


Forensic facial examiners vs. super-recognizers: Evaluating behavior beyond accuracy

Carina A. Hahn¹ 

Liansheng Larry Tang² 

Amy N. Yates¹ 

P. Jonathon Phillips¹ 

¹National Institute of Standards and Technology, Gaithersburg, MD, USA

²University of Central Florida, Orlando, FL, USA

Abstract

Forensic facial examiners and super-recognizers are highly accurate at comparing two faces to determine identity and outperform the general population. Typically, forensic facial examiners are highly trained, whereas super-recognizers are thought to rely on natural ability. Previous studies have compared the accuracy of facial examiners and super-recognizers but have not studied the detailed behavioral properties of their face matching performance. In this study, we further analyzed data from a previous study which tested facial examiners and super-recognizers on a challenging test of face matching, or facial comparison, ability. In that study, the two groups were equally accurate. Here, we further characterize their behavior. We found distinct behaviors between these two groups, independent of overall accuracy. For differences, we found: 1) Facial examiners took advantage of the full range of the 7-point identity judgment scale; super-recognizers had a preference for the extreme ends of the scale (highly confident decisions); 2) For facial examiners, identity judgments for same-identities and different-identities mirrored each other; those from super-recognizers did not; 3) We evaluated identity judgment agreement across participants. Facial examiners agreed with each other to a greater extent than super-recognizers. We next examined metacognitive awareness of their own ability. While there were qualitative differences in the use of the scale, both groups showed behavioral insight into their own accuracy: more confident people and those who rated the task to be easier tended to be more accurate. These findings suggest that studying facial comparisons may benefit from assessing more than accuracy. This deeper understanding allows us to better interpret judgments according to the nature of a person’s facial expertise and experience.

Keywords: biometrics, decision making, facial comparison, facial forensics, facial identification, facial recognition, forensic facial examiner, super-recognizer

1 Introduction

Stories of super-recognizers never forgetting a face and forensic facial examiners identifying criminals has sparked the interest of researchers, the press, and the public worldwide. Facial examiners receive extensive training and mentoring, and can testify in court (19, 20). Whereas, super-recognizers have an innate ability for face recognition (32). Because of this ability, law enforcement and security organizations employ super-recognizers (1, 14, 32, 36, 37). Consistently, findings show that facial examiners are highly accurate at facial comparisons (30, 46, 49, 50, 52); similarly, other studies found that super-recognizers are also highly accurate (1–3, 6–9, 14–16, 32, 33, 36–38, 40, 44). The first paper directly comparing these two groups reported that they are equally accurate to each other (34). Due to the consequential nature of their judgments and theoretical interest of understanding facial expertise, researchers are interested in the cognitive underpinnings of their superior face matching ability. Juxtaposing these highly, and equally, accurate groups of different backgrounds allows the unique opportunity to understand how they are the same and how they are different, independent of overall accuracy. In this study, we move beyond accuracy to investigate the properties of facial examiners’ and super-recognizers’ expertise.

Previous studies showed that face processing behaviors vary with different levels of overall accuracy (9, 25, 30, 44–46, 49). Behaviors tested include eye movements, reported reliance on different facial features,

and the use of facial vs. body information. Papers have examined behavioral differences between super-recognizers and other populations *or* facial examiners and other populations (1–4, 6–9, 14–16, 25, 30, 32–34, 36–38, 40, 44–46, 49, 50, 52). But none directly compare facial examiners and super-recognizers. For the first time, our data allows us to directly study similarities and differences between these two face specialists groups.

We build on Phillips et al. (34), which assessed the accuracy of facial examiners, super-recognizers, and control groups. For each facial comparison in Phillips et al. (34), the authors collected an identity judgment and a difficulty rating from participants. In the current study, we used these previously collected data to examine what the distribution of identity judgements looked like along a 7-point identity judgment scale. We assessed the consistency of identity judgments across participants. Finally, we examined how identity judgments and difficulty ratings correlated and how these measures correlated with a person’s own accuracy to assess metacognitive ability. These findings can inform whether these groups are the same, behaviorally and cognitively; if different strategies can underpin equal accuracy; and whether the nature of one’s expertise affects performance.

The comparison of these two groups revealed their approach to providing identity judgments on the identity judgment scale differed, but their metacognitive abilities were similar. First, we examined the distribution of judgments along the response scale. Facial examiners were more willing to conclude that an identity judgment could not be made¹, while super-recognizers clustered their judgments toward the extremes of the scale (i.e., highly confident judgments). We next evaluated identity judgment agreement. Ideally, the same facial comparison will result in the same conclusion across individuals. This is important for confidence in the facial comparison process. In line with this desire, we found greater consistency among facial examiners compared to super-recognizers.

Finally, we tested the relationship between confidence and difficulty ratings, and participants’ insights into their own ability. Members of both groups were more confident on comparisons they judged to be less difficult. Higher confidence and judging the task to be less difficult were associated with higher accuracy. Therefore, although these groups were different on the basis of their identity judgments, they showed harmony between their confidence and perceived difficulty as well as metacognitive insight into their own ability. This suggests that for facial examiners and super-recognizers, the identity judgment itself carries with it information beyond the judgments themselves.

This work fills a gap toward better interpreting judgments in the context of the nature of facial expertise. On an applied level, these findings support that the nature of one’s expertise is related to different underlying behaviors. In our study, to qualify as a super-recognizer, the participant could never have received any forensic training. The facial examiners in our study did receive training. Therefore, a history of training (or lack of training) was a defining difference in the backgrounds of facial examiners and super-recognizers. The role of training and/or professional experience could not be tested definitively in this study, but our conclusions suggest that the behavioral properties of examiners are consistent with the goals of forensic training. Usually, the effect of training is measured by changes in accuracy (47). This work suggests that the effect of training and/or professional experience may be assessed with other desired measures of training success. By directly measuring behaviors of interest, we can inform training programs to support forensic applications.

Investigating these underlying behaviors reveals how the nature of facial expertise can vary. On a theoretical level, these results show that rather than a single strategy supporting high accuracy, there are at least two ways to be highly accurate at face matching. From this, we can begin teasing apart the underlying cognitive processes supporting superior facial comparison ability.

2 Facial comparison task procedure

All data were collected in a previous study, reported in Phillips et al. (34). The Phillips et al. (34) study was designed to focus on facial comparison accuracy of identity judgments from a 7-point response scale. Participants compared two facial images and determined whether they depicted the same identity or different

¹This conclusion corresponds to selecting the midpoint option on the identity judgment scale. This response indicated that “the observations support neither that it is the same person nor that it is different persons.” We interpreted this as the participant concluding that an identity judgment could not be made. However, strictly speaking, a judgment itself is rendered.



- +3: The observations strongly support that it is the same person
- +2: The observations support that it is the same person
- +1: The observations support to some extent that it is the same person
- 0: The observations support neither that it is the same person nor that it is different persons
- 1: The observations support to some extent that it is not the same person
- 2: The observations support that it is not the same person
- 3: The observations strongly support that it is not the same person

Figure 1: An example facial comparison and the identity judgment scale from this study. Participants viewed pairs of faces and were instructed to determine identity on the 7-point scale shown above. For each comparison, participants also provided a difficulty rating (not shown in the figure). Facial examiners and super-recognizers were allowed 3 months to submit their responses and could use any tools and methods they had available. All images from the study and correct answers can be viewed in the supplementary information for Phillips et al. (34).

identities. Figure 1 shows an example comparison and the full *identity judgment scale*. There were 12 same-identity and 8 different-identity trials in the task. One of the objectives in the study was to measure facial comparison ability under conditions that mirrored real-world casework. To accomplish this, facial examiners and super-recognizers were allotted up to three months to submit their responses, and they were able to use any tools and methods of their choice when completing the task. These identity judgments were the basis of their accuracy, using the area under the receiver operating characteristic curve (AUC).

