The Effect of Facial Ageing on Forensic Facial Image Comparison.

Laura Sexton a, b, Reuben Moreton b, Eilidh Noyes c, Sergio Castro Martinez d & Sarah Laurence a

a School of Psychology & Counselling, The Open University, Milton Keynes, UK

b School of Psychology, University of Sunderland, Sunderland, UK

c School of Psychology, University of Huddersfield, Huddersfield, UK

d Comisaría General de Policía Científica, National Police, Spain

Corresponding Author: Sarah Laurence, School of Psychology & Counselling, The Open University, sarah.laurence@open.ac.uk

Conflict of interest statement: The authors declare no conflict of interest.

Data availability statement: The data that support the findings of this study are openly available in the OSF repository at https://osf.io/5szxr/?view_only=60a8e58ebc674d28b0b6279196ac964e.

Acknowledgements: This research was supported by an Economic and Social Research Council New Investigator Grant (ES/R005788/2) awarded to Sarah Laurence.

Abstract

Facial appearance changes over time as people age. This poses a challenge for individuals working in forensic settings whose role requires them to match the identity of face images. The present research aimed to determine how well an international sample of forensic facial examiners could match faces with a substantial age gap. We tested a sample of 60 facial examiners, 23 professional teams and 81 untrained control participants. Participants matched pairs of photographs with a 10–30-year age gap between the images. Participants also estimated the ages of the faces. On the matching task, individual professionals and teams outperformed controls and made fewer high confidence errors. On the age estimation task, there was no advantage for professionals relative to controls. Our results suggest that forensic facial examiners can tolerate substantial age differences between adult faces when performing comparisons, but this advantage does not extend to accurate age estimation.

Keywords: Face recognition, facial examiners, facial image comparison, facial ageing, age estimation

Introduction

Deciding that two or more face photographs depict the same unfamiliar person (i.e., unfamiliar face matching) is a challenging task (Bruce et al., 1999, 2001; Clutterbuck & Johnston, 2002, 2004; Megreya & Burton 2006, 2007). Different images of the same unfamiliar person can appear to belong to different people (Jenkins et al., 2011), and images of different people can appear to belong to the same person (Noyes & Jenkins, 2019). One reason why face matching is so difficult is because a single face photograph may not be a reliable indicator of a person's appearance (see Burton, 2013). Faces vary from moment to moment, from one day to the next (e.g., due to changes in hairstyle and makeup), and incidental changes in appearance (e.g., changes in viewpoint and expression) can also affect perceived identity (Jenkins et al., 2011).

There are also appearance changes that occur more slowly over time, such as those associated with ageing. Age-related changes may not be obvious in the short term but become more noticeable when looking at photographs taken years previously. For example, when looking at old family photographs, out-of-date photo identification or decades-old CCTV footage. As people age, they experience skeletal changes (Mendelson & Wong, 2020), changes to soft tissues, loss of facial volume, and skin changes such as wrinkling, pigmentation, and coarse texture (see Ko et al., 2017 for a review). The speed with which age-related changes take hold of facial appearance differs from person to person; both environmental (Rexbye et al., 2006) and genetic factors (Djordjevic et al., 2016) contribute to the ageing process. Given the significant appearance changes associated with facial ageing, it is unsurprising that ageing has a negative effect on face matching. Research has shown that matching faces taken on the same day is prone to error (e.g., Burton et al, 2010). Performance is worse when the gap between images is increased by months (Megreya et al., 2013), by a year (Davis & Valentine, 2009), or by decades (Mileva et al., 2020). Tests of face matching that involve a substantial change in age have therefore proved useful for discriminating good from poor recognisers (e.g.,

The Yearbook Task; Fysh et al., 2020, Stacchi et al., 2020, based on Bruck et al., 1991), including from a sample of police officers (Nador et al., 2022).

The research outlined so far involved untrained participants or police officers whose day-to-day role involves identification of faces. There are, however, trained professionals, known as facial examiners, who are involved in making face-matching decisions (also known as forensic facial comparisons) in forensic settings. Forensic examiners work in policing, for forensic service providers and in government departments and may provide expert evidence in court (European Network of Forensic Science Institutes, 2018). The decisions made by forensic examiners, therefore, have serious consequences and could, if incorrect, potentially lead to the incarceration of an innocent person.

There have been multiple demonstrations of superior face-matching ability in facial examiners relative to untrained controls (Norell et al., 2015; White et al., 2015; Towler et al., 2017; Phillips et al., 2018; see White, Towler & Kemp, 2021;). This advantage is believed to be due to the morphological analysis procedures applied by examiners, which involve systematic analysis and comparison of facial features (Towler et al., 2021; Moreton, 2021). Many examiners undergo years of training in morphological analysis (Moreton et al., 2021). Training novices to assess the similarity of individual facial features has been found to improve their matching performance (Towler et al., 2017). As has encouraging novices to focus on diagnostic features used by forensic examiners (e.g., Towler et al., 2021), such as moles (Fysh & Bindemann, 2022). However, the benefits of feature comparison associated with forensic examiners likely requires extensive training and mentoring (see Towler et al., 2019; Moreton et al., 2021)

In a recently published survey of 18 agencies that train forensic examiners, Moreton et al. (2021) found that in addition to training in morphological analysis, 16 agencies (89%)

provided training on the effects of ageing and the permanence of features. Despite many examiners being trained on the effects of ageing, little is known about how well forensic examiners can match faces that have undergone significant age-related changes. To our knowledge, only one study has examined facial comparison practitioner performance when comparing faces across a substantial change in age. Michalski et al. (2019) measured the ability of facial reviewers to match faces of children when the images in each pair could differ by up to 10 years in age. Facial reviewers are professionals trained to perform faster and less meticulous identifications than facial examiners that can be used to support the activities of law enforcement, e.g., generating leads in criminal cases.

