A forensic facial examiner and professional team advantage for masked face identification

Applied Cognitive Psychology

WILEY

RESEARCH ARTICLE OPEN ACCESS

A Forensic Facial Examiner and Professional Team Advantage for Masked Face Identification

Eilidh Noyes[1,2] | Reuben Moreton[3] | Peter J. B. Hancock[4] | Kay L. Ritchie[5] | Sergio Castro Martinez[6] | Katie L. H. Gray[7] | Josh P. Davis[8]

[1]School of Psychology, University of Leeds, Leeds, UK | [2]Department of Psychology, University of Huddersfield, Huddersfield, UK | [3]Reli Limited, Southampton, UK | [4]Psychology, Faculty of Natural Sciences, University of Stirling, Stirling, UK | [5]School of Psychology, Sport Science and Wellbeing, University of Lincoln, Lincoln, UK | [6]Sección de Técnicas Identificativas, Comisaría General de Policía Científica, National Police, Madrid, Spain | [7]School of Psychology and Clinical Language Sciences, University of Reading, Reading, UK | [8]School of Human Sciences, Institute of Lifecourse Development, University of Greenwich, London, UK

Correspondence: Eilidh Noyes (e.c.noyes@leeds.ac.uk)

Received: 27 January 2024 | **Revised:** 6 June 2025 | **Accepted:** 16 June 2025

Funding: The authors received no specific funding for this work.

Keywords: face masks | face matching | face recognition | facial examiners | facial image comparison

ABSTRACT

Face masks and coverings are often encountered by facial examiners ('examiners') in forensic case work. Examiners are skilled at unconcealed face identifications, but their accuracy for masked face identifications is unknown, yet can be used as evidence in court. Here we test performance of an international sample of 61 examiners, 39 professional teams, and 6 face identification algorithms for 20 image pairs. Pairs consisted of one unconcealed face image and one mask wearing face image. Examiners and professional teams outperformed controls, but professional teams made the least errors of all groups. The algorithms achieved high accuracy on the task. The findings back the notion that examiners use feature-based comparison strategies, and these are successful for matching images where one face wears a mask. Our results support the use of examiners for the identification of masked faces and suggest a role for teams and human-machine working in applied practice.

1 | Introduction

Trained facial examiners make high-risk, security critical, face-matching decisions in applied scenarios. They compare images (known in the forensic community as facial image comparison) to determine whether they are of the same person or different people, and the decision result can provide evidence in forensic investigations. In criminal investigations, some perpetrators use face masks to try to obscure facial regions which would typically be available for face-matching comparisons. Face masks became commonplace in everyday life due to the COVID-19 pandemic and there has since been an increase in mask wearing as a method to conceal identity in criminal activity (Babwin and Dazio 2020; Rawlinson 2021). While facial examiners can identify unconcealed faces with high accuracy (Phillips et al. 2018), their identification performance for masked faces is unknown. This is concerning because high-stakes forensic identifications may be based on the outcome of a facial examiner's decision for a masked face. Facial examiners can provide "expert testimony" in many jurisdictions (Edmond et al. 2021), yet there is currently no evidence that facial examiners are better at masked face comparisons than an untrained member of the public (e.g., a member of a jury). There is a pressing need for a scientific assessment of facial examiners' face matching performance for masked faces and to investigate methods of achieving the highest possible identification accuracy for masked faces in applied settings.

Face matching is the comparison of two face images to determine the identity of the subject (i.e., whether the images are of the same person or different people). This identification procedure is widely used in applied settings, including policing, border control, and forensics, and is most often applied to unfamiliar faces (faces that the identifier has not previously encountered). Unfamiliar face matching is a challenging task for typical observers (non-professionals) when face images are unconcealed (Burton et al. 2010), and the task is even harder for typical observers when one of the images for comparison depicts a person in a face mask (Carragher and Hancock 2020; Noyes, Davis, et al. 2021; Noyes, Parde, et al. 2021; Ritchie et al. 2024). In typical observers, faces tend to be processed using their features, the configural relations between features, and holistically (Maurer et al. 2002), whereby features are integrated into a non-decomposable whole (Maurer et al. 2002; Farah et al. 1998; Young et al. 1987). Holistic processing is thought to be fast and efficient and is particularly used at shorter presentation durations (Hole 1994). Evidence for holistic face processing has been provided by the composite face illusion; when the top half of one face is spatially aligned with the bottom half of another, the two face halves are perceptually fused and viewers' perception of the target region is changed (Young et al. 1987; Rossion 2013; Murphy et al. 2017). The effect disappears when the face halves are misaligned or presented upside down (Young et al. 1987), suggesting that intact faciotopy is important for holistic processing (Murphy et al. 2017). As face masks cover the nose and mouth regions of a face, this lower face half occlusion not only removes these features from the comparison, but holistic processing is also disrupted (Stajduhar et al. 2022).

The case work of facial examiners focuses on the comparison of unfamiliar faces, often for use as evidence in a court of law. Unlike typical observers, who are understood to compare unfamiliar faces based on featural, configural, and holistic face information, facial examiners are trained to compare faces on a featural basis (Carragher et al. 2022; Towler et al. 2017) according to documented procedures (e.g., European Network of Forensic Science Institutes 2018). This approach is known as morphological analysis and comparison (Steyn et al. 2018). Facial examiners' professional training in this method can last for several years (Moreton et al. 2021), and it can take many hours for an examiner to complete a case using the morphological analysis approach. Facial examiners outperform typical observers on unconcealed face matching tasks, and this advantage is most pronounced when facial examiners have access to their tools and follow normal working procedures (Phillips et al. 2018; White et al. 2015). Facial examiners' experience in feature comparison may aid them in the identification of masked faces. Perhaps they know which features are most informative for identification, or their experience in feature comparison might protect against the disruption to holistic processing that typical observers experience for masked faces. However, it is dangerous to assume an examiner advantage for masked faces without evidence. For example, one may intuitively expect that experience with face identification would predict face identification accuracy of passport officers. White et al. (2014) found no correlation between passport officer experience and face matching accuracy, and White et al. (2014, 2015) found limited difference in performance between examiners and untrained participants when examiners were unable to follow their normal working procedure.

In applied practice, facial examiners complete identification comparisons by working alone or by working as part of a team of face identification professionals. To date, research has focused primarily on the accuracy of facial examiners when they work alone (Phillips et al. 2018; White et al. 2015); however, recent research has shown that where examiners collaborate as a team, face matching accuracy is improved (Towler et al. 2023). The face identification literature on typical observers demonstrates that pairs consistently perform with higher face matching accuracy than individuals (Dowsett and Burton 2015; Jeckeln et al. 2018). The reasons behind the benefits of pair over individual decision making are unclear (Ritchie et al. 2022), but may be linked to relying on the more confident individual in the pair (Jeckeln et al. 2018; c.f. Ritchie et al. 2022). As similar advantages have been observed for professional teams for unconcealed faces, team working may provide a tangible route to improve face matching accuracy for masked face identifications in case work over that achieved by facial examiners who work independently.

In some applied settings, such as police investigations and security screening, face identification algorithms have been used to assist with face identification decisions (e.g., passport control e-gates, or 1:N searches on a police database). Many face identification algorithms are more accurate than typical observers for face matching tasks that consist of good-quality, unconcealed images (see Noyes and Hill 2021 for a review). Phillips et al. (2018) reported that a state-of-the-art algorithm performed with accuracy rates equal to that of individual facial examiners on an unconcealed face matching task. Furthermore, some algorithms outperform typical observers on face matching tasks that include superimposed face masks (Carragher and Hancock 2020). Ritchie et al. (2024) report that algorithm face identification performance is more accurate for superimposed mask face images than genuine mask images. Thus, tests that use superimposed mask images may not accurately predict algorithm performance for masked faces as encountered in applied settings. Given the high accuracy of face identification algorithms in some identification scenarios, it is possible that face identification algorithms could play a role in forensic face comparisons in the future (Jacquet and Champod 2020; Ruifrok et al. 2022). For face recognition algorithms to be considered as a method of achieving high identification for masked faces in forensic settings, it is necessary to compare human and algorithm performance for genuine masked faces, and specifically to compare the performance of face recognition algorithms against human professionals.