In addition to an identity judgment, participants rated each facial comparison’s difficulty on a 5-point *difficulty rating scale* (range: 1, Easy: The comparison was easier than most facial comparisons; 5, Not Possible: The comparison was virtually impossible, due to a lack of detail in the image(s)). This was included to evaluate if the perceived difficulty of each facial comparison provided independent information from the identity judgment. Further information regarding the scales, task, participants, and procedures can be found in the Methods section of this document (Section 7.2) and the original study (34).

3 Identity judgment distributions

We started with the basic question: do facial examiners and super-recognizers use the identity judgment scale differently? Visually inspecting the identity judgment distributions indicates that, yes, judgments between the two groups varied (Figure 2). Facial examiners’ judgments were distributed across the full scale. Super-recognizers, on the other hand, selected the scale extremes and avoided the midpoint 0 (“the observations support neither that it is the same person nor that it is different persons”). This was reflected in the modes of the distributions. The mode judgments for facial examiners were +2 and −2 for same- and different-identity trials, respectively. Super-recognizers, by contrast, selected +3 and −3 most frequently for same-identity and different-identity trials, respectively.

Subsequent statistical analyses supported these descriptive observations and visual inferences. For sta-

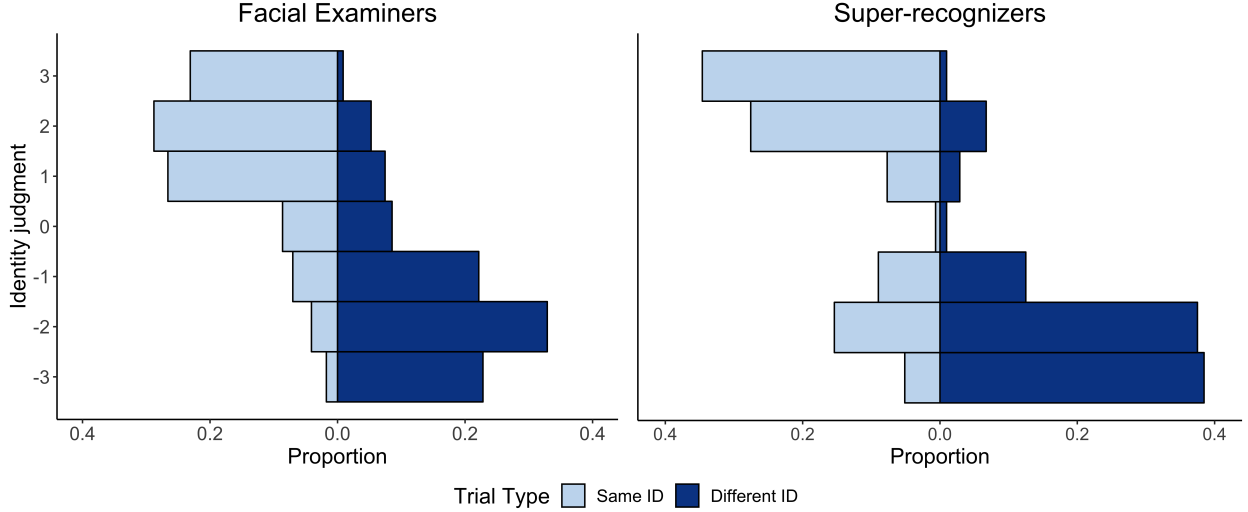


Figure 2: The back-to-back histograms above depict the proportion of judgments (x-axis) of a given identity judgment (y-axis) provided by facial examiners (left-side graph) and super-recognizers (right-side graph). For each graph, the left, light blue side shows the response distributions for same-identity trials. The right, dark blue side shows the distributions for different-identity trials.

tistical analysis, we compared the distribution of responses from facial examiners and super-recognizers with two-sample Kolmogorov–Smirnov (K-S) tests. We conducted two, separate tests for same- and different-identity trials. We found that for both trial types, facial examiners and super-recognizers used the scale differently (same-identity trials: $D = 0.17, p = .002$; different-identity trials: $D = 0.20, p = .002$).

The second basic question is whether participants in each group approached the scale similarly for the two trial types. As stated, the absolute value of the identity judgments were the same for each trial type, within a given group (facial examiners: $+2$ or -2 ; super-recognizers: $+3$ or -3). However, the overall distributions paint a more thorough picture of their response patterns. Based on a visual inspection of Figure 2, the same- and different-identity distributions appeared to be mirrored for facial examiners. This suggests they approach the scale similarly for comparisons of same- and different-identities. However, for super-recognizers, this mirroring was less apparent. To test this statistically, we mirrored different-identity judgments to reflect the direction of same-identity judgments by reversing the scale direction for different-identity trials (e.g., -3 was converted to $+3$ and vice versa). We then compared the distributions from same- and mirrored different-identity judgments. This process was applied to both participant groups. Results showed that for facial examiners, the distribution of judgments was equivalent for same-identity and mirrored different-identity trials (K-S test: $D = 0.04, p = 0.82$). For super-recognizers, the distribution of judgments differed depending on the trial type (K-S test: $D = 0.19, p = 0.02$).

We next tested how judgments varied at the level of the individual. We focused specifically on the scale extremes (highly confident judgments) and on the scale midpoint (0). First, we evaluated whether participants in the super-recognizer group were more likely to use the most extreme points on the scale ($+/-3$) compared to facial examiners. We computed the proportion of judgments corresponding $+/-3$ across all trials, separately for each participant. We found that the proportion of these high-confidence judgments was lower for facial examiners compared to super-recognizers (facial examiners: mean (M) = 0.24, standard deviation (SD) = 0.21; super-recognizers: $M = 0.40, SD = 0.25$; Mann-Whitney $U = 232, p = 0.036, r = 0.25$). Next, we replicated this analysis on judgments associated with the scale midpoint (0). The proportion of judgments for this response was higher for facial examiners than for super-recognizers (facial examiners: Mean (M) = 0.09, Standard Deviation (SD) = 0.09; super-recognizers: $M = 0.01, SD = 0.02$; Mann-Whitney $U = 178, p = 0.002, r = 0.37$)².

Therefore, equal accuracy was supported by different approaches to the identity judgment scale. We next

²See Supplemental Figure A.1 for corresponding visualizations.

tested whether these differences appeared in other identity judgments properties: agreement, the relationship between different behavioral measures, and insight into accuracy.

4 Identity judgment agreement

In forensic analysis, reliability is desired so that ideally multiple examiners comparing the same image pair will provide the same identity judgment – they will provide consistent answers. Measuring the consistency of judgments assesses the degree to which this ideal is met. To test the extent of agreement, we measured inter-rater reliability of the identity judgments. We modeled the case in which comparisons are completed independently by two examiners. To provide a benchmark, we repeated the analysis with super-recognizers.

