Received: 23 June 2022 | Revised: 4 November 2022 | Accepted: 7 December 2022

DOI: 10.1002/acp.4027

RESEARCH ARTICLE

WILEY

Facial comparison behaviour of forensic facial examiners

Jacky R. Claydon[1] | Matthew C. Fysh[1] | Jonathan E. Prunty[1] | Filipe Cristino[2] | Reuben Moreton[3] | Markus Bindemann[1]

[1]Psychology, University of Kent, Canterbury, Kent, UK

[2]Psychology, Nottingham Trent University, Nottingham, Nottinghamshire, UK

[3]Psychology, The Open University, Milton Keynes, Buckinghamshire, UK

Correspondence
Markus Bindemann, School of Psychology, Keynes Colleges, University of Kent, Canterbury CT2 7NP, Kent, UK.
Email: m.bindemann@kent.ac.uk

Funding information
Graduate Teaching Assistantship; Economic and Social Research Council, Grant/Award Number: ES/S010181/1

Abstract

Facial examiners make visual comparisons of face images to establish the identities of persons in police investigations. This study utilised eye-tracking and an individual differences approach to investigate whether these experts exhibit specialist viewing behaviours during identification, by comparing facial examiners with forensic fingerprint analysts and untrained novices across three tasks. These comprised of face matching under unlimited (Experiment 1) and time-restricted viewing (Experiment 2), and with a feature-comparison protocol derived from examiner casework procedures (Experiment 3). Facial examiners exhibited individual differences in facial comparison accuracy and did not consistently outperform fingerprint analysts and novices. Their behaviour was also marked by similarities to the comparison groups in terms of how faces were viewed, as evidenced from eye movements, and how faces were perceived, based on the made feature judgements and identification decisions. These findings further understanding of how facial comparisons are performed and clarify the nature of examiner expertise.

KEYWORDS
expertise, eye movements, face matching, facial examiners, facial image comparison, fingerprint analysts

1 | INTRODUCTION

Images of faces provide a common form of evidence in criminal investigations to identify the perpetrators of crime. The appearance of a person committing a crime captured on CCTV surveillance footage, for example, may be compared to photographs of possible suspects in an effort to identify the perpetrator. In applied settings, this task is referred to as facial image comparison and involves the side-by-side examination of two or more photographs of faces unfamiliar to the observer, to determine which of these depict the same individual. This is a rigorous and structured process that is carried out by forensic facial examiners, who are specialists employed by the police, government departments and forensic service providers (Moreton et al., 2019). These examiners produce written reports to inform investigations and legal proceedings (see, e.g., Moreton, 2021; Norell et al., 2015; Phillips et al., 2018). In these high-stakes environments, facial comparison is therefore a critically important task, whereby an incorrect decision could have potentially life-changing consequences such as the incrimination and subsequent incarceration of an innocent member of society, or a failure to apprehend a criminal.

In the psychological study of person identification, facial comparison is referred to as unfamiliar face matching and has been researched extensively in recent years. Experiments in this domain typically involve showing observers pairs of photographs of unfamiliar faces, which must then be classified as the same person (i.e., are an identity match) or different people (an identity mismatch). A coherent body of research now demonstrates the difficulty of face matching (see Bindemann, 2021). Under viewing conditions that facilitate this task by presenting standardised high-quality face photographs for comparison, identification errors are made on average on one in five decisions (see, e.g., Bindemann et al., 2012; Burton et al., 2010; Özbek & Bindemann, 2011). Performance declines further when additional variables are brought into play that are routinely present in applied settings, such as images that capture faces across a greater time interval (e.g., Fysh & Bindemann, 2018; Megreya et al., 2013) and under more variable viewing conditions (e.g., Bindemann & Sandford, 2011; Dowsett et al., 2016; Ritchie & Burton, 2017) or are obscured by disguise (Kramer & Ritchie, 2016; Wirth & Carbon, 2017). Variation in capture conditions, such as image resolution (Bindemann et al., 2013) and the distance of a person from the camera (Noyes & Jenkins, 2017), also negatively impact on face-matching accuracy.

Much of this work has been carried out with untrained lay observers, such as student participants, who have no professional training or experience of face matching. However, a volume of studies examining the face-matching accuracy of facial examiners is now also accumulating (see White, 2021). This body of work demonstrates that these professionals consistently outperform untrained observers in face matching at the group level (Norell et al., 2015; Phillips et al., 2018; Towler et al., 2017; White, Dunn, et al., 2015; White, Phillips, et al., 2015). For example, facial examiners have been shown to outperform untrained student participants in 1-to-1 face-matching tests that provide optimised conditions for identification, such as the comparison of same-day photos of a person that were also obtained under similar lighting, and challenging conditions, such as images captured under diverse ambient conditions and over longer time periods (Phillips et al., 2018; White, Phillips, et al., 2015). Superior accuracy for facial examiners over untrained participants has also been observed in one-to-many facial comparisons, which required identification of a target from arrays of faces (White, Dunn, et al., 2015). Furthermore, the advantage for facial examiners has also been observed against other control groups, such as fingerprint examiners and members of the general public who are held to have a high natural ability to identify faces and are regarded as 'super-recognisers' (see Phillips et al., 2018).

While the advantage that facial examiners demonstrate in these identification tasks might be expected on account of their professional status, it is remarkable that such superiority is not observed with other professional groups who also perform facial comparisons regularly.[1] Passport issuance officers (White et al., 2014; White, Dunn, et al., 2015), border control and police officers (Towler et al., 2019; Wirth & Carbon, 2017), bank tellers and notaries (Papesh, 2018), for example, perform at a comparable level to participants untrained in facial comparison, and are outperformed by facial examiners (Phillips et al., 2018; White, Dunn, et al., 2015). This indicates that facial examiners hold a unique performance advantage in face matching over a range of professional groups (for a review, see White et al., 2021).

The source of this expertise is not yet understood, but one element that sets facial examiners apart from other professional groups is the training that these specialists receive (see Heyer et al., 2011; Prince, 2012; Towler et al., 2019; Towler, Kemp, et al.,2021). Forensic facial examiners are trained intensively in morphological analysis, where faces are compared on a feature-by-feature basis, with comparisons taking hours or even days to complete (Steyn et al., 2018). This is in stark contrast to how untrained observers compare faces in a matter of only a few seconds (Bindemann et al., 2016; Fysh & Bindemann, 2017; Özbek & Bindemann, 2011; Wirth & Carbon, 2017), as well as the occupational demands of professionals such as passport control officers, who also must perform this task quickly. Training practices for facial examiners vary between different organisations, but commonly include lengthy mentorship, performance monitoring and feedback, to become proficient in morphological analysis (Moreton et al., 2021).

Evidence for the effectiveness of this training comes from the superior performance of forensic examiners compared to other observer groups, and also from data which suggest that examiners perform face-matching tasks in a qualitatively different way to other observers. For example, whereas untrained observers perform similarly to forensic examiners when only seconds are available to match faces, examiners perform substantially better when more time is available (White, Phillips, et al., 2015). In the same study, the performance of facial examiners was also relatively unimpaired in comparison to untrained observers when to-be-compared faces were turned upside down. A face inversion effect is thought to index the fast, holistic 'at-a-glance' perception that is typically applied to the processing of faces (Maurer et al., 2002). These findings therefore suggest that examiners extract feature-based identity information from faces via a slower and more systematic method than untrained observers, which would be consistent with the morphological training that these professionals receive. There is also evidence that examiners and novices rely on different sets of features when classifying faces. For example, when examiners and untrained observers are asked to rate the similarity of individual features in face pairs, the examiners exhibit greater sensitivity to diagnostic identity information from scars and blemishes (Towler et al., 2017).

These findings hint at a *qualitative* shift in how facial examiners match the identities of faces, compared to other observer groups. This appears to be related to an analytical process of feature comparison that forms the basis of their training. However, most of this evidence is indirect, reflecting outcome measures such as task accuracy and judgements of feature similarity. Consequently, it is not known how the differences between facial examiners and untrained observers manifest *during* the matching of faces. The aim of this study is to explore these differences with a method that enables the observation of facial examiners' viewing behaviour, by tracking their eye movements during the identity comparison of faces.

Eye-tracking has been utilised extensively to study a range of processes in the perception of faces, such as detection (e.g., Bindemann et al., 2010; Fletcher-Watson et al., 2008; Kelly et al., 2019), recognition (Arizpe et al., 2019; Henderson et al., 2005; Stacchi et al., 2019), and social interaction (Birmingham et al., 2008). This work indicates that the acquisition of information from faces is determined by how they are scanned, and that the viewing patterns employed to allocate attention to different face regions, as evidenced from the study of eye movements, are also likely to be an important component of the face-matching process. However, very limited research exists on eye-tracking and face matching (for some basic exceptions, see Bindemann et al., 2012; Bobak et al., 2017; Megreya et al., 2012; Özbek &

Bindemann, 2011) and to date, no investigations of the eye movement behaviours of facial examiners in comparison with other groups of observers have been conducted.

In this study, we report three experiments that provide such a comparison. We examine the face-matching accuracy of a group of five professional forensic facial examiners from a single organisation. Because of the small population size, we focus our analysis on the performance of individual facial examiners by adopting a method that allows for single-case comparisons against control groups. This approach makes good sense considering reports of substantial individual variation in face-matching accuracy between facial examiners (see Phillips et al., 2018), as well as among the wider population (see, e.g., Bindemann et al., 2012; Burton et al., 2010; Lander et al., 2018; McCaffery et al., 2018).

We contrast the accuracy of these facial examiners with two control groups. One of these comprises forensic fingerprint analysts from the same organisation. Similar to facial examiners, these professionals have received extensive specialist training and perform important visual comparisons on a daily basis, but these are focused on fingerprints rather than face images. The second group of controls is comprised of novice face matchers—student participants with no formal training or experience in forensic facial comparison or fingerprint analysis. The behaviour of these three types of observers was compared across three tasks, comprising self-paced (Experiment 1) and time-limited face matching (Experiment 2), and face matching with a feature-comparison protocol derived from examiner procedures utilised in casework (Experiment 3). These three tasks were completed successively in a single test session by the same participants and are reported here in the order in which they were administered.

Considering that facial examiners and fingerprint analysts typically perform intensive visual examinations of forensic material in occupational settings, we expected these professionals to perform longer inspections of the face pairs than novices, who normally make these identification decisions in a matter of seconds (see, e.g., Bindemann et al., 2016; Fysh & Bindemann, 2017; Özbek & Bindemann, 2011). In addition, considering the accuracy advantage that has been demonstrated for facial examiners over novices and fingerprint examiners (see Norell et al., 2015; Phillips et al., 2018; Towler et al., 2017; White, Phillips, et al., 2015), particularly when more viewing time is available (White, Dunn, et al., 2015), a similar superiority in face matching was also expected here.

2 | EXPERIMENT 1

In this experiment, forensic facial examiners, fingerprint analysts and novices were presented with pairs of faces, which had to be identified as depicting the same person (a match) or different people (a mismatch), while viewing time was unconstrained to encourage best-possible accuracy. Based on previous work, we expected an accuracy advantage in face matching for facial examiners (Norell et al., 2015; Phillips et al., 2018; Towler et al., 2017; White, Phillips, et al., 2015). The main aim of this experiment was to explore whether an accuracy advantage for facial examiners would be accompanied by distinct viewing behaviours during face matching, to determine whether these experts apply a fundamentally (i.e., a qualitatively) different approach to resolve this task, or view faces in a similar manner as untrained observers but for longer (i.e., a quantitative difference).