Michalski et al (2019) found that facial reviewers were better at matching images of older children than younger children, and performance was better with a smaller age gap between the images. However, age-related changes to appearance that occur during childhood, captured by Michalski et al.'s stimuli, are different to age-related changes that occur during adulthood (Enlow, 1982). The present research therefore aimed to extend previous research by examining forensic examiner ability to match face images that show ageing across the adult lifespan. As faces change extensively with age, this provides an important test of the extent to which morphological analysis is robust to ageing effects. We created a test in which facial examiners and untrained control participants compared/matched images of faces. In each comparison, one photo was a high-quality passport-style colour photograph taken within the last month and the second image (the comparison image) was taken between 10-30 years ago that either belonged to the same identity as the reference image (match trials) or a different identity (mismatch trials). Examiners receive formalised training in comparing facial images that can last from several months to over five years, and often includes instruction on the effects of ageing on facial appearance (Moreton et al. 2021). Therefore, we predicted that trained examiners would exhibit superior face-matching performance when comparing faces with a

substantial age gap. If examiners perform better than controls, then we expect them to make more accurate responses and fewer high confidence errors.

An additional aim was to explore whether forensic examiner expertise in face matching extends to other aspects of facial decision making, namely age estimation. People are typically relatively accurate at estimating the age of an unfamiliar face, with an average age estimation error of 8 years for faces aged between 7 – 70 years (Clifford et al., 2018). Similar results have been obtained for stimuli displaying a wider range of ages (Barthold Jones et al., 2019). According to classic models of face perception (e.g., Bruce & Young, 1986), face matching involves directed visual processing and facial examiner expertise is thought to be due to expertise via this route (Towler et al., 2021). Age perception is a form of visually derived semantic information that can be gleaned from unfamiliar faces (Bruce & Young, 1986) and is thought to be extracted early on in visual processing (Bruyer et al, 1991). If the route underpinning facial examiner expertise and the processes underpinning age perception represent partially distinct processes, examiners may not have an advantage relative to untrained controls – their expertise may not extend beyond face matching. Evidence to support this possibility comes from Chatterjee and Nakayama (2012), who found typical age perception in a group of people with developmental prosopagnosia, with severe face identity recognition deficits. Conversely, if expertise in face matching does extend to other aspects of facial processing, then practitioners may outperform controls both in matching and age judgements. It is possible that because examiners tend to be trained on the effects of ageing (Moreton et al. 2021), this might extend their expertise beyond face matching. There is some evidence that facial age estimation accuracy can be improved via training (Sörqvist, & Eriksson, 2007). Previous research has also found that people are more accurate at estimating the age of own-race faces than other-race faces (Dehon & Bredart, 2001; Thorley, 2021), suggesting that expertise may play a role in age estimation. To explore facial examiner ability to estimate age,

examiner and control participants completed an age estimation task in addition to the matching task.

Method

Participants

Facial examiners were recruited via adverts shared through professional and organisational mailing lists associated with the European Network of Forensic Science Institutes (ENFSI). Forty-one forensic organisations from 21 countries agreed to participate in the research. Facial examiners (n = 60) are individuals with extensive training in facial image comparison. Their decisions, which typically involve a rigorous and time-consuming process, are used to support legal cases, prosecutions, and expert testimony in court. Team responses (n = 23) were permitted where a particular agency's standard operating procedures for facial image comparison allow for this. Teams were diverse in nature, ranging in size from 2-12 individuals. Teams consisted either solely of examiners or a combination of examiners and other facial comparison professionals[1]. In order to preserve the anonymity of examiner and other professional participants, given the sensitivity of their work, it was not possible to collect demographic data for the Examiner or Teams groups.

The control group consisted of 81 untrained participants (43 female, 36 male, 2 other; mean age = 25.99, SD = 7.17) recruited online via Prolific (https://www.prolific.co/). All participants were fluent in English and had normal or corrected-to-normal vision. Control participants were compensated £2 for their time. All participants (forensic professionals and controls) provided online informed consent. Ethical approval for data collection was received from the Open University, UK.

[1] We also received responses from two other types of forensic professional. Reviewers (n = 13) are individuals trained to perform faster and less meticulous identifications (e.g., often used to support the activities of law enforcement). Police super-recognisers (n = 6) are law enforcement professionals recruited to specialist "super recognition teams" based on their interest and expertise in facial image comparison. Super-recognisers (SRs) engage in a variety of applied face recognition tasks, including the identification of live suspects or the recognition of individuals from CCTV footage (Robertson et al., 2016). As the results of this research centre on the performance of facial examiners, only results from examiners and teams are reported in the results section.

