Humans' extreme face recognition abilities challenge the well-established familiarity effect.

Gailt Yovel[1,2], Eden Bash[1], Sarah Bate[3]

[1] School of Psychological Sciences Tel Aviv University, Tel Aviv, Israel
[2] Sagol School of Neuroscience, Tel Aviv University, Tel Aviv, Israel
[3] Department of Psychology, Bournemouth University, UK

Corresponding author:

Galit Yovel
School of Psychological Sciences
Sagol School of Neuroscience
Tel Aviv University, Tel Aviv, Israel
gality@tauex.tau.ac.il
ORCID: 0000-0003-0971-2357

Running head: Humans with extreme face recognition abilities show no familiarity effect

Keywords: Face Recognition, Classification, Familiarity, Individual differences, Identification

The data can be viewed in this link:

https://osf.io/vm2fg/?view_only=6c38e3b3768a43d68ba5bfabfe374756

Acknowledgment

The study was funded by the Israeli Science Foundation 917/21 to GY. We would also like to thank Adva Shoham for help with data analysis.

Abstract

Classification performance is better for learned than unlearned stimuli. This was also reported for faces, where identity matching of unfamiliar faces is worse than for familiar faces. This familiarity advantage led to the conclusion that variability across appearances of the same identity is partly idiosyncratic and cannot be generalized from familiar to unfamiliar identities. Recent advances in machine vision challenge this claim by showing that the performance for untrained (unfamiliar) identities reached the level of trained identities as the number of identities that the algorithm is trained with increases. We therefore asked whether humans who reportedly can identify a vast number of identities, such as super recognizers, may close the gap between familiar and unfamiliar face classification. Consistent with this prediction, super recognizers classified unfamiliar faces just as well as typical participants who are familiar with the same faces, on a task that generates a sizable familiarity effect in controls. Additionally, prosopagnosics' performance for familiar faces was as bad as that of typical participants who were unfamiliar with the same faces, indicating that they struggle to learn even identity-specific information. Overall, these findings demonstrate that by studying the extreme ends of a system's ability we can gain novel insights into its actual capabilities.

Highlights

- Familiarity effect for faces suggested identity-specific representation.
- We found that super recognizers (SRs) show no face familiarity effect.
- This suggests no evidence for identity-specific representation.
- Prosopagnosics showed no generalization even for familiar identities.

1. Introduction

Successful classification relies on the ability to reveal the critical information that defines the boundary of a category. This information can be learned via a process known as category learning, in which a general classification rule is formed based on learned examples and is applied to new, unlearned examples (Goldstone et al., 2013; Raviv et al., 2022). However, information learned from trained examples may not fully generalize to unlearned examples, leading to better performance for learned than unlearned classes. Such advantage for learned classes is often reported for faces, where learning to classify different images of one set of familiar (learned) identities does not generalize well to unfamiliar (unlearned) identities (Jenkins et al., 2011a; Samuel et al., 2018; Young & Burton, 2018). This results in much poorer performance for identity matching of unfamiliar faces compared to familiar faces. These findings led to the proposal that within-person variability in facial appearance is composed of general variance that is common to all faces and enables generalization across them. But there is also an idiosyncratic component that is unique to each individual identity (Burton et al, 2016). Therefore, identity information that is learned from experience with familiar identities is not useful for identity decisions on unfamiliar identities (Jenkins & Burton, 2011; Kramer et al., 2018; Megreya et al., 2006; Burton et al, 2016; Ritchie, Burton, Burton, Burton, et al., 2017; Young & Burton, 2018).

The claim that within-person variability includes identity-specific information implies that performance for unfamiliar faces will be always inferior to familiar faces. However, recent advances in machine learning algorithms have shown that deep convolutional neural networks (DCNNs) that are optimized for face recognition can learn identity-general information from one set of identities and successfully apply it to the classification of unlearned identities (Blauch et al., 2020; Taigman et al., 2014). DCNNs' performance for untrained identities improves and reaches the level of trained identities when the algorithm is trained on a large number of identities (Blauch et al., 2021). This suggests that face identity can be represented by a general classification rule that can be applied for successful identification of unfamiliar faces. While it remains possible that the human face recognition system does not have the same capabilities as DCNNs (Young & Burton, 2020), an innovative way of testing this hypothesis is to test "super recognizers" (SRs) - humans who are proficient at face recognition and often report that they are able to identify large numbers of identities: (Russell et al., 2009) – and examine if they are able to learn identity-general representations, closing the gap in performance between familiar and unfamiliar face classification.

Humans show substantial individual differences in face recognition ability (Yovel et al., 2014). Whereas on the one end of the distribution are people with prosopagnosia who cannot recognize even familiar faces (Bate & Tree, 2017; Behrmann & Avidan, 2005; Duchaine & Nakayama, 2006; Susilo & Duchaine, 2013; White et al., 2017), on the other end of the distribution are SRs (Bate, Portch, Mestry, & Bennetts, 2019; Bobak et al., 2016; Davis et al., 2016; Dunn et al., 2020; Ramon, 2021), who find it extraordinarily easy to recognise unfamiliar individuals that they have only briefly seen before (see (Bate et al., 2021) for a discussion of the definition of SRs) and report that they can identify a very large number of identities (Russell et al., 2009). This anecdotal information raises the question of whether the ability to identify numerous familiar identities enables SRs to learn identity-general facial information that can be effectively applied to unfamiliar faces. Yet, whereas many studies have shown that SRs are better than controls at identity matching tasks that use unfamiliar faces (Bate & Dudfield, 2019; Bobak et al., 2016; Phillips, Yates, Hu, Hahn, Noyes, Jackson, Cavazos, Jeckeln, Ranjan, Sankaranarayanan, Chen, et al., 2018; Ramon, 2021; Russell et al., 2009; Stacchi et al., 2020), it is unknown whether they are as good at unfamiliar face matching as individuals who are familiar with the same identities. Such a finding would challenge the prevalent view that the human face recognition system cannot learn identity-general face information from experience with familiar faces, which can be effectively used for classification of unfamiliar faces.

To test whether SRs can classify unfamiliar faces as well as individuals who are familiar with them, we designed a computerized version of the elegant card sorting identification task developed by Jenkins and colleagues (Jenkins et al., 2011). In this task, typical participants were asked to sort a stack of 40 cards that displayed facial images of two identities into piles of the same identity. Intriguingly, individuals who were not familiar with the faces sorted them into 4-11 identity piles, whereas participants who were familiar with the faces easily sorted them into 2 piles. However, this, and other similar studies that have shown this familiarity advantage, only tested individuals with normal face recognition abilities (Mileva et al., 2019; Ritchie, Burton, Burton, & Burton, 2017). Ramon and colleagues did include the face identity card sorting task as part of their SR screening program (Ramon, 2021; Stacchi et al., 2020), but they only tested classification performance of unfamiliar identities. Thus, the question of whether SRs' performance equated with that of participants who were familiar with the faces at test remains unknown.

To that effect, we asked SRs to sort images that displayed celebrity faces from their own country (UK) and those from a different country (Israel). We also included typical participants from the two countries (UK and Israel) to replicate Jenkins et al.'s findings with our revised task. Further, by comparing SRs from the UK with Israeli controls on Israeli celebrity faces (that were familiar to the controls but unfamiliar to the SRs), we could assess if SRs close the large gap in performance between familiar and unfamiliar face classification on this challenging task. Finally, we also tested the performance of UK developmental prosopagnosics (DPs) on the same task. Whereas DPs' poor face matching abilities have been demonstrated in many studies in the past two decades (Bennetts et al., 2022; Stantić et al., 2022; White et al., 2017), their performance for familiar identities has never been directly compared to individuals who are unfamiliar with the same identities. Our design will therefore enable us to determine, for the first time, whether DPs show any benefit with familiar faces over and above participants who are unfamiliar with those identities.