Here we provide the first study to test face matching performance of international facial examiners and professional teams with masked faces (experiment 1) and compare their performance to that of computer algorithms (experiment 2). We also investigated the sensitivity and specificity of the different groups of human participants (experiment 3). We provide baseline accuracy scores for facial examiners, teams, and algorithms for genuine masked faces, and report how to achieve the most accurate identifications for masked faces in applied settings.

2 | Experiment 1. Facial Examiner and Team Accuracy for Masked Faces

2.1 | Methods

2.1.1 | Participants

Participants in the study were individual facial examiners ($N = 61$), professional teams ($N = 39$ teams) (mean age of professional participants = 42 years, 48% stated their gender as female, 52% stated their gender as male), and individual control participants ($N = 84$). All professional participants were members of the European Network of Forensic Science Institutes at the time of testing. Out of the control participants, 65% participants were aged 30–39, 27% were aged 40–49, and 7% were 50–65. For control participants, 50% people stated their gender as male, 49% as female, and 1% as other/not stated).

All of the facial examiners who took part in this study were practicing examiners at the time of the test and had professional facial image comparison experience and training. The length of professional experience and extent of training was not disclosed.

Teams were groups of professionals (at least two individuals) who performed the comparison tasks together. Teams were primarily comprised of facial examiners but could consist of any combination of facial examiners, facial reviewers (people who have experience in making high numbers of facial comparison decisions, usually through quicker identification methods than the examiner procedures), and police super recognisers (people who are naturally skilled at quick identifications and work within the police). Facial examiners and teams were recruited from 24 countries across Europe, North America, South America, Asia, and Oceania.

Control participants were members of the public from the same countries that professionals and teams were recruited from. Controls had not been trained in face identification and were recruited through the online platform Prolific (Prolific.co). Control participants were required to have achieved the highest approval rating on Prolific (95–100) to be eligible for recruitment in our study.

2.1.2 | Stimuli

The stimuli were 20 pairs of face images. One image in each pair was unconcealed (hereon referred to as the "reference image"). The other image depicted a person wearing a real face mask (referred to as "questioned image"). In case work, face matching comparisons involve the comparison of a good quality, unconcealed reference image, against a questioned image, hence our stimuli replicated this scenario. In our study there were 10 same identity image pairs, and 10 different identity pairs. The stimuli were carefully selected from a larger pool of face images to ensure that the images were challenging images for identification (in applied practice a Facial Examiner or team identification is typically only necessary for challenging cases).

2.1.2.1 | Image Selection Procedure. Images were collected from volunteers who responded to an advert for contributors on the University of Greenwich face recognition research database. From 218 contributing individuals, we created 60 identity pairings of similar-looking individuals (used elsewhere—Ritchie et al. 2024). All images were front-facing.

Three human reviewers rated the 120 image pairs (each identity seen as both same identity and different identity) as "easy" or "difficult." All images rated "easy" by two or more reviewers were removed, which reduced the database to 57 items. Participants in an online pilot study ($N = 50$, recruited through Prolific) completed a face matching task consisting of the 57 image pairs (these items included same and different identity trials). All image pairs were composed of an unconcealed image and a masked face image (see Figure 1). The participants' task was to respond "same" or "different" identity for each image pair. The most difficult 10 same identity image pairs, and most difficult 10 different identity image pairs (avoiding repeat identities) were used as the stimuli for the current study. Eight image pairs were of female faces, and 12 were male faces. Mean accuracy in the pilot study for our 20 most difficult image pairs was 51%.

2.1.3 | Procedure

The 20 test image pairs were distributed digitally to facial examiners and professional teams via the European Network of Forensic Science Institutes Digital Imaging Working Group (ENFSI DIWG). This allowed for the use of standard operating procedures and tools during the completion of the face matching test. Facial examiners and teams had 8 weeks to complete the task and could complete the face identifications in any order, returning to image pairs as they wished. Facial examiners completed the test either as an individual *or* as a member of a team. Individuals completed all 20 trials independently. Teams completed the trials according to team working practice for their agencies (we return to considerations of team protocol in the discussion). Control participants completed the test as an online face matching task via Qualtrics;

FIGURE 1 | All trials consisted of one reference image (unconcealed face image) (left) and a questioned image (masked face image) (right). The images shown are of the same identity. Images are representative of the experimental stimuli and depict someone who did not appear in the database but has given permission for their image to be used in this publication.

they completed all 20 trials in one sitting and viewed the images in a random order. There was no time limit per trial; however, an identification decision was required in order to progress to the next trial. All participants/teams were asked to record their identification decision for each image pair as a score on an 11-point rating scale (see Table 1). The scale reflects that used in forensic practice and encapsulates both identity decision and degree of support for the decision (ENFSI, 2018). The extremes of the scale denote greater support for a decision, and the zero response is used to denote that the decision is inconclusive.

2.2 | Results

The purpose of experiment 1 was to determine the overall accuracy of the different groups of human participants using the predefined conclusion scale. Results were calculated as the percentage of correct, incorrect, and no support decisions made by each participant group (facial examiners ["examiners"], professional teams ["teams"], and controls). Scores were recorded as correct if the rating fell on the correct side of the scale for the identification (+1 to +5 for a same identity pair, −1 to −5 for a different identity pair), incorrect if the score was on the wrong side of the scale (a + decision for a different identity trial, or a—decision for a same identity trial), and items rated as a 0 were scored as "no support." Due to the non-normal distribution of some of the data (particularly incorrect and no support decisions) a non-parametric Kruskal-Wallis test (one-way ANOVA on ranks) was used to compare performance between groups. Dunn's test was used for *post hoc* pairwise comparison of specific groups.

TABLE 1 | Identification rating scale.

+5	Extremely strong support same person
+4	Strong support same person
+3	Support same person
+2	Moderate support same person
+1	Weak support same person
0	No support
−1	Weak support different person
−2	Moderate support different person
−3	Support different person
−4	Strong support different person
−5	Extremely strong support different person

A Kruskal-Wallis test revealed a significant difference in the percentage of correct responses between groups ($\chi^2(2, 184) = 79.82$, $p < 0.001$, $\varepsilon^2 = 0.437$). Both teams (mean = 79.6%, median = 80, SD = 9.4) and examiners (mean = 74.8%, median = 75, SD = 10.5) scored a significantly higher percentage of correct responses than controls (mean = 57.5%, median = 60, SD = 13.3; both $p < 0.001$). There was no difference in the percentage of correct responses between teams and examiners ($p = 0.308$). There was also a significant difference in the number of incorrect responses across groups ($\chi^2(2, 184) = 115.87$, $p < 0.001$, $\varepsilon^2 = 0.633$). Both teams (mean = 9.2%, median = 5, SD = 7.5) and examiners (mean = 16.4%, median = 15, SD = 9.5) made significantly fewer incorrect responses than controls (mean = 39.1%, median = 40, SD = 13; both $p < 0.001$). Teams made significantly fewer incorrect responses than examiners ($p = 0.039$). This can be somewhat explained by use of the no support response, which differed across groups ($\chi^2(2, 184) = 29.14$, $p < 0.001$, $\varepsilon^2 = 0.159$). Teams (mean = 11.2%, median = 10, SD = 9.1) and examiners (mean = 8.8%, median = 5, SD = 11.9) used the no support response significantly more often than controls (mean = 3.4%, median = 0, SD = 6.6; control-teams $p < 0.001$, control-examiners $p = 0.005$), there was no significant difference in "no support" decisions between teams and examiners ($p = 0.052$). These results are illustrated in Figure 2.