For this purpose, we examined several aspects of observers' eye movements. We first explored whether general viewing differences exist between facial examiners, fingerprint analysts and novices, by examining the direction of eye movements (scanning), the spatial exploration of face areas (foraging), and viewing sequences (dependencies). We then investigated whether facial examiners and the comparison groups allocate their attention differently to the main features of a face, comprising the eyes, nose, and mouth. These features have been the primary focus of eye-tracking research in psychology (e.g., Althoff & Cohen, 1999; Bindemann et al., 2009; Blais et al., 2008), and provide a comparison of these observer groups at the next level of detail. We then examined eye movements to faces in further detail, by dissecting each face in a stimulus pairing into 15 areas to reflect the wider set of interest regions that are employed by facial examiners during identification (see Moreton, 2021). We compared the extent to which each of these regions was fixated, to understand how facial examiners distribute their attention across faces during identity matching in comparison with fingerprint analysts and untrained observers.

3 | METHOD

3.1 | Participants

Five forensic facial examiners from a UK-based organisation took part in this experiment (mean age = 34.4 years, *SD* = 1.8, range 32–37 years, 1 male), with mean experience in facial comparison of 34.8 months (*SD* = 31.6, range 6–84 months). All five forensic facial examiners had received formal training in facial image comparison, including recommended topics in anatomy, image science and processing, and methods of comparison (see Moreton et al., 2021), and had either completed or were undergoing a 6-month period of workplace mentoring with a more senior examiner. In addition, eight fingerprint analysts from the same organisation (mean age = 41.3 years, *SD* = 7.1, range 32–50 years, 3 males) with mean experience in fingerprint comparison of 162 months (*SD* = 35.6, range 108–204 months), served as a forensically-trained control group. Both groups undertook the experiment in a quiet office or laboratory van in a basement car park at their usual place of work. A further control group of 30 university students (mean age = 21.7 years, *SD* = 7.9, range 18–54 years, 4 males), who were untrained and inexperienced in forensic facial comparison, participated as novices in return for course credit. All participants reported normal or corrected-to-normal eyesight and provided informed consent to take part. This research was approved by the University of Kent Ethics Committee (Ethics ID 20181534456681508).

3.2 | Stimuli

The stimuli in this experiment consisted of 20 challenging face pairs from the KFMT (Fysh & Bindemann, 2018), comprising 10 identity

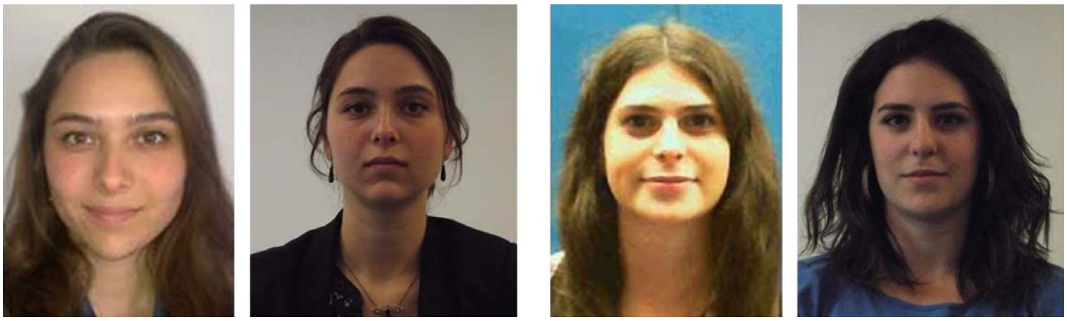


FIGURE 1 Examples of an identity match (left) and mismatch (right) face pair

matches and 10 mismatches.[2] For each face pair, the image on the right side of the screen comprised a controlled image of the target with a neutral expression, which had been taken against a plain background with even illumination. This was scaled to a size of 467 (W) × 550 (H) pixels. The image on the left side consisted of an unconstrained image taken from a student ID photograph and was re-scaled to a size of 406 (W) × 550 (H) pixels. Both images were presented onscreen at a resolution of 66 ppi and there was a minimum gap of 3 months between the capture of both images. In forensic facial examination casework, unknown or questioned images are typically taken in unconstrained conditions and are often poor quality. Compared to these unknown images, the known or reference images are typically constrained with uniform lighting, expression and image quality (Moreton, 2021). Unknown images can come from a wide range of sources, including identity documents, CCTV and social media. Therefore, the stimuli from the KFMT can be considered representative of typical forensic facial examination casework. Example match and mismatch pairs are shown in Figure 1.

3.3 | Procedure

The face stimuli were displayed using SR-Research Experiment Builder software (Version 1.1.0) on a 21-inch colour monitor connected to an EyeLink 1000 eye-tracking system running at 1000 Hz sample rate. The viewing distance was fixed at 60 cm with a chin rest. The participant's left eye was tracked although viewing was binocular. Prior to the experiment, the eye tracker was calibrated by participants fixating a nine-point sequence on the monitor, using the standard EyeLink calibration procedure. This was validated by successful fixation of a further nine targets. The procedure was repeated during the experiment if the participant changed their seating position or took a break.

At the beginning of each trial, participants fixated a dot in the centre of the display, which allowed for drift correction. A face pair was then presented until an identification response was made, with participants indicating whether a stimulus pairing depicted the same person or different people by pressing 'S' or 'D' on a standard computer keyboard. Match trials were interspersed with mismatch trials in a pre-randomised order, which was maintained for all participants to support comparison of individual differences. Accuracy was emphasised and viewing time was unrestricted.

4 | RESULTS

4.1 | Data preparation[3]

Participants' button responses were converted into percentage accuracy and response times for correct trials. For the analysis of the eye-tracking data, all eye movements were pre-processed by merging fixations of less than 80 ms with the preceding or following fixation if it fell within one degree of visual angle (for similar approaches, see e.g., Attard & Bindemann, 2014; Bindemann, 2010; Bindemann et al., 2009, 2010). In addition, any fixations that fell outside the dimensions of the display monitor or that were obscured by blinking were excluded.

For the analysis of eye movements to different face regions, each face was coded to define 15 different regions of interest (ROI). These ROIs were based on 12 main items (hair, face shape, forehead, brows, eyes, ears, nose, cheeks, mouth area, mouth, chin, and neck) from the feature list used in published guidelines for forensic facial image comparison (FISWG, 2018). In addition, the ears, eyebrows and eyes were coded separately to reflect those displayed on the left and right sides of the face. This created 15 ROIs for each face, and a total of 30 ROIs for each face pair. An illustration of these ROIs is provided in Figure 2.

4.2 | Response-based data

4.2.1 | Accuracy

To analyse face matching performance, the percentage accuracy scores were computed for each of the facial examiners, fingerprint analysts and novices. These data are illustrated in Figure 3. The accuracy of each facial examiner was then compared with the groups of fingerprint analysts and novices via a series of modified *t* tests (two-tailed) for single case comparisons (Crawford et al., 2010). These data are summarised in Table 1 and show a numerical advantage for all facial examiners over both control groups. However, this advantage

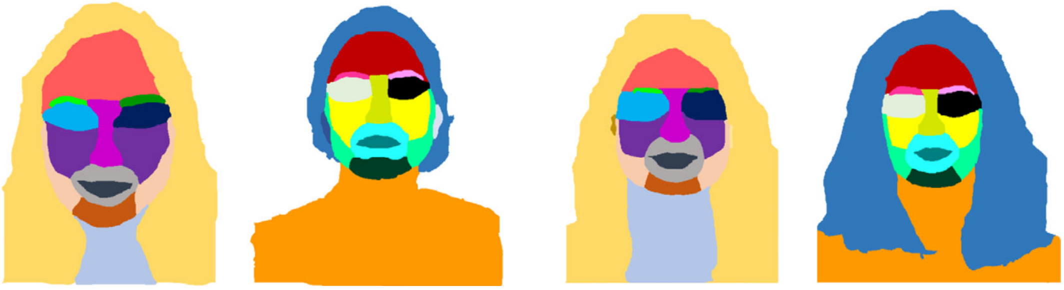


FIGURE 2 Illustration of the colour-coded regions of interest (ROI) for the example face pairs depicted in Figure 1

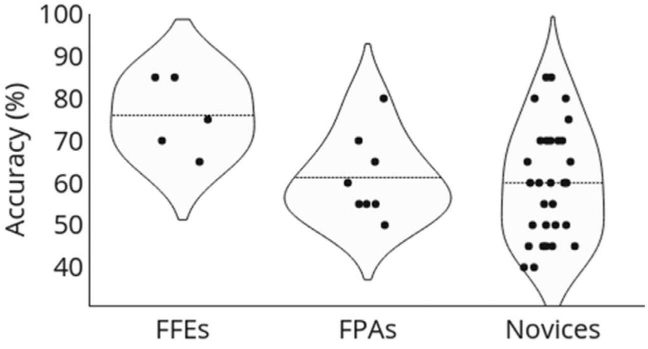


FIGURE 3 Face matching accuracy of individual facial examiners, fingerprint analysts and novices. Violin plots indicate distribution of scores and the horizontal line the group mean

was only significant for FFE1 and FFE2, and only in comparison with the fingerprint analyst (FPA) group, but not the novices.

4.2.2 | *d*′ and criterion

Accuracy was also converted into signal detection measures of sensitivity and bias (*d*′ and *criterion*) using the loglinear method to overcome extreme hit and false alarm rates (Hautus, 1995; see also, Stanislaw & Todorov, 1999), which were subsequently compared for each facial examiner with the groups of fingerprint analysts and novices via a series of modified *t* tests (two-tailed) for single case comparisons (Crawford et al., 2010). Similarly to the accuracy data, this reveals a significant advantage in sensitivity in FFE1 and FFE2, which was present relative to both the fingerprint analysts and the novices groups (see Table 1). There were no significant differences in criterion between individual facial examiners in comparison with the fingerprint analysts and novices.

4.2.3 | Response times

Viewing time was unrestricted in this experiment and the mean response time was therefore also analysed for correct matching decisions. In contrast to the accuracy data, a more consistent difference between individual facial examiners and the novice group was found, whereby four of the five facial examiners viewed faces for substantially longer than novices (FFE1, FFE2, FFE4, FFE5; see Table 1). However, average viewing times for fingerprint analysts were also long and comparable to those of the facial examiners. In combination with the accuracy and signal detection data, this implies that differences in the accuracy of individual facial examiners (specifically, FFE1 and FFE2) cannot be explained simply by the speed with which decisions were made.

4.3 | Eye-tracking data[4]

4.3.1 | Scanning

In a first step of the eye movement analysis, participants' general scanning behaviour of face pairs was investigated. Example scanpaths for each of the five facial examiners and a matching set of individuals from both comparison groups are shown in Figure 4.[5] These scanpaths reveal similarities across all observers, whereby face viewing is marked by horizontal saccades that traverse between both faces in a pair, and gradually shift vertically down the faces. To analyse these fixations, we performed correlations of the vertical fixation coordinates, in the sequence that these fixations were made during face viewing, with the vertical image coordinates. Thus, if face pairings were scanned systematically from top to bottom, then correlations should emerge between the vertical fixation and image coordinates. This analysis revealed positive correlations for each of the facial examiners (FFE1: $r = .711$; FFE2: $r = .738$; FFE3: $r = .624$; FFE4: $r = .404$; FFE5: $r = .642$), as well as the group of fingerprint analysts (mean $r = .618$, min $r = .472$, max $r = .668$), and novices (mean $r = .339$, min $r = .023$, max $r = .650$). This indicates that the facial examiners scanned faces similarly to the two comparison groups, following a pattern that systematically scanned faces from top-to-bottom. However, a comparison of these correlations using modified *t* tests (two-tailed) for single case comparisons (Crawford et al., 2010) showed that these vertical viewing patterns were more pronounced in FFE1, FFE2, FFE3 and FFE5 than the novices (see Table 2: Scanning).