2. Methods

2.1 Participants: A total of 113 participants completed the task, of which 25 were SRs (average age: 44 years, range 30-59), 14 were DPs (average age: 49 years, range 31-60), 45 were UK controls (average age: 46 years, range: 30-60) and 31 were Israeli controls (average age: 40 years, range: 31-55). UK and Israeli control participants were recruited on the Prolific platform. SRs and DPs were recruited based on initial face recognition screening tests (Bate, Bennetts, Gregory, et al., 2019; Bate, Bennetts, Tree, et al., 2019). Participants were asked to read a consent form. By clicking OK they approve that they are willing to participate in the experiment and continued to the next stage. The familiarity effect size reported by Jenkins et al. (2011) was d = 1.4. Power analysis ($p < 0.05$, two-sided with a power = 0.8) indicates that such an effect size can be detected with a sample of 7 participants.

<u>Super recognizer screening tests</u>. The SRs were screened based on their performance on the following challenging face recognition tests: (1) the Cambridge Face Memory Test – long form (CFMT+, 102 items;(Russell et al., 2009) average = 94.2), (2) the Models Memory Test (MMT, 90 items, (Bate, Bennetts, Hasshim, et al., 2019; Bate et al., 2018)average = 75.5), and (3) the Pairs Matching Test (PMT, 48 items,(Bate, Bennetts, Hasshim, et al., 2019; Bate et al., 2018) average = 40). In line with existing criteria (Bate et al., 2021), participants had to outperform controls on at least two of these tests to be included in the study (see Supplementary Table 1 for individuals scores and cutoff scores for each test).

<u>Developmental prosopagnosia screening tests</u>. All DPs reported severe difficulties with face recognition in daily life. Their face recognition skills were assessed with three dominant screening

tasks: (1) the Cambridge Face Memory Test (CFMT, 72 items, (Duchaine & Nakayama, 2006)range: 28-45, average = 38), (2) the Cambridge Face Perception Test (CFPT, upright items only, (Duchaine et al., 2007), error range: 38-80, average = 60), and (3) a famous faces test (Bate, Bennetts, Gregory, et al., 2019)range: 19-67, average=44) (see Supplementary Table 2 for individuals scores and cutoff scores for each test). In addition, all DPs self-reported everyday difficulties with face recognition, and declared no history of brain injury or concurrent psychiatric, developmental or intellectual conditions.

2.2 Stimuli: We selected 10 face images of each of two male and two female people who are famous in the UK, and corresponding images of four celebrities from Israel, via a Google image search. To increase the difficulty of the task, we selected two male and two female identities that were visually similar, (see Figure 1 and supplementary material). Pilot testing with five Israeli and five UK participants ensured that the faces were familiar within the congruent country and unfamiliar in the incongruent country. In the experiment, we excluded trials at the participant level where this was not the case (see Data Analysis section).

Figure 1: **An example trial of Israeli celebrity faces**. Left: 20 images of two identities are presented in two rows in a random order at the bottom of the screen. Right: Participants are asked to sort the images into piles of different identities, by moving them on the screen to form identity clusters. The example shows the correct classification into two different identities of two Israeli famous faces.

2.3 Procedure: We created a computerized version of the card sorting task. All participants performed the task online. The task included four trials. In each trial, 20 face images, ten of each of two identities were intermixed at the bottom of the screen in two rows (Figure 1, left). After completing demographic information about gender and age, participants were presented with written instructions where they were asked to sort the images into piles of different identities by dragging them with the mouse. That is, participants were instructed to cluster images of the same identity in the same pile, and those of different identities into their respective piles (see Figure 1, right). The participant was allowed to continue to the next trial only after all the images had been moved by the mouse. Participants were not told the number of identities in advance. UK participants were presented first with the male and female Israeli celebrities, and then with the male and female UK celebrities; Israeli participants were presented with the UK celebrities first and then the Israeli celebrities. This ensured that their successful performance with the familiar identities would not inform them about the number of identities for the unfamiliar faces. Within each country of origin, the order of the male and female faces was randomized. At the end of the task, participants were presented with one image of each of the eight identities that were presented during the study and were asked whether they are familiar or unfamiliar with those individuals. DPs were also asked whether they were familiar with the name of the UK famous identities. This data was then used to exclude trials in which participants were unexpectedly unfamiliar with the particular celebrities from their own country, or where they happened to be familiar with a face from the different country (see Data Analysis).

2.4 Data Analysis

Familiarity based exclusions: We excluded trials in which at least one of the unfamiliar identities was unintentionally familiar to participants (UK faces for Israeli participants and Israeli faces for UK participants), as well as trials in which at least one of the familiar identities turned out to be unknown (UK participants for UK faces and Israeli participants for Israeli faces). This resulted in the exclusion of four unfamiliar trials and four familiar trials in SRs, 10 familiar trials and two unfamiliar trials in Israeli controls, and 16 familiar trials and eight unfamiliar trials in UK controls. For the DPs, we included trials in which DPs were familiar with the name of the celebrity faces even if they did not recognize their face. A total of four DPs were not familiar with at least one of the familiar UK identities based on their name which resulted in exclusion of 5 unfamiliar trials for the DPs.

Classification performance measures: We computed the following measures: (1) number of clusters, (2) number of generalization errors and (3) number of separation errors, and (4) The area under the ROC curve based on the distance between each pair of face images that were sorted into piles on the screen. *Number of clusters* refers to the number of piles that were generated by the participants when asked to sort the images based on their identities. *Generalization errors* refer to the number of different identity images that were clustered with the wrong identities. To count generalization errors, we identified the two clusters with the maximum number of images of each of the two identities, and counted the number of images of the other identity that were included in each of these clusters. *Separation errors* are the number of same identity images that were separated into different identities. To count separation errors, we identified the two clusters with the maximum number of images of each of the two identities, and counted the number of images that were not included in any of these two clusters.

Our computerized task enabled us to compute a fourth performance measure: a ROC (Receiver Operating Curve) and to compute the area under the curve (AUC) as another measure of classification performance. To do so, we calculated the Euclidean distance (between each pair of images based on the location (x, y coordinates) that the participants placed them on the screen. This distance was used to calculate the correct classification for different thresholds for same and different identity pairs to generate a ROC curve. We then computed the area under the curve (AUC) for each participant on each trial. It is noteworthy that participants were instructed to sort face images into different piles based on their perceived identity, and not to arrange them on the screen according to their perceptual similarity. Thus, this measure does not reflect a continuous perceptual similarity measure, but a discrete measure of identity classification. Thus, this measure determines identity classification based on image proximity and not pre-determined cluster assignment.

To test the statistical significance of between-group differences we used a mixed ANOVA with group (DP, Control UK. Control IL, SR) as between subject factor and familiarity (familiar, unfamiliar) as a within subject factor on each of the four dependent measures, followed by post-hoc comparisons. Because some of the comparisons that were central to our hypothesis predicted a null effect (e.g., no differences between SRs sorting unfamiliar faces in comparison to Israeli controls who were familiar with them) we also performed Bayesian analysis to assess if the data favors the alternative over the null hypothesis. To ensure that the interaction between Group and Familiarity was not due to the inclusion of the IL Control group, which shows reversed pattern of familiarity due to nationality effect, we ran an ANOVA with only UK participants.