We measured identity judgment agreement on all 20 facial comparisons separately for each participant group. Agreement was assessed using a standard measure of inter-rater reliability, Cohen’s Kappa, weighted for ordinal data ($\hat{\kappa}'_c$) (12, 21, 22). The ordinal weighting was applied to reflect the identity judgment response scale. The possible range of $\hat{\kappa}'_c$ between two participants was +1 (indicating perfect agreement) to −1 (indicating complete disagreement). Agreement was measured between all possible unique pairs of facial examiners. With 57 facial examiners in our sample, this produced 1,596 $\hat{\kappa}'_c$ values. Between the 13 super-recognizers in our study, there were 78 unique pairs of participants and corresponding $\hat{\kappa}'_c$ values. Figure 3 shows the distributions of pairwise agreement with a violin plot and box-and-whisker plot overlay.

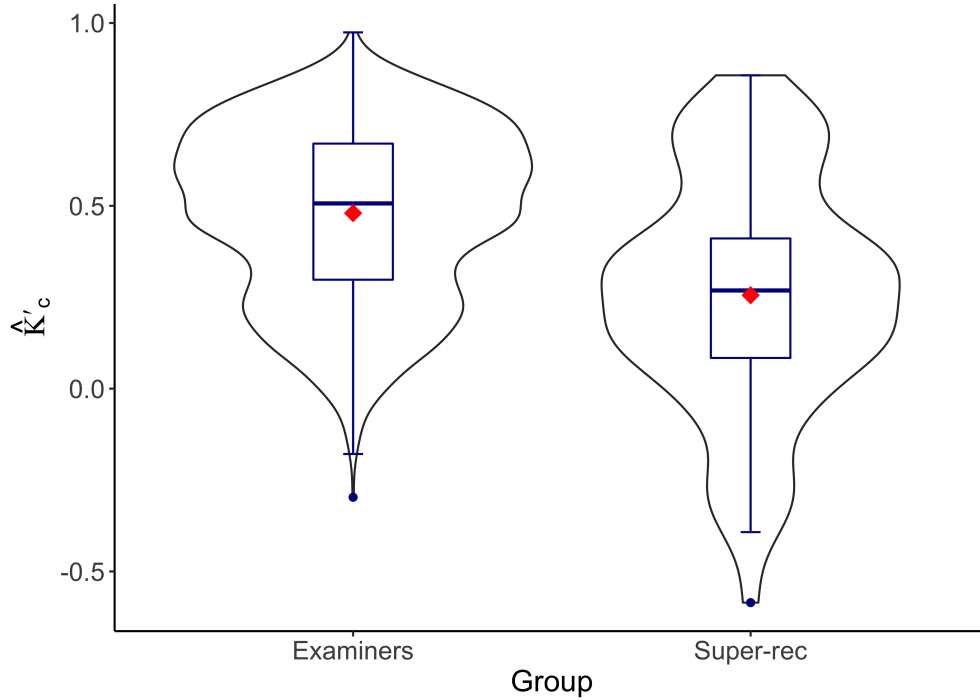


Figure 3: A violin plot with a box-and-whisker plot overlay shows the distribution of agreement between all possible pairs of participants within each group. The violin plot represents the range and density of different levels of agreement ($\hat{\kappa}'_c$). A box-and-whisker plot overlay shows the median, interquartile ranges, and the full range of agreement for each group. Points represent outliers. Red diamonds show the mean for each group.

Average pairwise agreement was nominally higher among facial examiners ($M = 0.48, SD = 0.24$) compared to agreement among super-recognizers ($M = 0.26, SD = 0.32$). The rule of thumb from Fleiss et al. (21) suggests that $\hat{\kappa}'_c \geq 0.75$ indicates excellent agreement; $\hat{\kappa}'_c$ between 0.75 and 0.40 indicates fair agreement, and $\hat{\kappa}'_c \leq 0.40$ indicates low agreement. Although context guides these interpretations, this rule of thumb suggest facial examiners show a fair amount of agreement, and super-recognizers show poor agreement.

We tested whether the observed agreement between facial examiners and super-recognizers was statisti-

cally significant. Because $\hat{\kappa}'_c$ values were measured between all possible pairs of participants within a group, each person contributed to multiple $\hat{\kappa}'_c$ values. This lack of independence between measures precluded the use of standard statistical tests because the variance of the statistic is not readily available for such dependent measures. Therefore, we implemented a bootstrap resampling procedure (e.g., (24)) to compare the two groups. For each bootstrap iteration, 20 trials (facial comparisons) were selected randomly with replacement to replicate the number of trials presented in the original experiment. For these selected trials, we measured $\hat{\kappa}'_c$ between all possible, unique pairs of individuals within each participant group, repeating the original procedure described above. Next, we computed the difference between the average $\hat{\kappa}'_c$ values for facial examiners and for super-recognizers on that iteration. We refer to this as the $\hat{\kappa}'_c$ difference. We repeated this process for 1,000 iterations which produced a distribution of 1,000 $\hat{\kappa}'_c$ differences between the two groups (see Figure A.2 in supplemental materials). We determined if agreement was significantly different between the two groups according to a 95% bootstrap confidence interval (CI) around the top and bottom 2.5% of the distribution. If the 95% CI did not intersect 0 (indicating no difference), the differences between the two groups were judged to be statistically significant.

The 95% CI for this distribution of $\hat{\kappa}'_c$ differences was [0.10, 0.34]. This indicates that agreement differed significantly across the two groups. While the lower bound of this CI determined significance, the upper bound characterized the probable range of agreement differences between the two groups. Therefore, with 95% confidence, our estimate of the agreement difference range between examiners and super-recognizers extended from a small (0.10) to a large difference (0.34).

5 Relationship between behavioral measures

The previous analyses focused on differences in identity judgment scale use and agreement between facial examiners and super-recognizers. Next, we conducted a deeper analysis into the behavioral measures obtained in this task: the confidence of the identity judgments and difficulty ratings. This provides insight into the shared and unique information across different behavioral measures. We collected both an identity judgment on a 7-point scale and a difficulty rating for each trial (see Section 2 for procedure details). We obtained confidence from the absolute value of the identity judgment. For example, a judgment of +3 or -3 implied high confidence. We expect that a high confidence judgment would be associated with a lower difficulty rating. Likewise, low absolute values in the identity judgment (i.e., 0) implied low confidence, and we predicted that these decisions would be associated with greater rated difficulty. By measuring the extent to which these are related, we addressed the question of whether these ratings provided distinct or overlapping information. Note that confidence does not necessarily indicate accuracy: a judgment might be highly confident but erroneous.

Considering the accuracy of judgments was the second part of this analysis. Identity judgments were the basis of accuracy as measured by the area under the receiver operating characteristic curve (AUC). We measured metacognitive ability: the extent to which people were aware of their own ability. To examine the level of this behavioral insight, we tested whether accuracy was associated with confidence and rated difficulty.

5.1 Confidence and difficulty

We evaluated the relationship between rated difficulty and confidence at two levels of analyses: 1) at the level of the trials themselves to address the question: how much does the rated difficulty of a facial comparison correspond to the confidence in that identity judgment? and 2) at the level of the individual to address: does a person who rated the task to be more difficult overall also respond with less confidence overall? The effect of rated difficulty on performance itself has been well-studied (see Scasserra (41) for a review). However, to our knowledge, the relationship between the rated difficulty and confidence of a given response has not been studied. For this analysis, we only considered the confidence itself, not the accuracy of the judgment.

