected familiarity ratings because our chosen celebrities may still have been recognised by some of the (US) individuals. As in Experiment 1, participants' previous familiarity associated with 'same person' trials was the familiarity rating given to that identity, whereas the familiarity associated with 'different people' trials was the average familiarity rating derived from the values given to the two featured identities. Again, estimating the model using a disaggregated familiarity rating for each trial resulted in the same pattern of results and so we report only the simpler model below. Figure 4 summarises participants' responses prior to modelling.

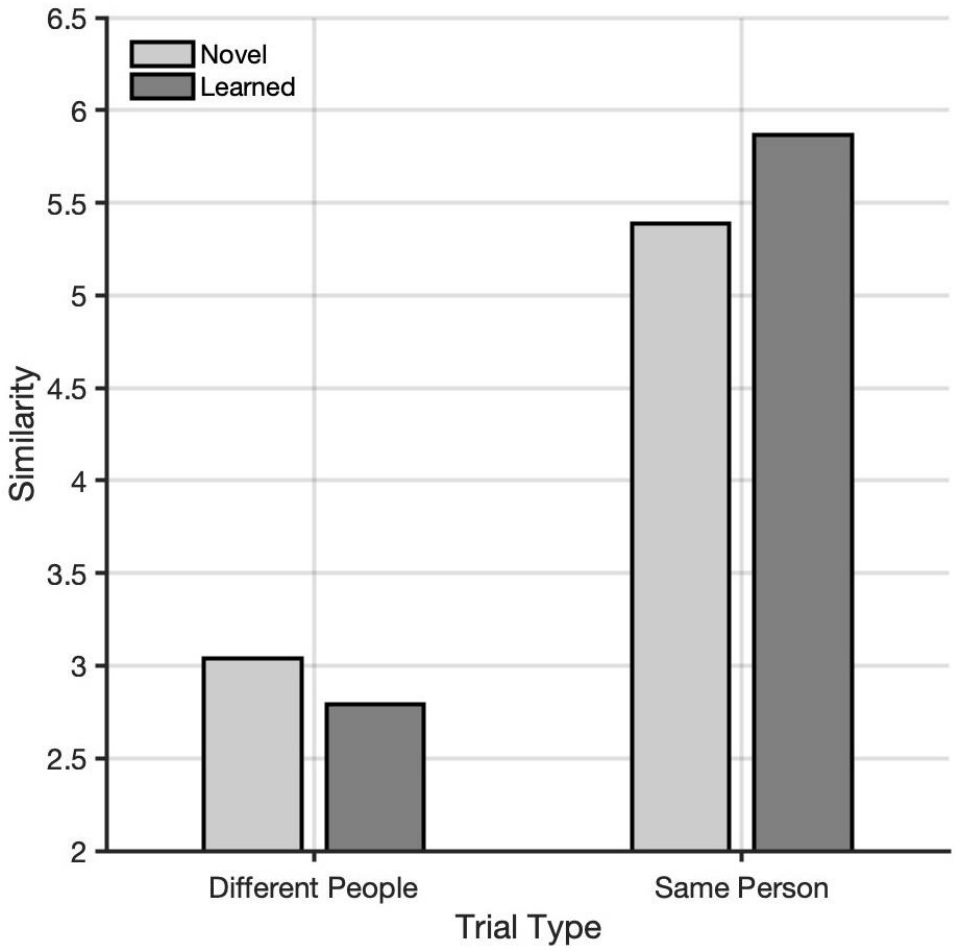

Figure 4. The mean similarity rating for each level of familiarity and trial type. These means ignore the hierarchical structure of the data, as well as pre-experiment familiarity with the identities, and

therefore only illustrate the pre-modelled responses. (No error bars are presented since inferences should not be made from the data in this form.)

Again, we employed Bayesian inference to estimate a hierarchical linear mixed regression. Similarity ratings were predicted from trial type (again a dummy-coded variable, with 'same person' trials coded as one, and 'different people' trials coded as zero), familiarity (a dummy-coded variable with trials featuring learned identities coded as one, and trials featuring novel identities coded as zero), and the interaction between these variables. We included previous familiarity (termed 'previous experience' to avoid confusion) as an additional covariate, allowing the model to learn the effects of this previous experience on the interaction between trial type and (experimentally manipulated) familiarity. As such, when making predictions about the expected similarity rating for both levels of familiarity and trial type, the model held constant the effect of previous experience at a rating of one.[1] Group specific effects included only the random intercepts for each participant. Our model was thus analogous to a repeated-measures ANOVA.

The prior structure of the model followed the same convention as in Experiment 1, with weakly-informative priors set on all model parameters, and a Gaussian likelihood. The intercept and coefficients representing trial type, familiarity, and their interaction had a Gaussian distribution prior with a mean of zero and a standard deviation of ten, as did the coefficient representing previous experience, and a half-Gaussian distribution with a standard deviation of three was used for the error variance. The group-specific effect for the participant intercept was a Gaussian distribution with a mean of zero and a standard deviation of one. Four Markov Monte Carlo chains were run, with each having 3,000 tuning steps and 4,000 samples drawn from the posterior. The model converged, and all parameters had an $\hat{R}$ of 1.

[1] As an alternative, we carried out analyses after excluding all similarity ratings for trials in which participants reported any familiarity with the featured identity or identities (i.e., by responding above one on the 1-7 scale). In this case, the same pattern of results was found.

Model interpretation

We predicted a specific pattern of results, such that the average similarity ratings for trials featuring learned identities would be higher than those featuring novel identities for the 'same person' trial type. In contrast, we predicted that similarity ratings would be lower for trials featuring learned identities than those featuring novel identities for the 'different people' trial type. To test this, we used the model to predict the conditional means within each of the four conditions (i.e., learned and novel identities for the 'same person' trials, as well as learned and novel identities for the 'different people' trials). Given that these means were distributions themselves, within each trial type, we subtracted the estimate of the novel identities' trials from the learned identities' trials. These contrast distributions allow for the straightforward derivation of the probability of individual directional hypotheses – whether the difference is positive for 'same person' trials, and negative for 'different people' trials – as well as the joint probability of these differences being in the predicted direction (Kruschke, 2014; Kruschke, 2018; Kruschke & Liddell, 2018). As in Experiment 1, we used 94% HDI and posterior probabilities.

Results

After estimating the model, we predicted the expected similarity rating under each combination of trial type and familiarity, holding previous experience constant at a rating of one. These predictions are shown in Figure 5. Similarity ratings were unsurprisingly higher for 'same person' trials (learned, M = 5.76, [5.52, 6.01]; novel, M = 5.33, [5.11, 5.54]), in comparison with 'different people' trials (learned, M = 2.70, [2.47, 2.94]; novel, M = 3.00, [2.78, 3.21]). Overall, the model explained 60.2%, [56.4%, 64.0%], of the variance in similarity ratings.