4.3.2 | Foraging

The example scanpaths suggest that facial examiners and fingerprint analysts made more eye movements and fixations during face

TABLE 1 Individual case analyses comparing the performance of individual facial examiners against the group (mean) performance of the fingerprint analysts and novices in Experiment 1 in terms of accuracy, *d*′, criterion and response times. Parentheses show standard deviation of the means.

	Group mean	FFE1	FFE2	FFE3	FFE4	FFE5
Accuracy (%)	76.0 (8.9)	85.0	85.0	65.0	75.0	70.0
FPAs	61.0 (9.9)	***t* = 2.29, *p* = .05**	***t* = 2.29, *p* = .05**	*t* = 0.38, *p* = .72	*t* = 1.33, *p* = .22	*t* = 0.86, *p* = .42
Novices	60.0 (13.4)	*t* = 1.84, *p* = .07	*t* = 1.84, *p* = .07	*t* = 0.37, *p* = .72	*t* = 1.10, *p* = .28	*t* = 0.73, *p* = .47
d′	1.53 (0.50)	1.84	2.16	0.87	1.33	1.46
FPAs	0.57 (0.53)	***t* = 2.26, *p* = .03**	***t* = 2.83, *p* = .01**	*t* = 0.53, *p* = .31	*t* = 1.35, *p* = .11	*t* = 1.58, *p* = .08
Novices	0.56 (0.71)	***t* = 1.77, *p* = .04**	***t* = 2.22, *p* = .02**	*t* = 0.43, *p* = .34	*t* = 1.07, *p* = .15	*t* = 1.25, *p* = .11
Criterion	0.50 (0.42)	−0.17	0.61	0.66	0.43	0.96
FPAs	0.35 (0.63)	*t* = 0.27, *p* = .40	*t* = 0.39, *p* = .35	*t* = 0.46, *p* = .33	*t* = 0.12, *p* = .45	*t* = 0.91, *p* = .20
Novices	0.27 (0.53)	*t* = 0.19, *p* = .43	*t* = 0.63, *p* = .27	*t* = 0.73, *p* = .24	*t* = 0.30, *p* = .38	*t* = 1.28, *p* = .11
Response times (ms)	27,431 (11,507)	28,925	32,747	7605	37,213	30,665
FPAs	25,329 (12,405)	*t* = 0.27, *p* = .79	*t* = 0.56, *p* = .59	*t* = 1.35, *p* = .22	*t* = 0.90, *p* = .40	*t* = 0.41, *p* = .70
Novices	7269 (5088)	***t* = 4.19, *p* < .001**	***t* = 4.93, *p* < .001**	*t* = 0.07, *p* = .95	***t* = 5.79, *p* < .001**	***t* = 4.52, *p* < .001**

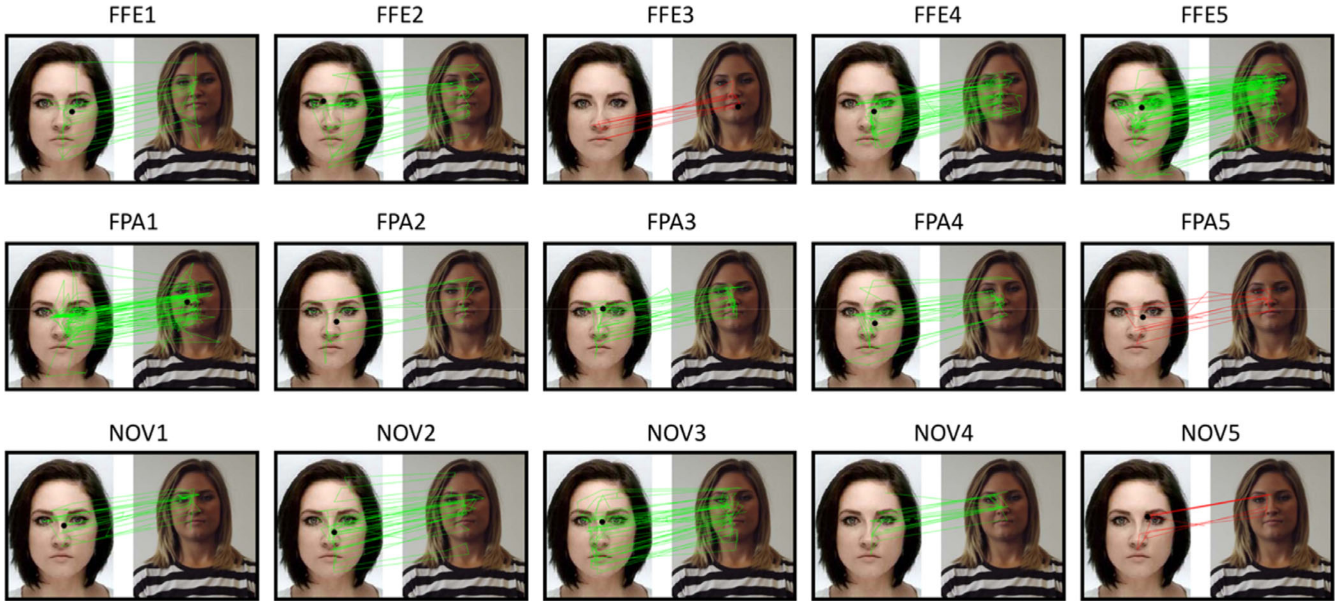


FIGURE 4 Example scan paths of facial examiners (top row) for FFE1 to FFE5, and for five fingerprint analysts (row 2) and novices (row 3) in Experiment 1. Green scan paths indicate correctly classified face pairs, red scan paths indicate identification errors, black dots indicate the first fixation on any of the ROIs

matching compared to the novices. To quantify these differences, we examined whether forensic facial examiners search faces more exhaustively for visual information than fingerprint analysts and novices, by calculating the percentage of ROIs that were fixated at least once on any trial. The single-case analysis of these data shows that none of the facial examiners viewed significantly more of the available features than the fingerprint analysts, but FFE1, FFE2 and FFE4 viewed substantially more of the available features than novices (see Table 2: Foraging). This did not, however, translate to higher accuracy in these FFEs compared to Novices (see Table 1: Accuracy). As such, there does not appear to be a consistent link between the foraging data and the accuracy data in facial examiners.

4.3.3 | Dependencies

We also measured whether the viewing of a face region depends on what has been viewed with the immediately preceding fixation, to investigate whether facial examiners might examine a fixated face region more thoroughly than the comparison groups. For this purpose, we calculated Markov probabilities to compare the conditional probabilities that observers view the same region within the same face on successive fixations compared to the same region in the other face in a pair, and the probabilities that a different face region is viewed next, both in the same and the other face. These probabilities show that FFE1, FFE2 and FFE4 were more likely than novices to fixate the

TABLE 2 Individual case analyses comparing the viewing behaviour of each facial examiner against the performance of the fingerprint analysts and novices in Experiment 1. Parentheses show standard deviation of the means.

Scanning	**Group mean**	**FFE1**	**FFE2**	**FFE3**	**FFE4**	**FFE5**
Correlations	0.624 (0.132)	0.711	0.738	0.624	0.404	0.642
FPAs	0.618 (0.063)	*t* = 1.39, *p* = .10	*t* = 1.80, *p* = .06	*t* = 0.90, *p* = .47	***t* = 3.20, *p* = .01**	*t* = 0.36, *p* = .37
Novices	0.339 (0.157)	***t* = 2.33, *p* = .01**	***t* = 2.50, *p* = .01**	***t* = 1.79, *p* = .04**	*t* = 0.41, *p* = .34	***t* = 1.90, *p* = .03**
Foraging	**Group mean**	**FFE1**	**FFE2**	**FFE3**	**FFE4**	**FFE5**
Features scanned (%)	56.4 (8.9)	63.6	60.4	41.1	58.9	53.0
FPAs	56.4 (10.2)	*t* = 0.67, *p* = .53	*t* = 0.37, *p* = .36	*t* = 1.42, *p* = .20	*t* = 0.23, p = .83	*t* = 0.32, p = .76
Novices	31.7 (12.5)	***t* = 2.50, *p* = .02**	***t* = 2.25, *p* = .02**	*t* = 0.73, *p* = .47	***t* = 2.13, p = .04**	*t* = 1.67, p = .11
Dependencies	**Group mean**	**FFE1**	**FFE2**	**FFE3**	**FFE4**	**FFE5**
Same area, same face	0.222 (0.047)	0.278	0.247	0.202	0.226	0.155
FPAs	0.236 (0.052)	*t* = 0.75, *p* = .24	*t* = 0.20, *p* = .42	*t* = 0.62, *p* = .28	*t* = 0.18, *p* = .43	*t* = 1.47, *p* = .09
Novices	0.135 (0.041)	***t* = 3.43, *p* = .001**	***t* = 2.69, *p* = .01**	*t* = 1.61, *p* = .06	***t* = 2.18, *p* = .02**	*t* = 0.48, *p* = .32
Same area, diff. face	0.162 (0.021)	0.150	0.150	0.152	0.162	0.200
FPAs	0.137 (0.033)	*t* = 0.37, *p* = .36	*t* = 0.37, *p* = .36	*t* = 0.43, *p* = .34	*t* = 0.71, *p* = .25	*t* = 1.80, *p* = .06
Novices	0.163 (0.074)	*t* = 0.17, *p* = .43	*t* = 0.17, *p* = .43	*t* = 0.15, *p* = .44	*t* = 0.13, *p* = .50	*t* = 0.49, *p* = .32
Diff. area, same face	0.234 (0.027)	0.249	0.228	0.265	0.235	0.194
FPAs	0.239 (0.037)	*t* = 0.26, *p* = .40	*t* = 0.28, *p* = .39	*t* = 0.66, *p* = .26	*t* = 0.10, *p* = .46	*t* = 1.50, *p* = .15
Novices	0.228 (0.069)	*t* = 0.30, *p* = .39	*t* = 0.00, *p* = .50	*t* = 0.53, *p* = .30	*t* = 0.10, *p* = .46	*t* = 0.49, *p* = .32
Diff. area, diff. face	0.381 (0.046)	0.322	0.376	0.381	0.377	0.450
FPAs	0.389 (0.027)	***t* = 2.34, *p* = .03**	*t* = 0.45, *p* = .33	*t* = 0.28, *p* = .39	*t* = 0.42, *p* = .34	***t* = 2.13, *p* = .04**
Novices	0.474 (0.066)	***t* = 2.27, *p* = .02**	*t* = 1.46, *p* = .08	*t* = 1.39, *p* = .08	*t* = 1.45, *p* = .08	*t* = 0.36, *p* = .36
Key features	**Group mean**	**FFE1**	**FFE2**	**FFE3**	**FFE4**	**FFE5**
Eyes fixations (%)	30.0 (6.4)	25.4	38.8	26.0	24.9	34.9
FPAs	26.5 (2.9)	*t* = 0.35, *p* = .73	***t* = 2.89, *p* = .01**	*t* = 0.16, *p* = .87	*t* = 0.52, *p* = .62	***t* = 2.71, *p* = .04**
Novices	33.0 (13.2)	*t* = 0.56, *p* = .58	*t* = 0.43, *p* = .34	*t* = 0.52, *p* = .61	*t* = 0.60, *p* = .55	*t* = 0.14, *p* = .88
Nose fixations (%)	22.6 (3.4)	24.0	21.8	17.2	26.1	23.8
FPAs	26.4 (4.9)	*t* = 0.46, *p* = .66	*t* = 0.39, *p* = .35	*t* = 1.76, *p* = .13	*t* = 0.06, *p* = .96	*t* = 0.50, *p* = .64
Novices	31.9 (4.9)	*t* = 1.59, *p* = .12	*t* = 1.27, *p* = .11	***t* = 2.95, *p* = .01**	*t* = 1.16, *p* = .25	*t* = 1.63, *p* = .12
Mouth fixations (%)	19.1 (5.5)	20.3	12.5	27.5	19.0	16.4
FPAs	16.0 (2.9)	*t* = 1.39, *p* = .22	*t* = 0.65, *p* = .27	***t* = 3.71, *p* = .01**	*t* = 0.97, *p* = .37	*t* = 0.13, *p* = .90
Novices	14.3 (10.6)	*t* = 0.56, *p* = .58	*t* = 0.17, *p* = .43	*t* = 1.23, *p* = .23	*t* = 0.44, *p* = .66	*t* = 0.20, *p* = .85
All features		**FFE1**	**FFE2**	**FFE3**	**FFE4**	**FFE5**
FFE correlation	versus FPAs	***r* = .955, *p* < .001**	***r* = .878, *p* < .001**	***r* = .761, *p* < .001**	***r* = .920, *p* < .001**	***r* = .880, *p* < .001**
	versus Novices	***r* = .935, *p* < .001**	***r* = .852, *p* < .001**	***r* = .713, *p* < .001**	***r* = .927, *p* < .001**	***r* = .938, *p* < .001**