Materials and Stimuli

Images used in the final test were donated by consenting adult individuals. One hundred and six models were recruited via Prolific and through a word-of-mouth campaign. Models were compensated £5 for the donation of their images. Each model submitted at least two photographs of themselves. The first image (the reference image) was a high-quality passport-style colour photograph taken within the last month, in good lighting, against a light-coloured background. Models were advised that their eyes should be open and visible in the image and that their face should not be obstructed by hair. The second image (the comparison image) was a clear photograph of the model's face taken between 10-30 years ago. The majority of the submitted comparison images were of slightly lower quality than the reference images and were reflective of photos typically shared on social media. We requested that both images be unaltered by software and filters and free from red eye. Figure 1 (below) represents an example of the type of images used in the test. Ethical approval for the acquisition of the photographs was granted by the Open University, UK.

Figure 1. Example images that represent the stimuli used in this experiment. In the left-hand image (the reference image), the subject is 35 years old, while in the right-hand image (the questioned image), the subject is 18. (Due to restrictions on the reuse and distribution of the test images, it is not possible to reproduce them in this paper).

The donated images (minus any that were deemed to be of insufficient quality) were placed into 210 pairs by two members of the team, each pair containing one reference and one questioned image. Half of the pairs were matches (where the reference and questioned image featured the same person), and half were non-matches (where the reference and questioned image featured different but similar-looking people). Each reference image featured in one matching and one non-matching pair. For non-matches, each reference image was paired with the questioned image that bore its greatest similarity. Similarity judgements were based on gender, ethnicity and the general likeness of facial features. In each pair, the age gap between the two facial images ranged from 10-30 years.

Two members of the research team who were not involved in pairing up the images then reduced the stimulus set to 85 pairs by removing trials in which the correct answer (match or mismatch) was overtly obvious. The remaining image pairs were compared by 50 untrained participants recruited online from Prolific. Participants were asked to respond if they thought the images featured the same person or two different people. Accuracy rates ranged from 26-94% for match trials (M = 69.67%) and 24-96% (M = 62.41%) for non-match. The twenty most difficult trials (10 match, 10 non-match) were selected for inclusion in the final face-matching task. Ten trials (3 match, 7 non-match) featured female models, and ten (7 matches, 3 non-matches) featured males. The age range of the reference images included in the final test was 31-70 years and the mean age gap between the faces in each pair was 15.7 years (12.4 years for match trials and 17.6 years for mismatch trials). Accuracy rates on the selected trials ranged from 26-70% for match trials (M = 54%, SD = 14.54%) and 24-68% for non-match (M =

55.8%, SD = 13.18%). A one-sample t-test revealed that recognition accuracy for the selected stimuli set was above chance, $t(49) = 2.79$, $p < 0.01$, $d = 0.39$.

Procedure

Forensic Professionals

Participants completed 20 experimental trials (10 matches, 10 mismatches). In each trial, participants compared two images – an older reference image and a younger questioned image – and were asked to decide if the two images featured the same person or two different people.

Forensic professionals were provided with the test image files in jpeg format, at a size of 300 x 450 pixels. They were requested to complete the face-matching test using their own organisation's standard operating procedures, with the exception that their test responses should be made on an ENFSI-recommended 11-point scale ranging from +5 to -5 (see Figure 3). According to this scale, a response of +5 indicates that the results of the conducted examination provide extremely strong support that the same person is depicted in the questioned and the reference images. A score of -5 indicates that the results of the conducted examination provide extremely strong support that two different people are depicted in the questioned and reference images. A response of zero (0) can be used when the results of the examination are inconclusive, i.e., they do not support either hypothesis. For each point on the scale, participants are shown the relationship between the results under hypothesis that the images are the same person and the results under the hypothesis that the images are different people. Intervals for an equivalent numerical likelihood ratio connected to each verbal response are also provided.

Prior to completing the face-matching task, forensic professionals were requested to conduct an image quality assessment on each of the forty test images. If a participant felt that an image was of unsuitable quality for comparison, they were requested to leave that particular trial blank. Professionals were also asked to estimate the age of the person in each image in years. Responses were submitted via an online survey hosted on Qualtrics (https://www.qualtrics.com/).

Control Participants

Control participants completed a slightly modified version of the task. Stimuli were presented online using the Qualtrics survey platform. The two images in each pair were presented simultaneously on the screen, with each image sized at 300 x 450 pixels. Participants responded using a simplified version of the conclusion scale, which did not provide likelihood ratios. There was no time limit to respond, and participants were encouraged to take as long as necessary to make an accurate response. Control participants could not navigate back and forth between the image pairs. After completing the face-matching task, control participants were shown the 40 task images (20 reference, 20 questioned) again and asked to estimate the age of the person in each photograph in years.

Results

Face Matching Task

To examine face matching accuracy, we calculated the Area Under the Receiver Operating Characteristic Curve (AUC) for each participant. AUC provides a measure of how well a participant's answers predicted whether the images were a match or a non-match. This analysis

incorporates no support (0) decisions. AUC scores range from 0 to 1, with a score of 0.5 representing chance and a score of 1 representing perfect performance.

Due to the data violating the assumption of normality, between-group differences were assessed using the non-parametric Kruskal Wallis H test. The results revealed significant between-group differences in AUC scores, $\chi^2(2, 164) = 96.12$, $p < 0.001$, $\eta^2 = 0.59$. Pairwise comparisons were performed using Dunn's (1964) procedure with a Bonferroni correction for multiple comparisons. This analysis revealed that the examiners and teams both performed better than controls (both *p*s < 0.001), but there was no difference between examiners and teams ($p = 0.43$).