First, to analyze the relationship between difficulty ratings and confidence at the level of the trials themselves, we plotted every combination of difficulty rating and identity judgment observed across all trials (Figure 4). A visual inspection shows that as the difficulty rating increased, the confidence associated with the identity judgment decreased, for both correct and incorrect judgments. For example, there were no cases in which the lowest difficulty rating was provided (1) with an identity judgment at the midpoint (0).

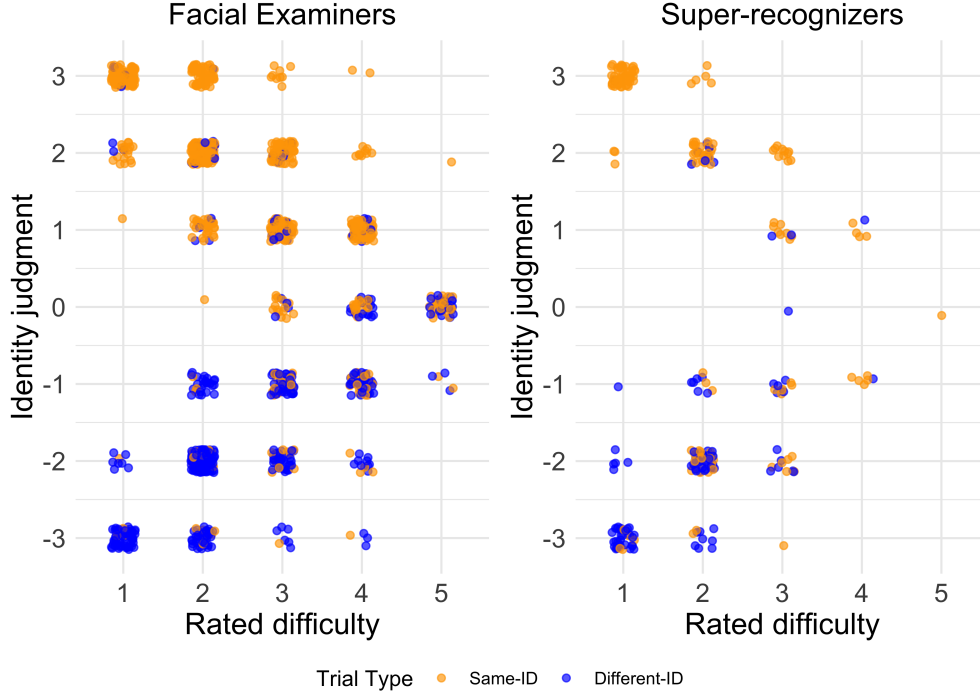


Figure 4: The relationship between difficulty ratings and identity judgments for facial examiners (left-side graph) and super-recognizers (right-side graph). Each point represents one judgment. Point color indicates trial type. Points are jittered at each grid intersection to show overlapping points. For identity judgments, responses with lower absolute values indicated lower confidence; responses with higher absolute values indicated higher confidence. For difficulty ratings, lower values indicated greater ease than higher values.

Likewise, there were no cases in which a high difficulty rating (5) was provided with a highly confident (+/-3) judgment.

We examined this relationship statistically with a Kendall’s τ correlation. We first converted the raw identity judgments to a measure of confidence (the absolute value of the identity judgment). This retained the confidence of the identity judgment regardless of the direction of the decision (i.e., judged to be the same identity or different identities). Results showed that as difficulty ratings increased, confidence decreased for both groups (facial examiners: Kendall’s $\tau = -.67, p < .001$; super-recognizers: $\tau = -.76, p < .001$).

This trial-based analysis provided a full-test view of the relationship between confidence and difficulty. However, this does not address whether *people* who are more confident rate the task to be less difficult. To test this, we computed each participant’s average confidence and difficulty rating across the whole task. We then computed the correlation between their average confidence and their average difficulty rating. For both groups, there was a negative correlation between confidence and difficulty: the more difficult they found the task, the lower their confidence (facial examiners: $\tau = -0.56, p < 0.001$, super-recognizers: $\tau = -0.70, p = 0.001$, Figure 5).

5.2 Behavioral insight: Confidence, difficulty, and accuracy

We tested the extent to which face specialists have insight into their own accuracy. Previous studies have measured the extent to which super-recognizers can estimate their own accuracy using self-report questionnaires to estimate face recognition ability (2, 10). Instead of self-estimated ability, our measure of insight was the degree to which a person’s confidence and difficulty ratings were correlated with their accuracy on the test. For each participant, we computed their mean confidence and difficulty rating across the entire

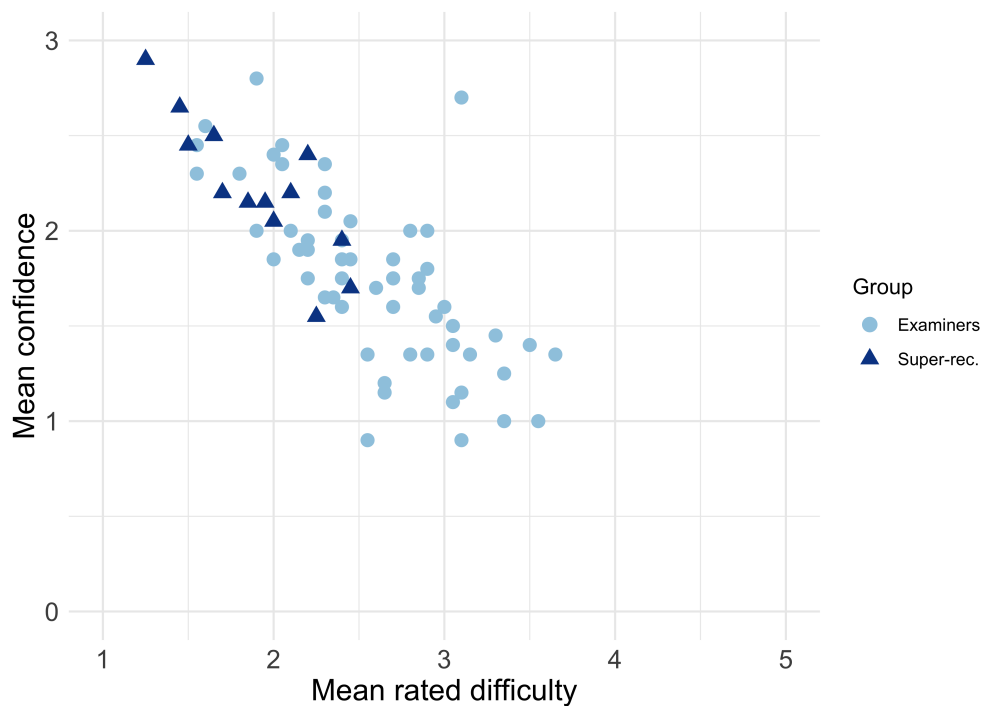


Figure 5: The relationship between rated difficulty and confidence at the level of the individuals. Each point represents one participant. A point’s position on the x-axis indicates a participant’s mean rated difficulty across the whole task. The position on the y-axis indicates the participant’s mean confidence across the task. Point color and shape indicate participant group: light blue circles represent facial examiners; dark blue triangles represent super-recognizers. For confidence, lower values indicate a lower level of confidence; higher values reflect higher confidence. For difficulty ratings, lower values indicate ease; higher values indicate greater difficulty.