3. Results

Table 1 reports the mean performance and range for the DPs and SRs, and the UK and Israeli control groups, based on the number of clusters, AUC, generalization errors and separation errors (see data analysis section for details about each measure).

Table 1: Performance mean (range) for DP and SR UK participants and in UK and Israeli (IL) control participants.

	No. of Clusters		**Generalization Errors**		**Separation Errors**		**AUC**	
Face Country	**IL**	**UK**	**IL**	**UK**	**IL**	**UK**	**IL**	**UK**
DP UK	4.4 (2-10.5)	3.8 (2.0-7.5)	3.82 (0-8)	3.96 (0-7)	5.46 (0-14.5)	5.50 (0-14)	0.63 (0.50-0.91)	0.60 (0.48-0.77)
Controls IL	2.2 (1.5-4)	4.8 (1.0-8.0)	0.39 (0-5)	2.77 (0-10)	0.50 (0-4.5)	5.96 (0-12)	0.86 (0.55-1.0)	0.62 (0.47-0.83)
Controls UK	4.5 (1.5-10)	2.8 (2.0-6.0)	1.72 (0-10)	1.08 (0-4)	6.08 (0-14)	1.70 (0-14)	0.71 (0.48-1.00)	0.82 (0.59-1.0)
SR UK	2.6 (2-10.5)	2.1 (2.0-6.0)	0.07 (0-0.5)	0.13 (0-1.5)	1.63 (0-8)	0.17 (0-3)	0.90 (0.71-1.0)	0.91 (0.54-1.0)

We first examined whether our findings with the computerized task replicated those of Jenkins et al. (2011) by performing an analysis for UK and Israeli familiar faces in UK and Israeli control participants.

3.1 A familiarity advantage in controls: Replication of Jenkins et al. (2011)

Number of clusters: We first analyzed the number of clusters that participants created for celebrity faces from their own country (familiar) versus those from the other country (unfamiliar). A mixed ANOVA with the participants' country (Israel, UK) as a between subjects factor and the country where the face is familiar (Israel, UK) as a within subject factor revealed a significant interaction ($F(1,66) = 153.54$, $p < .001$, $\eta_p^2 = 0.56$) with no significant main effects of participant country ($F(1,66) = 0.41$, $p = .72$, $\eta_p^2 = 0.002$). As shown in Figure 2, Israeli and UK participants classified the faces into a larger number of identities for the unfamiliar UK and Israeli faces, respectively, whereas Israeli and UK participants correctly classified the images to 2 identities for the familiar Israeli and UK faces, respectively. Overall, these findings show that our computerized task replicates Jenkins et al.'s (2011) card sorting task, despite the fact that we used a smaller set of face images (20 in our task and 40 in Jenkins' task) showing that participants classify an array of 20 images of 2 individuals into an average of 4-5 identities if they are not familiar with them, and correctly classify them into two identities if they are familiar with them. Post hoc comparisons shows a significant familiarity advantage both in UK controls ($t(39) = 6.03$, $p < .001$) and in Israeli controls ($t(27) = 7.96$, $p_{bonf} < 0.001$).

Generalization errors: We then counted the number of different identity images that were clustered as the same identity as a dependent measure. Figure 3 (top) shows more errors for unfamiliar than familiar faces in both UK and Israeli participants. A mixed ANOVA with the participants' country (Israel, UK) as a between subjects factor and the country where the face is familiar (Israel, UK) as a within subject factor revealed a highly significant interaction ($F(1,66) = 27.53$, $p < .001$, $\eta_p^2 = 0.29$) with no significant main effects of participant country ($F(1,66) = 0.43$, $p = .51$, $\eta_p^2 = 0.006$). Post hoc comparisons reveal a strong familiarity advantage in Israeli controls ($t(27) = 5.9$, $p_{bonf} < 0.001$). We found no familiarity advantage in the generalization measure for the UK controls ($t(39) = 1.89$, $p = .19$). It should be noted that previous studies that used this task did not find generalization errors, only separation errors. In our task we reveal generalization errors because we intentionally selected visually similar identities to make the task more challenging, particularly for the SRs.

Separation errors: We counted the number of same identity images that were separated into different identities as a dependent measure. Figure 3 (bottom) shows more separation errors for unfamiliar than familiar faces. A mixed ANOVA with participant country (Israel, UK) as a between subjects factor and the country where the face is familiar (Israel, UK) as a within subject factor revealed a highly significant interaction ($F(1,66) = 103.64$, $p < .001$, $\eta_p^2 = 0.61$) with no significant main effect of participant country ($F(1,66) = 0.83$, $p = .36$, $\eta_p^2 = 0.013$). Post hoc comparisons reveal a strong familiarity advantage in both Israeli controls ($t(27) = 8.17$, $p_{bonf} < 0.001$) and UK controls ($t(39) = 7.51$, $p_{bonf} < 0.001$).

AUC: We then ran the same analysis for the classification accuracy measure (AUC). Since participants were not asked to arrange the faces on the screen based on their relative similarity, but to sort them into different piles based on their perceived identity, this measure reflects the accuracy of classification of images into different piles. Figure 4 shows better performance for familiar than unfamiliar faces in both UK and Israeli typical participants. A mixed ANOVA with the participants' country (Israel, UK) as a between subjects factor and the country where the face is familiar (Israel, UK) as a within subject factor revealed a highly significant interaction (F(1,66) =

103.64, $p < .001$, $\eta_p^2 = 0.61$) with no significant main effects of participant country ($F(1,66) = 0.84$, $p = .36$, $\eta_p^2 = 0.01$). Classification performance was better for UK faces than Israeli faces in UK participants and vice versa for Israeli participants. Post hoc comparisons show a significant familiarity advantage in both UK controls ($t(39) = 4.16$, $p_{bonf} < .001$) and in Israeli controls ($t(27) = 7.59$, $p_{bonf} < 0.001$).

3.2 No familiarity advantage in SRs and DPs

To examine how DPs and SRs perform the task and to compare them to controls we added the DPs and SRs to the analysis and performed ANOVAs on the same dependent measures. Figures 2-4 display the DP performance in the two left bars and the SR performance in the two right bars.

Number of clusters: Unlike the control participants who showed strong familiarity advantage, DPs and SRs show no familiarity effect (see Figure 1). A mixed ANOVA with Group (DPs, UK controls, Israeli controls, SRs) and face country (UK, Israel) reveal a main effect of group $F(3,99) = 5.18$, $p < 0.002$, $\eta_p^2 = 0.14$) and a significant interaction of group and face country $F(3,99) = 34.05$, $p < 0.001$, $\eta_p^2 = 0.51$). Post hoc comparisons show no familiarity advantage in DPs ($t(12) = 1.1$, $p_{bonf} = 1$, BF10 = 0.41) or SRs ($t(21) = 1.2$, $p_{bonf} = 1$, BF10 = 2.31), in contrast to the familiarity advantage in controls reported above. Given the superb performance of SRs for unfamiliar (Israeli) faces, our main question was whether they are as good as the Israeli participants who are familiar with these faces. Post hoc comparison and Bayesian analysis revealed no difference between SRs and Israeli controls for the Israeli faces that were unfamiliar to the SRs ($t(48) = 0.93$, $p_{bonf} = 1$, BF10 = 1.13). We also found that DP performance was as poor as Israeli controls for the UK faces ($t(39) = 1.91$, $p_{bonf} = 1$, BF10 = 0.78), which were unfamiliar to the controls but familiar to the DPs.