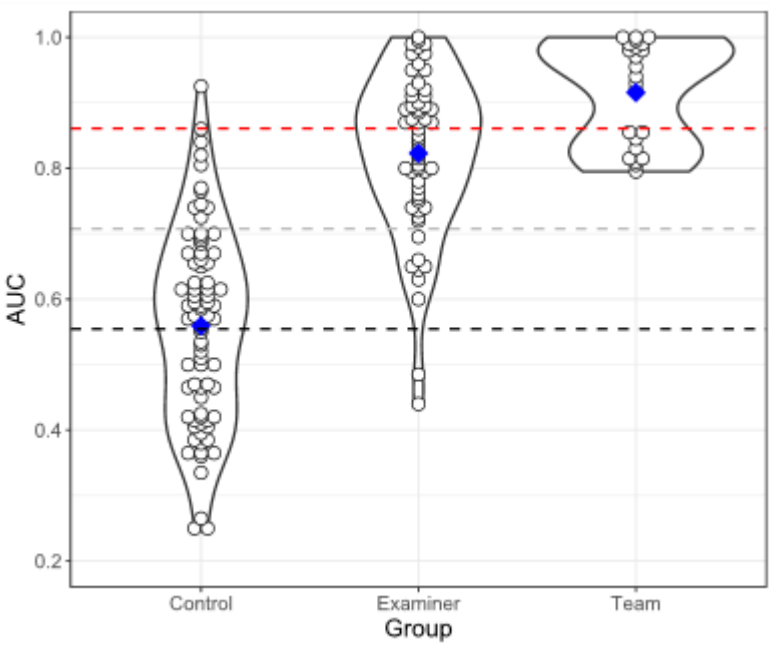

Figure 2. AUC scores for controls examiners and teams. The blue dot shows the mean group performance. The dashed lines show mean control performance, 1 SD above the control mean, and 2 SD above the control mean.

Crawford Howell single-case t-tests (Crawford et al., 2009) were used to determine how many of the individual participants in each group were statistically superior to the control group using

a two-tailed test at the 95% confidence level. Twenty-seven out of 60 individual examiners (45%) and 14 out of 23 teams (61%) were superior to controls.

In addition to examining diagnostic accuracy, we also assessed the percentage of high-confidence errors made by each group. We defined a high confidence error as a response of ≤ -4 for a match trial or ≥ +4 for a non-match trial. The mean percentage of high-confidence errors in each group is shown in Figure 3.

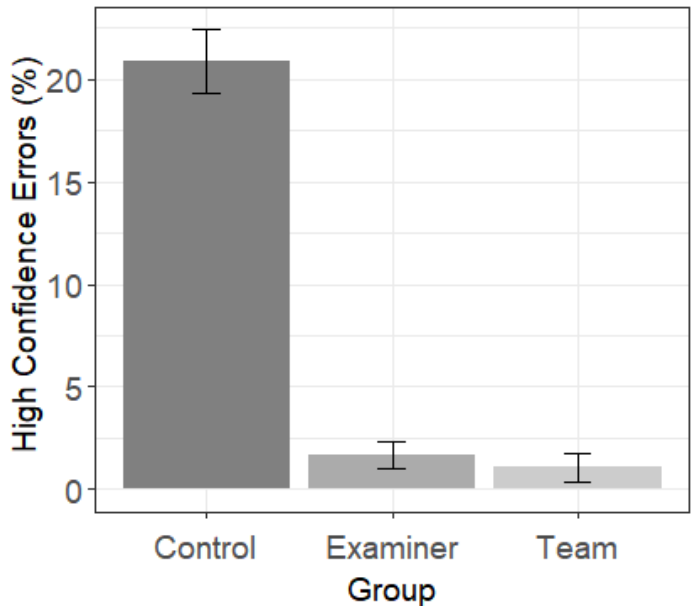

Figure 3. Mean percentage of high confidence errors. Error bars represent the standard error of the mean.

The results of a Kruskal Wallis H test revealed significant between-group differences in the percentage of high confidence errors, $\chi^2(2, 164) = 104$, $p < 0.001$, $\eta^2 = 0.63$. Pairwise comparisons with a Bonferroni correction revealed that the control group made significantly

more high-confidence errors than examiners and teams (all p values > 0.01). No other significant pairwise comparisons were observed.

We also ranked the trials that each group found the most challenging. This is visualised in Figure 4, which shows the percentage of incorrect responses for each trial, ranked from smallest to largest for each group. Figure 4 shows that there are differences in the trials that the professional and control groups found the most challenging. Interestingly, trial 15M was the most challenging trial for the examiners but was one of the least challenging for controls. Additionally, controls performed better than the individual examiners on trial 15M (26% incorrect versus 38% incorrect respectively). To quantify the level of association between the three participant groups, the trial rankings were examined using Spearman's rank order correlations. The results revealed a significant association between the two professional groups, in terms of which trials proved the most challenging, r (18) = 0.73, $p < 0.001$. However, there was no significant association between the trial rankings of the control group and either the individual examiners, r (18) = 0.12, $p = 0.62$, or teams, r (18) = 0.03, $p = 0.89$.