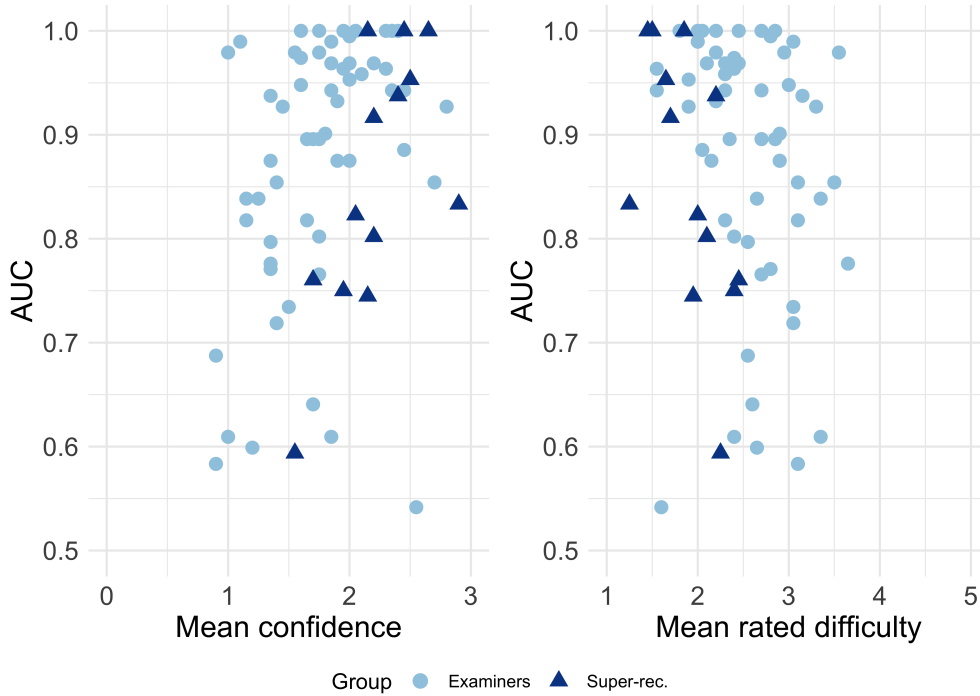


Figure 6: These two scatter plots show the relationship between behavioral measures and accuracy (AUC). Each point represents one participant. Point color and shape reflect participant group: light blue circles represent facial examiners; dark blue triangles represent super-recognizers. A point’s position on the x-axis indicates the participant’s mean confidence (left-side graph) or mean difficulty rating (right-side graph). Position on the y-axis indicates the participant’s accuracy. For confidence, lower values indicate a lower level of confidence; higher values reflect higher confidence. For difficulty ratings, lower values indicate ease; higher values indicate greater difficulty.

task³. Accuracy was measured using AUC.

First, we measured the relationship between participants’ mean confidence and accuracy with Kendall’s τ correlation. This was computed separately for facial examiners and super-recognizers. For both groups, participants who were more confident on average tended to be more accurate (facial examiners: $\tau = 0.29, p = 0.002$; super-recognizers: $\tau = 0.54, p = 0.01$, Figure 6). Therefore, those in both face specialist groups demonstrated insight into their own accuracy. The relationship between mean rated difficulty and accuracy was consistent with those of mean confidence: people who found the task more difficult overall tended to be less accurate (facial examiners: $\tau = -0.22, p = 0.02$; super-recognizers: $\tau = -0.48, p = 0.023$, Figure 6).

6 Discussion

Over a series of analyses, we characterized the similarities and differences between facial examiners and super-recognizers. With the exception of Phillips et al. (34), previous studies have measured the accuracy and qualitative behavioral properties of facial examiners or super-recognizers, separately (2, 4, 6, 9, 14–16, 25, 30, 32, 33, 37, 38, 40, 44–46, 49, 52). Our work directly compared the behavioral properties of these two face specialist groups. The interest in these groups is largely spurred by a desire to investigate the role of training and the cognitive underpinnings of superior face recognition. Often, these evaluations have focused on differences in accuracy across viewing and testing conditions. Here, we show the importance of

³Detailed analyses of difficulty ratings can be found in the supplemental materials.

evaluating underlying behaviors themselves to produce a fuller characterization of people with superior facial comparison ability and the behaviors supporting this high performance.

Both groups showed metacognitive awareness into their own ability: confidence and rated difficulty moderately predicted accuracy. This extends similar findings observed in the general population (11, 43, 48) to face specialists. There are few studies examining the metacognitive insight for face specialists, and these used self-report questionnaires Bate and Dudfield (2), Bobak et al. (10). Both studies found that super-recognizers self-assessed their ability better than the general population. In these studies, participants completed a questionnaire in which they estimated their own face recognition ability. Overall, self-assessments moderately predicted or did not strongly predict actual ability on objective face recognition tests. There are two key differences between their studies and the current one. First in the previous studies, metacognition was based on explicitly estimated, self-assessed ability. In our study, metacognitive abilities were assessed via the confidence of their identity judgments and their difficulty ratings relative to their accuracy. Second, they measured self-reported estimates of global ability based on everyday facial recognition tasks (e.g., “I can spot familiar people in unexpected contexts.”). Our measures were obtained on a trial-by-trial basis, within this specific facial comparison task. Therefore, both self-assessed measures from previous studies and the implicit, decision-based measures in this study all resulted in low to moderate relationships to actual ability. For the first time, we show this relationship for both facial examiners and super-recognizers. For face specialist recruitment, future work can directly compare the utility of self-assessed ability estimates to implicit, decision-based estimates. This comparison can measure whether one is more effective, or whether a combination of both is helpful to identify highly accurate individuals. It is promising that the relationship between confidence and difficulty to accuracy was stable across both groups, despite the behavioral differences underlying their performance. We discuss the implications of their differences below.

Based on the groups’ differences, we conclude that interpretations of judgments may need to be tailored per group. The overall response distribution differences between facial examiners and super-recognizers call to attention that a given judgment may carry different meaning, depending on who produced it. For example, a response of $+/-3$ is inferred to reflect high confidence. However, it is unclear whether a given level of confidence carries the same weight across participant groups. Phillips et al. (34) addressed this question with the same data in this study by measuring the probability of incorrect, highly confident responses. They found that high confidence errors were rare for both facial examiners and super-recognizers. However, super-recognizers were more likely to judge images of same person as being different people with high confidence (-3 responses for same-identity trials). In this study, we expanded this knowledge to show that super-recognizers’ overall scale use differed for same- and mirrored different-identity trials. For law enforcement and forensic testimony, the potential for different meanings is consequential for those making sense of the judgments. Future work may test whether one meaning can be construed from conclusions for all groups or whether it is more informative and meaningful to tailor interpretations.

Similarly, providing different conclusion scales for each group may be beneficial. One way to evaluate the appropriateness of a conclusion scale is to assess whether the judgments carry consistent meaning across individuals. This consistency supports the ability to compare and interpret conclusions across different individuals. We found that facial examiners agreed with each other to a higher degree than super-recognizers did. The scale in this study was developed in consultation with the forensic community to reflect those used in real-world applications. It is possible that if super-recognizers had greater familiarity and/or training with this scale, they would also show high consistency. How their familiarity with the scale may have influenced consistency is unknown, but this finding is promising for the desire that forensic judgments be consistent across individuals.