Note: All inferential statistics are based on two-tailed single-case comparisons of individual facial examiners against the group means.

same facial region on successive fixations, pointing to more systematic processing of information (see Table 2: Dependencies). FFE1 and FFE5 were also less likely to switch attention to a different region on the other face on successive fixations.

4.3.4 | Key features[6]

We examined whether facial examiners might differ from fingerprint analysts and novices in their usage of specific facial features. The eyes, nose and mouth regions typically receive the most attention during the viewing of faces. The percentage of eye fixations on these regions was therefore calculated for facial examiners and the comparison groups for correct trials (see Table 2: Key features). For this purpose, fixation data from the eyes and eyebrows were combined into a single score, as were data from the mouth and mouth region. These data show that FFE2 and FFE5 directed more fixations at the eye regions than the fingerprint analysts, whereas FFE3 directed fewer fixations at the nose than novices, but more fixations at the mouth than fingerprint analysts. None of the other comparisons between individual facial examiners and the comparison groups were significant. Overall, the fixation data of the facial examiners to the eyes,

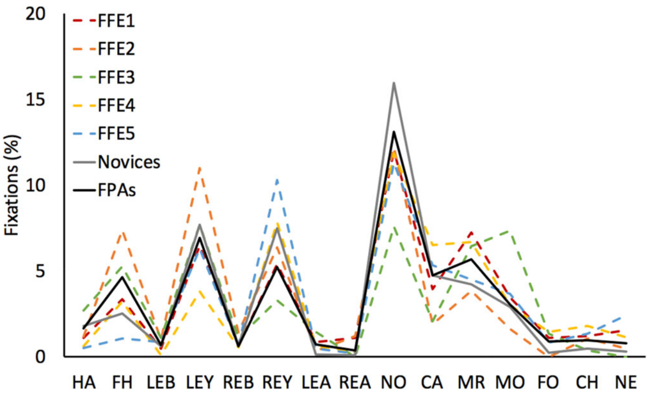


FIGURE 5 Comparison of mean percentage fixations to each ROI in Experiment 1, for each of the facial examiners (FFEs) and the fingerprint analyst and novice groups. HA = hair, FH = forehead, LEB = left eyebrow, LEY = left eye, REB = right eyebrow, REY = right eye, LEA = left ear, REA = right ear, NO = nose, CA = cheek area, MR = mouth region, MO = mouth, FO = face outline, CH = chin, NE = neck

nose and mouth regions therefore appear to be marked by similarities with fingerprint analysts and novices, rather than differences.

4.3.5 | All features

In a final step of the analysis, we examined similarities between individual examiner performance with the fingerprint analyst and novice groups across the full set of 30 ROIs. These data are shown in Figure 5, collapsed across the left and right face in each stimulus pair. We correlated these percentage fixations to assess the relationship between individual facial examiners with the group data for the fingerprint analysts and novices. This revealed strong positive correlations between each of the facial examiners and the fingerprint expert and novice groups (see Table 2: All features). Thus, attention to features followed a similar pattern in the examiners and the comparison groups, suggesting similar viewing of the faces.

5 | DISCUSSION

This experiment examined the face-matching performance of five facial examiners against a control group of fingerprint analysts and a group of novices, comprising of student observers, who were untrained and inexperienced in forensic facial comparison. All facial examiners exhibited a numerical face-matching advantage over the control groups. This accuracy advantage for facial examiners converges with previous research findings of superior performance by these types of experts (e.g., Norell et al., 2015; Towler et al., 2017; White, Dunn, et al., 2015; White, Phillips, et al., 2015). However, previous research has only examined whether this advantage is robust at a group level. The current study focused on individual performance and showed that this advantage was only reliable in two of the five examiners (FFE1 and FFE2) in sensitivity (*d′*), and in terms of percentage accuracy only in comparison with the fingerprint analysts. Notably, the two best-performing facial examiners also made several identification errors. This finding also aligns with previous studies, which have observed a range in face-matching ability among facial examiners (Phillips et al., 2018).

The question of main interest was whether these facial examiners matched the identities of faces with different viewing strategies than observers who do not share their professional expertise. Compared to novices, the facial examiners took substantially longer to make identification decisions, indicating that these observers studied the face pairs in greater depth. However, fingerprint analysts also displayed similarly long response times to facial examiners, but without an increase in face-matching accuracy over the forensically untrained novices (mean overall accuracy for fingerprint analysts 61% vs. 60% for novices). This indicates that response times are not a defining difference of the expertise of facial examiners, nor can this explain the identification advantage of FFE1 and FFE2 here considering FFE4 and FFE5 exhibited similarly long response times but not a reliable advantage in face identification.

To investigate the basis of the facial examiners' advantage in more depth, we analysed the eye movement data of all observers. This reveals some differences between the facial examiners and the comparison groups. For example, some of the examiners studied faces more exhaustively (foraging; FFE1, FFE2, FFE4), more systematically (scanning; FFE1, FFE2, FFE3, FFE5) and in more depth (dependencies; FFE1, FFE2, FFE4) than novices. Some examiners also fixated the eye (FFE2, FFE5) or mouth (FFE3) regions more frequently than fingerprint analysts. Overall, however, these differences were largely inconsistent across facial examiners and do not appear to map onto an accuracy advantage at the individual level in a straightforward manner. This indicates that these behaviours do not reflect a shared, systematic viewing strategy in facial examiners. Moreover, the data were generally marked by similarities rather than differences in how this task appeared to be solved by facial examiners and the comparison groups. For example, all facial examiners appeared to view faces similarly to the two comparison groups, following a pattern that systematically scanned faces from top to bottom. And the pattern of fixations to the full set of face regions under investigation here correlated consistently and strongly for each of the facial examiners with both control groups. Generally, the eye movement data showed little evidence of systematic differences in face viewing that can explain the accuracy advantage of facial examiners, and suggest that the novices and fingerprint analysts exercised comparable viewing behaviours to the facial examiners during face matching.

6 | EXPERIMENT 2

Experiment 1 provides no systematic evidence that the face-matching accuracy of facial examiners is accounted for by systematic differences in viewing processes compared to fingerprint analysts and novices. However, the data were marked by substantial differences in viewing time between observer groups. As such differences were anticipated (see, e.g., Phillips et al., 2018), Experiment 2 examines accuracy under conditions that control viewing time during face matching across the observer groups. This experiment was conducted in the same testing session as Experiment 1, but participants undertook the face-matching task with new stimuli, whilst viewing time was equated to 30 s for each face pair before an identification decision was made. This exposure duration has previously been found to differentiate the accuracy of examiners and non-expert controls (White, Phillips, et al., 2015), and for this reason was applied to the face-matching tasks in Experiment 2. Similarly, response times for most FFEs in Experiment 1 were around 30 s and longer than those of FPAs (M = 25.3) and Novices (M = 7.3). Constraining viewing time in this manner therefore ensured that all groups had equal time to examine the face pairs. If the behavioural and eye movement findings are replicated under these more equal conditions, then this will increase confidence in the similarities and differences that were observed between facial examiners and the comparison groups in Experiment 1.

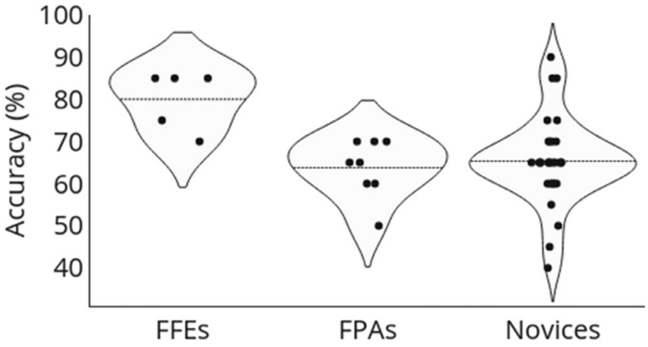


FIGURE 6 Face matching accuracy of individual facial examiners, fingerprint analysts and novices. Violin plots indicate distribution of scores and the horizontal line the group mean

7 | METHOD

7.1 | Participants, stimuli and procedure

All participants from Experiment 1 took part in Experiment 2 on the same day, following a short break. The stimuli consisted of 20 new pairs of faces from the KFMT, with equal numbers of identity matches and mismatches. The procedure was identical to Experiment 1 except that stimulus presentation was limited to 30 s. The face pairs were then removed from view and participants were prompted to make a same- or different-identity matching decision.

8 | RESULTS

8.1 | Response-based data

8.1.1 | Accuracy

To analyse face matching performance, the percentage accuracy scores were computed for each of the facial examiners, fingerprint analysts and novices. These data are illustrated in Figure 6. The accuracy of each facial examiner was then compared with the group means for the fingerprint analysts and novices via a series of modified t tests for single case comparisons (Crawford et al., 2010). These data are summarised in Table 3 and again show a numerical advantage for all facial examiners over both control groups. However, this advantage was only significant for FFE1, FFE2 and FFE4, and only in comparison with the fingerprint analyst (FPA) group but not the novices.

8.1.2 | *d′* and criterion

Similar to the accuracy data, analysis of sensitivity (d') revealed a significant advantage in FFE1, FFE2 and FFE4 in comparisons with the

TABLE 3 Individual case analyses comparing the performance of individual facial examiners against the group (mean) performance of the fingerprint analysts and novices in Experiment 2 in terms of accuracy, d', criterion and response times. Parentheses show standard deviation of the means.