Team	% Incorrect
2M	0.00
8M	0.00
16M	0.00
9NM	0.00
5M	4.76
12M	4.76
20M	4.76
10NM	4.76
19M	9.52
4NM	9.52
13NM	9.52
6M	14.29
14M	14.29
15M	14.29
1NM	14.29
3NM	14.29
7NM	19.05
18NM	19.05
17NM	23.81
11NM	33.33

Examiner	% Incorrect
16M	1.67
2M	3. 33
10NM	5
9NM	8. 33
20M	11. 67
13NM	11. 67
19M	13. 33
8M	15
4NM	18. 33
18NM	18. 33
5M	20
1NM	20
7NM	20
6M	21. 67
3NM	25
17NM	28. 33
12M	31. 67
11NM	33. 33
14M	35
15M	38. 33

Control	% Incorrect
13NM	20.99
15M	25.93
6M	28.40
8M	30.86
20M	30.86
5M	32.10
16M	33.33
17NM	35.80
19M	38.27
18NM	38.27
9NM	39.51
4NM	41.98
7NM	45.68
11NM	46.91
10NM	48.15
1NM	50.62
12M	56.79
14M	58.02
2M	66.67
3NM	67.90

Team	% Incorrect
2M	0.00
8M	0.00
16M	0.00
9NM	0.00
5M	4.76
12M	4.76
20M	4.76
10NM	4.76
19M	9.52
4NM	9.52
13NM	9.52
6M	14.29
14M	14.29
15M	14.29
1NM	14.29
3NM	14.29
7NM	19.05
18NM	19.05
17NM	23.81
11NM	33.33

Figure 4. Each trial ranked by percentage incorrect for examiners, teams and controls.

Age Estimation Task

Ten participants (6 examiners and 4 teams) made their responses to the age estimation task in the form of an age range (e.g., 30-40 years). These participants were subsequently excluded from the age-determination analysis.

For each individual response, a deviation score was calculated, which represented the difference between the true chronological age of the model in the image and the age estimation made by the participant. For example, if the true chronological age of a face was 50 years old and a participant estimated that said face was 40 years old, then the deviation score for that example would be 10. For each participant, a mean deviation score was calculated, which represented the average difference between the participant's age estimations and the true age of the stimuli. Thus, a lower deviation score is indicative of a more accurate ability to estimate facial age. The mean deviation scores for each participant group are presented in Figure 5.

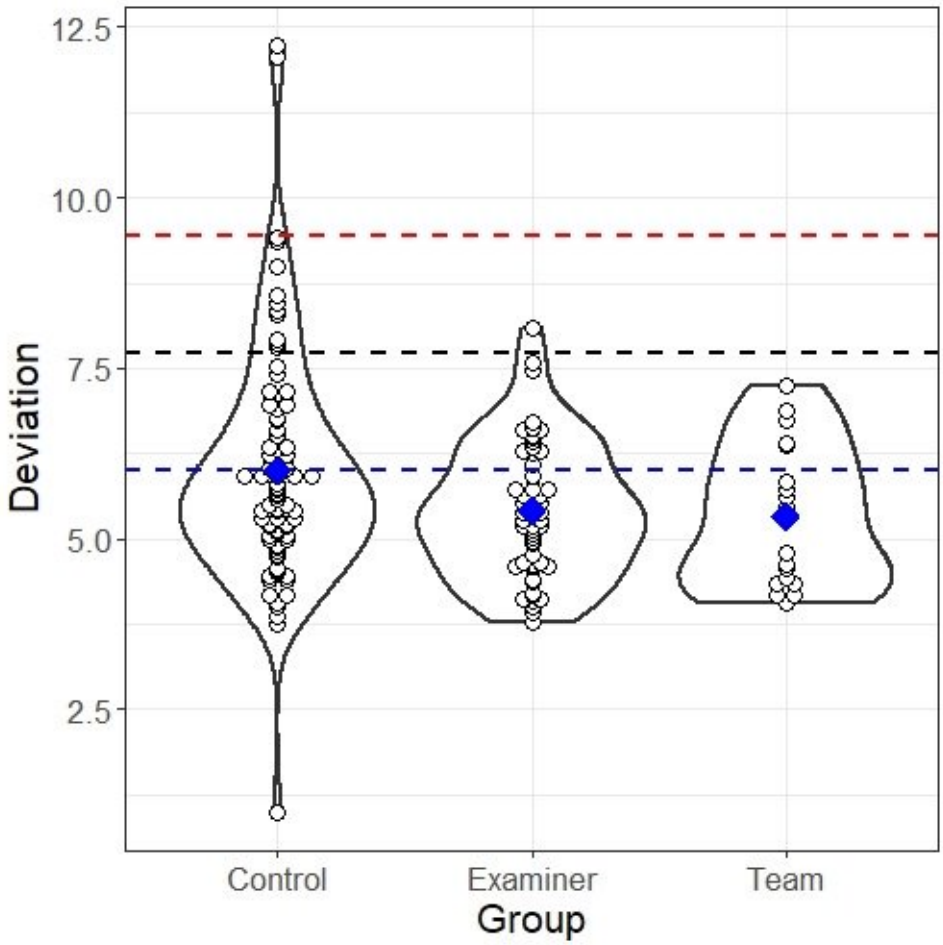

Figure 5. Mean performance on the age estimation task. Scores represent deviation from the stimulus age (in years). The blue dot shows the mean group performance. The dashed lines show mean control performance, 1 SD above the control mean, and 2 SD above the control mean.