Another property of scale use was that facial examiners’ judgments for same- and different-identity trials mirrored each other in this study. A common assumption is that facial examiners are more conservative from a signal detection theory perspective. In other words, the assumption is that facial examiners are less likely to make highly confident responses when judging faces to be of the same person compared to judging faces to be of different people. This is possibly because the consequences of falsely stating that two people are the same is thought to be greater than falsely judging the same person as being different people. Our identity judgment scale prevented the binarization of the judgments themselves in this manner. However, their high accuracy allows us to interpret judgments from same- and different-identity trials in a similar fashion. Their distributions for same- and mirrored different-identity trials indicates that a response bias toward judging faces to be of different people (regardless of the correct answer) is not likely to be driving

factor of performance in the real-world conditions in which facial examiners operate. In a perceptual test in which participants were timed and unable to use their tools and methods, Hu et al. (25) found facial examiners trended toward a more conservative response bias: they were more inclined to judge faces as being different. In considering their findings alongside ours, we can conclude that in restricted conditions, facial examiners trended toward a bias of judging faces to be different. In our study modeling real-world judgments, we found facial examiners distributed responses across the full scale regardless of whether the faces were of the same person or different people. They did so in a manner consistent with a measured, cautious, consistent approach toward determining identity.

We found this overall cautious approach only among facial examiners. Super-recognizers’ identity judgments clustered at the scale extremes. Norell et al. (30) found that facial examiners provided judgments with less confidence on challenging comparisons relative to untrained, less accurate undergraduate studies. It was previously unknown whether this strategy was coupled with overall accuracy. By comparing identity judgments to equally accurate super-recognizers, findings show this is not the case. Therefore, this behavior is perhaps a product of forensic training and experience, rather than an overall optimal strategy for highly accurate judgments.

We found for both groups, confidence, rating difficulty, and accuracy were all correlated. This indicates that additional information comes embedded with judgments themselves. This facilitates the ability to gain insight into judgments from experts who often operate as a closed-box⁴, unable to accurately describe the process in which they form conclusions (18). Additional insight into judgments themselves may enhance trust in forensic judgments. The current gap between forensic explanations of conclusions and a lay person’s interpretation produces misunderstandings and confusion (26). Simplified evidence in support of conclusions may facilitate interpretation of, and possibly trust in, forensic judgments by broad audiences. Importantly for these inferences, the moderate relationships between measures in our study were found with responses from the full experiment: we cannot necessarily infer a person’s overall accuracy from any given judgment. Future work is needed to test the extent to which probing underlying behaviors on proficiency tests is helpful to better understand decision processes and predictors of accuracy. Existing evaluations of forensic training have focused primarily on accuracy. The findings from this study highlight the importance of directly measuring the role of forensic training on behaviors of interest. To fully understand how training influences facial comparison decision making, we need more robust evaluations of its potential power.

The role of training was explored in Towler et al. (45). They propose a “two routes” hypothesis to facial comparison expertise: one for facial examiners and one for super-recognizers. This theory proposes that facial examiners achieve expertise through training, and super-recognizers do so through natural ability. They base their theory on experiments which compared facial examiners to control populations or super-recognizers to control populations. They then triangulated this evidence to characterize differences between facial examiners and super-recognizers. By directly comparing facial examiners and super-recognizers, we provide evidence directly supporting the “two routes” hypothesis. We found qualitative differences between facial examiners and super-recognizers in their scale use and response consistency. Their metacognitive similarities do not discount a two-routes to expertise hypothesis. Both routes to expertise may retain metacognitive abilities.

In regards to this expertise, the findings from this study indicate that considering other group-level factors apart from accuracy reveals patterns of performance associated with the nature of one’s expertise itself, not just one’s accuracy alone. Delving into facial comparison behaviors provided deeper insight into two face specialist groups with superior facial comparison ability. From a theoretical standpoint, this study highlights different, and equally accurate, approaches toward the task of identifying faces in challenging situations. Our results showed that response scale distributions, agreement, and difficulty judgments can be added to evaluations of training in order to more completely test its influence. Proficiency tests’ focus on accuracy can miss behaviors influenced by training and/or professional experience. Pivoting toward directly measuring, and having a standard for, the behavior of interest will be important to understand the role of training and the nature of facial expertise. To answer why these observed behaviors emerged, we need long-term experiments targeted toward this question to separate the role of nature (natural ability), nurture (training/professional experience), or a combination of both. It is important that researchers and practitioners continue working together to achieve these goals to ensure real-world validity and generalizability (36). This work shows that forward movement in these efforts will require a more thorough knowledge of facial comparison behaviors.

⁴These have been referred to as “black-boxes” in the forensics community.

The community needs to consider training and tests of facial ability beyond accuracy.

7 Methods

All data were obtained from Phillips et al. (34). The National Institute of Standards and Technology Institutional Review Board reviewed and approved the protocol for this project and all subjects provided informed consent in accordance with 15 CFR 27, the Common Rule for the Protection of Human Subjects. De-identified data can be obtained by signing a data transfer agreement with NIST⁵. All facial comparisons in this study appear in the supplemental materials of the original study from which the data was obtained (34). Images are covered under the license for the Forensic Facial Examiner Study Data Set from the University of Notre Dame⁶. In the the current document, we reiterated and summarized the overlapping relevant methodological information for the sake of clarity and to keep the current document self contained. More detailed information is included in the original study (34).

7.1 Participants

In the study from which this data was obtained (34), the authors measured the accuracy of forensic facial examiners, forensic facial reviewers, super-recognizers, fingerprint examiners, undergraduate students, and face-recognition algorithms on a facial comparison task. In the current study, we examined a subset of these *participant groups*. We compared two highly skilled face specialist groups: *forensic facial examiners*, or *facial examiners*, ($n = 57$, 28 females) and *super-recognizers* ($n = 13$, 8 females). Facial examiners were defined as those who performed one-to-one facial comparisons as part of their profession. In these types of facial comparisons, only two images are evaluated at a time (as opposed to comparing a face to a gallery of images to judge identity). Super-recognizers were defined as those who had completed a face recognition test which categorized them as a super-recognizer⁷ or were actively employed as a super-recognizer. Super-recognizers in this study never received any forensic comparison training. Because of this, no one could qualify as both a facial examiner and super-recognizer.

7.2 Facial comparison task

In the task, participants determined whether pairs of face images were of the same person or of different people. An example of a test trial appears in the introduction of this document (Figure 1). The test contained 20 image pairs in total. Twelve pairs were of the same identity, and 8 pairs were of different identities. This imbalance prevented participants from applying a process of elimination strategy to determine identity. All face images were sent electronically to the participants. This allowed them to perform the task under their preferred viewing and working conditions, using their tools and methods. Comparisons could be completed in any order. Participants were allowed up to three months to submit their responses.

Identity judgments were obtained using a 7-point *identity judgment scale*: -3 , the observations strongly support that it is not the same person; -2 , the observations support that it is not the same person; -1 , the observations support to some extent that it is not the same person; 0 , the observations support neither that it is the same person nor that it is different persons; $+1$, the observations support to some extent that it is the same person; $+2$, the observations support that it is the same person; $+3$, the observations strongly support that it is the same person. This identity judgment scale was developed in consultation with members of the forensic face community and is based on scales used in forensic work (e.g., see Conclusion Scale from (30)).