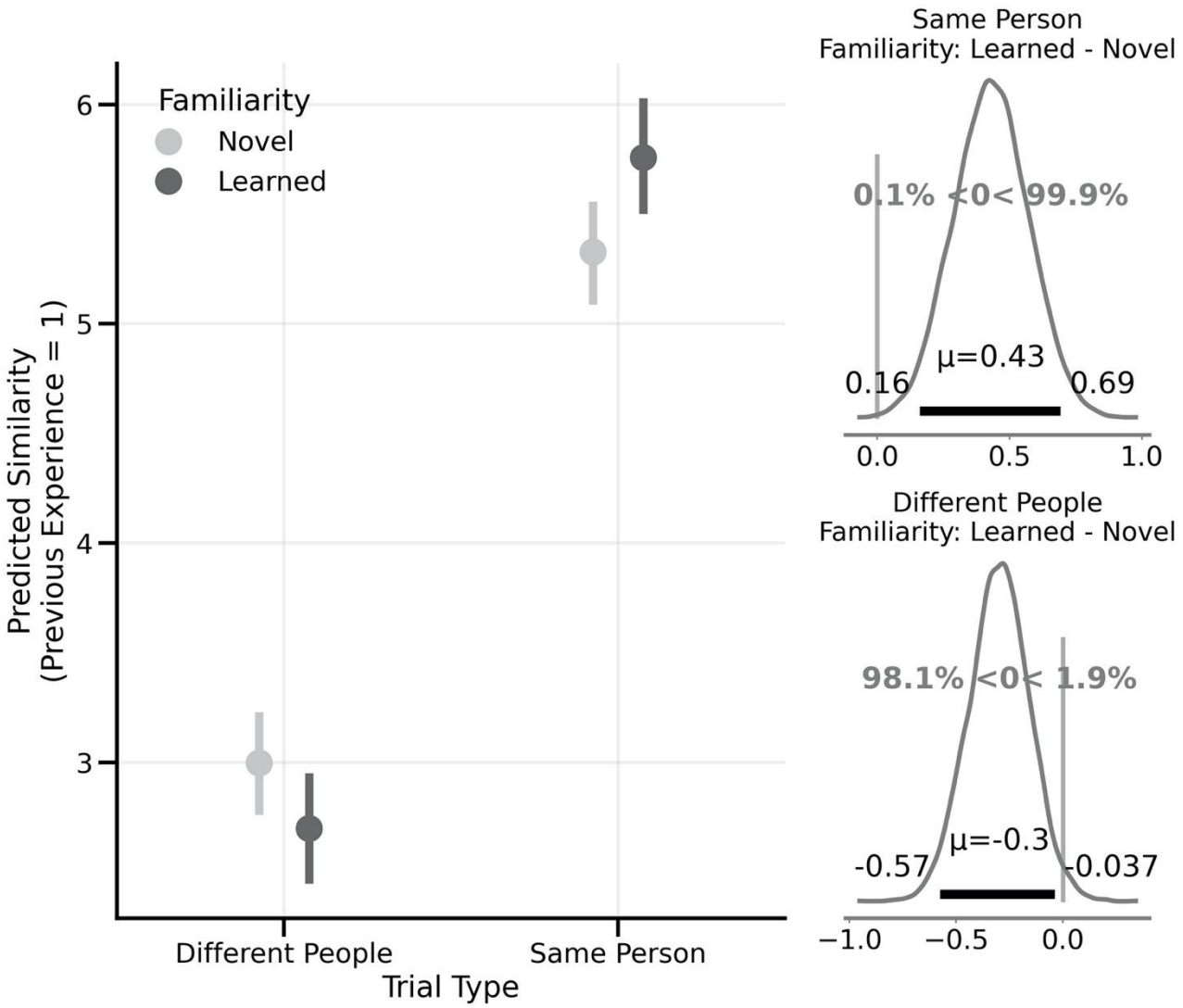

Figure 5. Left panel – model estimates of mean similarity for each level of familiarity and trial type, including 94% HDI. Right panels – posterior distributions of the contrasts between the two levels of familiarity for 'different people' and 'same person' trial types, highlighting the probability of the effects being in the specified direction.

Taking the difference between the two levels of familiarity within each trial type showed that, for 'same person' trials, the mean similarity rating was higher under learned trials compared to novel trials; mean difference = 0.43, [0.16, 0.69], $p(\boldsymbol{\theta} > 0) = 99.9\%$. For 'different people' trials, the mean similarity rating was lower under learned trials compared to novel trials; mean difference = -0.30, [-0.57, -0.04], $p(\boldsymbol{\theta} < 0) = 98.1\%$. These difference distributions are shown in Figure 5. Finally,

the probability of the joint hypothesis test – that the 'same person' trial difference was positive *and* the 'different people' trial difference was negative – was 98%, supporting our predictions.

Experiment 3: Computational modelling

Both Experiments 1 and 2 provided evidence that increasing familiarity resulted in a change in perceived similarity between image pairs. In addition, the results from both experiments suggested that familiarity exerts a stronger influence on within-person, in comparison with between-person, representational distances (White et al., 2022).

In this final experiment, we investigated the effect of familiarity on distance in face space using a simple computational model. Our approach utilised principal components analysis (PCA), followed by linear discriminant analysis (LDA), with these techniques being well established in the field (e.g., Kramer et al., 2018). Although the use of LDA was expected to minimise within-identity distances while maximising between-identity distances for trained images, our focus here was on how novel instances were represented as familiarity increased. In other words, did prior familiarity (through training) alter face space to accommodate the recognition of new instances? In addition, following on from Experiments 1 and 2, we considered whether familiarity more strongly influenced within-person distances in comparison with those between people.

Method

Stimuli

We used the image set featured in Kramer et al. (2018), which comprised 4,154 photographs of 335 different identities. The number of images per identity ranged from a single image (for 161 identities) to 159 images, with M = 22.16 images, SD = 26.20 images. This set was collected using

Google Images searches for actors, athletes, etc., with each image depicting the identity's unobscured face in colour, posing within approximately ± 30° from full face to facilitate the placement of landmarks (for further details, see Kramer et al., 2018). Otherwise, the images were unconstrained regarding pose, expression, age, lighting, and camera conditions, and were cropped to include only the head.

Model specification

Following the general approach used in previous work (e.g., Burton et al., 2016; Kramer, Jenkins, et al., 2017; Kramer, Young, et al., 2017; Kramer et al., 2018), all images were first landmarked (via the semi-automatic placement of 82 fiducial points, e.g., corners of eyes, corners of mouth, etc.; see Burton et al., 2016) and then shape-standardised by morphing each of them to a template derived from the average shape of the entire set. This resulted in a vector of 40,755 numbers (95 pixels wide x 143 pixels high x 3 RGB layers) that represented each image.

For those (normalised) images/identities that represented prior knowledge (i.e., the training set), we carried out PCA to reduce the image vectors without significant loss of variability, resulting in their representation within a 335-dimensional space. These highest 335 principal components (PCs; explaining 97.7% of the variance in the original RGB information) were retained as this was the minimum number required for the LDA due to the number of identities involved. The images' projections on these PCs were then entered into an LDA, where each class represented an identity. This resulted in a reshaped space comprising 334 dimensions (the number of identities minus 1). Again, with the goal of reducing the number of dimensions without significant loss in performance, we retained the first 215 components, which accounted for 95.0% of the 'discriminability' from the overall LDA space. To be clear, this process was applied to the training set images only, and the specific images included in this set were subject to minor variations detailed below (e.g., the number of images depicting identities which we chose to vary in their levels of familiarity).

To investigate how an increase in familiarity might affect identity differentiation within our modelled face space, we selected two identities from our image set (Ryan Gosling – 101 images; Ryan Reynolds – 104 images) who demonstrated the same general appearance (White, male, Hollywood actors of a similar build, height, and age) and for whom a large number of images had been collected. For these two men, we considered the distances between two novel images of the same man, and one novel image of each man, in order to address the predictions supported by the results of Experiments 1 and 2.

Model justification

We utilised PCA, followed by LDA, to model prior knowledge of faces. This approach has previously been shown to simulate key findings within the face perception literature, including the benefits of familiarity when matching faces, resisting the effects of image degradation, recognising novel instances of trained faces, and better recognising internal (in comparison with external) facial features (Kramer et al., 2018). Such models have also been used to simulate the identification of faces as they change across the lifespan (Mileva et al., 2020), or vary in social categories like sex and race (Kramer, Young, et al., 2017). To be clear, we are not arguing that the brain is also carrying out a combination of PCA + LDA, or that this approach produces levels of performance comparable with humans or state-of-the-art algorithms. Instead, the benefits of implementing this type of model are that it uses relatively straightforward, transparent, and well understood procedures, and that its properties are easy to manipulate/interrogate (see Young & Burton, 2021).