Age estimation scores were not normally distributed, therefore, inter-group differences in age estimation ability were analysed using a Kruskal Wallis H test. The results revealed no significant differences in age estimation accuracy between participant groups $\chi^2(2, 154) = 5.1$, $p = 0.08$, $\eta^2 = 0.02$.

Discussion

This study assessed the ability of forensic examiners to match faces across a substantial change in age. The results revealed superior performance in individual examiners and teams relative to controls, lending support to numerous other demonstrations of increased examiner accuracy (e.g., Norell et al., 2015; White et al., 2015; Towler et al., 2017; Phillips et al., 2018; Claydon et al, 2023). In addition, examiners made fewer high confidence errors than controls, complementing findings from Phillips et al., (2018) who also found that forensic professionals tend not to make high confidence decisions (also see Norell et al., 2015). This is also in contrast to super-recognisers who tend to make more high confidence errors and use the response scale differently to forensic examiners (also see Towler et al., 2023 Hahn et al., 2022). One possibility is that examiners have a better understanding of how to use the response scale effectively, potentially because of the consequences associated with decisions in professional settings (White et al., 2021). An interesting direction for future research will be to determine how mock jurors perceive the use of the response scale. For example, do jurors comprehend the scale in the way in which they are intended by facial examiners when making their evaluation (see Eldridge, 2019 for a discussion).

Performance by control participants on the face-matching task was little better than chance, indicating that recognising the same person across a 10–30-year age gap using unconstrained images can be an extremely challenging task. This is considerably lower than Mileva et al (2020), who measured untrained participants' ability to match faces across a 20- and 40-year age gap. There are, however, some important differences between our study and Mileva et al. that could explain differences in accuracy. Importantly, our task was designed to be challenging and went through two stages of piloting to ensure that only the most difficult trials were included in the final test. The average age gap between the image pairs was 15.7 years, more than the 10-year validity period of a passport. Our findings suggest that

morphological analysis used by forensic examiners does confer some advantage, even for images with a substantial age difference.

We also found that there was poor agreement between the professional groups and controls in terms of which trials were the most challenging. For comparison in particular 15M the controls actually did better than the individual examiners (26% incorrect versus 38% incorrect). Although this is only one comparison, it indicates there are limitations to facial examiner capabilities for certain faces. It also suggests that untrained participants are likely making their decisions in a qualitatively different way to examiners, perhaps using more holistic perceptual processes in contrast to the systematic analysis and comparison of facial features by examiners (Towler et al., 2021; Moreton, 2021). It is worth noting that the face pairs used in the test were determined by pilot data from an untrained sample of participants and not professionals, and this could point to a potential limitation with how the test was constructed.

While at a group level we found superior performance for examiners and teams, there were individual differences within both professional groups, with some individuals/teams performing better than others. Similar findings have been observed in previous research. For example, Phillips et al (2018) found that some facial examiners performed below the mean of their student control group. An interesting direction for future research will be to determine the reason for these individual differences in forensic examiners. One possibility is that the matching test used in the present research does not reflect the nature of the day-to-day tasks completed by some examiners. For example, upon completion, anecdotal reports from some examiners revealed that the images were better quality than they were used to, whereas others reported that the images were lower quality, suggesting some heterogeneity in the types of face images that different examiners compare. Our results may, therefore, over/underestimate examiner performance.

Interestingly, neither individual examiners nor teams showed an advantage in age estimation relative to controls. This is despite the effects of facial ageing typically forming part of their training (Moreton et al., 2021). The results indicate that face matching and age estimation are distinct tasks and that expertise in matching does not extend to other aspects of facial decision making, possibly because they represent distinct abilities (also see Dagovitch & Ganel, 2010). Indeed, past research has found that manipulations that impair face recognition do not affect age estimation to the same extent (George & Hole, 2000). It also suggests that the advantage found on the matching task cannot wholly be attributed to an effect of motivation. There are mixed results for effects of motivation on face matching. Some studies have found that increasing motivation (e.g., via food or financial incentives) improves face matching (Moore & Johnston, 2013), particularly for other-race faces (Susa et al., 2016), however other studies have found limited or no effects of motivation (e.g., Bobak et al., 2016; Kemp et al., 1997). It has been argued that motivation is unlikely to explain the differences between facial examiners and controls (White et al., 2021). For example, examiners have been found to outperform other highly motivated control groups (White et al., 2015; Phillips et al., 2018). Our findings lend further support to this. If superior examiner performance was solely a result of increased motivation, then it would be reasonable to expect more accurate age estimation (i.e., increased motivation led to better performance across both matching and age estimation). There are some limitations of the approach used to analyse age estimation in the present research. Examiners were asked to estimate the age of each face in years. While examiners tend to be trained in the effects of ageing, it is possible that this would lead to an advantage in estimating the age gap between images (a task that requires comparison of images) rather than the absolute age of each individual face image. Future research should therefore determine whether examiners have an advantage for estimating the age gap between images, or whether

ageing can explain the difference between two images, rather than the absolute age of a of an individual face image.