Next, to investigate the types of incorrect decisions that were made, we broke down the percentage of incorrect responses by trial type (same and different identity trials).

There was a main effect of the percentage of errors made on same identity trials across the participant groups ($\chi^2(2, 184) = 67.32$, $p < 0.001$, $\varepsilon^2 = 0.368$). For same identity trials, controls made more

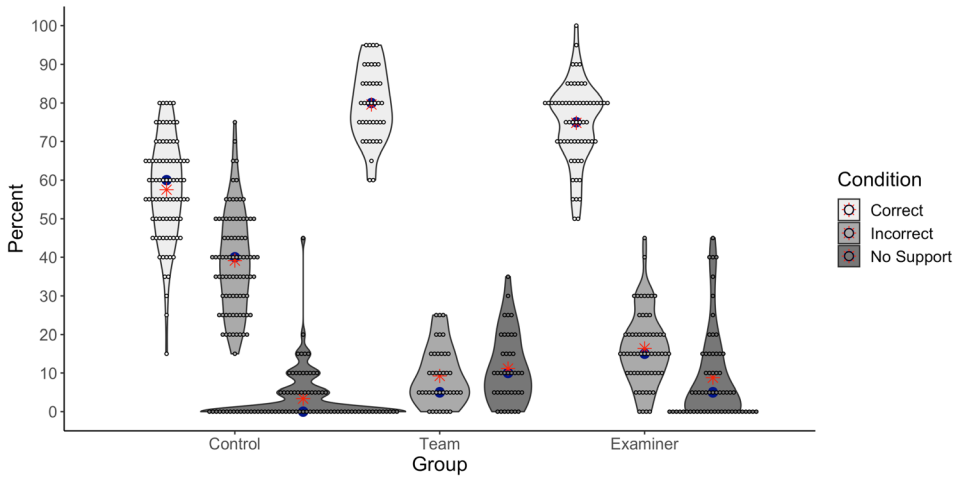

FIGURE 2 | Percentage of correct (light grey), incorrect (mid grey) and no support responses (dark grey) for each participant group. Each dot depicts the response of one participant, the blue dot on each violin represents the median, and the red asterisk represents the mean.

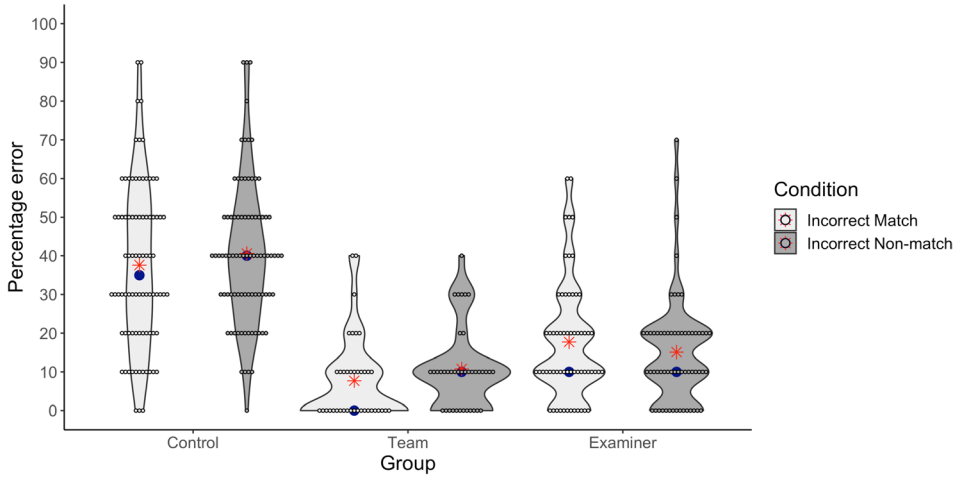

FIGURE 3 | Percentage of incorrect same identity (light grey) and incorrect different identity (dark grey) decisions for each participant group. Each dot depicts the response of one participant, the blue dot on each violin represents the median, and the red asterisk represents the mean.

errors (mean = 37.6%, median = 35, SD = 21.1) than both teams ($p < 0.001$) and examiners (mean = 17.7%, median = 10, SD = 15.4) ($p < 0.001$), and teams (mean = 7.7%, median = 0, SD = 10.9) made fewer incorrect responses than examiners ($p = 0.010$), meaning that teams were less likely than examiners to respond that images of the same identity were different identities.

For different identity trials, there was a main effect for group ($\chi^2(2, 184) = 8.43$, $p < 0.001$, $\varepsilon^2 = 0.483$), with examiners (mean = 15.1%, median = 10, SD = 14.2) and teams (mean = 10.8%, median = 10, SD = 10.6) making significantly fewer errors than controls (mean = 40.6%, median = 40, SD = 18.7) (both $p < 0.001$). There was no significant difference in the percentage of errors made by teams and examiners for different identity trials. These effects are visualized in Figure 3.

An analysis of no support decisions for same identity trials revealed a main effect ($\chi^2(2, 184) = 15.63$, $p < 0.001$, $\varepsilon^2 = 0.085$), with examiners (mean = 8.2%, median = 0, SD = 12.4) ($p = 0.025$) and teams (mean = 9.7%, median = 10, SD = 10.9) ($p < 0.001$) using the no support decision more frequently than controls (mean = 3.1%, median = 0, SD = 6.2), but no significant difference between teams and examiners. There was also a main effect of group for different identity trials ($\chi^2(2, 184) = 25.61$, $p < 0.001$, $\varepsilon^2 = 0.140$). Examiners (mean = 9.3%, median = 0, SD = 13.6) ($p = 0.017$) and teams (mean = 12.6%, Median = 10, SD = 11.4) ($p < 0.001$) made significantly more no support decisions than controls (mean = 3.7%, median = 0, SD = 8.6). Teams were also more likely to respond no support for different identity trials than examiners ($p = 0.047$).

In order to visualize the different behaviors in decision making between the three groups, Figure 4 shows the performance of the three groups for each facial image pair ranked by percentage correct for that group. For the control group, as percentage accuracy declines, the percentage of incorrect decisions steadily increases with little change in no support decisions. Whereas, for examiners and teams, as the percentage of correct responses decreases, there is an increased use of no support decisions. Compared to controls (60.9% correct or no support decisions), examiners (83.6%) and teams (90.8%) are more likely to make a correct or no support response than to make an error for the majority of facial image pairs, and this is most pronounced for teams over examiners. Teams appear to have the greatest accuracy/error trade-off of all groups.

We report sensitivity and specificity for each group as part of the analysis in Experiment 3.

2.3 | Discussion

The results of Experiment 1 show that facial examiners and professional teams are better than controls at facial comparisons that involve an unconcealed and masked face. Performance of control participants is low compared to previous studies on masked face identification (e.g., Carragher and Hancock 2020; Noyes, Davis, et al. 2021; Noyes, Parde, et al. 2021; Ritchie et al. 2024), which reflects the difficulty of the items (see procedure—we used the 20 most challenging image pairs). There was no statistical difference in the percentage of correct responses between examiners and professional teams, however, professional teams made fewer incorrect decisions than examiners. This is explained by greater use of the 'no support' option amongst professional teams than examiners. In forensic scenarios, a 'no support' decision is preferable to an incorrect identification; therefore, our finding that teams make fewer identification errors is important for applied practice.