In addition to providing an identity judgment, participants rated the difficulty of the comparison. The 5-point *difficulty rating scale* was as follows: 1, Easy: The comparison was easier than most facial comparisons; 2, Moderate: The comparison was a typical facial comparison; 3, Difficult: The comparison was more difficult than most facial comparisons; 4, Very Difficult: The comparison was unusually difficult, involving significant

⁵Contact the corresponding author for more information on how to obtain a data transfer agreement.

⁶Forensic Facial Examiner Study Data Set: <https://cvrl.nd.edu/projects/data/#forensic-facial-examiner-study-data-set>

⁷Tests commonly used to assess super-recognizer performance can be found in Noyes and O'Toole (31), Dunn et al. (17), and Ramon (35).

photometric illumination, or pose changes, other red flags; 5, Not Possible: The comparison was virtually impossible, due to a lack of detail in the image(s).

7.3 Selected analyses

All analyses and visualizations were completed using R version 3.5.1 and corresponding packages (5, 13, 23, 27–29, 42, 51) and executed in RStudio version 1.1.456 (39)⁸.

The measures in this study were defined as follows. *Accuracy* was measured using the area under the receiver operating characteristic curve (AUC) for each individual. The inputs to compute the AUC were the identity judgments themselves, labeled by trial type. *Confidence* was defined as the absolute value of an identity judgment (see Section 7.2 for the full identity judgment scale). This yielded a possible range of 0 – 3. Difficulty ratings were the responses on the difficulty rating scale (see Section 7.2 for the full scale).

Each section of results contains its pertinent methodological information. In what follows, we expand on select analyses for clarity and replicability.

7.3.1 Identity judgment distributions

From all identity judgments provided in the test, we produced a distribution of the responses to reflect the frequency of each judgment. First, we computed the number of judgments associated with a given point on the scale relative to the total number of trials, separately for same- and different-identity trials and for each participant group. The result of this was the proportion of responses associated with each scale option across all participants within a participant group and trial type (Figure 2). For statistical comparisons, we converted the frequencies of raw identity judgments into cumulative response distributions. These cumulative distributions were submitted to Kolmogorov-Smirnov (K-S) tests. For each distribution comparison, we conducted a two-sample Kolmogorov-Smirnov (K-S) test between the two distributions of interest. We conducted four comparisons in total: 1) facial examiners vs. super-recognizers for same-identity trials, 2) facial examiners vs. super-recognizers for different-identity trials, 3) same- vs. different-identity trials within facial examiners, and 4) same- vs. different identity trials within super-recognizers. To compare the frequency of judgments across same- and different-identity trials within each participant group, the judgments associated with different-identity trials were reversed, or *mirrored*, such that responses of –3 were coded as +3, and vice versa. This reverse coded different-identity distribution was submitted to the analysis and compared to the same-identity distribution.

To compare the frequencies of responses at the scale extremes (+/–3) and the scale midpoint (0), we computed the frequency of each judgment for each participant relative to the total number of trials. For each participant, this gives us the proportion of judgments corresponding to a particular decision. Statistical comparisons were conducted with non-parametric, Mann-Whitney *U* tests with the effect size, *r*, defined as

$$r = Z/\sqrt{N}$$

where *Z* is the *Z* score of the test statistic, and *N* is the number of observations.

7.3.2 Identity judgment agreement

To evaluate the consistency of responses across facial examiners and super-recognizers, we measured the inter-rater reliability of the identity judgments across all possible unique pairs of individuals within each participant group. For facial examiners, this produced $n = 1,596$ unique pairs of participants. For super-recognizers, this produced $n = 78$ unique pairs of participants. Inter-rater reliability was computed using Cohen’s Kappa, weighted for ordinal data ($\hat{\kappa}_c'$)(12, 22).

We tested for group differences with a bootstrap resampling procedure (e.g., (24)). We computed a total of $n = 1,000$ bootstrap iterations. For each iteration, 20 trials (image pairs) were randomly selected with replacement. The following steps were done separately for each participant group. Pairwise agreement ($\hat{\kappa}_c'$) for the selected sample of image pairs was computed across all possible, unique pairs of participants within

⁸Certain commercial equipment, instruments, or materials are identified in this paper in order to specify the procedures adequately. Such identification is not intended to imply recommendation or endorsement by NIST, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

a group. Next, we computed the average $\hat{\kappa}'_c$ coefficient across these pairwise combinations. Finally, we computed the difference between these average $\hat{\kappa}'_c$ coefficient for facial examiners and super-recognizers. We repeated this procedure for the 1,000 bootstrap iterations. This process produced a distribution of 1,000 $\hat{\kappa}'_c$ mean differences between the two groups. See Figure A.2 in the supplemental material for a visualization of this distribution. The 95% confidence interval around this distribution was bounded by the top and bottom 2.5% of the distribution. The lower bound of this interval determined whether the distributions were significantly different. We determined that agreement between facial examiners and super-recognizers were different if the lower bound did not intersect 0. The upper bound served to characterize the probable range of agreement level differences between the two groups.

Acknowledgments and Disclaimers

We thank Shanée Dawkins, Eilidh Noyes, and Craig Watson for improving the paper with their helpful feedback, and Nina Ning for her contributions to data analysis and visualizations.

The opinions, recommendations, findings, and conclusions in this publication do not necessarily reflect the views or policies of NIST or the United States Government.

Author L.T.'s work was supported in part by Award No. 2019-DU-BX-0011 awarded by the National Institute of Justice, Office of Justice Programs, US Department of Justice. The opinions, findings, and conclusions or recommendations expressed in this publication do not necessarily reflect those of the US Department of Justice.

References

- [1] S. Bate, "The consistency of super face recognition skills in police officers," *Applied Cognitive Psychology*, 2019.
- [2] S. Bate and G. Dudfield, "Subjective assessment for super recognition: an evaluation of self-report methods in civilian and police participants," *PeerJ*, vol. 7, p. e6330, 2019.
- [3] S. Bate, C. Frowd, R. Bennetts, N. Hasshim, E. Murray, A. K. Bobak, H. Wills, and S. Richards, "Applied screening tests for the detection of superior face recognition," *Cognitive Research: Principles and Implications*, vol. 3, no. 1, p. 22, 2018. [Online]. Available: <https://cognitiveresearchjournal.springeropen.com/articles/10.1186/s41235-018-0116-5>
- [4] S. Bate, R. Bennetts, N. Hasshim, E. Portch, E. Murray, E. Burns, and G. Dudfield, "The limits of super recognition: An other-ethnicity effect in individuals with extraordinary face recognition skills," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 45, no. 3, pp. 363–377, 2019.
- [5] H. Bengtsson, "R.utils: Various Programming Utilities," 2019.
- [6] A. K. Bobak, R. J. Bennetts, B. A. Parris, A. Jansari, and S. Bate, "An in-depth cognitive examination of individuals with superior face recognition skills," *Cortex*, vol. 82, pp. 48–62, 2016. [Online]. Available: <http://dx.doi.org/10.1016/j.cortex.2016.05.003>
- [7] A. K. Bobak, A. J. Dowsett, and S. Bate, "Solving the border control problem: Evidence of enhanced face matching in individuals with extraordinary face recognition skills," *PLoS ONE*, vol. 11, no. 2, pp. 1–13, 2016.
- [8] A. K. Bobak, P. J. Hancock, and S. Bate, "Super-recognisers in Action: Evidence from Face-matching and Face Memory Tasks," *Applied Cognitive Psychology*, vol. 30, no. 1, pp. 81–91, 2016.
- [9] A. K. Bobak, B. Parris, N. Gregory, R. Bennetts, and S. Bate, "Eye-movement strategies in developmental prosopagnosia and "super" face recognition," *Quarterly Journal of Experimental Psychology*, vol. 70, no. 2, pp. 201–217, 2017.