Analysis restricted to UK participants: To assure that the significant interaction was not due to the reversed familiarity effect of Israeli control participants, we tested the same effect only in UK participants. A mixed ANOVA with Group (DPs, UK controls, SRs) and face country (UK, Israel) reveals a main effect of group $F(2,72) = 6.82$, $p = 0.002$, $\eta_p^2 = 0.16$) and a significant interaction of group and face country $F(2,72) = 4.6$, $p = 0.01$, $\eta_p^2 = 0.11$).

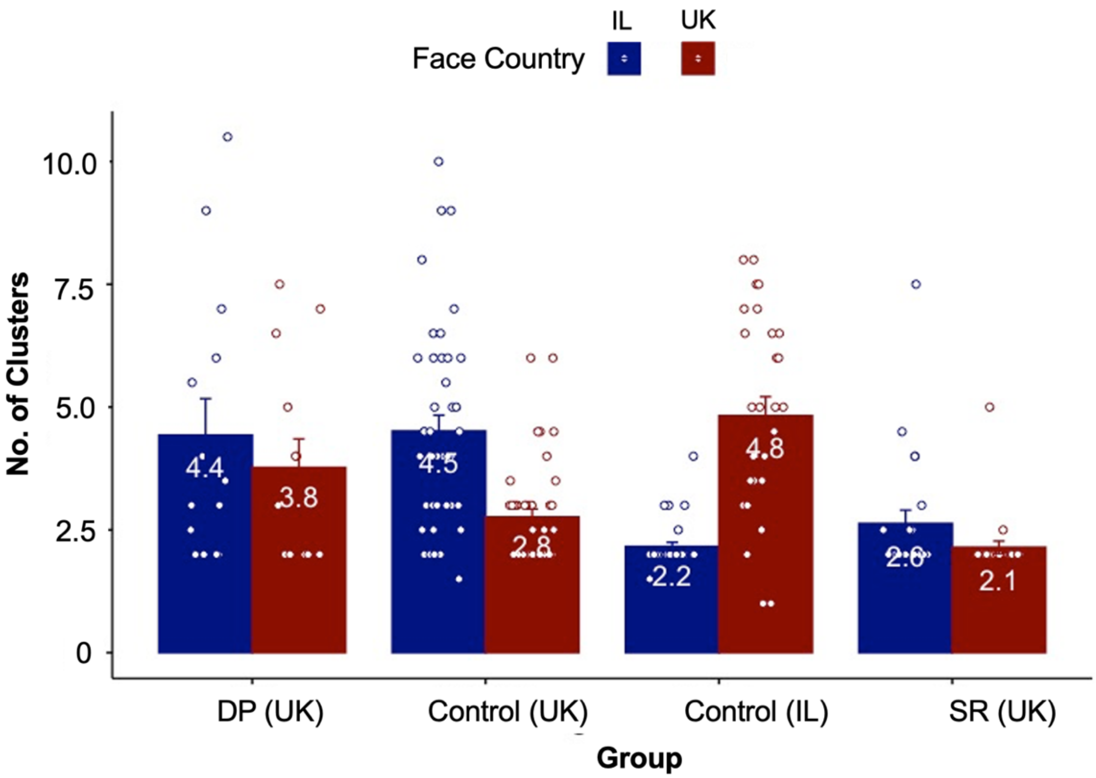
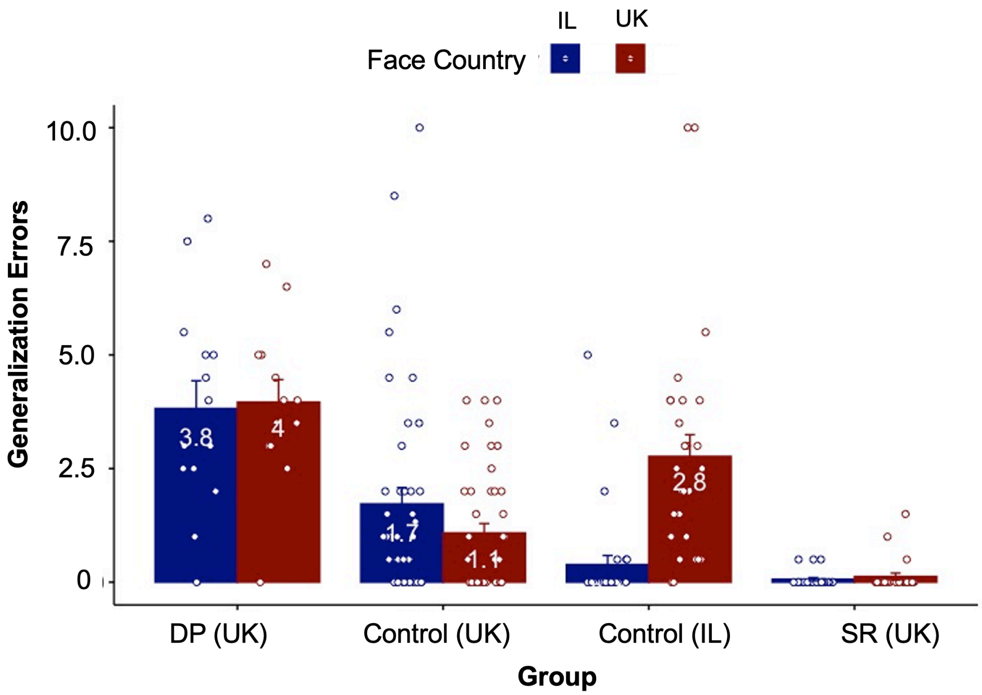


Figure 2: **Number of clusters for UK and Israeli (IL) celebrity faces**. The number of clusters that were generated for UK and Israeli (IL) celebrities by UK and Israeli (IL) controls and DPs and SRs from the UK. Each dot is a participant, and the error bars indicate the S.E.M.

Generalization errors: SRs showed very low number of generalization errors for both the UK and Israeli faces, indicating superb performance for the unfamiliar Israeli faces. DPs, on the other hand, showed a large number of generalization errors for both the UK and Israeli faces, indicating similarly poor performance for familiar and unfamiliar faces (Figure 4). A mixed ANOVA with Group (DPs, UK controls, Israeli controls, SRs) and face country (UK, Israel) revealed a main effect of Group $F(3,99) = 20.19$, $p < 0.001$, $\eta_p^2 = 0.38$) and a significant interaction $F(3,99) = 24.99$, $p < 0.001$, $\eta_p^2 = 0.26$). Post hoc comparisons and Bayesian analysis showed no familiarity advantage in DPs ($t(12) = 0.4$, $p_{bonf} = 1$, $BF10 = 0.29$) or SRs ($t(21) = 0.15$, $p_{bonf} = 1$, $BF10 = 0.29$). In addition, we found no difference between SRs and Israeli controls for the Israeli faces that were unfamiliar to the SRs ($t(48) = 0.73$, $p_{bonf} = 1$, $BF10 = 0.6$), and that DP performance was as poor as Israeli controls for the UK faces ($t(39) = 2.05$, $p_{bonf} = 0.37$, $BF10 = 0.80$), which were unfamiliar to the controls but familiar to the DPs.