3 | Experiment 2. Algorithm Performance for Masked Faces

Face identification algorithms have achieved substantial accuracy gains in recent years (latest results of NIST testing regularly updated, for latest at time of submission see Grother 2022). Whilst algorithms are not commonly used to assist facial examiners in their face matching comparisons, if their usage would improve accuracy, then it is possible that a role for face identification algorithms in forensic face matching may become more commonplace in the future. The purpose of Experiment 2 was to test the performance of six face identification algorithms on our unconcealed to masked image face matching task. Algorithm performance was then compared against the performance of

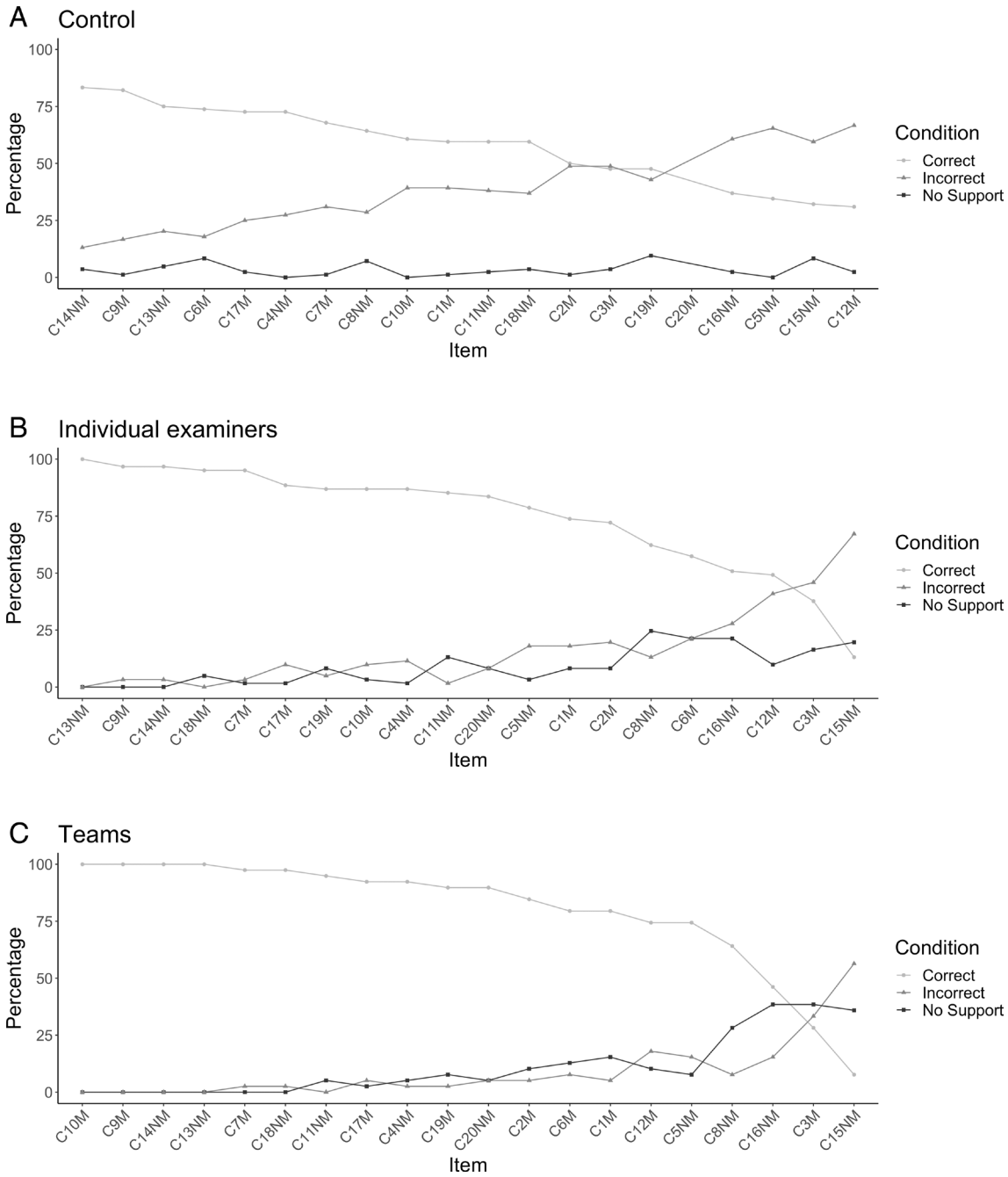

FIGURE 4 | Panel A (Controls), B (facial examiners), C (professional teams). Graphs show the performance breakdown for each image pair ranked by percentage of correct responses for that group.

facial examiners, teams, and controls using a threshold independent measure of accuracy to overcome the differences between the conclusion scale used by human participants and the score-based metrics produced by the algorithms.

3.1 | Methods

3.1.1 | Stimuli

The stimuli for this experiment were the 20 image pairs from Experiment 1.

3.1.2 | Algorithms

Six face identification algorithms were tested in this study: FaceNet (Schroff et al. 2015), VGG2 (Cao et al. 2018), ARC face (available through the FACER2VM project (Deng et al. 2019)), Surrey Face Identification System, the Imperial College Face identification algorithm, and Adaface (Kim et al. 2022). These algorithms were selected based on our access to these systems, and the variation in publication date of these algorithms allows us to test the performance of older and newer DCNNs. None were specifically designed to work with masked faces. The algorithms all work by processing a single face image to

give an output of a set of numbers, typically 512, that characterize that face. Face matching is performed by comparing the output numbers for the two faces, for example, by the cosine of the angle between them or the Euclidean distance, to generate a match score. This match score typically goes from 1, for a perfect match, to, in principle, −1 for a complete mismatch, though scores do not often go much below 0. The algorithm designers specify a threshold, above which a match is declared. A higher threshold reduces the chance of a false match at the risk of missing true matches. Here, we report performance using Area Under the Receiver Operating Characteristic Curve (AUC) score. If all the true mismatches have a similarity score lower than all of the true matches, then a threshold in the gap between them will separate them perfectly, and the AUC score is 1. If mismatch and match similarity scores overlap, the AUC will fall, with a value of 0.5 representing chance performance. This allows a direct comparison of the performance of different algorithms with each other, and against the human participants tested in Experiment 1.

3.2 | Results

All six algorithms achieved very high AUC scores on our face matching test (FaceNet AUC = 0.97, VGG2 AUC = 0.96, ArcFace AUC = 0.88, Surrey Face Identification System AUC = 1, Imperial College face identification algorithm AUC = 1, Adaface AUC = 1).

Collectively, the algorithms achieved a median AUC score of 0.98. For human participants tested in Experiment 1 (but here analysed as AUC score to allow for direct comparison with algorithm accuracy and for analysis of sensitivity and specificity) Teams achieved the highest median AUC (0.95), followed by Examiners (0.88) and then controls (0.65) A Kruskal-Wallis test confirmed that there was a significant difference in AUC scores across the groups, $\chi^2(5, 189) = 132.54$, $p < 0.001$, $\varepsilon^2 = 0.663$. Algorithms, Teams, and Examiners outperformed controls (all p values < 0.001), with no difference in performance between algorithms and Teams ($p = 1$), algorithms and Examiners ($p = 0.276$), or Teams and Examiners ($p = 0.174$).

Group level statistics can mask individual differences within groups that can provide important information about the performance of individual group members (Noyes et al. 2018). Single case t-tests were used to compare individual algorithms, teams, and examiners to the control participants' AUC distribution. Results revealed that 5 out of 6 algorithms (83%) were statistically superior to controls at the 95% confidence level for a two-tailed test (> 2 SDs above the mean). 65% of teams (25 out of 39) were superior to controls in contrast to only 36% of individual examiners (22 out of 61) (Figure 5). Spread of performance at the individual level is evident within all groups (Figure 5). All other individual comparisons to controls were not significant at the 95% confidence level for a two-tailed test.