As in previous work, we operationalised familiarity as the number of different training images used, motivated by the behavioural findings that increased exposure to face variability produced improvements in face learning, matching, and searching (e.g., Baker et al., 2017; Corpuz & Oriet, 2022; Matthews & Mondloch, 2018; Menon et al., 2015; Mileva & Burton, 2019; Murphy et al., 2015; Ritchie & Burton, 2017).

PCA is an unsupervised, 'bottom-up' technique for representing the faces within a lower-dimensional space, based solely on the statistical properties of the image set. In contrast, LDA is a supervised, 'top-down' method which attempts to cluster the images by identity, minimising intra-class and maximising inter-class differences. As such, LDA represented the learning of identities in that each image was given an identity label, with these providing the information necessary for deriving dimensions that maximised identity discrimination.

Procedure

Our PCA + LDA face space had prior knowledge of a core 333 identities (3,949 images), varying in their levels of familiarity (i.e., the number of images in the training set per identity). To investigate the effects of familiarity on the distances between images in this space, we manipulated the levels of familiarity with two additional identities (Gosling and Reynolds).

For one iteration of our simulation, the model was initially trained (PCA + LDA) with an image set containing the core 333 identities, along with 15 images each of Gosling and Reynolds. These were chosen at random from the images available for the two identities. In addition, three images of each man were randomly selected (again, from all available images but excluding those used for training) to serve as test images. These six images were projected into the face space that resulted from training the model.

After training, we calculated the Euclidean distances between 1) two images of Gosling; 2) two images of Reynolds; and 3) the third (remaining) image of each man. These same three distances, using the same six images, were then calculated as we increased the familiarity of both identities within the training set. That is, the number of images of both men was increased from 15 images each, in steps of 15, up to 90 images each.[2] These were chosen at random from the images

[2] We began with 15 rather than zero images because a model with no familiarity is qualitatively different from one with some familiarity – the estimated face space would have to change to accommodate two new identities. As such, we focussed here on a continuous increase in familiarity rather than incorporating this initial step from 'none' to 'some' familiarity. (We thank an anonymous reviewer for highlighting this issue.)

available (excluding the test images) while building on the previous step (i.e., the first 15 images in the training set remained while an additional 15 were added in, and so on). Each time, the model was trained and the three distances were calculated. This process is illustrated in Figure 6, along with our predictions for the effects of increased familiarity.

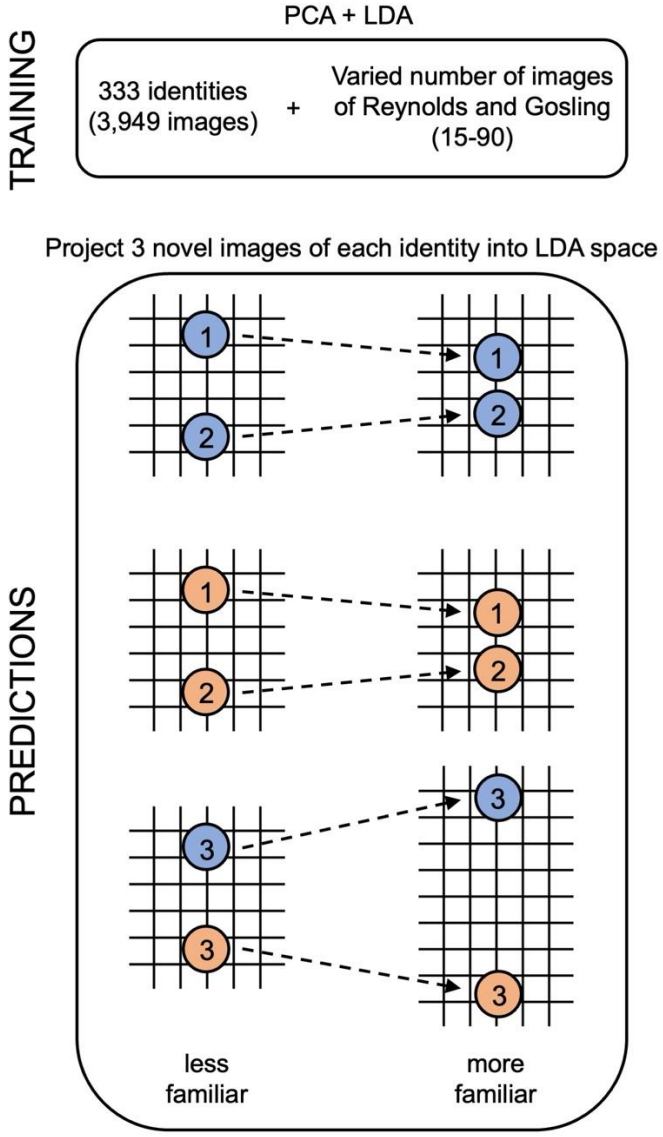

Figure 6. A summary of the model's training, where familiarity with two identities was systematically increased from 15 to 90 images. The predictions illustrate how the distances between the three test images of each identity (1-3) are expected to change as familiarity with the two identities (blue and orange) is increased.

This training and testing process represented a single iteration of our simulation. In total, we carried out 100 iterations, each time selecting the training and test images at random from our original database as described above to avoid image-specific effects. Therefore, for each iteration and each level of familiarity (ranging from 15 to 90 in steps of 15 images), we calculated the distances for the three pairs of images.

Analytic strategy

To describe the results of our simulations, we again relied on a Bayesian hierarchical model-based approach. We predicted Euclidean distance (i.e., similarity) from increasing familiarity and a dummy-coded variable that represented the three identity conditions (the between-identities image pair, which served as the reference category, and two predictors indexing the within-Gosling and within-Reynolds pairs) and the interaction between these variables. We also included a group-specific effect of a random intercept for each iteration of the simulation. More simply, our model simultaneously identified the association between familiarity and Euclidean distance for each image pair condition, accounting for variability in results over simulation iterations.

The prior structure of the model followed the same convention as in Experiment 1, with weakly-informative priors set on all model parameters and a Gaussian likelihood, but the scale of the priors was altered to reflect the (arbitrary) scale of the Euclidean distance metric. The intercept and coefficients representing identity condition, familiarity, and their interaction had a Gaussian distribution prior, with a mean of zero and a standard deviation of 10,000, with a half-Gaussian distribution with a standard deviation of 5,000 used for the error variance. The group-specific effect for the simulation was also Gaussian with a mean of zero and a standard deviation of 10,000. Four Markov Monte Carlo chains were run, with each having 3,000 tuning steps and 4,000 samples drawn from the posterior. The model converged, and all parameters had an $\hat{R}$ of < 1.01.

Model interpretation

We again predicted a specific pattern of results. For both the within-Gosling and within-Reynolds pairs, we expected negative slopes, such that increasing familiarity decreased the Euclidean distance. In contrast, we predicted that the between-identities pair would have a positive slope – increasing familiarity would yield greater distance. These estimates were recovered from the model by the simple addition of the familiarity slope (which represents the slope of the between-identities pair) to the two interaction coefficients, which (before addition) represented the difference between the between-identities slope and the slope for the within-Gosling and within-Reynolds pairs, respectively.

Results

After estimating the model, we predicted the expected Euclidean distance for each image pair and level of familiarity (see Figure 7). Recovering the familiarity slope for each image pair showed negative associations for both the within-Reynolds and within-Gosling pairs, b = -4.77, [-7.71, -1.80], $p(\boldsymbol{\theta} < 0)$ = 99.9%, and b = -5.62, [-8.56, -2.69], $p(\boldsymbol{\theta} < 0)$ = 100%, respectively. In contrast, the slope for the between-identities pair was positive, b = 5.93, [2.99, 8.84], $p(\boldsymbol{\theta} > 0)$ = 100%. Given these certain probabilities, the joint hypothesis – that both within-identity pairs were negatively associated with increased familiarity while the between-identities pair was positively associated – was 99.8%.