	Group mean	FFE1	FFE2	FFE3	FFE4	FFE5
Accuracy (%)	80.0 (7.1)	85.0	85.0	70.0	85.0	75.0
FPAs	63.8 (6.9)	**t = 2.87, p = .02**	**t = 2.87, p = .02**	t = 0.82, p = .44	**t = 2.87, p = .02**	t = 1.50, p = .18
Novices	65.3 (10.5)	t = 1.88, p = .07	t = 1.88, p = .07	t = 0.47, p = .64	t = 1.88, p = .07	t = 0.94, p = .36
d'	1.57 (0.40)	1.84	1.84	0.98	1.84	1.33
FPAs	0.72 (0.34)	**t = 3.11, p = .009**	**t = 3.11, p = .009**	t = 0.72, p = .25	**t = 3.11, p = .009**	t = 1.69, p = .07
Novices	0.81 (0.62)	t = 1.63, p = .06	t = 1.63, p = .06	t = 0.27, p = .40	t = 1.63, p = .06	t = 0.83, p = .21
Criterion	0.17 (0.22)	−0.17	0.17	0.26	0.17	0.43
FPAs	0.21 (0.49)	t = 0.07, p = .47	t = 0.07, p = .47	t = 0.10, p = .46	t = 0.07, p = .47	t = 0.42, p = .34
Novices	0.10 (0.45)	t = 0.15, p = .44	t = 0.15, p = .44	t = 0.35, p = .37	t = 0.15, p = .44	t = 0.72, p = .24
Response times (ms)	31,126 (301)	31,606	31,074	31,173	30,979	30,800
FPAs	31,081 (518)	t = 0.96, p = .37	t = 0.01, p = .99	t = 0.17, p = .87	t = 0.19, p = .86	t = 0.51, p = .63
Novices	30,995 (327)	t = 1.84, p = .08	t = 0.24, p = .81	t = 0.54, p = .60	t = 0.05, p = .96	t = 0.59, p = .56

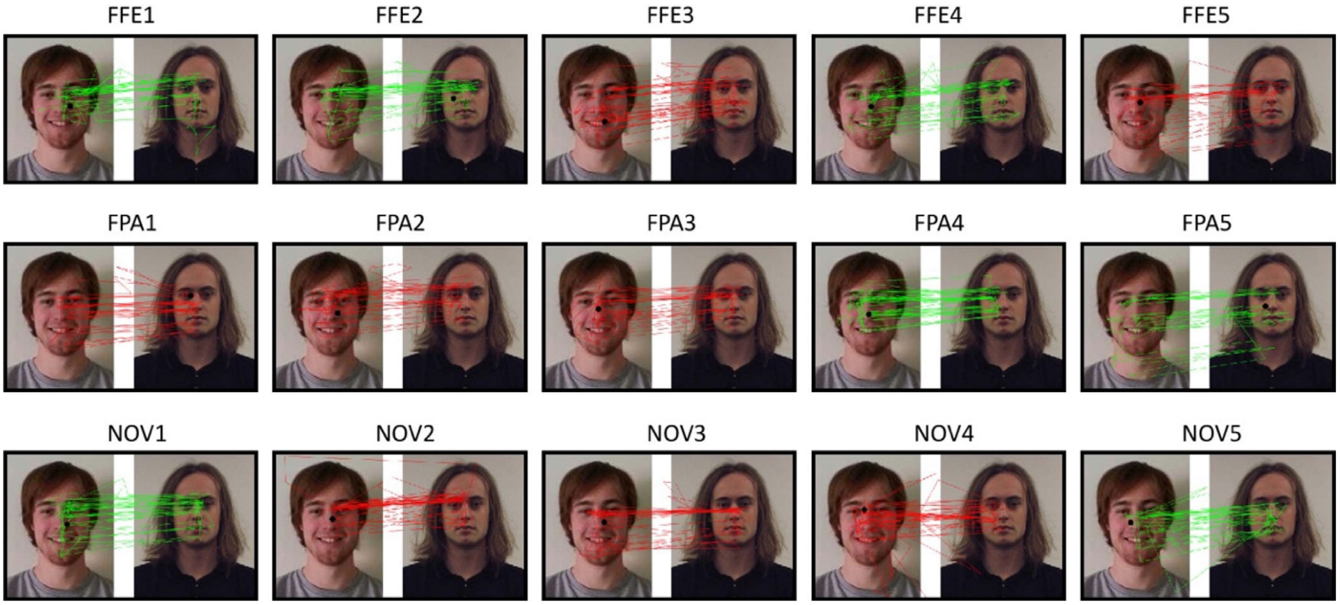


FIGURE 7 Example scan paths of facial examiners (top row) for FFE1 to FFE5, and for five fingerprint analysts (row 2) and novices (row 3) in Experiment 2. Green scan paths indicate correctly classified face pairs, red scan paths indicate identification error, black dots indicate the first fixation on any of the ROIs

fingerprint analysts only (see Table 3). There were no significant differences in criterion between individual facial examiners in comparison with the fingerprint analysts and novices.

8.1.3 | Response times

Although viewing time was controlled in this experiment, the mean response time for correct trials was analysed for completeness. These data revealed no difference in the average response times of facial examiners compared with the fingerprint analysts and novices (see Table 3).

8.2 | Eye-tracking data

8.2.1 | Scanning

Once again, participants' general scanning behaviour of face pairs was marked by eye movements that systematically traversed from the top to bottom of the face pairs (see Figure 7). Correlations of the fixation coordinates and the vertical image coordinates were observed for all of the facial examiners (FFE1: $r = .753$; FFE2: $r = .741$; FFE3: $r = .668$; FFE4: $r = .749$; FFE5: $r = .781$), as well as the group of fingerprint analysts (mean $r = .738$, min $r = .618$, max $r = .794$), and novices (mean $r = .573$, min $r = .365$, max $r = .704$). A comparison of these correlations showed that these vertical viewing patterns were more pronounced in FFE1, FFE2, FFE4 and FFE5 in comparison with the novices, but not compared to the fingerprint analysts (see Table 4: Scanning).

8.2.2 | Foraging

To examine whether forensic facial examiners searched faces more exhaustively for visual information than fingerprint analysts and novices, we calculated the percentage of ROIs that were fixated at least once on any trial. The single-case analysis of these data shows that none of the facial examiners viewed significantly more of the available features than the fingerprint analysts or novices (see Table 4: Foraging).

8.2.3 | Dependencies

We also measured Markov probabilities to examine whether facial examiners differ in whether they view the same region within the same face on successive fixations compared to the same region in the other face in a pair, or a different region. These probabilities show that FFE1 and FFE4 were more likely than novices to fixate the same facial region of a face on successive fixations, whereas FFE3 and FFE5 were more likely to switch to the same area in the other face in a pair than novices (see Table 4: Dependencies). In addition, FFE4 and FFE5 were less likely to view a different area in the same face, and FFE1, FFE3 and FFE4 were less likely to view a different area in a different face on successive fixations than novices. Overall, while these dependencies show a somewhat complex pattern, there is evidence that facial examiners were significantly *more* likely to view the *same* face area on successive fixations (FFE1 and FFE4 in the same face, FFE3 and FFE5 in the other face), and were *less* likely to view a *different* face area (FFE4 and FFE5 in the same face, FFE1, FFE3 and FFE4 in the other face). However, all of these effects were observed in comparison with the novice group, against which any differences in

TABLE 4 Individual case analyses comparing the viewing behaviour of each facial examiner against the performance of the fingerprint analysts and novices in Experiment 2. Parentheses show standard deviation of the means.

Scanning	**Group mean**	**FFE1**	**FFE2**	**FFE3**	**FFE4**	**FFE5**
Correlation (Pearson's *r*)	0.738 (0.042)	0.753	0.741	0.668	0.749	0.781
FPAs	0.738 (0.054)	*t* = 0.26, *p* = .40	*t* = 0.05, *p* = .48	*t* = 1.22, *p* = .13	*t* = 0.19, *p* = .43	*t* = 0.75, *p* = .24
Novices	0.573 (0.091)	***t* = 1.95, *p* = .03**	***t* = 1.82, *p* = .04**	*t* = 1.03, *p* = .16	***t* = 1.90, *p* = .03**	***t* = 2.25, *p* = .02**
Foraging	**Group mean**	**FFE1**	**FFE2**	**FFE3**	**FFE4**	**FFE5**
Features scanned (%)	67.9 (4.6)	68.8	63.0	72.6	63.3	71.7
FPAs	67.3 (4.2)	*t* = 0.33, *p* = .75	*t* = 0.97, *p* = .37	*t* = 1.19, *p* = .27	*t* = 0.90, *p* = .40	*t* = 0.99, *p* = .36
Novices	61.9 (5.6)	*t* = 1.21, *p* = .24	*t* = 0.19, *p* = .85	*t* = 1.88, *p* = .07	*t* = 0.25, *p* = .81	*t* = 1.72, *p* = .10
Dependencies	**Group mean**	**FFE1**	**FFE2**	**FFE3**	**FFE4**	**FFE5**
Same area, same face	0.243 (0.042)	0.277	0.230	0.234	0.290	0.183
FPAs	0.261 (0.065)	*t* = 0.23, *p* = .41	*t* = 0.45, *p* = .33	*t* = 0.39, *p* = .35	*t* = 0.42, *p* = .34	*t* = 1.13, *p* = .15
Novices	0.187 (0.036)	***t* = 2.46, *p* = .01**	*t* = 1.75, *p* = .13	*t* = 1.28, *p* = .10	***t* = 2.82, *p* = .01**	*t* = 0.11, *p* = .46
Same area, diff. face	0.151 (0.024)	0.131	0.138	0.153	0.140	0.191
FPAs	0.129 (0.030)	*t* = 0.60, *p* = .48	*t* = 0.28, *p* = .39	*t* = 0.75, *p* = .24	*t* = 0.35, *p* = .37	*t* = 1.95, *p* = .05
Novices	0.112 (0.023)	*t* = 0.81, *p* = .20	*t* = 1.12, *p* = .14	***t* = 1.75, *p* = .05**	*t* = 1.20, *p* = .12	***t* = 3.80, *p* = .001**
Diff. area, same face	0.233 (0.020)	0.260	0.237	0.240	0.215	0.211
FPAs	0.241 (0.033)	*t* = 0.54, *p* = .30	*t* = 0.11, *p* = .46	*t* = 0.30, *p* = .50	*t* = 0.74, *p* = .24	*t* = 0.86, *p* = .21
Novices	0.267 (0.028)	*t* = 0.25, *p* = .40	*t* = 1.10, *p* = .15	*t* = 0.95, *p* = .18	***t* = 1.83, *p* = .04**	***t* = 1.97, *p* = .03**
Diff. area, diff. face	0.374 (0.033)	0.331	0.394	0.373	0.356	0.415
FPAs	0.369 (0.038)	*t* = 0.94, *p* = .19	*t* = 0.62, *p* = .28	*t* = 0.10, *p* = .46	*t* = 0.32, *p* = .38	*t* = 1.41, *p* = .15
Novices	0.434 (0.035)	***t* = 3.90, *p* < .001**	*t* = 1.12, *p* = .14	***t* = 1.72, *p* = .05**	***t* = 2.20, *p* = .02**	*t* = 0.53, *p* = .30
Key features	**Group mean**	**FFE1**	**FFE2**	**FFE3**	**FFE4**	**FFE5**
Eyes fixations (%)	31.3 (6.9)	25.2	38.3	25.4	28.4	39.1
FPAs	26.7 (4.3)	*t* = 0.33, *p* = .75	***t* = 2.54, *p* = .04**	*t* = 2.85, *p* = .78	*t* = 0.37, *p* = .72	***t* = 2.72, *p* = .03**
Novices	31.7 (9.7)	*t* = 0.65, *p* = .52	*t* = 0.67, *p* = .51	*t* = 0.63, *p* = .53	*t* = 0.33, *p* = .74	*t* = 0.76, *p* = .46
Nose fixations (%)	20.4 (4.5)	19.8	21.9	16.5	27.3	16.3
FPAs	22.6 (7.8)	*t* = 0.39, *p* = .75	*t* = 0.09, *p* = .94	*t* = 0.74, *p* = .49	*t* = 0.57, *p* = .59	*t* = 0.76, *p* = .47
Novices	23.5 (3.9)	*t* = 0.93, *p* = .36	*t* = 0.40, *p* = .69	*t* = 1.77, *p* = .08	*t* = 0.96, *p* = .35	*t* = 1.82, *p* = .08
Mouth fixations (%)	18.4 (3.6)	19.3	14.6	24.2	17.8	16.3
FPAs	15.5 (4.7)	*t* = 0.76, *p* = .47	*t* = 0.18, *p* = .86	*t* = 1.75, *p* = .12	*t* = 0.46, *p* = .66	*t* = 0.16, *p* = .88
Novices	15.3 (6.4)	*t* = 0.62, *p* = .54	*t* = 0.11, *p* = .92	*t* = 1.37, *p* = .18	*t* = 0.38, *p* = .70	*t* = 0.15, *p* = .88
All features		**FFE1**	**FFE2**	**FFE3**	**FFE4**	**FFE5**
FFE correlation	versus FPAs	***r* = .909, *p* < .001**	***r* = .898, *p* < .001**	***r* = .870, *p* < .001**	***r* = .891, *p* < .001**	***r* = .836, *p* < .001**
	versus Novices	***r* = .915, *p* < .001**	***r* = .957, *p* < .001**	***r* = .872, *p* < .001**	***r* = .919, *p* < .001**	***r* = .902, *p* < .001**