3.3 | Discussion

All six algorithms achieved high AUC scores on our unconcealed to mask facial comparison test. Three algorithms (Surrey Face Identification System, Imperial college face identification algorithm and Adaface) made no errors on the task. The median algorithm AUC score is similar to that of human professional teams.

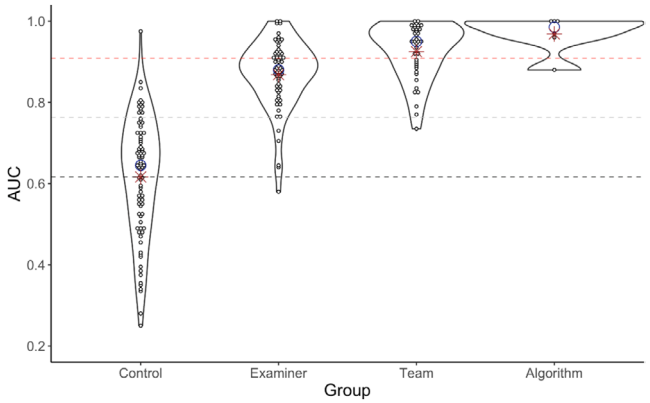

FIGURE 5 | Accuracy (AUC) scores for all participant groups (Controls, Examiners, Teams and Algorithms). Each dot depicts the response of one participant; the blue circle on each violin represents the median, and the red asterisk represents the mean. The red line indicates 2SDs above the control mean.

While the six algorithms that we tested performed with high accuracy, NIST tests of masked face matching found that some algorithms that they tested on mask to masked face comparisons performed with very low accuracy (Ngan et al. 2022). If algorithms are to be considered for applied use, then it is crucial to understand the accuracy of the specific algorithms for the specific applied task. The high accuracy of face identification algorithms on this task suggests that there may be a role for algorithms to work alongside humans (especially professional teams) in forensic face matching comparisons in the future, even for the challenging task of unconcealed to masked face matching; however, it is yet to be seen how this would work in practice.

4 | Experiment 3. Sensitivity and Specificity of Facial Examiners, Teams, and Controls

In Experiment 1 percentage accuracy was calculated using the predefined conclusion scale, with the "No support" decision acting as a threshold for determining correct and incorrect decisions. It was expected that the forensic examiner and team participants may be more familiar and adept at using the conclusion scale compared to novice controls, giving a potential advantage to the professional participants. The purpose of Experiment 3 was to determine the performance of the different groups of human participants by calculating the optimal thresholds on the conclusion scale for each individual participant, which should overcome any issues caused by a participant not understanding or not being adept in the use of the predefined conclusion scale.

4.1 | Methods

4.1.1 | Stimuli

The stimuli for this experiment were the twenty image pairs from Experiment 1.

4.1.2 | Procedure

The receiver operating characteristic (ROC) curve was used to calculate the threshold on the conclusion scale that resulted in the highest values in both sensitivity (the true positive rate, in this case responding support that the images depict the same person for a same identity trial) and specificity (the true negative rate, in this case responding support that the images depict different people for a different identity trial) for each human participant, at the top left of the curve. Where more than one threshold was given for a participant, the threshold that favored sensitivity was chosen. Threshold values were rounded to the nearest point on the 11-point verbal conclusion scale. The sensitivity and specificity of each participant were then calculated using the new thresholds.

4.2 | Results

Figure 6 shows the distribution of the new thresholds by group. As expected, the median thresholds for each group were close to 0 or "No support". A Kruskal-Wallis test confirmed that there was no significant difference in the distribution of new thresholds between the three groups ($\chi^2(2, 184) = 1.97$, $p = 0.373$, $\varepsilon^2 = 0.011$), however both individual examiners and teams showed a smaller range of thresholds, which were closer to the original central 'No support' decision, indicating that the professional groups were more adept at using the conclusion scale than novices.

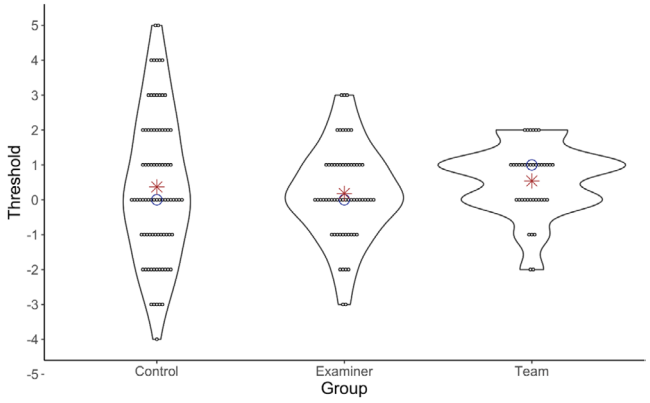

FIGURE 6 | The distribution of optimal thresholds for human participants by group with red star showing the mean, and blue circle the median sensitivity threshold.

Summary data for sensitivity and specificity are shown in Table 2. Teams had the highest overall sensitivity and specificity, followed by individual examiners. A Kruskal-Wallis test confirmed that there was a significant difference in sensitivity across the groups, $\chi^2(2, 189) = 59.64$, $p < 0.001$, $\varepsilon^2 = 0.326$. Teams and examiners outperformed controls (all p values < 0.001) and teams were observed to have significantly higher sensitivity compared to individual examiners ($p = 0.027$). A Kruskal-Wallis test confirmed that there was a significant difference in specificity across the groups, $\chi^2(2, 189) = 81.11$, $p < 0.001$, $\varepsilon^2 = 0.443$. Teams and examiners outperformed controls (all p values < 0.001) with no significant difference between teams and examiners ($p = 0.427$).

4.3 | Discussion

The findings of Experiment 3 mirror those found for percentage accuracy for same identity and different identity trials in Experiment 1 and further demonstrate that both Teams and Examiners are better at correctly resolving both same identity trials and different identity trials compared to controls. The findings did indicate that use of the predefined conclusion scale was more diverse in control participants compared to Examiners and Teams, but this did not affect the overall pattern of results. Teams also outperformed Examiners on same identity trials.

5 | General Discussion

Facial examiners and professional teams encounter concealed and masked face images in their applied work, but prior to this study, the accuracy of their identifications for such faces was unknown, which is concerning given the high stakes of forensic identifications. Here we assessed the performance of facial examiners and professional teams on an unconcealed to masked face image comparison task. In Experiment 1, we tested the largest international sample of practicing face identification practitioners in any published face identification study to date and compared the performance of facial examiners, professional teams, and control participants. We found that facial examiners and professional teams outperformed control participants. This result supports the use of trained professionals for facial image comparisons that include masked faces in legal and applied settings. In Experiment 2, face matching performance of six face identification algorithms was calculated for the images used in Experiment 1. Most of the algorithms

TABLE 2 | Summary data for sensitivity and specificity by group using optimal threshold.

	Group	*N*	Mean	SD	SEM	Max	Median	Min
Control	Sensitivity	84	0.68	0.15	0.02	1.00	0.70	0.10
	Specificity		0.66	0.12	0.01	0.90	0.70	0.30
Examiner	Sensitivity	61	0.81	0.13	0.02	1.00	0.80	0.40
	Specificity		0.83	0.11	0.01	1.00	0.80	0.50
Team	Sensitivity	39	0.89	0.11	0.02	1.00	0.90	0.60
	Specificity		0.87	0.09	0.01	1.00	0.90	0.70

performed with extremely high accuracy on the task, and the median accuracy score for algorithms was comparable to that seen for professional teams. Notably, the algorithms that did not achieve perfect performance were algorithms from or pre the year 2019, and the newer algorithms achieved perfect performance. The high accuracy scores achieved by algorithms for masked faces facilitate discussion on the potential future use of face identification algorithms to support facial examiners with comparison in case work.