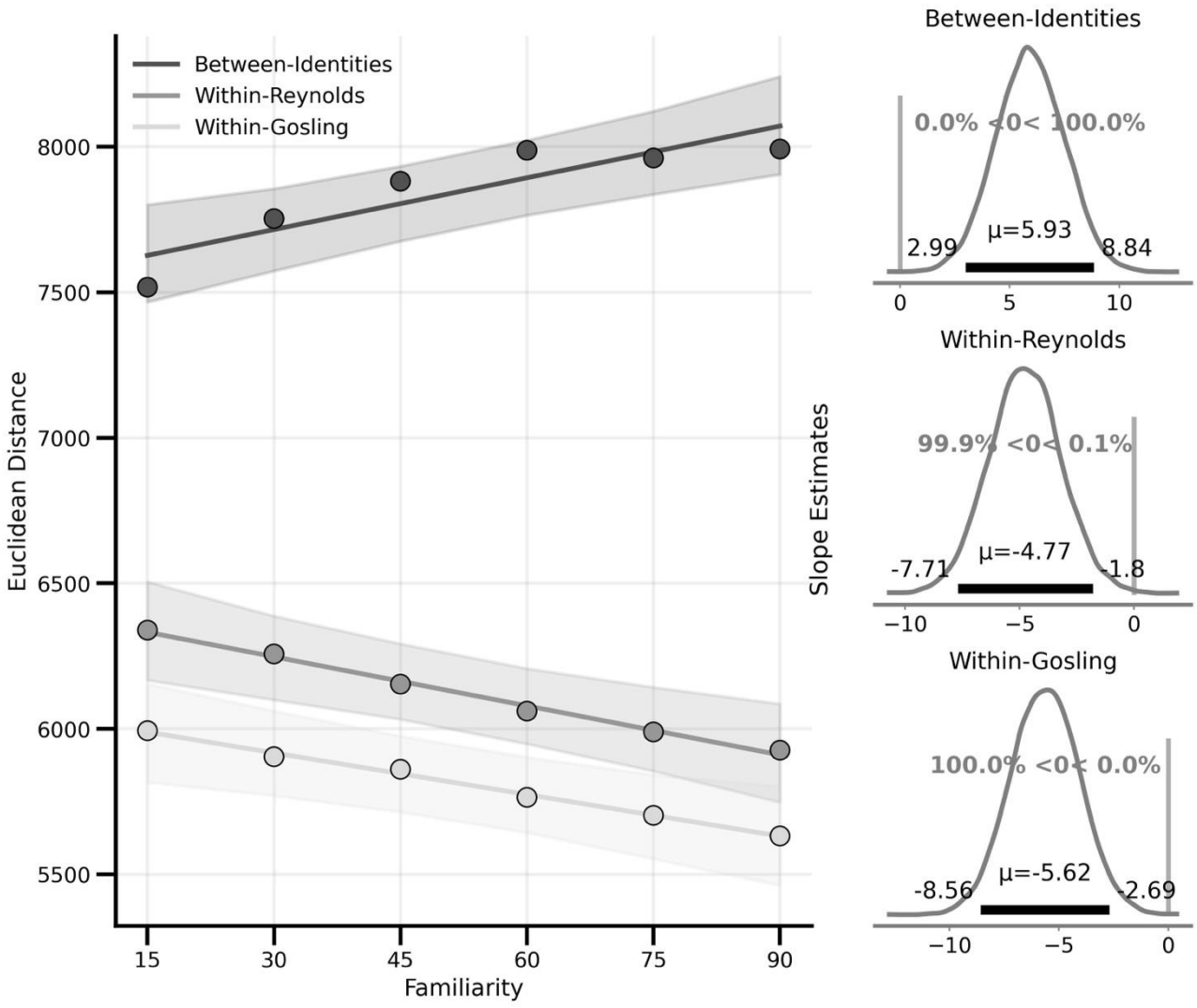

Figure 7. Left panel – the association between each image pair and familiarity, extracted from the simulation data by the hierarchical model. Points represent the average Euclidean distance calculated from the raw data at that familiarity level, though the model was estimated from disaggregated data. Right panels – slope estimates of the association between each image pair and familiarity, with associated posterior probabilities.

Ruling out a possible confound

We have shown that increased familiarity produced a decrease in the distances between images of the same identity, while increasing the distances between images of different identities. However, by considering within- and between-identity distances for the two men being learned, it

may be that these measures have been confounded. In other words, since the within-person distances decreased with increasing familiarity, could this be causing the apparent increase in distance between images of the two men? If novel images of Reynolds are represented closer to each other *and* the same is true for the novel images of Gosling, perhaps this explains why one novel image of each man appear further from each other.

To address this possibility, during the above simulations, we collected additional data. Alongside calculating the distance between one novel image of each man (see the bottom of Figure 6), we also calculated the distances from both of these images to representations of two trained (already familiar) men. To this end, we selected two identities from the trained image set of 333 identities – Brandon Beemer (97 images) and Christian Oliver (97 images). Both men shared the same general appearance as Reynolds and Gosling (i.e., White, male, actors of a similar build, height, and age).

During each iteration, after the model was trained with the varying number of images of Gosling and Reynolds, we calculated the positions of both Beemer's and Oliver's centroids (i.e., the average location across all images for that identity). We then calculated the distances from each of the novel images of Gosling and Reynolds to these two centroids. Throughout the training process, the number of images of Beemer and Oliver did not vary (remaining at 97 for each man). As such, if the distance from a novel image of Reynolds to Beemer's centroid, for instance, increased with increasing familiarity with Reynolds then this could not be explained by a decrease in the within-identity distances of Beemer's images and a resulting increase in distance to one of his images.

To analyse these distances, as before, we employed a Bayesian regression model that predicted Euclidean distance from increasing familiarity, a dummy-coded variable representing the two test identities (Gosling and Reynolds), and a dummy-coded variable representing the novel identities (Beemer and Oliver), and allowed all of these variables to interact. Our model thus estimated the association between familiarity and distance within each combination of test identity

and trained identity simultaneously, allowing us to test the hypothesis that these slopes were positive.

The prior structure of this model followed the same convention as above, with weakly-informative priors on all parameters and a Gaussian likelihood, scaled to reflect the Euclidean distance metric. The intercept and coefficients, including interactions, had wide Gaussian priors with a mean of zero and a standard deviation of 10,000, and the error variance had a half-Gaussian with a standard deviation of 5,000. Four chains were run, with 3,000 tuning steps and 4,000 samples drawn from the posterior. The model converged with all parameters having an $\hat{R}$ of 1.

We examined the slopes within each combination of test and trained identity to confirm a positive association, showing clear relationships between familiarity and distance – Gosling to Beemer, b = 5.25, [3.65, 6.81]; Gosling to Oliver, b = 10.78, [9.19, 12.31]; Reynolds to Beemer, b = 4.42, [2.84, 5.95]; and Reynolds to Oliver, b = 8.22, [6.68, 9.79]. These are illustrated in Figure 8 and clearly confirm the prediction that increased familiarity resulted in increasing distances between identities, even when considering the distances to trained, already familiar identities.