Overall, the results of the present research show that matching faces that vary substantially in age is a challenging task. We found that facial examiners and teams were more accurate and made fewer high confidence errors than controls, but the advantage does not extend to age estimation. Morphological analysis may therefore provide some advantage for face matching, even for faces that differ in age. However, an important direction for future research will be to shed light on individual differences between facial comparison practitioners.

References

Barthold Jones, J. A. B., Nash, U. W., Vieillefont, J., Christensen, K., Misevic, D., & Steiner, U. K. (2019). The AgeGuess database, an open online resource on chronological and perceived ages of people aged 5–100. *Scientific Data*, *6*(1), 246. doi: https://doi.org/10.1038/s41597-019-0245-9

Bobak, A. K., Dowsett, A. J., & Bate, S. (2016). Solving the border control problem: Evidence of enhanced face matching in individuals with extraordinary face recognition skills. *PloS one*, *11*(2), e0148148. doi: https://doi.org/10.1371/journal.pone.0148148

Bruce, V., Henderson, Z., Greenwood, K., Hancock, P. J., Burton, A. M., & Miller, P. (1999). Verification of face identities from images captured on video. *Journal of Experimental Psychology: Applied*, *5*(4), 339 - 360. doi: https://doi.org/10.1037/1076-898X.5.4.339

Bruce, V., Henderson, Z., Newman, C., & Burton, A. M. (2001). Matching identities of familiar and unfamiliar faces caught on CCTV images. *Journal of Experimental Psychology: Applied*, *7*(3), 207. doi: 10.1037/1076-898X.7.3.207

Bruce, V., & Young, A. (1986). Understanding face recognition. *British journal of psychology*, *77*(3), 305-327. doi: https://doi.org/10.1111/j.2044-8295.1986.tb02199.x

Bruck, M., Cavanagh, P., & Ceci, S. J. (1991). Fortysomething: Recognizing faces at one's 25th reunion. *Memory & Cognition, 19*(3), 221–228. https://doi.org/10.3758/BF03211146.

Bruyer, R., Lafalize, A., & Distefano, M. (1991). Age decisions on familiar and unfamiliar faces. *Behavioural Processes*, *24*(1), 21-35. doi: 10.1016/0376-6357(91)90084-D

Burton, A. M. (2013). Why has research in face recognition progressed so slowly? The importance of variability. *Quarterly Journal of Experimental Psychology*, *66*(8), 1467-1485. doi: 10.1080/17470218.2013.800125

Burton, A. M., White, D., & McNeill, A. (2010). The Glasgow face matching test. *Behavior research methods*, *42*(1), 286-291. doi: https://doi.org/10.3758/BRM.42.1.286

Chatterjee G, & Nakayama K. (2012) Normal facial age and gender perception in developmental prosopagnosia. *Cognitive Neuropsychology, 29*(5-6), 482-502. doi: 10.1080/02643294.2012.756809. PMID: 23428082

Claydon, J. R., Fysh, M. C., Prunty, J. E., Cristino, F., Moreton, R., & Bindemann, M. (2023). Facial comparison behaviour of forensic facial examiners. *Applied Cognitive Psychology*, *37*(1), 6-25. doi: https://doi.org/10.1002/acp.4027

Clifford, C. W., Watson, T. L., & White, D. (2018). Two sources of bias explain errors in facial age estimation. *Royal Society Open Science*, *5*(10), 180841. doi: https://doi.org/10.1098/rsos.180841

Clutterbuck, R., & Johnston, R. A. (2002). Exploring levels of face familiarity by using an indirect face-matching measure. *Perception*, *31*(8), 985-994. doi: 10.1068/p3335

Clutterbuck, R., & Johnston, R. A. (2004). Matching as an index of face familiarity. *Visual Cognition*, *11*(7), 857-869. doi: https://doi.org/10.1080/13506280444000021

Crawford, J. R., Garthwaite, P. H., & Howell, D. C. (2009). On comparing a single case with a control sample: an alternative perspective. *Neuropsychologia*, *47*(13), 2690-2695.

Dagovitch, Y., & Ganel, T. (2010). Effects of Facial Identity on Age Judgments. *Experimental Psychology, 57 (5).* doi: 10.1016/j.neuropsychologia.2009.04.011

Davis, J. P., & Valentine, T. (2009). CCTV on trial: Matching video images with the defendant in the dock. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, *23*(4), 482-505.

Dehon, H., & Brédart, S. (2001). An 'other-race' effect in age estimation from faces. *Perception*, *30*(9), 1107-1113.

Djordjevic, J., Zhurov, A. I., Richmond, S., & Visigen Consortium. (2016). Genetic and environmental contributions to facial morphological variation: a 3D population-based twin study. *PloS one*, *11*(9), e0162250.

Eldridge, H. (2019). Juror comprehension of forensic expert testimony: A literature review and gap analysis. *Forensic Science International: Synergy*, *1*, 24-34.

Enlow, D. (1982). Handbook of Facial Growth. Philadelphia: W. B. Saunders.

European Network of Forensic Science Institutes. (2018). ENFSI Best Practice Manual for Facial Image Comparison (Vol. 01). Retrieved May 18, 2023 from: https://enfsi.eu/wp-content/uploads/2017/06/ENFSI-BPM-DI-01.pdf

Fysh, M. C., & Bindemann, M. (2022). Molistic processing in facial image comparison. *Applied Cognitive Psychology*, *36*(4), 830-841.