Previous tests of facial examiner performance have focused on unconcealed faces (Phillips et al. 2018; Towler et al. 2023; Sexton et al. 2024). Facial examiners outperform typical observers for unconcealed faces, but in case work, faces may be concealed by masks or other face coverings. Our result that facial examiners and professional teams are more accurate than controls for unconcealed to masked faces demonstrates that their skills extend beyond facial comparison of unconcealed faces. Therefore, the training and techniques used by examiners for unconcealed faces are useful for comparison of unconcealed to concealed (masked) faces.

We believe that facial examiners, following normal working practice, use different strategies to compare faces than typical observers. Typical observers tend to use various processing strategies, including feature-by-feature, configural, and holistic processing (Bindemann and Burton 2021; Maurer et al. 2002), all of which are disrupted when a mask covers part of the face. However, there is currently not a clear theoretical account of how individuals perform face matching tasks (Bindemann and Burton 2021). The facial examiners were likely to be using a feature comparison strategy, which focuses on the comparison of specific features of the face. This processing strategy can still be applied to masked images, as comparisons can be made for the features that are visible in both images. Carragher et al. (2022) found that the face matching performance of typical observers for masked faces improved when the typical observers were trained to use feature comparison methods. Taking the results of our study and previous work together, it can be argued that facial examiners use the feature comparison strategy for comparisons, and that the feature comparison strategy is beneficial for both unconcealed and masked face comparisons (Carragher et al. 2022; Carragher and Hancock 2020; Moreton et al. 2021; Noyes, Davis, et al. 2021; Noyes, Parde, et al. 2021). Facial examiners' experience with the feature comparison strategy may put them at an advantage for masked face identification over control participants. Or perhaps facial examiners are more aware of which of the visible features are most informative for identifications. Our study compared the identification of examiners, teams, and controls in scenarios that match the real-world comparison situation. A more controlled study would be needed to disentangle the mechanisms used by professionals and controls.

Similar to Towler et al. (2023) we found a performance advantage for professional teams over individual examiners. While both facial examiners and professional teams outperformed controls on the facial comparison task, professional teams made significantly fewer incorrect identifications than facial examiners and were more likely to use the no support option than facial examiners. In case work, a no support decision is preferable to an incorrect identification. An incorrect decision in a high-stakes or security-critical forensic investigation means that two different people are mistaken to be the same person or that the same person is mistaken to be two different people—potentially leading to the arrest of an innocent person or excess time spent to find a second individual who does not exist. Professional team decision making could reduce the number of incorrect identifications in high-risk identification scenarios.

The benefit of professional team working is clear; teams reduce incorrect identifications. However, little is known about how professional teams reach a decision. For typical observers, the mechanisms behind the team advantage are not thought to be linked to the confidence of individuals within a team, response times, nor the content of discussion between individuals (Ritchie et al. 2022). It is unknown whether these factors contribute to enhanced decision making for professional teams. The professional participants in our study completed the responses *either* as an individual facial examiner *or* as part of a professional team. This was determined by normal working practice for the agency (if examiners usually completed identifications as an individual then they completed the tests as an individual, if they usually operated as teams, then they returned a team response). It is possible that the participants who completed the test as part of a team were individually better identifiers than those who completed the task as individuals. This is an unlikely explanation of the results, given the previously shown team advantage for typical observers, and our result that teams acted differently from individual examiners (increased use of the no support response for difficult trials). Professional teams were instructed to complete the task as they would in normal working practice. Therefore, different teams may have benefited from different strategies. For example, some teams may have discussed identifications with one another during the feature comparison process; others may have fused individual results of team members after each member of the team had made their individual identification; or professional teams may have divided up the image comparisons between individuals in the team. Past work has shown benefits of both collaborative working (Dowsett and Burton 2015) and blind fusion in typical observers (Jeckeln et al. 2018; Ritchie et al. 2022); alternatively, a divide and conquer approach to identifications may have reduced fatigue effects (Alenezi et al. 2015). More research is needed to understand the nature of the professional team advantage in applied practice.

In Experiment 2 we found that the six face recognition algorithms that we tested performed with high accuracy on the unconcealed to masked face image comparison test. Many algorithms are now extremely accurate at face recognition, and it is interesting that the algorithms that we tested performed well on the unconcealed to masked face task, despite not being specifically designed to do so. This suggests that these face recognition algorithms can make face identifications even if one of the comparison images is concealed. The algorithms that we tested in this study achieved a mean AUC score comparable to that of professional teams. If algorithms can complete the task with very high accuracy (as observed in our study) then there may be a role for face recognition algorithms to assist human professionals with their case work. It is unclear how this would work in practice, specifically in terms of the role of both the human and the algorithm, and how responses could be combined. There are many possible methods of use. Currently, algorithms

are used to aid humans in the decision-making process, generally by narrowing down potential matches from candidate lists. The human then performs the one-to-one facial image comparison. If algorithms were to be used in one-to-one comparisons, thought is needed over the order of operation of decisions. Is the decision first made by the human, or first made by the algorithm? Either order has the potential to influence the final response, although a decision initially made by an algorithm may have greater influence over the final response (Howard et al. 2020). If the decisions of the human and algorithm are to be combined, how would this work in practice? In published papers, comparisons of human and algorithm accuracy are typically made using AUC scores. AUC scores provide a convenient method of comparing the overall performance of humans and machines for a set of face images because it is threshold independent. However, performance for individual case comparisons (a specific image pair) in forensic case work requires a threshold in order to make an identification decision. It is possible to implement threshold-dependent methods of algorithm identifications that can be applied to single comparison cases; however, it is unclear how this would map onto the decision scale used by examiners, not least because algorithms do not have a no support decision which is available to examiners. Researchers have proposed likelihood ratios derived from algorithm similarity scores as a way of combining human examiner and algorithm decisions (Macarulla Rodriguez et al. 2020); however, current automated face recognition systems are not designed to support forensic examinations in this way (Bollé et al. 2020). Answers to these questions must be carefully considered before algorithms are implemented into the facial comparison process. Tests of both algorithm and human performance remain crucial to ensure that each component of the decision chain is skilled at identifying the image type at hand.

It is worth noting that while patterns of overall accuracy for facial examiners, professional teams, controls, and algorithms in our study were clearly visible at the group level, individual differences in performance were observed for members of all groups. Starting with facial examiners, the observed spread in performance at the individual level could be linked to individual differences in facial comparison ability, differences in training, on-the-job experience, or time spent on comparisons (Moreton et al. 2021). The spread in team performance across teams could also be linked to any of these factors, or perhaps by how the different teams approached the task (e.g., discussion/blind fusion/division of images). The largest variation in individual performance was observed for control participants. This is a typical result for studies on face recognition (e.g., Phillips et al. 2018; White et al. 2015). We made efforts to match the demographics of our control participants to our professional groups (control participants were recruited from the same age bracket, gender demographics, and countries as professional participants). Control participants are untrained in facial comparison methods and must therefore rely on their natural face recognition abilities to complete the task. Face recognition ability varies drastically from person to person (Bobak et al. 2016; Fysh et al. 2020; Wilmer 2017). Participants with extremely high scores may have been 'super-recognisers', meaning that they are people who are naturally extremely good at recognizing faces (for a review see Noyes et al. 2017). Or they may have paid more attention and put more effort into the task than other controls. Likewise, low performers may have been influenced by low natural face recognition ability, and/or paid less attention to or been less motivated by the task. We also found differences in the performance of different algorithms. Just as performance can vary for human participants within a participant group, performance also differs across different algorithms (Grother 2022). Most work on individual differences in the performance of humans and algorithms has come from research on unconcealed faces (e.g., Bobak et al. 2016; Grother 2022; Noyes et al. 2018; Wilmer 2017). Our study demonstrates that these differences are also observed for concealed faces. Whether the same examiners/algorithms who are good at face recognition for unconcealed faces are good at unconcealed faces to masked faces is a different matter that needs to be addressed. We expect that the answer will be complicated, with some examiners/algorithms performing at similar levels on both tasks, and others showing differences in performance across tasks.