Note: *p* values for *t* tests are two-tailed and based on single-case comparison of individual facial examiners against the group means.

the viewing strategies of FFEs were not associated with corresponding effects in accuracy for any of the facial examiners.

8.2.4 | Key features

We examined whether facial examiners differed from fingerprint analysts and novices in their usage of specific facial features when viewing time was constrained, by calculating the percentage of fixations to the eyes, nose and mouth on correct trials. These data show that FFE2 and FFE5 directed significantly more fixations at the eye regions than the fingerprint analysts, but not compared to the novices (see Table 4: Key features). None of the other comparisons between the facial examiners and the other groups were significant. Therefore, the fixation data of the facial examiners to the eyes, nose and mouth are generally similar to the comparison groups.

8.2.5 | All features

Finally, we examined similarities between individual examiner performance with the fingerprint analysts and novice groups across the full set of ROIs (see Figure 8), by correlating the percentage fixations across these ROIs for facial examiners with the group data of the fingerprint analysts and novices. This revealed strong positive correlations between each of the facial examiners and the fingerprint expert and novice groups (see Table 4: All features). Thus, attention to features followed a similar pattern across the facial examiners and the comparison groups, suggesting similar viewing of the faces.

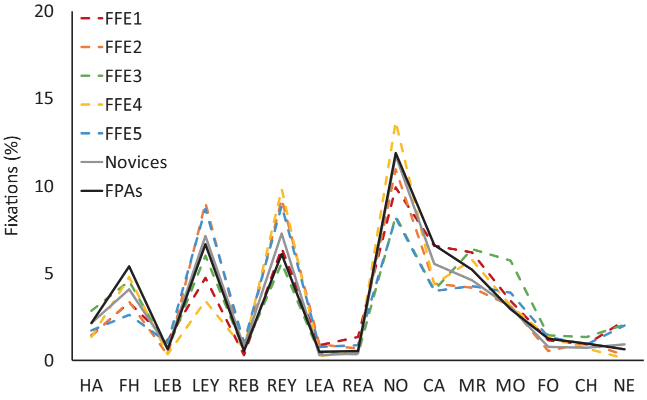


FIGURE 8 Comparison of mean percentage fixations to each ROI in Experiment 2, for each of the facial examiners (FFEs) and the fingerprint analyst and novice groups. HA = hair, FH = forehead, LEB = left eyebrow, LEY = left eye, REB = right eyebrow, REY = right eye, LEA = left ear, REA = right ear, NO = nose, CA = cheek area, MR = mouth region, MO = mouth, FO = face outline, CH = chin, NE = neck

9 | DISCUSSION

This experiment replicates the main findings of Experiment 1, but under conditions that equated viewing time of face pairs across observers. All facial examiners continued to demonstrate a numerical advantage in face-matching accuracy and sensitivity (*d′*) over fingerprint analysts and novices. This advantage was reliable in FFE1, FFE2 and FFE4, but only in comparison with the mean accuracy of the fingerprint analysts. The two best-performing facial examiners in Experiment 1 (FFE1 and FFE2) therefore sustained their accuracy advantage in Experiment 2, suggesting some intra-individual consistency in their performance across the two experiments. Although the performance of FFE4 was not above the control means in Experiment 1, accuracy in Experiment 2 was superior to that of the fingerprint analysts, reflecting a less consistent performance advantage in this examiner.

The main focus of this experiment was to compare the viewing behaviours of facial examiners with the comparison groups when the viewing time allowed for each face pair was constrained. In line with this manipulation, response times for the examiners now were matched with both comparison groups. In eye movements, there were substantial similarities between the facial examiners and the comparison groups. For example, the percentage of fixations that were allocated to key features and all features was generally comparable, nor were there differences in the foraging of face regions, suggesting that the extent to which all parts of the faces were viewed were broadly similar.

However, some systematic differences seem to exist in Experiment 2 in how this information is acquired. For example, although facial examiners generally scanned faces similarly to the two comparison groups, following a pattern that systematically scanned faces from

Classify each feature as: Similar (S), Dissimilar (D), or Can't compare (C)

Face shape:

Forehead:

Eyebrows:

Eyes:

Ears:

Cheek area:

Nose:

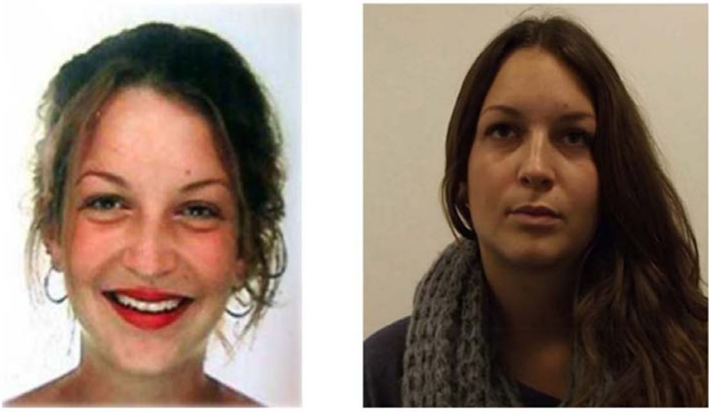
Mouth:

Mouth area:

Jawline:

Chin:

Scars and blemishes:



Now classify whole faces as: Same identity (S) or Different identity (D)

FIGURE 9 Example of a face pair with feature list and instructions to participants in Experiment 3

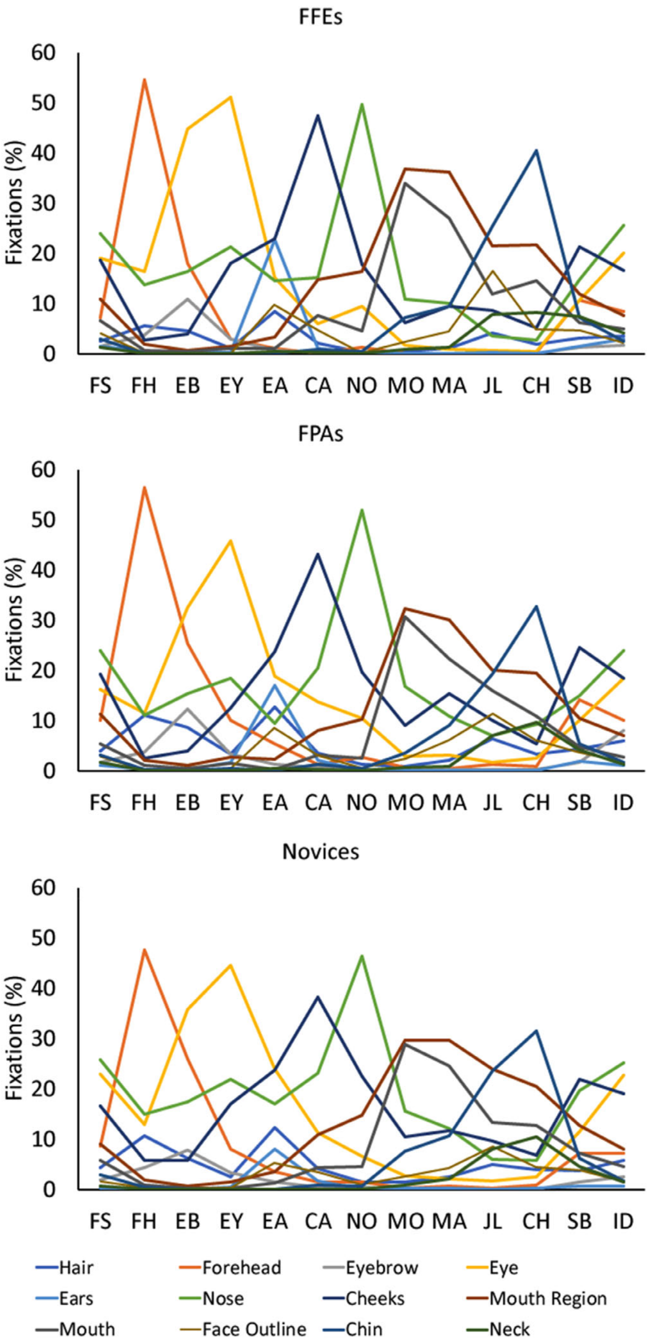


FIGURE 10 The distribution of fixations across ROIs during the interval of each feature judgement (FS, FH, EB, EY, EA, CA, NO, MO, MA, JL, CH, and SB) and the final identification decision (ID) for facial examiners (top), fingerprint analysts (middle), and novices (bottom)

top-to-bottom, this behaviour was more pronounced in the examiners. In addition, analysis of dependencies between successive fixations show that four of the facial examiners (FFE1, FFE3, FFE4, FFE5) were significantly *more* likely to view the *same* face area on successive fixations and *less* likely to view a *different* face. This indicate that facial examiners are particularly systematic in how faces are scanned, moving gradually across faces and inspecting features in more detail. However, this behaviour was evident particularly in comparison with the novices, against which significant advantages in identification accuracy were not found. Overall, these data therefore converge with Experiment 1 to suggest that the facial examiners generally approached this task similarly to the comparison groups, and where differences in viewing behaviour are found, these do not seem to account for differences in identification accuracy.