Analysis restricted to UK participants: A mixed ANOVA with group (DPs, UK controls, SRs) and face country (UK, Israel) reveals a main effect of group $F(2,72) = 33.74$, $p < 0.001$, $\eta_p^2 = 0.49$) and no significant interaction of group and face country $F(2,72) = 1.39$, $p = 0.25$, $\eta_p^2 = 0.037$).



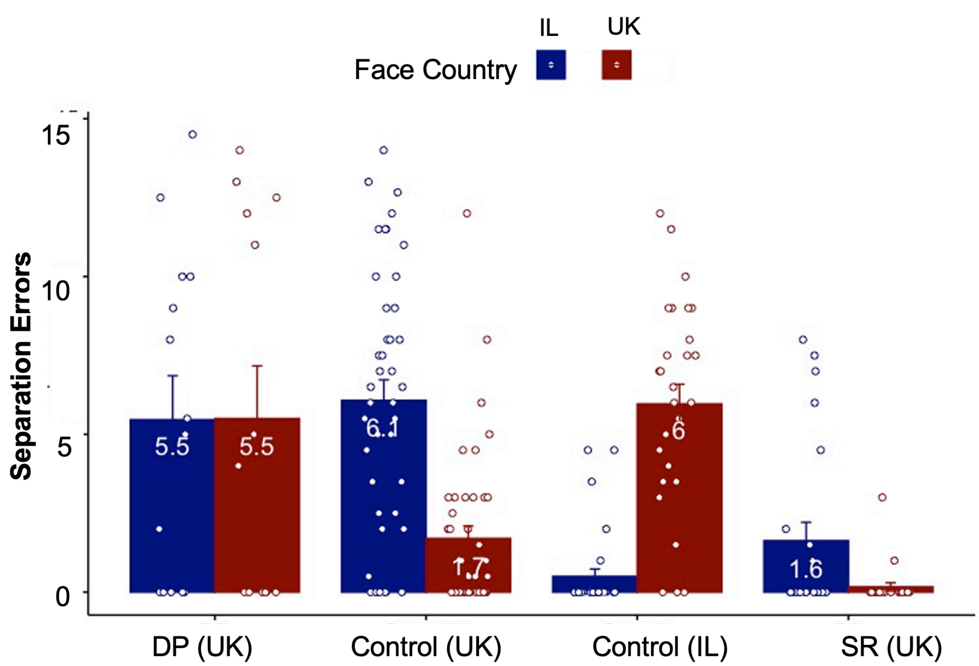


Figure 3: **Generalization and separation errors for UK and Israeli celebrity faces. Top**: The number of generalization errors that were made for face images by UK and Israeli controls and for DPs and SRs from the UK. **Bottom**: The number of separation errors that were made for face images by by UK and Israeli controls and for DPs and SRs from the UK. Each dot is a participant, and the error bars indicate the S.E.M.

Separation errors: Unlike controls who showed a much larger number of separation errors for unfamiliar than familiar faces, DPs showed a similarly large number of separation errors for familiar and unfamiliar faces; SRs showed almost no separation errors for both familiar and unfamiliar faces (Figure 3). A mixed ANOVA with Group (DPs, UK controls, Israeli controls, SRs) and face country (UK, Israel) revealed a main effect of Group ($F(3,99) = 7.96$, $p < 0.001$, $\eta_p^2 = 0.19$) and a significant interaction $F(3,99) = 41.76$, $p < 0.001$, $\eta_p^2 = 0.56$). Post hoc comparisons and Bayesian analysis show no familiarity advantage in DPs ($t(12) = 0.39$, $p_{bonf} = 1$; BF10 = 0.30) or SRs ($t(21) = 1.76$, $p_{bonf} = 0.97$; BF10 = 2.32). In addition, we found no difference between SRs and Israeli controls for the Israeli faces that were unfamiliar to the SRs ($t(48) = 0.99$, $p_{bonf} = 1$, BF10 = 1.38), and that DP performance was as poor as Israeli controls for the UK faces ($t(39) = 0.41$, $p_{bonf} = 1$, BF10 = 0.34), which were unfamiliar to the controls but familiar to the DPs.

Analysis restricted to UK participants: A mixed ANOVA with group (DPs, UK controls, SRs) and face country (UK, Israel) reveals a main effect of group $F(2,72) = 9.88$, $p < 0.001$, $\eta_p^2 = 0.22$) and a significant interaction of group and face country $F(2,72) = 9.69$, $p < 0.001$, $\eta_p^2 = 0.21$).

Accuracy (AUC): Analysis of AUC also reveals that, unlike the control participants who showed a clear familiarity advantage, DPs and SRs show no such benefit (see Figure 4). A mixed ANOVA with group (DPs, UK controls, Israeli controls, SRs) and face country (UK, Israel) reveal a main effect of group $F(3,99) = 20.14$, $p < 0.002$, $\eta_p^2 = 0.38$) and a significant interaction $F(3,99) = 50.23$, $p < 0.001$, $\eta_p^2 = 0.60$). Post hoc comparisons and Bayesian analysis show no familiarity advantage in DPs ($t(12) = 0.16$, $p_{bonf} = 1$, BF10 = 0.28) or SRs ($t(21) = 1.23$, $p_{bonf} = 1$, BF10 = 2.3). In addition, we found no difference between SRs and Israeli controls for the Israeli faces that were unfamiliar to the SRs ($t(48) = 0.08$, $p_{bonf} = 1$, BF10 = 0.3), and that DP performance was as poor as Israeli controls for the UK faces ($t(39) = 0.45$, $p_{bonf} = 1$, BF10 = 0.35), which were unfamiliar to the controls but familiar to the DPs.

Analysis restricted to UK participants: A mixed ANOVA only on UK participants, with group (DPs, UK controls, SRs) and face country (UK, Israel) reveals a main effect of group $F(2,72) = 37.82$, p

< 0.001, η_p^2= 0.51) and a significant interaction of group and face country $F(2,72) = 5.58$, $p = 0.006$, η_p^2= 0.13).

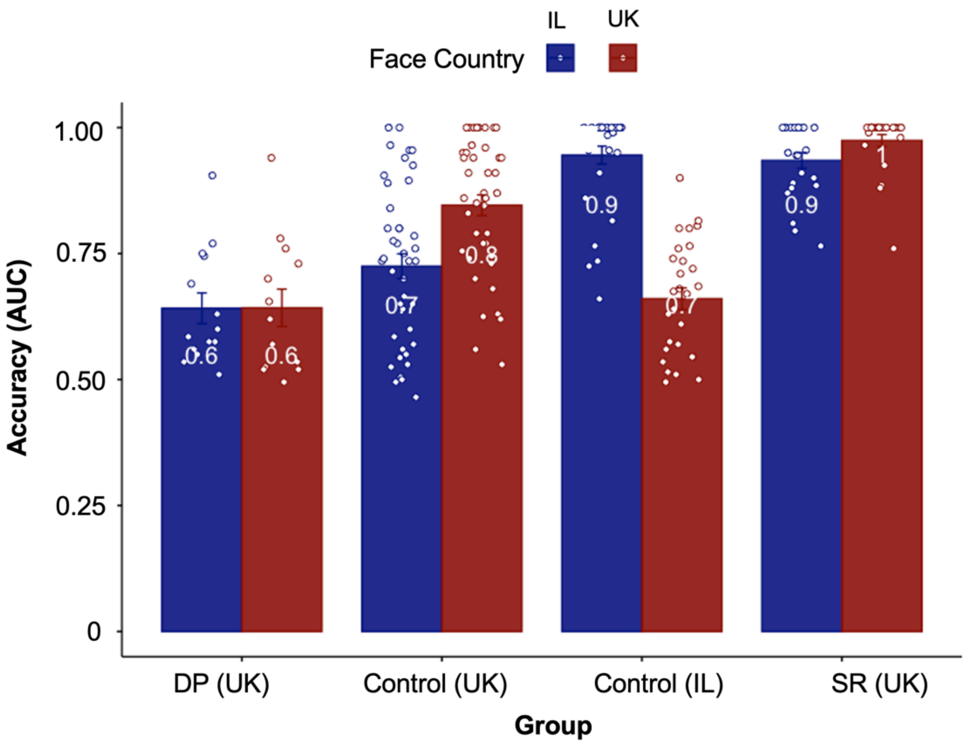


Figure 4: **Classification accuracy for UK and Israeli (IL) celebrity faces.** The area under the curve (AUC) of an ROC was computed for UK and Israeli controls and for DPs and SRs from the UK based on the distance between the face images that were sorted into piles on the screen. Each dot is a participant, and the error bars indicate the S.E.M.