Our study focuses on the applied scenario of comparing the comparisons made by a facial examiner under conditions which match working practice, against quicker decisions made by an untrained individual, such as a member of a jury's assessment of images. While we did not match the time allowed for controls to complete the task with the time allowed for facial examiners, as this was not the purpose of our study, previous research shows that additional time does not benefit the performance of controls (White et al. 2015). Superior performance of facial examiners is likely due to a combination of factors including training, tool use, document use, procedures, and adequate time to complete the facial comparison using the methods which they are trained in.

The current study focused specifically on unconcealed to masked face image comparisons—a scenario that examiners encounter in their case work, and that has become more frequent since the covid-19 pandemic. There are of course other methods of concealing appearance, from simple measures such as sunglasses (e.g., Noyes, Davis, et al. 2021; Noyes, Parde, et al. 2021; Kramer and Ritchie 2016) to more complicated forms of disguise (e.g., Noyes and Jenkins 2019; Noyes, Davis, et al. 2021; Noyes, Parde, et al. 2021). We suggest a need for future research to investigate examiner performance for other types of concealment.

6 | Conclusions

Facial examiners and professional teams both compare images of unconcealed to concealed (masked) faces with higher accuracy than controls. Professional teams made fewer errors than individual examiners; this is a particularly important result as errors can lead to serious consequences in forensic investigations. Facial examiners and professional teams are much better than the average person (e.g., a jury member) at face identifications for masked faces. Our results support the notion that facial examiners use feature-based strategies for face identification as a result of their training. Recent computer algorithms also performed very highly. There is work to be done to establish a possible future role of face recognition algorithms to assist facial examiners in their facial image comparisons.

Author Contributions

Eilidh Noyes: conceptualisation, investigation, methodology, data curation, writing – original draft, project administration, supervision. **Reuben Moreton:** conceptualisation, investigation, methodology, data curation, writing – original draft, analysis. **Peter J. B. Hancock:** conceptualisation, methodology, data curation, writing – original draft, analysis. **Kay L. Ritchie:** conceptualisation, methodology, data curation, writing – review and editing. **Sergio Castro Martinez:** data curation, project administration. **Katie L. H. Gray:** conceptualisation, methodology, data curation, writing – review and editing. **Josh P. Davis:** data curation, writing – review and editing.

Acknowledgments

The authors have nothing to report.

Ethics Statement

This study was approved by the School of Human and Health Sciences Ethics Committee at the University of Huddersfield.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are available in OSF at https://osf.io/4g8mx/?view_only=089ea94cfdca49818b8f33cdc259da43.

References

Alenezi, H. M., M. Bindemann, M. C. Fysh, and R. A. Johnston. 2015. "Face Matching in a Long Task: Enforced Rest and Desk-Switching Cannot Maintain Identification Accuracy." *PeerJ* 3: e1184. https://doi.org/10.7717/peerj.1184.

Babwin, D., and S. Dazio. 2020. "Coronavirus Masks a Boon for Crooks Who Hide Their Faces." AP News. https://apnews.com/article/f97b4914b4159dec0c98359fac123d52.

Bindemann, M., and M. Burton. 2021. "Steps Towards a Cognitive Theory of Unfamiliar Face Matching." In *Forensic Face Matching: Research and Practice*, edited by M. Bindemann, 38–61. Oxford University Press.

Bobak, A. K., P. Pampoulov, and S. Bate. 2016. "Detecting Superior Face Recognition Skills in a Large Sample of Young British Adults." *Frontiers in Psychology* 7: 1378. https://doi.org/10.3389/fpsyg.2016.01378.

Bollé, T., E. Casey, and M. Jacquet. 2020. "The Role of Evaluations in Reaching Decisions Using Automated Systems Supporting Forensic Analysis." *Forensic Science International: Digital Investigation* 34: 301016. https://doi.org/10.1016/j.fsidi.2020.301016.

Burton, A. M., D. White, and A. McNeill. 2010. "The Glasgow Face Matching Test." *Behavior Research Methods* 42, no. 1: 286–291. https://doi.org/10.3758/BRM.42.1.286.

Cao, Q., L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. 2018. "VGGFace2: A Dataset for Recognising Faces Across Pose and Age." 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). Xi'an. https://doi.org/10.1109/fg.2018.00020.

Carragher, D. J., and P. J. B. Hancock. 2020. "Surgical Face Masks Impair Human Face Matching Performance for Familiar and Unfamiliar Faces." *Cognitive Research: Principles and Implications* 5, no. 1: 59. https://doi.org/10.1186/s41235-020-00258-x.

Carragher, D. J., A. Towler, V. R. Mileva, D. White, and P. J. B. Hancock. 2022. "Masked Face Identification Is Improved by Diagnostic Feature Training." *Cognitive Research: Principles and Implications* 7, no. 1: 30. https://doi.org/10.1186/s41235-022-00381-x.

Deng, J., J. Guo, N. Xue, and S. Zafeiriou. 2019. "Arcface: Additive Angular Margin Loss for Deep Face Recognition." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4690–4699. IEEE.

Dowsett, A. J., and A. M. Burton. 2015. "Unfamiliar Face Matching: Pairs Out-Perform Individuals and Provide a Route to Training." *British Journal of Psychology* 106, no. 3: 433–445. https://doi.org/10.1111/bjop.12103.

Edmond, G., D. White, A. Towler, M. San Roque, and R. Kemp. 2021. "Facial Recognition and Image Comparison Evidence: Identification by Investigators, Familiars, Experts, Super-Recognisers and Algorithms." *Melbourne University Law Review* 45: 99.

European Network of Forensic Science Institutes. 2018. "ENFSI Best Practice Manual for Facial Image Comparison." Vol. 01, Issue January. https://enfsi.eu/wp-content/uploads/2017/06/ENFSI-BPM-DI-01.pdf.

Farah, M. J., K. D. Wilson, M. Drain, and J. N. Tanaka. 1998. "What Is "Special" About Face Perception?" *Psychological Review* 105, no. 3: 482–498. https://doi.org/10.1037/0033-295X.105.3.482.

Fysh, M. C., L. Stacchi, and M. Ramon. 2020. "Differences Between and Within Individuals, and Subprocesses of Face Cognition: Implications for Theory, Research and Personnel Selection." *Royal Society Open Science* 7: 200233. https://doi.org/10.1098/rsos.200233.

Grother, P. 2022. "Face Recognition Vendor Test (FRVT) Part 8: Summarizing Demographic Differentials." NIST Interagency Report, NIST IR 8429 Ipd. https://doi.org/10.6028/NIST.IR.8429.ipd.

Hole, G. J. 1994. "Configurational Factors in the Perception of Unfamiliar Faces." *Perception* 23: 65–74.

Howard, J. J., L. R. Rabbitt, and Y. B. Sirotin. 2020. "Human-Algorithm Teaming in Face Recognition: How Algorithm Outcomes Cognitively Bias Human Decision-Making." *PLoS One* 15, no. 8: e0237855. https://doi.org/10.1371/journal.pone.0237855.

Jacquet, M., and C. Champod. 2020. "Automated Face Recognition in Forensic Science: Review and Perspectives." *Forensic Science International* 307: 110124. https://doi.org/10.1016/j.forsciint.2019.110124.