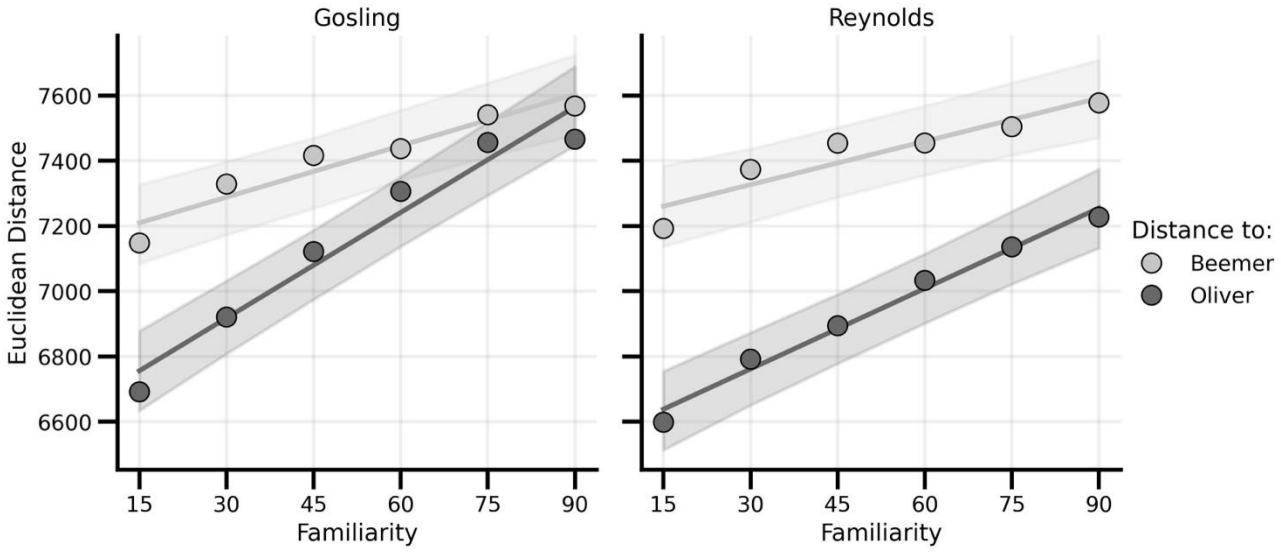

Figure 8. The distances between novel images of our test identities (Gosling and Reynolds) and the centroids of two trained (familiar) identities (Beemer and Oliver). Shaded areas represent 94%

credible intervals, while points represent the observed average distances calculated from the raw data.

Confirming these patterns are identity-specific

Our results have shown that increasing familiarity with an identity produced a decrease in the distances between images of that identity, while increasing the distance to other identities. However, this increase in familiarity also (necessarily) involved an increase in the number of images in the training set as a whole. We must therefore consider whether this increase in model complexity (i.e., the total number of trained images) would produce our pattern of results for any identity in the model and not just those which are increasing in familiarity.

To address this possibility, during the above simulations, we collected additional data. Alongside calculating the distances described above for novel images of Gosling and Reynolds, we also considered six novel (untrained) images of Beemer and Oliver (three of each identity). Specifically, we calculated the Euclidean distances between 1) two images of Beemer; 2) two images of Oliver; and 3) the third (remaining) image of each man. Since the familiarity of both men remained constant throughout our simulations (i.e., 97 images of each in the training set), we predicted no decrease in within-identity distances, and no increase in between-identity distance, for these novel images. Mirroring the process with Gosling/Reynolds, the training and test images of these two men were selected at random from the original set of 100 images for each identity at the start of each iteration to avoid image-specific effects.

Model estimation followed the same analysis strategy as with the novel images of Gosling and Reynolds (above). We predicted the Euclidean distances for the within-Oliver and within-Beemer image pairs, as well as the between-identities pair, for each familiarity level. The familiarity slopes for within-Oliver and within-Beemer pairs were both close to zero, $b = 0.74$, [-1.30, 2.95], $p(\boldsymbol{\theta} < 0) = 26.2\%$, and $b = 0.14$, [-2.05, 2.23], $p(\boldsymbol{\theta} < 0) = 44.9\%$, respectively. The slope

for the between-identities pair was positive and small in magnitude, although still likely to be positive, b = 1.97, [-0.18, 4.12], $p(\boldsymbol{\theta} > 0)$ = 95.8%. This is illustrated in Figure 9.

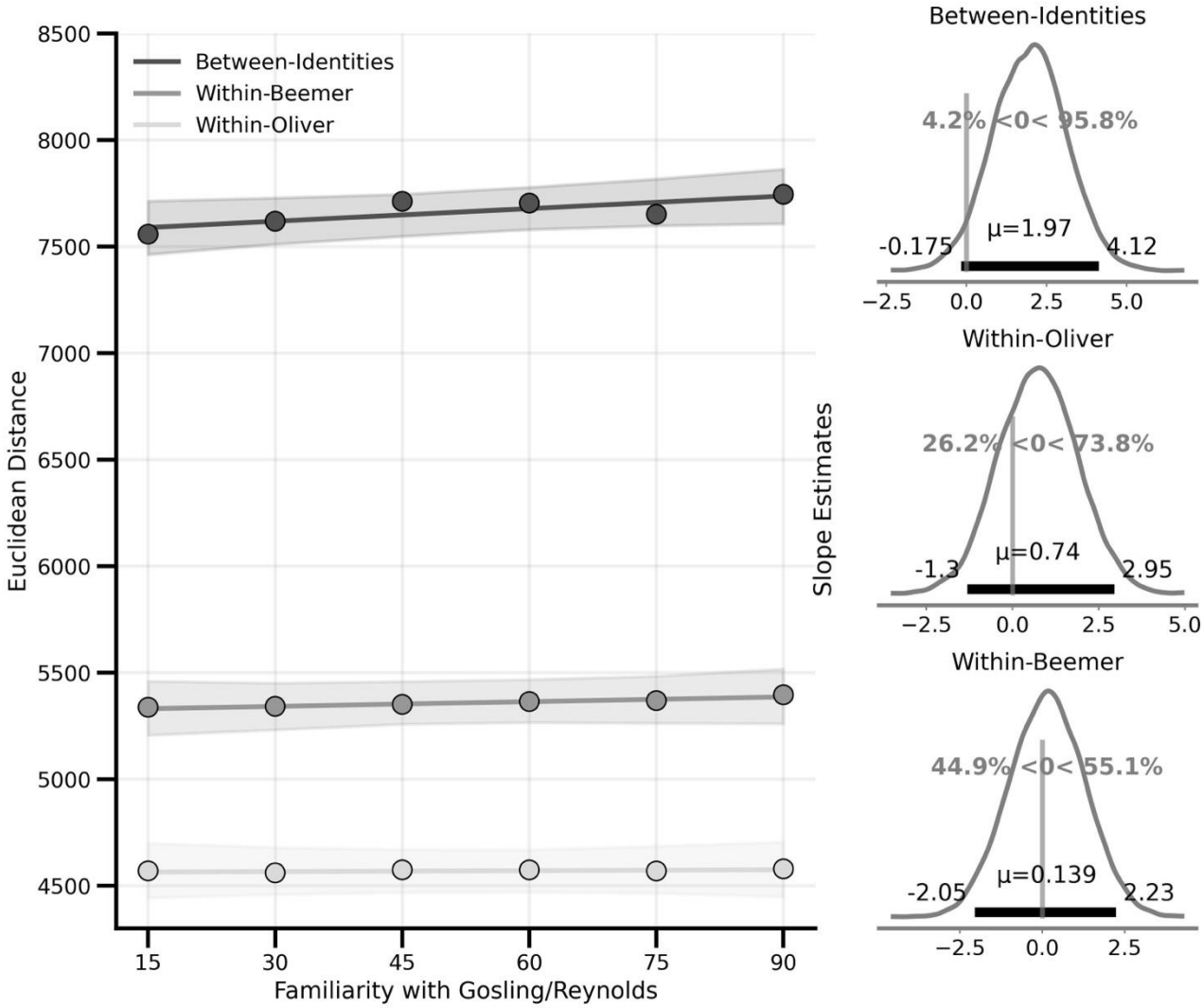

Figure 9. Left panel – the association between each image pair as familiarity with Gosling/Reynolds increased, extracted from the simulation data by the hierarchical model. Points represent the average Euclidean distance calculated from the raw data at that familiarity level, though the model was estimated from disaggregated data. Right panels – slope estimates of the association between each image pair and Gosling/Reynolds familiarity, with associated posterior probabilities.