10 | EXPERIMENT 3

While Experiment 1 and 2 reveal a significant accuracy advantage for specific examiners, there is no clear pattern in the data to explain the accuracy advantage of individual examiners. However, these experiments only examined the visual information that is fixated during face matching, but not how this information is *evaluated*. In a final experiment, we investigated whether an accuracy advantage for facial examiners remains under viewing conditions that entail an evaluative assessment of facial feature similarity. For this purpose, we tested the observer groups under conditions that more closely resemble the working practices of facial examiners. Guidelines for facial image comparison suggest an analytical comparison of features (FISWG, 2018). While the implementation of this process differs both across and within organisations (Moreton, 2021), it typically involves the listwise examination of facial features to determine similarities and differences (see, e.g., Towler et al., 2017). In this final task, we apply an abbreviated version of the feature list that is typically utilised by the facial examiners who participated in this study in their day-to-day work. Participants were required to work methodically through this list, by rating whether each feature in a face pair is indicative of depicting the same person or different people before an overall identification decision was made.

This approach allowed us to examine qualitative differences between observer groups in a different way, by looking at the pattern of feature evaluations. This approach constrains eye movements during the listwise evaluation of features, which should homogenise these data across observer groups. For example, when asked to judge the similarity of the eye regions, all observers would be expected to view this facial feature predominantly. In turn, however, this experiment can provide insight into whether an identification advantage for facial examiners remains when the comparison groups are required to evaluate the features of a face in the same way.

11 | METHOD

11.1 | Participants, stimuli and procedure

Experiment 3 was conducted on the same day as Experiments 1 and 2. All previous participants therefore took part in Experiment 3, following a short break after Experiment 2. The stimuli in this experiment consisted of 20 new face pairs from the KFMT (10 matches and 10 mismatches). The left face in each pair was displayed onscreen at a

TABLE 5 Correlations of fixations across features during all feature judgements (FS, FH, etc.) and the identification decision (ID) in Experiment 3

		FS	FH	EB	EY	EA	CA	NO	MO	MA	JL	CH	SB	ID
FPAs versus Novices		.954	.980	.954	.951	.900	.984	.986	.986	.981	.965	.977	.885	.901
FPAs versus	FFE1	.938	.860	.589	.693	.874	.838	.708	.836	.815	.838	.918	.679	.771
	FFE2	.879	.983	.892	.944	.785	.976	.979	.980	.920	.960	.908	.893	.794
	FFE3	.849	.987	.756	.896	.682	.537	.956	.890	.838	.841	.928	.684	.708
	FFE4	.859	.977	.669	.784	.843	.954	.970	.860	.759	.801	.648	.696	.742
	FFE5	.837	.879	.652	.906	.792	.933	.988	.920	.917	.702	.863	.930	.771
Novices versus	FFE1	.924	.866	.742	.816	.831	.871	.801	.888	.854	.844	.916	.709	.876
	FFE2	.932	.959	.913	.956	.756	.943	.971	.985	.950	.964	.910	.931	.741
	FFE3	.797	.966	.751	.906	.773	.614	.963	.919	.888	.906	.928	.711	.758
	FFE4	.913	.932	.693	.814	.719	.921	.943	.858	.773	.729	.720	.714	.814
	FFE5	.776	.936	.709	.905	.651	.896	.978	.918	.934	.682	.856	.919	.891

Note: All correlations are significant at p < .01.

size of 300 × 406 pixels, and the right face at a size of 346 × 406 pixels, alongside a list of 12 facial features (see Figure 9). These reflected face shape (FS), forehead (FH), eyebrows (EB), eyes (EY), ears (EA), cheek area (CA), nose (NO), mouth (MO), mouth area (MA), jawline (JL), chin (CH) and scars and blemishes (SB). Participants were asked to classify each feature as similar or dissimilar by pressing two buttons ('S' and 'D') on a standard computer keyboard, so that these decisions correspond to the binary response format (i.e., match or mismatch) that was required in Experiment 1 and 2. A 'cannot compare' option was also available should a feature be obscured or unclear in the presented images. An illustration of this visual display can be seen in Figure 9.

Each feature needed to be rated in the order presented in the list, with answers displayed next to each feature name on screen. Participants were unable to amend any previous ratings to ensure that all would work through the feature list in the same systematic order, by focusing on each feature in turn. Having completed the feature rating, participants were then required to identify each face pair as depicting the same or different identity. Faces remained onscreen for the duration of the trial and responses were self-paced. As in Experiments 1 and 2, eye movements of the participants were tracked during the face-matching tasks.

12 | RESULTS

12.1 | Eye-tracking data

We report the analysis of the eye-tracking data first to demonstrate adherence to task demands. This analysis focused on the percentage fixations on facial regions during the feature evaluations. These 30 face regions were collapsed across the left and right face in each stimulus pair, and across the left and right ears, eyebrows and eyes in each face to form a single region for each of these features. This yielded 12 different regions reflecting the hair, forehead, eyebrows, eyes, ears, nose, cheeks, mouth region, mouth, face shape, chin and neck. The fixations on the regions were then calculated for all correct trials, broken down by time intervals that corresponded to the decision intervals for each of the 12 facial features that participants were asked to judge (face shape, forehead, eyebrows, eyes, ears, cheek area, nose, mouth, mouth area, jawline, chin, scars and blemishes). A final decision interval corresponded to the time window in which an identification for each face pair was made. This captures the fixations between the final feature judgement (scars and blemishes/SB) and the identification decision (same/different identity). Because these fixations were analysed to demonstrate adherence to task demands, we simplify illustration of these data by collapsing across the five facial examiners. Figure 10 illustrates these data for the examiners and the two comparison groups.

Inspection of these data reveals similar fixation patterns across the three observer groups, and shows that the fixation data map onto the corresponding face regions during the evaluation of features. For example, during classification of the forehead the majority of fixations land on this face region, and a similar pattern is observed for the eyes, nose, mouth, cheeks, and chin. To analyse these data, Pearson correlations were performed separately for the pattern of fixations around the 12 facial regions during the classification of each feature, and during the overall identification decision, to compare each facial examiner with the comparison groups. This pattern of correlations is summarised in Table 5 and shows consistently strong correlations for the pattern of fixations across the facial features during each of the feature judgements between each of the facial examiners and the comparison groups.

These high correlations demonstrate adherence to the task demands during feature classification (e.g., participants *should* look at the forehead when instructed to judge the similarity of this face region), but the same correlations also remained high during the final identity decision to each face pair even though eye movements were not constrained by instructions at this point. Overall, these data therefore indicate that the facial examiners and the comparison groups adhered similarly to the task instructions throughout the evaluation of

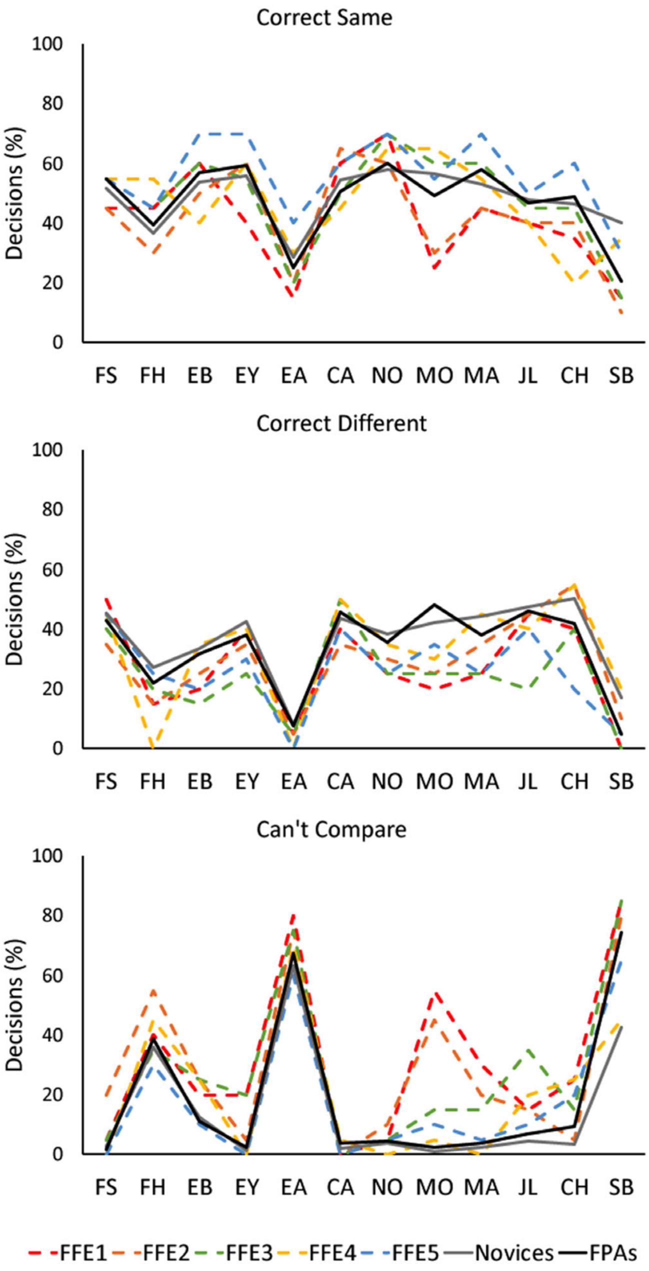


FIGURE 11 Comparison of the mean percentage of correct ‘same’, correct ‘different’ and ‘cannot compare’ responses to features by group in Experiment 3. The features comprised of face shape (FS), forehead (FH), eyebrows (EB), eyes (EY), ears (EA), cheek area (CA), nose (NO), mouth (MO), mouth area (MA), jawline (JL), chin (CH) and scars and blemishes (SB)

the face pairs, and also divided attention similarly between facial features during the final identity judgement.

12.2 | Response-based data

12.2.1 | Feature ratings

The next step of the analysis focused on how the facial features were evaluated by the facial examiners and the comparison groups. For each of the 12 evaluated facial features, participants rated whether they were the ‘same’, ‘different’ or ‘cannot compare’ across each pair of faces. For each group, the mean percentage of these responses was calculated separately for face pairs that were classified correctly as ‘same’ and ‘different’, as well as the proportion of ‘cannot compare’ responses. These data are displayed in Figure 11 for each of the facial examiners and the comparison groups.

These graphs show that the feature judgements followed a similar pattern across all observer groups. These ratings were then used to compute Pearson correlations to assess the relationship in these feature ratings between individual facial examiners and the comparison groups. A summary of these correlations is provided in Table 6 and shows consistent correlations between individual facial examiners and the observer groups. This indicates that facial examiners, fingerprint analysts and novices evaluated the facial features in a similar manner.

12.2.2 | Accuracy

The percentage accuracy scores for each of the facial examiners, fingerprint analysts and novices are illustrated in Figure 12. To determine whether an identification advantage for facial examiners remains when the comparison groups are required to evaluate the features of a face in the same way, the accuracy of each facial examiner was compared to the mean accuracy of the comparison groups via modified *t* tests for single case comparisons. These data are provided in Table 7 and show that only FFE4 performed reliably above the control means of the fingerprint analysts and the novices.

12.2.3 | *d′* and criterion

Similar to the accuracy data, analysis of sensitivity revealed a significant advantage in FFE4 only, in comparison with both the fingerprint analysts and novices (see Table 7). There were no significant differences in criterion between individual facial examiners and the comparison groups.

12.2.4 | Response times

Finally, response times were pooled across the feature evaluation and identification decisions and showed that only FFE4 displayed significantly longer response times than the novice group. No other comparisons were significant.