- [10] A. K. Bobak, V. R. Mileva, and P. J. Hancock, "Facing the facts: Naive participants have only moderate insight into their face recognition and face perception abilities," *Quarterly Journal of Experimental Psychology*, p. 174702181877614, 2018.
- [11] V. Bruce, Z. Henderson, K. Greenwood, P. J. B. Hancock, A. M. Burton, and P. Miller, "Verification of face identities from images captured on video." *Journal of Experimental Psychology: Applied*, vol. 5, no. 4, pp. 339–360, 1999. [Online]. Available: <http://doi.apa.org/getdoi.cfm?doi=10.1037/1076-898X.5.4.339>
- [12] J. Cohen, "Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit," *Psychological bulletin*, vol. 70, no. 4, p. 213, 1968.
- [13] R. Core Team, "R: A Language and Environment for Statistical Computing," Vienna, Austria, 2018.
- [14] J. P. Davis, K. Lander, R. Evans, and A. Jansari, "Investigating Predictors of Superior Face Recognition Ability in Police Super-recognisers," *Applied Cognitive Psychology*, vol. 30, no. 6, pp. 827–840, 2016.
- [15] J. P. Davis, L. D. Bretfelean, E. Belanova, and T. Thompson, "Assessing the long-term face memory of highly superior and typical-ability short-term face recognisers," *PsyArXiv*, 2019. [Online]. Available: <https://psyarxiv.com/var4m/>
- [16] J. P. Davis, A. Maigut, and C. Forrest, "The wisdom of the crowd: A case of post- to ante-mortem face matching by police super-recognisers," *Forensic Science International*, 2019. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S0379073819303226>
- [17] J. D. Dunn, S. Summersby, A. Towler, J. P. Davis, and D. White, "UNSW Face Test: A screening tool for super-recognizers," *PLoS ONE*, vol. 15, no. 11, p. e0241747, 2020. [Online]. Available: <http://dx.doi.org/10.1371/journal.pone.0241747>
- [18] G. Edmond, A. Towler, B. Gowns, G. Ribeiro, B. Found, D. White, K. Ballantyne, R. A. Searston, M. B. Thompson, J. M. Tangen, R. I. Kemp, and K. Martire, "Thinking forensics: Cognitive science for forensic practitioners," *Science and Justice*, vol. 57, no. 2, pp. 144–154, 2017. [Online]. Available: <http://dx.doi.org/10.1016/j.scijus.2016.11.005>
- [19] FISWG, "Guide for mentorship of facial comparison trainees in role based facial comparison, version 1.0," 2019. [Online]. Available: https://fiswg.org/FISWG_Mentorship_for_Facial_Comparison_Trainees_v1.0_20190510.pdf
- [20] FISWG, "Guide for Role-Based Training in Facial Comparison, version 1.0," 2020. [Online]. Available: https://fiswg.org/fiswg_guide_for_role-based_training_in_facial_comparison_v1.0_20200717.pdf
- [21] J. L. Fleiss, B. Levin, and C. P. Myunghee, *Statistical Methods for Rates and Proportions*, 3rd ed., D. J. Balding, N. A. C. Cressie, N. I. Fisher, I. M. Johnstone, J. Kadane, L. M. Ryan, D. W. Scott, A. F. Smith, and J. L. Teugels, Eds. John Wiley & Sons, Inc., 2003.
- [22] K. Gwet, *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC, 2014.
- [23] K. L. Gwet, "Appendix B.1: Software Solutions, The R Software," in *Handbook of Inter-Rater Reliability*, 4th ed., K. L. Gwet, Ed. Gaithersburg, MD: Advanced Analytics, LLC, 2014, ch. Appendix B, pp. 375–386.
- [24] J. J. Harden, "A bootstrap method for conducting statistical inference with clustered data," *State Politics and Policy Quarterly*, vol. 11, no. 2, pp. 223–246, 2011.
- [25] Y. Hu, K. Jackson, A. Yates, D. White, P. J. Phillips, and A. J. O'Toole, "Person recognition: Qualitative differences in how forensic face examiners and untrained people rely on the face versus the body for identification," *Visual Cognition*, vol. 25, no. 4-6, pp. 492–506, 2017. [Online]. Available: <http://dx.doi.org/10.1080/13506285.2017.1297339>

- [43] R. G. Stephens, C. Semmler, and J. D. Sauer, “The effect of the proportion of mismatching trials and task orientation on the confidence-accuracy relationship in unfamiliar face matching,” *Journal of Experimental Psychology: Applied*, vol. 23, no. 3, pp. 336–353, 2017.
- [44] J. Tardif, X. Morin Duchesne, S. Cohan, J. Royer, C. Blais, D. Fiset, B. Duchaine, and F. Gosselin, “Use of Face Information Varies Systematically From Developmental Prosopagnosics to Super-Recognizers,” *Psychological Science*, pp. 1–9, 2018. [Online]. Available: <http://journals.sagepub.com/doi/10.1177/0956797618811338>
- [45] A. Towler, R. I. Kemp, and D. White, “Can face identification ability be trained? Evidence for two routes to expertise,” in *Forensic face matching: Research and practice*, M. Bindemann, Ed. Oxford University Press.
- [46] A. Towler, D. White, and R. I. Kemp, “Evaluating the feature comparison strategy for forensic face identification,” *Journal of Experimental Psychology: Applied*, vol. 23, no. 1, pp. 47–58, 2017.
- [47] A. Towler, R. I. Kemp, A. M. Burton, J. D. Dunn, T. Wayne, R. Moreton, and D. White, “Do professional facial image comparison training courses work?” *PLoS ONE*, vol. 14, no. 2, p. e0211037, 2019.
- [48] D. White, R. I. Kemp, R. Jenkins, and A. M. Burton, “Feedback training for facial image comparison,” *Psychonomic Bulletin & Review*, vol. 21, no. 1, pp. 100–106, 2014. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pubmed/23835616>
- [49] D. White, P. J. Phillips, C. A. Hahn, M. Hill, and A. J. O. Toole, “Perceptual expertise in forensic facial image comparison,” *Proceedings of the Royal Society B: Biological Sciences*, vol. 282, p. 20151292, 2015.
- [50] D. White, A. Towler, and R. I. Kemp, “Expertise in unfamiliar face matching,” pp. 1–33, 2020.
- [51] H. Wickham, “tidyverse: Easily Install and Load the ‘Tidyverse’,” 2017. [Online]. Available: <https://cran.r-project.org/package=tidyverse>
- [52] C. Wilkinson and R. Evans, “Are facial image analysis experts any better than the general public at identifying individuals from CCTV images?” *Science & justice : journal of the Forensic Science Society*, vol. 49, no. 3, pp. 191–6, 9 2009. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pubmed/19839418>