Taken together, these results suggest that simply increasing model complexity (i.e., by increasing the number of images in the training set) had little effect on within-identity distances for novel images when the familiarity of those identities remained constant. However, between-identities distance showed a slight increase, even for these 'constant familiarity' identities, suggesting the potential influence of the increasing number of training images (of other identities) in the model. Importantly, this slope was far shallower in comparison with the Gosling/Reynolds slope (see Figure 7), where familiarity with those identities varied.

Replicating our findings with a minority group

It is worth noting that our training set (333 identities) was well-balanced in terms of gender (168 women and 165 men) but not ethnicity (269 White, 52 Black, 12 other). Therefore, it may be the case that the above results are specific to the learning of White identities (i.e., Gosling and Reynolds), given that the majority of the training set shared this ethnicity. To investigate further, we repeated the above analyses but considered a new pair of identities – Kelly Rowland and Beyoncé Knowles-Carter. For these women, we collected and landmarked images in line with the creation of the original image database (100 images of Rowland, 101 images of Knowles). The simulation steps, as well as the model and prior specification, were identical to the previous analyses.

As before, after model estimation, we predicted the Euclidean distances for the within-Rowland and within-Knowles image pairs, as well as the between-identities pair, for each familiarity level. The familiarity slopes for within-Rowland and within-Knowles pairs were both negative and similar in magnitude to the original model implementation, b = -5.45, [-8.20, -2.80], $p(\boldsymbol{\theta} < 0)$ = 100%, and b = -5.73, [-8.43, -2.97], $p(\boldsymbol{\theta} < 0)$ = 100%, respectively. The slope for the between-identities pair, however, was close to zero, b = -0.20, [-2.89, 2.52], $p(\boldsymbol{\theta} > 0)$ = 44.5%. Taken together, the joint hypothesis that both within-identity slopes were negative while the between identity-pair slope was positive was 44.5%. This is illustrated in Figure 10.

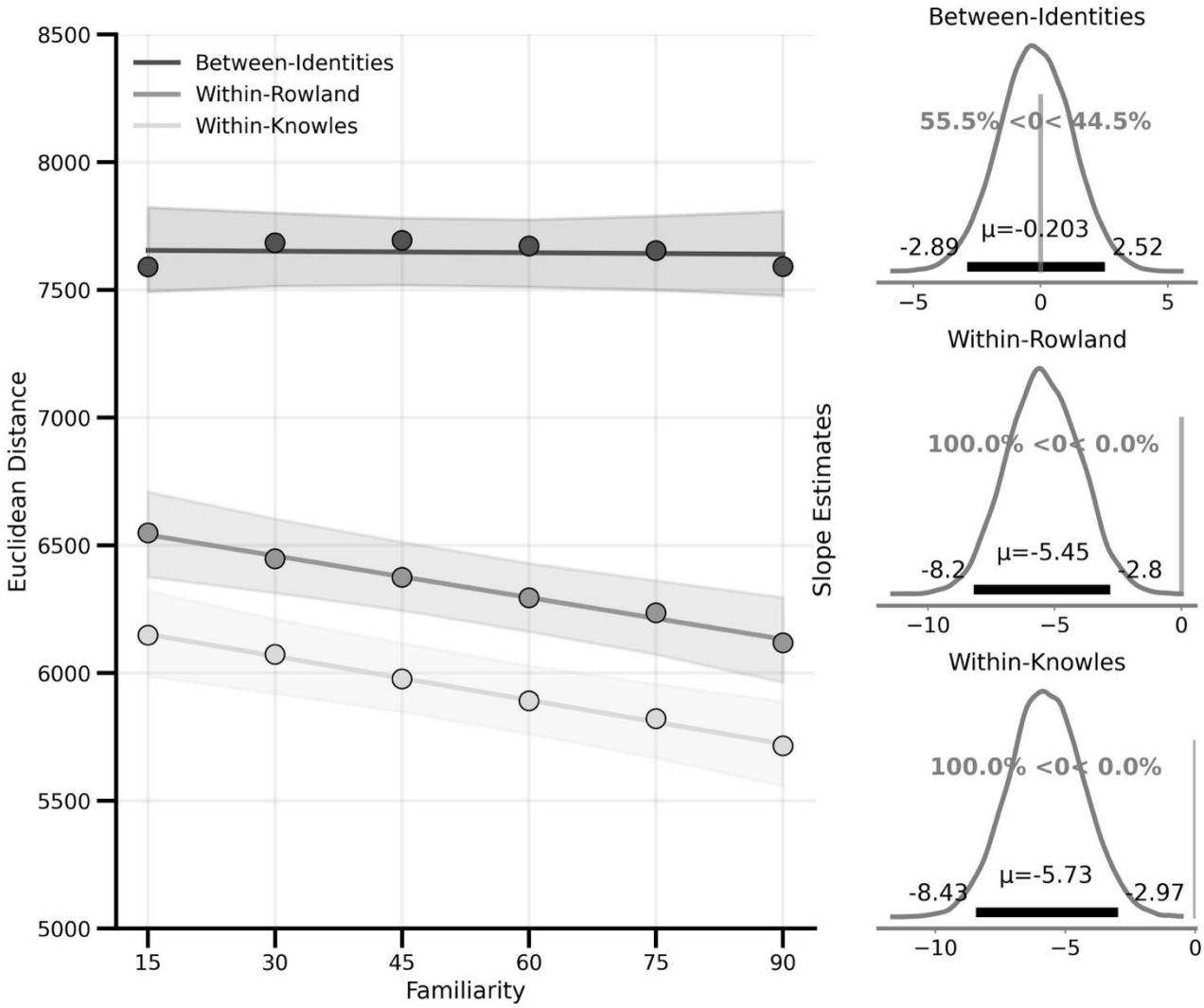

Figure 10. Left panel – the association between each image pair and familiarity, extracted from the simulation data by the hierarchical model. Points represent the average Euclidean distance calculated from the raw data at that familiarity level, though the model was estimated from disaggregated data. Right panels – slope estimates of the association between each image pair and familiarity, with associated posterior probabilities.

These results suggest that there may be a type of other-race effect present in our model, whereby the distance between Black identities (who represent the minority in our training set) did not increase alongside an increase in their familiarity. However, the decrease in within-identity distances appears more robust and was again present in these data.

General Discussion

Across three experiments, we have provided support for the existence of two representational changes that underlie face familiarity. Higher levels of familiarity with a face resulted in 1) a perceived increase in similarity between different instances of that face, as well as 2) a perceived decrease in similarity between different faces. These findings were evident in both the continua of pre-existing familiarities with celebrities (Experiment 1) and the learned familiarity of previously novel identities (Experiment 2). For Experiment 3, we modelled the effects of familiarity in face space using a simple, supervised classification approach (LDA). We found that novel images (i.e., ones not included within the training set) of increasingly familiar identities also demonstrated these two patterns of results if we considered the distance between images as a proxy for similarity. However, we uncovered some caveats to this result, which we discuss below.

The results of Experiments 1 and 2 suggested that the influence of familiarity was smaller for between-person, in comparison with within-person, similarity. In other words, increasing familiarity with a face may produce a more coherent representation of that person predominantly via decreasing the distances between their instances in face space. Although we found evidence that between-person similarity also decreased with increasing familiarity, suggesting larger distances between different people in face space, the strength of this effect was smaller. This result mirrors the findings of White and colleagues (2022), whose investigation focussed on the similarities between identity averages, as well as between instances and these averages. That both approaches suggest a lesser role for between-person representational distances is in contrast with early models of face space, which concentrated primarily on this transformation while ignoring the importance of within-person variability (e.g., Valentine et al., 2016).