Fysh, M. C., Stacchi, L., & Ramon, M. (2020). Differences between and within individuals, and subprocesses of face cognition: Implications for theory, research and personnel selection. *Royal Society Open Science*, *7*(9), 200233.

George, P. A., & Hole, G. J. (2000). The role of spatial and surface cues in the age-processing of unfamiliar faces. *Visual Cognition*, *7*(4), 485-509.

Hahn, C. A., Tang, L. L., Yates, A. N., & Phillips, P. J. (2022). Forensic facial examiners versus super-recognizers: Evaluating behavior beyond accuracy. *Applied Cognitive Psychology, 36*(6), 1209-1218.

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

Kemp, R., Towell, N., & Pike, G. (1997). When seeing should not be believing: Photographs, credit cards and fraud. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, *11*(3), 211-222.

Ko, A. C., Korn, B. S., & Kikkawa, D. O. (2017). The aging face. *Survey of Ophthalmology*, *62*(2), 190-202.

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

Megreya, A. M., & Burton, A. M. (2007). Hits and false positives in face matching: A familiarity-based dissociation. *Perception & psychophysics*, *69*, 1175-1184.

Megreya, A. M., Sandford, A., & Burton, A. M. (2013). Matching face images taken on the same day or months apart: The limitations of photo ID. *Applied Cognitive Psychology*, *27*(6), 700-706.

Mendelson, B., & Wong, C. H. (2020). Changes in the facial skeleton with aging: implications and clinical applications in facial rejuvenation. *Aesthetic Plastic Surgery*, *44*(4), 1151-1158.

Michalski, D., Heyer, R., & Semmler, C. (2019). The performance of practitioners conducting facial comparisons on images of children across age. *Plos one*, *14*(11), e0225298.

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

Moore, R. M., & Johnston, R. A. (2013). Motivational incentives improve unfamiliar face matching accuracy. *Applied Cognitive Psychology*, *27*(6), 754-760.

Moreton, R. (2021). Forensic Face Matching. *Forensic face matching: Research and practice*, 144.

Nador, J. D., Vomland, M., Thielgen, M. M., & Ramon, M. (2022). Face recognition in police officers: Who fits the bill?. *Forensic Science International: Reports*, *5*, 100267 Noyes, E., & Jenkins, R. (2019). Deliberate disguise in face identification. *Journal of Experimental Psychology: Applied*, *25*(2), 280.

Norell, K., Läthén, K. B., Bergström, P., Rice, A., Natu, V., & O'Toole, A. (2015). The effect of image quality and forensic expertise in facial image comparisons. *Journal of Forensic Sciences*, *60*(2), 331-340.

Phillips, P. J., Yates, A. N., Hu, Y., Hahn, C. A., Noyes, E., Jackson, K., ... & O'Toole, A. J. (2018). Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Proceedings of the National Academy of Sciences*, *115*(24), 6171-6176.

Rexbye, H., Petersen, I., Johansens, M., Klitkou, L., Jeune, B., & Christensen, K. (2006). Influence of environmental factors on facial ageing. *Age and Ageing*, *35*(2), 110-115.

Sörqvist, P., & Eriksson, M. (2007). Effects of training on age estimation. *Applied Cognitive Psychology*, *21*(1), 131-135.

Stacchi, L., Huguenin-Elie, E., Caldara, R., & Ramon, M. (2020). Normative data for two challenging tests of face matching under ecological conditions. *Cognitive research: principles and implications*, *5*, 1-17.

Susa, K. J., Gause, C. A., & Dessenberger, S. J. (2019). Matching faces to ID photos: the influence of motivation on cross-race identification. *Applied Psychology in Criminal Justice*, *15*(1), 86-96.

Thorley, C. (2021). How old was he? Disguises, age, and race impact upon age estimation accuracy. *Applied Cognitive Psychology*, *35*(2), 460-472.

Towler, A., Dunn, J. D., Castro Martínez, S., Moreton, R., Eklöf, F., Ruifrok, A., Kemp, R. I. & White, D. (2023). Diverse types of expertise in facial recognition. *Scientific Reports, 13*(1), 11396.

Towler, A., Kemp, R. I., Burton, A. M., Dunn, J. D., Wayne, T., Moreton, R., & White, D. (2019). Do professional facial image comparison training courses work?. *PloS one*, *14*(2), e0211037.

Towler, A., Kemp, R. I., & White, D. (2021a). Can Face Identification Ability Be Trained?. *Forensic face matching: Research and practice*, 89.

Towler, A., Keshwa, M., Ton, B., Kemp, R. I., & White, D. (2021b). Diagnostic feature training improves face matching accuracy. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *47*(8), 1288.

Towler, A., White, D., & Kemp, R. I. (2017). Evaluating the feature comparison strategy for forensic face identification. *Journal of Experimental Psychology: Applied*, *23*(1), 47.

White, D., Phillips, P. J., Hahn, C. A., Hill, M., & O'Toole, A. J. (2015). Perceptual expertise in forensic facial image comparison. *Proceedings of the Royal Society B: Biological Sciences*, *282*(1814), 20151292.

White, D., Towler, A., & Kemp, R. (2021). Understanding professional expertise in unfamiliar face matching. *Forensic face matching: Research and practice*, 62-88.