4. Discussion

The better performance that humans show for familiar compared to unfamiliar faces on identity matching tasks is a robust and well-established finding (Jenkins et al., 2011b; Ritchie et al., 2015; Young & Burton, 2017a). Consistent with these findings, our study also found that individuals with normal face recognition abilities showed significantly worse performance for unfamiliar than familiar identities. This familiarity advantage led to the suggestion that within-person face variability is partly idiosyncratic and cannot be fully generalized from familiar to unfamiliar faces (Jenkins et al., 2011a; Young & Burton, 2017b, 2018). Here, we challenge this prevalent view by showing that individuals with superb face recognition abilities can in fact apply such identity-general facial information to unlearned faces. We show, for that first time, that SRs can successfully classify different images of unfamiliar faces just as well as participants who are familiar with them. This finding indicates that human face recognition of unfamiliar faces is not limited by idiosyncratic identity information. Instead, individuals with superb face recognition abilities can learn identity-general information that can be successfully applied to unfamiliar faces. Remarkably, SRs closed the substantial gap in performance between unfamiliar and familiar face recognition ability that is typically found in controls (Figures 2-4).

Our findings also show no familiarity advantage in DPs, a unique population with extremely poor face recognition abilities. DPs showed similarly poor identity classification performance for familiar faces as typical individuals who are unfamiliar with the same faces. Whereas many previous studies with DPs and SRs have shown their respectively poor and superb abilities for unfamiliar faces our study is the first to directly compare this to the performance of individuals who are familiar or unfamiliar with the same faces. This provided us with an upper and lower bound of performance on this task. Overall, our findings show how examination of the extreme ends of the face recognition spectrum can reveal previously unknown capabilities of the system.

Deep learning algorithms have recently reached and even surpassed human-level performance on classification of untrained face images, a task that is parallel to human's classification performance of unfamiliar faces (Phillips, Yates, Hu, Hahn, Noyes, Jackson, Cavazos, Jeckeln, Ranjan, Sankaranarayanan, Chen, et al., 2018). Blauch et al (2020) showed that performance for untrained faces improves and reach the level of trained images as the number of identities that the algorithm is trained with increases (Blauch et al., 2020). These algorithms therefore show that DCNNs trained with a large enough data set can learn an identity-general representation from one set of identities and generalize well to other sets of identities. Nevertheless, such an ability may be super-human and requires certain computations and extensive training that humans do not possess (Young & Burton, 2020). SRs who can identify a very large number of identities were therefore an ideal case to test this hypothesis. We speculated that the very large number of familiar identities that SRs can reportedly identify (Russell et al., 2009) provides them with enough training examples to learn a general classification rule which extracts identity-relevant facial information and ignores identity-irrelevant facial information. This may also account for SRs' ability to recognize celebrity faces from their images before they were famous, a task that is very challenging for controls (Russell et al., 2009). It is also consistent with recent findings that show that SRs better ignore identity-irrelevant changes in facial features, which enables them to better focus on identity-relevant features that are critical for face identification (Abudarham et al., 2021).

Another explanation for SRs' superb abilities with unfamiliar faces is that they may have an innate perceptual machinery that efficiently extracts identity-general facial features that are critical for the identification of all faces, regardless of the vast number of familiar faces they are able to identify. However, the critical role of experience with faces in SRs is evident in their better performance for own- compared to other-race faces (Bate, Bennetts, Hasshim, et al., 2019). Caucasian SRs were found to be worse than Asian controls in an identity matching task using Asian faces. This other-race effect suggests that experience is needed to reach the superb recognition abilities that SRs show for own-race faces, perhaps because other-race faces depend on a different set of facial features that require specific experience in the classification of different races. Consistent with this claim, DCNNs also show worse performance for faces of races that are not included in the training set (Cavazos et al., 2020) indicating that the identity-general information that they learn during training does not generalize to faces of races they are not trained with. Notably, Bate and colleagues also found that SRs were better than Caucasian controls in the Asian face matching and memory tasks. This suggests that SRs can better generalize from their experience with one race to other races, or that they have an innate efficient face classification machinery that enables them to identify the relevant features for classification which is facilitated by relevant experience. These two accounts are not mutually exclusive and both may explain SRs' remarkable performance.

The question of whether individual differences in face recognition, and in particular, at the extreme ends of the distribution, reflect qualitative or quantitative differences is still unresolved. Here we suggest a possible quantitative account, where the number of identities an individual can identify determines classification performance for untrained identities. However, it is not clear what type of mechanism underlies SRs ability to store and identify such an exceptional number of identities and whether a deficit in the same mechanism is what prevents DPs to store and identify familiar faces. Furthermore, whereas the difference between the number of familiar faces that SRs and DPs can identify is a well-known characteristic of their face recognition abilities, a formal test to measure that has not been conducted yet. Such a measure has been proposed and applied by Jenkins and colleagues (2021) in normal individuals. Future studies may use this test to quantify and validate DPs and SRs' reports on the number of familiar identities they can identify and whether it is correlated with performance on unfamiliar face classification task.

To assess performance for familiar and unfamiliar faces we used a computerized version of the face sorting task (Jenkins et al., 2011), in which face images are sorted into piles by moving them with the mouse on the screen. Whereas this task generates a sizable familiarity effect that is abolished in SRs, performance for familiar identities on this task is close to ceiling. Thus, it is still possible that by using a more challenging task or a different measure that is not provided by this task (e.g., reaction time, memory), SRs may show lower performance for unfamiliar faces than individuals who are familiar with the same sets of faces. Notably, the face sorting task is very challenging for unfamiliar faces, and the fact that SRs were able to close this substantial gap is remarkable, suggesting that they can generate an identity-general representation and successfully apply it to unfamiliar faces. Nonetheless, replicating and extending our findings with a more challenging familiar face recognition task in future studies will be valuable.

Our study also reveals novel findings with respect to DPs, who also showed no familiarity advantage on the identity classification task. Whereas DPs' poorer performance for familiar faces is not surprising as this is inherent to their condition, this has never been directly compared to the performance of individuals who are unfamiliar with the same faces on the same task. Our findings show that not only do DPs show similarly poor face classification performance for familiar and unfamiliar faces, but their performance for familiar faces is not better than individuals who are unfamiliar with them. Note that we included in the analysis only trials in which DPs were familiar with the celebrities. Thus, DPs do not benefit from the experience that they have with familiar identities and are unable to reveal the relevant facial information that is required to classify different face images into different identities even for familiar faces. This suggests that they extract very limited information from faces, even if they have had extensive exposure to them. This may also account for the failure to develop an effective training program to improve face recognition in DPs (Bate & Bennetts, 2014; Degutis et al., 2014). It is unknown whether this deficit is evident at the perceptual level or in the operation of their face learning system, such that they cannot form an effective classification rule even for faces that they have extensive experience with.