Our computational models, presented in Experiment 3, also provided evidence of a difference in the two representational changes under investigation. Increasing familiarity resulted in a robust

decrease in within-identity distances that could not be explained by a simple increase in the number of training images in the overall model, and this change was present for both White and Black identities. However, we found that the increase in between-identities distances may have, in part, been due to training set size in addition to familiarity. For two identities that remained constant in terms of familiarity, we detected a slight increase in the distance between their representations as the number of training images (of other identities) increased within the model. Although requiring further investigation, there is behavioural evidence to support this idea. Balas and Saville (2015, 2017) found that more limited early experience with faces was associated with poorer face processing. In our model, simply incorporating more images within the face space, no matter who they depict, may improve identity discrimination in general. Therefore, we suggest future studies might explore this line of questioning further.

In addition, the increase in between-identities distance was absent for identities representing ethnic minorities. This result mirrors older research using autoassociative networks (trained on one image per identity), where similarity was higher for novel images of minority race faces in comparison with those representing the majority of the training set (O'Toole et al., 1991). Here, the dimensions of the face space model may be more attuned to describing variability within and between images of the majority race and, as such, were poorer at differentiating between (by locating further apart) minority race faces. This is in contrast with the decrease in within-identity distances as familiarity increased, which was also present for minority race faces, perhaps supporting previous findings of a stronger role for this transformation in face discrimination (White et al., 2022).

Additional modelling in Experiment 3 suggested that the increase in the representational distance between identities (at least, for Gosling and Reynolds) was not simply a by-product of within-identity images becoming more tightly clustered with increasing familiarity (by considering the distances to already familiar identities' centroids). However, we acknowledge that this approach still may not have entirely ruled out the possibility, and to do so completely may be challenging.

While the change in distance/similarity between identities may play a lesser role in familiarity's influence on face representations (White et al., 2022), our results across all three experiments suggest this change is still present in refining perceptions of increasingly familiar faces (although perhaps may feature more heavily for own-race faces).

Experiment 2 sought to directly manipulate familiarity through 5 minutes of exposure using video clips. Recent evidence has suggested that this length of time may be sufficient to produce a robust face representation (Popova & Wiese, 2023), and future research could investigate the minimum amount of exposure required to detect noticeable changes in similarity perceptions. Indeed, we might predict a continuous change whereby perceptions of similarity increase (or decrease, as applicable) steadily as familiarity with a new face increases. Further, our manipulation of familiarity was limited to the viewing of videos, unlike the real-world social interactions we experience every day. Following the methods of Popova and Wiese (2023), researchers might incorporate a more realistic learning paradigm to investigate the resulting shift in our internal representations of newly-learned faces.

Our experiments investigated the effects of familiarity on our identity representations overall, but it may be worth considering identity-specific differences in this process. Within- and between-identity similarities must also depend on the identities involved, and so the benefits of familiarity for recognition (through alterations to face space) will likely vary as a result. For instance, how individuals with particularly invariant features (e.g., Bono's tinted sunglasses; Parde et al., 2017) are represented may result in a lesser benefit of familiarity. To date, research has little considered face-level differences when it comes to learning and familiarity.

Along similar lines, the variability in previously experienced/trained images may play a role in similarity perceptions. For instance, if such images were highly variable then initial learning of an identity may be more challenging, but the resulting representation might be more robust, and therefore better able to deal with representing new images closer together in face space (presumably facilitating recognition). In contrast, low variability images may produce simpler learning for a

given identity, but the representation formed from these might lead to representing new images in a less tightly clustered manner. Previous research has shown that learning an identity from high- rather than low-variability images resulted in benefits during subsequent speeded name verification and matching tasks (Ritchie & Burton, 2017). Therefore, future research might investigate the influence of variability during learning rather than quantity (i.e., familiarity) alone.

While our investigation focussed on the change in perceived similarity between images of the same person, we were partly limited by logistical constraints (e.g., the time to complete the experiment). As a result, we chose to include a range of identities but only two images per identity. Although it would certainly be interesting to explore similarity perceptions for multiple pairs of images of the same identity, we were also concerned by the necessary exposure required. If participants were shown several pairs of images, we expect that they would start to develop some familiarity with the identity during the task (much like a 'two-sort' card sorting task; Andrews et al., 2015). Since this would be a particularly problematic by-product of using several image pairs, we chose to limit our design to only one image pair per identity.

Finally, a recently proposed signal detection-based model (Fitousi, 2023) has attempted to provide an integrated account of various findings regarding unfamiliar face matching. However, by simply considering familiarity as a shift in perceived similarity between images (as we have demonstrated here), this model is no longer limited in scope to unfamiliar face perception. Future research can begin to apply such models to a broader array of contexts with the goal of deriving a more unified theory of face perception.

In conclusion, our previous understanding of how we internally represent faces has focussed on differentiating between different people (e.g., Valentine et al., 2016). Far less is known about the challenge of 'telling people together' (Burton, 2013). In other words, how we deal with, and indeed utilise, the variability of each face for identification (Burton et al., 2016). Familiarity plays a substantial role in this process, and yet the changes underlying its beneficial effects in face recognition are not well understood. Recent work has identified potential transformations (Blauch et

al., 2021; Kramer et al., 2018) and provided some initial support for these (White et al., 2022). Here, we present further evidence of both hypothesised representational changes, with the aim of driving forwards our understanding in this field.

Funding

This work was supported by the Israel Science Foundation grant to Daniel Fitousi (ISF-1498/21).

References

Andrews, S., Jenkins, R., Cursiter, H., & Burton, A. M. (2015). Telling faces together: Learning new faces through exposure to multiple instances. *Quarterly Journal of Experimental Psychology, 68*(10), 2041-2050.

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods, 52*(1), 388-407.

Baker, K. A., Laurence, S., & Mondloch, C. J. (2017). How does a newly encountered face become familiar? The effect of within-person variability on adults' and children's perception of identity. *Cognition, 161*, 19-30.

Balas, B., Sandford, A., & Ritchie, K. (2023). Not the norm: Face likeness is not the same as similarity to familiar face prototypes. *i-Perception, 14*(3), 1-12.

Balas, B., & Saville, A. (2015). N170 face specificity and face memory depend on hometown size. Neuropsychologia, 69, 211-217.

Balas, B., & Saville, A. (2017). Hometown size affects the processing of naturalistic face variability. *Vision Research, 141*, 228-236.

Blauch, N. M., Behrmann, M., & Plaut, D. C. (2021). Computational insights into human perceptual expertise for familiar and unfamiliar face recognition. *Cognition, 208*, 104341.

Bruce, V., Henderson, Z., Newman, C., & Burton, A. M. (2001). Matching identities of familiar and unfamiliar faces caught on CCTV images. *Journal of Experimental Psychology: Applied, 7*(3), 207-218.

Burton, A. M. (2013). Why has research in face recognition progressed so slowly? The importance of variability. *The Quarterly Journal of Experimental Psychology, 66*(8), 1467-1485.

Burton, A. M., Wilson, S., Cowan, M., & Bruce, V. (1999). Face recognition in poor-quality video: Evidence from security surveillance. *Psychological Science, 10*(3), 243-248.

Burton, A. M., Kramer, R. S. S., Ritchie, K. L., & Jenkins, R. (2016). Identity from variation: Representations of faces derived from multiple instances. *Cognitive Science, 40*(1), 202-223.

Burton, A. M., White, D., & McNeill, A. (2010). The Glasgow face matching test. *Behavior Research Methods, 42*(1), 286-291.