Jeckeln, G., C. A. Hahn, E. Noyes, J. G. Cavazos, and A. J. O'Toole. 2018. "Wisdom of the Social Versus Non-Social Crowd in Face Identification." *British Journal of Psychology* 109, no. 4: 724–735. https://doi.org/10.1111/bjop.12291.

Kim, M., A. K. Jain, and X. Liu. 2022. "AdaFace: Quality Adaptive Margin for Face Recognition." In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. https://doi.org/10.1109/cvpr52688.2022.01819.

Kramer, R. S., and K. L. Ritchie. 2016. "Disguising Superman: How Glasses Affect Unfamiliar Face Matching." *Applied Cognitive Psychology* 30, no. 6: 841–845. https://doi.org/10.1002/acp.3261.

Macarulla Rodriguez, A., Z. Geradts, and M. Worring. 2020. "Likelihood Ratios for Deep Neural Networks in Face Comparison." *Journal of Forensic Sciences* 65, no. 4: 1169–1183. https://doi.org/10.1111/1556-4029.14324.

Maurer, D., R. Le Grand, and C. J. Mondloch. 2002. "The Many Faces of Configural Processing." *Trends in Cognitive Sciences* 6, no. 6: 255–260. https://doi.org/10.1016/S1364-6613(02)01903-4.

Moreton, R., C. Havard, A. Strathie, and G. Pike. 2021. "An International Survey of Applied Face-Matching Training Courses." *Forensic Science*

International 327: 110947. https://doi.org/10.1016/j.forsciint.2021.110947.

Murphy, J., K. L. H. Gray, and R. Cook. 2017. "The Composite Face Illusion." *Psychonomic Bulletin & Review* 24: 245–261. https://doi.org/10.3758/s13423-016-1131-5.

Ngan, M., P. Grother, and K. Hanoaka. 2022. "Ongoing Face Recognition Vendor Test (FRVT), Part 6B: Face Recognition Accuracy With Face Masks Using Post COVID-19 Algorithms. NIST 8331 Draft Supplement." https://doi.org/10.6028/NIST.IR.8331. (nist.gov) on 09/11/2023.

Noyes, E., J. P. Davis, N. Petrov, K. L. H. Gray, and K. L. Ritchie. 2021. "The Effect of Face Masks and Sunglasses on Identity and Expression Recognition With Super-Recognizers and Typical Observers." *Royal Society Open Science* 8, no. 3: 201169. https://doi.org/10.1098/rsos.201169.

Noyes, E., and M. Q. Hill. 2021. "Forensic Face Matching: Research and Practice." In *Automatic Recognition Systems and Human Computer Interaction in Face Matching*, edited by M. Bindemann, 193–215. Oxford University Press.

Noyes, E., M. Q. Hill, and A. J. O'Toole. 2018. "Face Recognition Ability Does Not Predict Person Identification Performance: Using Individual Data in the Interpretation of Group Results." *Cognitive Research: Principles and Implications* 3, no. 1: 23. https://doi.org/10.1186/s41235-018-0117-4.

Noyes, E., and R. Jenkins. 2019. "Deliberate disguise in face identification." *Journal of Experimental Psychology: Applied* 25: 280–290. https://doi.org/10.1037/xap0000213.

Noyes, E., C. Parde, I. Colon, et al. 2021. "Seeing Through Disguise: Getting to Know You With a Deep Convolutional Neural Network." *Cognition* 211: 104611. https://doi.org/10.1016/j.cognition.2021.104611.

Noyes, E., P. J. Phillips, and A. J. O'Toole. 2017. "What Is a Super-Recogniser?" In *Face Processing: Systems, Disorders and Cultural Differences*. Nova Science Publishers Inc.

Phillips, P. J., A. N. Yates, Y. Hu, et al. 2018. "Face Recognition Accuracy of Forensic Examiners, Superrecognizers, and Face Recognition Algorithms." *Proceedings of the National Academy of Sciences of the United States of America* 115, no. 24: 6171–6176. https://doi.org/10.1073/pnas.172135511.

Rawlinson, K. 2021. "Rise in Suspects Using Face Coverings to Mask Identity, Say Kent Police." The Guardian. https://www.theguardian.com/uk-news/2021/apr/16/rise-in-suspects-using-face-coverings-to-mask-identity-say-kent-police.

Ritchie, K. L., D. J. Carragher, J. P. Davis, et al. 2024. "Face Masks and Fake Masks: The Effect of Real and Superimposed Masks on Face Matching With Super-Recognisers, Typical Observers, and Algorithms." *Cognitive Research: Principles and Implications* 9, no. 1: 5.

Ritchie, K. L., T. R. Flack, E. A. Fuller, C. Cartledge, and R. S. S. Kramer. 2022. "The Pairs Training Effect in Unfamiliar Face Matching." *Perception* 51, no. 7: 477–495. https://doi.org/10.1177/030100662210969.

Rossion, B. 2013. "The Composite Face Illusion: A Whole Window Into Our Understanding of Holistic Face Perception." *Visual Cognition* 21, no. 2: 139–253. https://doi.org/10.1080/13506285.2013.772929.

Ruifrok, A. C. C., P. Vergeer, and A. M. Rodrigues. 2022. "From Facial Images of Different Quality to Score Based LR." *Forensic Science International* 332: 111201. https://doi.org/10.1016/j.forsciint.2022.111201.

Schroff, F., D. Kalenichenko, and J. Philbin. 2015. "FaceNet: A Unified Embedding for Face Recognition and Clustering." In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. https://doi.org/10.1109/cvpr.2015.7298682.

Sexton, L., R. Moreton, E. Noyes, S. Castro Martinez, and S. Laurence. 2024. "The Effect of Facial Ageing on Forensic Facial Image Comparison." *Applied Cognitive Psychology* 38, no. 1: e4153. https://doi.org/10.1002/acp.4153.

Stajduhar, A., T. Ganel, G. Avidan, R. S. Rosenbaum, and E. Freud. 2022. "Face Masks Disrupt Holistic Processing and Face Perception in School-Age Children." *Cognitive Research: Principles and Implications* 7, no. 1: 9. https://doi.org/10.1186/s41235-022-00360-2.

Steyn, M., M. Pretorius, N. Briers, N. Bacci, A. Johnson, and T. M. R. Houlton. 2018. "Forensic Facial Comparison in South Africa: State of the Science." *Forensic Science International* 287: 190–194. https://doi.org/10.1016/j.forsciint.2018.04.006.

Towler, A., J. D. Dunn, S. Castro Martínez, et al. 2023. "Diverse Types of Expertise in Facial Recognition." *Scientific Reports* 13, no. 1: 11396. https://doi.org/10.1038/s41598-023-28632-x.

Towler, A., D. White, and R. I. Kemp. 2017. "Evaluating the Feature Comparison Strategy for Forensic Face Identification." *Journal of Experimental Psychology. Applied* 23, no. 1: 47–58. https://doi.org/10.1037/xap0000108.

White, D., R. I. Kemp, R. Jenkins, M. Matheson, and A. M. Burton. 2014. "Passport Officers' Errors in Face Matching." *PLoS One* 9, no. 8: e103510.

White, D., P. J. Phillips, C. A. Hahn, M. Hill, and A. J. O'Toole. 2015. "Perceptual Expertise in Forensic Facial Image Comparison." *Proceedings of the Royal Society B: Biological Sciences* 282, no. 1814: 20151292. https://doi.org/10.1098/rspb.2015.1292.

Wilmer, J. B. 2017. "Individual Differences in Face Recognition: A Decade of Discovery." *Current Directions in Psychological Science* 26, no. 3: 225–230. https://doi.org/10.1177/0963721417710693.

Young, A. W., D. Hellawell, and D. C. Hay. 1987. "Configurational Information in Face Perception." *Perception* 16, no. 6: 747–759. https://doi.org/10.1177/096372141771069.