Sorting images of familiar faces according to their identities can be performed in two ways: either by comparing each image to a representation of the familiar person in memory and identifying it regardless of its similarity to the other images, or by judging the perceptual similarity of the different images and clustering together those that are perceptually similar. For unfamiliar faces, only perceptual similarity can be used for classification. The similar performance that DPs show for familiar and unfamiliar faces implies that they use perceptual similarity judgements for both familiar and unfamiliar faces. Interestingly, their classification ability based on perceptual similarity is not worse than individuals who are unfamiliar with the faces. This may suggest that the deficit in DPs on this task stems from their inability to recognize some of the familiar face images. It further suggests that such recognition may be the mechanism that underlies the familiarity advantage on this task for controls.

The poor performance that humans show for unfamiliar faces also has important implications for face recognition in applied settings. Recent studies have suggested that these individuals should be screened based on their performance on identity matching tasks (Bate, Frowd, Bennetts, Hasshim, et al., 2019; Mayer & Ramon, 2023; White et al., 2014). Our findings further support these suggestions, showing the great differences that individuals with different recognition abilities may demonstrate on these tasks. This may also have implications for the question of human machine collaboration (Phillips, Yates, Hu, Hahn, Noyes, Jackson, Cavazos, Jeckeln, Ranjan, Sankaranarayanan, & others, 2018) and highlight the fact that such collaborations should be evaluated at the individual level rather than based on human averaged performance.

In summary, our study shows for the first time no familiarity advantage in face recognition in SRs and DPs – groups of individuals who are at the extreme ends of the distribution of face recognition abilities. We speculated that this may occur in two different ways. First, a system that can store and identify a large number of learned (familiar) classes can form an identity-general classification rule that can be applied to unlearned (unfamiliar) examples, to the level of performance of individuals who are familiar with them. Second, when a classification rule is not effectively learned due to a certain deficit in the system, it may similarly fail for both learned and unlearned examples. These effects may not be limited to faces and should be tested in future studies in other domains. Overall, our findings indicate that by studying the extreme ends of a system's ability, we can gain novel insights into its operations.

5. References

Abudarham, N., Bate, S., Duchaine, B., & Yovel, G. (2021). Developmental prosopagnosics and super recognizers relyCavazos, J. G., Phillips, P. J., Castillo, C. D., & O'Toole, A. J. (2020). Accuracy comparison across face recognition algorithms: Where are we on measuring race bias? *IEEE Transactions on Biometrics, Behavior, and Identity Science*, *3*(1), 101–111.

on the same facial features used by individuals with normal face recognition abilities for face identification. Neuropsychologia, 160(July), 107963. https://doi.org/10.1016/j.neuropsychologia.2021.107963

Bate, S., Bennetts, R., Hasshim, N., Portch, E., Murray, E., Burns, E., & Dudfield, G. (2019). The limits of super recognition: An other-ethnicity effect in individuals with extraordinary face recognition skills. Journal of Experimental Psychology: Human Perception and Performance, 45(3), 363–377. https://doi.org/10.1037/xhp0000607

Bate, S., & Bennetts, R. J. (2014). The rehabilitation of face recognition impairments: A critical review and future directions. In Frontiers in Human Neuroscience. https://doi.org/10.3389/fnhum.2014.00491

Bate, S., Bennetts, R. J., Gregory, N., Tree, J. J., Murray, E., Adams, A., Bobak, A. K., Penton, T., Yang, T., & Banissy, M. J. (2019). Objective patterns of face recognition deficits in 165 adults with self-reported developmental prosopagnosia. Brain Sciences. https://doi.org/10.3390/brainsci9060133

Bate, S., Bennetts, R. J., Tree, J. J., Adams, A., & Murray, E. (2019). The domain-specificity of face matching impairments in 40 cases of developmental prosopagnosia. Cognition, 192, 1–52. https://doi.org/10.1016/j.cognition.2019.104031

Bate, S., & Dudfield, G. (2019). Subjective assessment for super recognition: An evaluation of self-report methods in civilian and police participants. PeerJ. https://doi.org/10.7717/peerj.6330

Bate, S., Frowd, C., Bennetts, R., Hasshim, N., Murray, E., Bobak, A. K., Wills, H., & Richards, S. (2018). Applied screening tests for the detection of superior face recognition. Cognitive Research: Principles and Implications. https://doi.org/10.1186/s41235-018-0116-5

Bate, S., Frowd, C., Bennetts, R., Hasshim, N., Portch, E., Murray, E., & Dudfield, G. (2019). The consistency of superior face recognition skills in police officers. Applied Cognitive Psychology. https://doi.org/10.1002/acp.3525

Bate, S., Portch, E., & Mestry, N. (2021). When two fields collide: Identifying “super-recognisers” for neuropsychological and forensic face recognition research. Quarterly Journal of Experimental Psychology, 74(12), 2154–2164. https://doi.org/10.1177/17470218211027695

Bate, S., Portch, E., Mestry, N., & Bennetts, R. J. (2019). Redefining super recognition in the real world: Skilled face or person identity recognizers? In British Journal of Psychology (Vol. 110, Issue 3). https://doi.org/10.1111/bjop.12392

Bate, S., & Tree, J. J. (2017). The definition and diagnosis of developmental prosopagnosia. In Quarterly Journal of Experimental Psychology (Vol. 70, Issue 2, pp. 193–200). Psychology Press Ltd. https://doi.org/10.1080/17470218.2016.1195414

Behrmann, M., & Avidan, G. (2005). Congenital prosopagnosia: Face-blind from birth. Trends in Cognitive Sciences, 9(4), 180–187. https://doi.org/10.1016/j.tics.2005.02.011

Bennetts, R. J., Johnson Humphrey, P., Zielinska, P., & Bate, S. (2022). Face masks versus sunglasses: limited effects of time and individual differences in the ability to judge facial identity and social traits. Cognitive Research: Principles and Implications, 7(1). https://doi.org/10.1186/s41235-022-00371-z

Blauch, N. M., Behrmann, M., & Plaut, D. C. (2020). Computational insights into human perceptual expertise for familiar and unfamiliar face recognition. Cognition. https://doi.org/10.1016/j.cognition.2020.104341

Blauch, N. M., Behrmann, M., & Plaut, D. C. (2021). Deep learning of shared perceptual representations for familiar and unfamiliar faces: Reply to commentaries. Cognition, 208. https://doi.org/10.1016/j.cognition.2020.104484

Bobak, A. K., Pampoulov, P., & Bate, S. (2016). Detecting superior face recognition skills in a large sample of young British adults. Frontiers in Psychology, 7(SEP). (Cavazos et al., 2020)

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

Cavazos, J. G., Phillips, P. J., Castillo, C. D., & O'Toole, A. J. (2021). Accuracy Comparison Across Face Recognition Algorithms: Where Are We on Measuring Race Bias? IEEE Transactions on Biometrics, Behavior, and Identity Science, 3(1), 101–111. https://doi.org/10.1109/TBIOM.2020.302726