Capretto, T., Piho, C., Kumar, R., Westfall, J., Yarkoni, T., & Martin, O. A. (2020). *Bambi: A simple interface for fitting Bayesian linear models in Python*. arXiv. https://doi.org/10.48550/arXiv.2012.10754

Clutterbuck, R., & Johnston, R. A. (2002). Exploring levels of face familiarity using an indirect face-matching measure. *Perception, 31*, 985-994.

Clutterbuck, R., & Johnston, R. A. (2004). Matching as an index of face familiarity. *Visual Cognition, 11*, 857-869.

Clutterbuck, R., & Johnston, R. A. (2005). Demonstrating how unfamiliar faces become familiar using a face matching task. *European Journal of Cognitive Psychology, 17*, 97-116.

Collins, E., & Behrmann, M. (2020). Exemplar learning reveals the representational origins of expert category perception. *Proceedings of the National Academy of Sciences, 117*(20), 11167-11177.

Corpuz, R. L., & Oriet, C. (2022). Within-person variability contributes to more durable learning of faces. *Canadian Journal of Experimental Psychology, 76*(4), 270-282.

de Fockert, J., & Wolfenstein, C. (2009). Rapid extraction of mean identity from sets of faces. *The Quarterly Journal of Experimental Psychology, 62*, 1716-1722.

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science, 6*(3), 274-290.

Etz, A., Gronau, Q. F., Dablander, F., Edelsbrunner, P. A., & Baribault, B. (2018). How to become a Bayesian in eight easy steps: An annotated reading list. *Psychonomic Bulletin & Review, 25*(1), 219-234.

Faerber, S. J., Kaufmann, J. M., Leder, H., Martin, E. M., & Schweinberger, S. R. (2016). The role of familiarity for representations in norm-based face space. *PLOS ONE, 11*(5), e0155380.

Fitousi, D. (2023). A signal detection–based confidence–similarity model of face matching. *Psychological Review*. Advance online publication.

Fysh, M. C., & Bindemann, M. (2018). The Kent face matching test. *British Journal of Psychology, 109*(2), 219-231.

Gelman, A., Simpson, D., & Betancourt, M. (2017). The prior can often only be understood in the context of the likelihood. *Entropy, 19*(10), 10.

Hauser, D. J., & Schwarz, N. (2016). Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods, 48*(1), 400-407.

Jenkins, R., White, D., Van Montfort, X., & Burton, A. M. (2011). Variability in photos of the same face. *Cognition, 121*(3), 313-323.

Johnston, R. A., & Edmonds, A. J. (2009). Familiar and unfamiliar face recognition: A review. *Memory, 17*(5), 577-596.

Koca, Y., & Oriet, C. (2023). From pictures to the people in them: Averaging within-person variability leads to face familiarization. *Psychological Science, 34*(2), 252-264.

Kramer, R. S. S., Jenkins, R., & Burton, A. M. (2017). InterFace: A software package for face image warping, averaging, and principal components analysis. *Behavior Research Methods, 49*, 2002-2011.

Kramer, R. S. S., Ritchie, K. L., & Burton, A. M. (2015). Viewers extract the mean from images of the same person: A route to face learning. *Journal of Vision, 15*(4):1, 1-9.

Kramer, R. S. S., Young, A. W., & Burton, A. M. (2018). Understanding face familiarity. *Cognition, 172*, 46-58.

Kramer, R. S. S., Young, A. W., Day, M. G., & Burton, A. M. (2017). Robust social categorization emerges from learning the identities of very few faces. *Psychological Review, 124*(2), 115-129.

Kruschke, J. K. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan* (2nd ed.). Academic Press.

Kruschke, J. K. (2018). Rejecting or accepting parameter values in Bayesian estimation. *Advances in Methods and Practices in Psychological Science, 1*(2), 270-280.

Kruschke, J. K., & Liddell, T. M. (2018). The Bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review, 25*(1), 178-206.

Makowski, D., Ben-Shachar, M. S., Chen, S. H. A., & Lüdecke, D. (2019). Indices of effect existence and significance in the Bayesian framework. *Frontiers in Psychology, 10*, 2767.

Matthews, C. M., & Mondloch, C. J. (2018). Improving identity matching of newly encountered faces: Effects of multi-image training. *Journal of Applied Research in Memory and Cognition, 7*(2), 280-290.

Megreya, A. M., & Burton, A. M. (2006). Unfamiliar faces are not faces: Evidence from a matching task. *Memory & Cognition, 34*, 865-876.

Menon, N., White, D., & Kemp, R. I. (2015). Variation in photos of the same face drives improvements in identity verification. *Perception, 44*(11), 1332-1341.

Mileva, M., & Burton, A. M. (2019). Face search in CCTV surveillance. *Cognitive Research: Principles and Implications, 4*, 37.

Mileva, M., Kramer, R. S., & Burton, A. M. (2019). Social evaluation of faces across gender and familiarity. *Perception, 48*(6), 471-486.

Mileva, M., Young, A. W., Jenkins, R., & Burton, A. M. (2020). Facial identity across the lifespan. *Cognitive Psychology, 116*, 101260.

Murphy, J., Ipser, A., Gaigg, S. B., & Cook, R. (2015). Exemplar variance supports robust learning of facial identity. *Journal of Experimental Psychology: Human Perception and Performance, 41*(3), 577-581.

Neumann, M. F., Schweinberger, S. R., & Burton, A. M. (2013). Viewers extract mean and individual identity from sets of famous faces. *Cognition, 128*, 56-63.

O'Toole, A. J., Deffenbacher, K., Abdi, H., & Bartlett, J. (1991). Simulating the "other-race effect" as a problem in perceptual learning. *Connection Science, 3*(2), 163-178.

Parde, C. J., Castillo, C., Hill, M. Q., Colon, Y. I., Sankaranarayanan, S., Chen, J. C., & O'Toole, A. J. (2017). Face and image representation in deep CNN features. In *Proceedings of the 12th IEEE International Conference on Automatic Face and Gesture Recognition* (pp. 673-680).

Popova, T., & Wiese, H. (2023). How quickly do we learn new faces in everyday life? Neurophysiological evidence for face identity learning after a brief real-life encounter. *Cortex, 159*, 205-216.

Ramon, M., & Gobbini, M. I. (2018). Familiarity matters: A review on prioritized processing of personally familiar faces. *Visual Cognition, 26*(3), 179-195.

Ritchie, K. L., & Burton, A. M. (2017). Learning faces from variability. *The Quarterly Journal of Experimental Psychology, 70*(5), 897-905.

Ritchie, K. L., Kramer, R. S. S., & Burton, A. M. (2018). What makes a face photo a 'good likeness'? *Cognition, 170*, 1-8.

Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review, 21*, 301-308.

Salvatier, J., Wiecki, T. V., & Fonnesbeck, C. (2016). Probabilistic programming in Python using PyMC3. *PeerJ Computer Science*, *2*, e55.

Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion, and race in face recognition. *The Quarterly Journal of Experimental Psychology*, *43*(2), 161-204.

Valentine, T., Lewis, M. B., & Hills, P. J. (2016). Face-space: A unifying concept in face recognition research. *The Quarterly Journal of Experimental Psychology*, *69*(10), 1996-2019.

Welsch, R., von Castell, C., & Hecht, H. (2020). Interpersonal distance regulation and approach-avoidance reactions are altered in psychopathy. *Clinical Psychological Science*, *8*(2), 211-225.

White, D., Wayne, T., & Varela, V. P. (2022). Partitioning natural face image variability emphasises within-identity over between-identity representation for understanding accurate recognition. *Cognition, 219*, 104966.

Young, A. W., & Burton, A. M. (2021). Insights from computational models of face recognition: A reply to Blauch, Behrmann and Plaut. *Cognition, 208*, 104422.