Davis, J. P., Lander, K., Evans, R., & Jansari, A. (2016). Investigating Predictors of Superior Face Recognition Ability in Police Super-recognisers. Applied Cognitive Psychology, 30(6). https://doi.org/10.1002/acp.3260

Degutis, J., Cohan, S., & Nakayama, K. (2014). Holistic face training enhances face processing in developmental prosopagnosia. Brain. https://doi.org/10.1093/brain/awu062

Duchaine, B., Germine, L., & Nakayama, K. (2007). Family resemblance: Ten family members with prosopagnosia and within-class object agnosia. Cognitive Neuropsychology, 24(4), 419–430. https://doi.org/10.1080/02643290701380491

Duchaine, B., & Nakayama, K. (2006). The Cambridge Face Memory Test: Results for neurologically intact individuals and an investigation of its validity using inverted face

stimuli and prosopagnosic participants. Neuropsychologia. https://doi.org/10.1016/j.neuropsychologia.2005.07.001
Dunn, J. D., Summersby, S., Towler, A., Davis, J. P., & White, D. (2020). UNSW Face Test: A screening tool for super-recognizers. PLoS ONE, 15(11 November). https://doi.org/10.1371/journal.pone.0241747
Goldstone, RL., Kersten, A., & Carvalho, PF. (2013). Concepts and categorization. In Handbook of Psychology (Second Edi, pp. 607–630).
Jenkins, R., & Burton, A. M. (2011). Stable face representations. Philosophical Transactions of the Royal Society B: Biological Sciences, 366(1571), 1671–1683. https://doi.org/10.1098/rstb.2010.0379
Jenkins, R., White, D., Van Montfort, X., & Mike Burton, A. (2011). Variability in photos of the same face. Cognition, 121(3), 313–323. https://doi.org/10.1016/j.cognition.2011.08.001
Kramer, R. S. S., Towler, A., Reynolds, M. G., & Burton, A. M. (2018). Familiarity and Within-Person Facial Variability : The Importance of the Internal and External Features. https://doi.org/10.1177/0301006617725242
Mayer, M., & Ramon, M. (2023). Improving forensic perpetrator identification with Super-Recognizers. Proceedings of the National Academy of Sciences of the United States of America, 120(20). https://doi.org/10.1073/pnas.2220580120
Megreya, A. M., Burton, A. M., Burton, M. A., & Burton, A. M. (2006). Unfamiliar faces are not faces: Evidence from a matching task. Memory & Cognition, 34(4), 865–876. https://doi.org/10.3758/BF03193433
Mike Burton, A. (2013). Why has research in face recognition progressed so slowly? The importance of variability. Quarterly Journal of Experimental Psychology, 66(8), 1467–1485. https://doi.org/10.1080/17470218.2013.800125
Mileva, M., Young, A. W., Kramer, R. S. S., & Burton, A. M. (2019). Understanding facial impressions between and within identities. Cognition. https://doi.org/10.1016/j.cognition.2019.04.027
Phillips, P. J., Yates, A. N., Hu, Y., Hahn, C. A., Noyes, E., Jackson, K., Cavazos, J. G., Jeckeln, G., Ranjan, R., Sankaranarayanan, S., & others. (2018). Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. Proceedings of the National Academy of Sciences, 115(24), 6171–6176.
Ramon, M. (2021). Super-Recognizers – a novel diagnostic framework, 70 cases, and guidelines for future work. In Neuropsychologia (Vol. 158). Elsevier Ltd. https://doi.org/10.1016/j.neuropsychologia.2021.107809

Raviv, L., Lupyan, G., & Green, S. C. (2022). How variability shapes learning and generalization. In Trends in Cognitive Sciences (Vol. 26, Issue 6, pp. 462–483). Elsevier Ltd. https://doi.org/10.1016/j.tics.2022.03.007

Ritchie, K. L., Burton, A. M., Burton, M. A., & Burton, A. M. (2017). Learning faces from variability. Quarterly Journal of Experimental Psychology, 70(5), 897–905. https://doi.org/10.1080/17470218.2015.1136656

Ritchie, K. L., Burton, M. A., Burton, A. M., Burton, M. A., Burton, A. M., & Burton, M. A. (2017). Learning faces from variability. Quarterly Journal of Experimental Psychology, 70(5), 897–905. https://doi.org/10.1080/17470218.2015.1136656

Ritchie, K. L., Smith, F. G., Jenkins, R., Bindemann, M., White, D., & Burton, A. M. (2015). Viewers base estimates of face matching accuracy on their own familiarity: Explaining the photo-ID paradox. Cognition, 141, 161–169. https://doi.org/10.1016/j.cognition.2015.05.002

Russell, R., Duchaine, B., & Nakayama, K. (2009). Super-recognizers: People with extraordinary face recognition ability. Psychonomic Bulletin and Review, 16(2), 252–257. https://doi.org/10.3758/PBR.16.2.252

Samuel, R. S., William, A., Michael, A., Kramer, R. S. S., Young, A. W., & Burton, A. M. (2018). Understanding face familiarity. Cognition. https://doi.org/10.1016/j.cognition.2017.12.005

Stacchi, L., Huguenin-Elie, E., Caldara, R., & Ramon, M. (2020). Normative data for two challenging tests of face matching under ecological conditions. Cognitive Research: Principles and Implications, 5(1). https://doi.org/10.1186/s41235-019-0205-0

Stantić, M., Ichijo, E., Catmur, C., & Bird, G. (2022). Face memory and face perception in autism. Autism, 26(1), 276–280. https://doi.org/10.1177/13623613211027685

Susilo, T., & Duchaine, B. (2013). Advances in developmental prosopagnosia research. Current Opinion in Neurobiology, 23(3), 423–429. https://doi.org/10.1016/j.conb.2012.12.011

Taigman, Y., Yang, M., Ranzato, M., & Wolf, L. (2014). DeepFace: Closing the gap to human-level performance in face verification. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. https://doi.org/10.1109/CVPR.2014.220

White, D., Kemp, R. I., Jenkins, R., Matheson, M., & Burton, A. M. (2014). Passport officers' errors in face matching. PLoS ONE, 9(8). https://doi.org/10.1371/journal.pone.0103510

White, D., Rivolta, D., Burton, A. M., Al-Janabi, S., & Palermo, R. (2017). Face matching impairment in developmental prosopagnosia. Quarterly Journal of Experimental Psychology, 70(2), 287–297. https://doi.org/10.1080/17470218.2016.1173076

https://doi.org/10.1080/17470218.2016.1173076

Young, A. W., & Burton, A. M. (2017a). Are We Face Experts ? Trends in Cognitive Sciences, xx, 1–11. https://doi.org/10.1016/j.tics.2017.11.007

Young, A. W., & Burton, A. M. (2017b). Recognizing Faces. Current Directions in Psychological Science. https://doi.org/10.1177/0963721416688114

Young, A. W., & Burton, A. M. (2018). Are We Face Experts? In Trends in Cognitive Sciences. https://doi.org/10.1016/j.tics.2017.11.007

Young, A. W., & Burton, A. M. (2020). Insights from computational models of face recognition : A reply to Blauch , Behrmann and Plaut. Cognition, June, 104422. https://doi.org/10.1016/j.cognition.2020.104422

Yovel, G., Wilmer, J. B., & Duchaine, B. (2014). What can individual differences reveal about face processing? Frontiers in Human Neuroscience, 8(AUG). https://doi.org/10.3389/fnhum.2014.00562