13 | DISCUSSION

In this experiment, participants were required to work methodically through the face pairs, by rating whether each of a set of features was indicative of depicting the same person or different people before an overall identification decision was made. This approach constrains

		Correct	Incorrect	Cannot compare
FPAs versus	FFE1	$r = .761$, $p = .004$	$r = .835$, $p < .001$	$r = .851$, $p < .001$
	FFE2	$r = .867$, $p < .001$	$r = .820$, $p = .001$	$r = .894$, $p < .001$
	FFE3	$r = .952$, $p < .001$	$r = .790$, $p = .002$	$r = .941$, $p < .001$
	FFE4	$r = .533$, $p = .074$	$r = .781$, $p = .003$	$r = .893$, $p < .001$
	FFE5	$r = .945$, $p < .001$	$r = .875$, $p < .001$	$r = .980$. $p < .001$
Novices versus	FFE1	$r = .626$, $p = .030$	$r = .863$, $p < .001$	$r = .796$, $p = .002$
	FFE2	$r = .752$, $p = .005$	$r = .925$, $p < .001$	$r = .876$, $p < .001$
	FFE3	$r = .847$, $p < .001$	$r = .781$, $p = .003$	$r = .881$, $p < .001$
	FFE4	$r = .578$, $p = .049$	$r = .869$, $p < .001$	$r = .938$, $p < .001$
	FFE5	$r = .790$. $p = .002$	$r = .810$, $p = .001$	$r = .923$, $p < .001$

TABLE 6 Correlations between mean responses to features by group and decision in Experiment 3

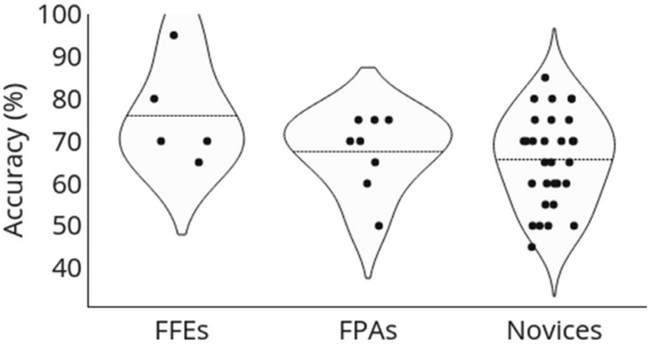


FIGURE 12 Face matching accuracy of individual facial examiners, fingerprint analysts and novices. Violin plots indicate distribution of scores and the horizontal line the group mean

eye movements during the listwise evaluation of features, in an attempt to homogenise the viewing and evaluation of faces across observers. This manipulation was successful in equating viewing behaviour, as the pattern of fixations to the various regions of the faces correlated strongly and consistently across the facial examiners and comparison groups across all stages of the task. Similarly, the by-item feature judgements also correlated strongly for facial examiners, fingerprint analysts and novices. In the context of these similarities, we sought to examine whether an accuracy advantage for facial examiners is maintained even when viewing and evaluation of faces is equated. A numerical accuracy advantage in face matching was observed for four of the five facial examiners over fingerprint analysts and novices. In contrast to the preceding experiments, however, this advantage was only significant for one of the facial examiners (FFE4).

This pattern could be explained by the application of the feature evaluation method raising the accuracy of the comparison groups. We performed an exploratory analysis to investigate this question, by comparing performance at a group level across Experiment 1 and Experiment 3. This analysis suggests that accuracy for the facial examiner group was comparable between Experiment 1 and 3 (both means of 76.0%, $t(4) = 0.00$, $p = 1.00$), but showed modest increases in accuracy for novices (60.0% in Experiment 1 vs. 65.7% in Experiment 3, $t(29) = 2.73$, $p < .05$) and fingerprint analysts (61.3% vs. 67.5%, $t(7) = 2.38$, $p < .05$). For both comparison groups, overall accuracy in Experiment 2 fell in-between these experiments, and did not differ from either in novices (Exp. 1 [60.0%] vs. Exp. 2 [65.3%], $t(29) = 2.01$,

TABLE 7 Individual case analyses comparing the performance of individual facial examiners against the group (mean) performance of the fingerprint analysts and novices in Experiment 3 in terms of accuracy, d', criterion and response times. Parentheses show standard deviation of the means.

	Group mean	FFE1	FFE2	FFE3	FFE4	FFE5
Accuracy (%)	76.0 (11.9)	70.0	65.0	70.0	95.0	80.0
FPAs	67.5 (8.9)	$t = 0.27$, $p = .80$	$t = 0.27$, $p = .80$	$t = 0.27$, $p = .80$	**$t = 2.91$, $p = .02$**	$t = 1.32$, $p = .23$
Novices	65.7 (10.8)	$t = 0.39$, $p = .70$	$t = 0.06$, $p = .95$	$t = 0.39$, $p = .70$	**$t = 2.67$, $p = .01$**	$t = 1.30$, $p = .20$
d'	1.41 (0.83)	0.98	0.75	0.98	2.79	1.57
FPAs	0.85 (0.43)	$t = 0.29$, $p = .78$	$t = 0.22$, $p = .83$	$t = 0.29$, $p = .78$	**$t = 4.25$, $p = .004$**	$t = 1.58$, $p = .16$
Novices	0.85 (0.60)	$t = 0.21$, $p = .83$	$t = 0.16$, $p = .87$	$t = 0.21$, $p = .83$	**$t = 3.18$, $p = .003$**	$t = 1.18$, $p = .25$
Criterion	0.09 (0.32)	−0.26	0.37	−0.26	0.30	0.31
FPAs	0.04 (0.44)	$t = 0.48$, $p = .65$	$t = 0.71$, $p = .50$	$t = 0.48$, $p = .65$	$t = 0.58$, $p = .60$	$t = 0.58$, $p = .58$
Novices	0.02 (0.48)	$t = 0.49$, $p = .62$	$t = 0.72$, $p = .48$	$t = 0.49$, $p = .62$	$t = 0.57$, $p = .57$	$t = 0.59$, $p = .58$
Response times (ms)	6804 (5065)	5985	5220	4132	15,642	3040
FPAs	7999 (4198)	$t = 0.45$, $p = .67$	$t = 0.62$, $p = .55$	$t = 0.87$, $p = .42$	$t = 1.72$, $p = .13$	$t = 1.11$, $p = .30$
Novices	3164 (2880)	$t = 0.96$, $p = .34$	$t = 0.70$, $p = .49$	$t = 0.33$, $p = .74$	**$t = 4.26$, $p < .001$**	$t = 0.04$, $p = .97$

$p = .05$; Exp. 2 [65.3%] vs. Exp. 3 [65.7%], $t(29) = 0.15$, $p = .89$) and fingerprint analysts (Exp. 1 [61.3%] vs. Exp. 2 [63.8%], $t(7) = 0.71$, $p = .50$; Exp. 2 [63.8%] vs. Exp. 3 [67.5%], $t(7) = 0.94$, $p = .38$). We interpret these data as evidence that the systematic feature evaluation method raised the accuracy of fingerprint analysts and novices.

14 | GENERAL DISCUSSION

This study investigated the accuracy and eye movements of forensic facial examiners during the identity-matching of faces, and contrasted these with control groups of fingerprint analysts and untrained novice observers. The main aim was to examine whether facial examiners exhibit qualitatively different viewing behaviours during face matching that can explain their superior identification accuracy, and whether the accuracy of comparison groups who are untrained in facial comparison becomes more similar to facial examiners when faces are evaluated in a similar method during identification. Because of the small population of facial examiners in the organisation under investigation, and in light of the individual differences that have been documented in examiner performance (Phillips et al., 2018), we focused our analysis on individual facial examiners throughout, by adopting a method that allows for single-case comparisons against control groups.

All facial examiners demonstrated a numerical advantage in face-matching accuracy over fingerprint analysts and novices in Experiment 1 and 2. This finding converges with previous work showing an accuracy advantage for facial examiners over untrained observers and fingerprint analysts, as well as other comparison groups (Norell et al., 2015; Phillips et al., 2018; Towler et al., 2017; White, Dunn, et al., 2015; White, Phillips, et al., 2015). At an individual level, however, only two facial examiners—FFE1 and FFE2—performed reliably above the mean performance of the control groups in Experiment 1, both with an accuracy of 85%. And in Experiment 2, the same two observers (again, both with an accuracy of 85%) as well as FFE4 (85%) reliably outperformed the control group means. These findings suggest some intra-individual consistency in examiner performance across the two experiments, but show also that some examiners do not perform with significantly better accuracy (FFE3 and FFE5) than comparison groups who are untrained in facial image comparison. This finding converges with other reports that have demonstrated such differences in these experts (Phillips et al., 2018).

The question of primary interest for this study was whether facial examiners solve identity matching via qualitatively different strategies to the control groups, by virtue of their extensive training and on-the-job experience in facial image comparison. To explore this question with the response data, we examined general aspects of face viewing, such as the direction of eye movements (scanning) and spatial exploration of face areas (foraging), the repeated viewing of face features in successive fixations (dependencies), how attention is allocated to main features of a face (comprising the eyes, nose, and mouth), and the extent to which different regions are viewed when faces are dissected into much greater detail. Across most of these measures, the eye movements of facial examiners were marked by similarities, rather than differences to that of the fingerprint analysts and untrained novices, suggesting similar viewing of faces during matching. And when differences between facial examiners and the comparison groups were found—for example, in the more extensive foraging of features (Experiment 1), a more pronounced tendency to scan faces systematically along the vertical plane (Experiment 1 and 2), and the repeated viewing of the same face areas in successive fixations (Experiment 1 and 2)—these differences did not map onto the accuracy data in a clear manner, suggesting limited explanatory power for the face-matching performance of the examiners. Overall, the results of Experiment 1 and 2 therefore suggest that these experts did not apply a fundamentally—or qualitatively—different approach to successfully resolve the current tasks, but viewed faces in a similar manner to observers who are untrained in forensic facial image comparison.

These findings contrast with qualitative differences in the face-matching abilities of experts and non-experts that have been observed in previous research. In these studies, forensic facial image experts were more accurate than novices in determining that two high quality face images were of the same person, and were also more cautious when faced with lower quality images (Norell et al., 2015). Examiners were also more accurate at both short (2 s) and long (30 s) exposures to face pairs, and showed less impairment when face images were inverted (White, Dunn, et al., 2015; White, Phillips, et al., 2015). Similar differences in cognitive and perceptual processes have also been observed between super-recognisers and forensic facial examiners, with examiners slower to respond during face comparisons and demonstrating a lower likelihood of biased decisions and misidentification errors (Towler, Dunn, et al., 2021).

However, several factors make a direct comparison across studies difficult. Some previous studies have, for example, focused predominantly on *average* examiner performance, by reporting mean identification accuracy across individuals (e.g., Norell et al., 2015; White, Dunn, et al., 2015; White, Phillips, et al., 2015). In contrast, the current study focused on analysis of individual differences as facial examiners are not a homogenous group (see Phillips et al., 2018). The content, quality, and benefits of facial examiner training programmes also varies considerably and, despite widespread adoption of FISWG guidance, local working practices are also likely to differ (Towler et al., 2019). Studies of facial examiners also vary in their sample populations, with some focusing on European (Norell et al., 2015) and international examiners (White et al., 2018), whereas we have tested British participants. In addition, these studies also employed different stimulus sets, which also affects mapping of group and individual face-matching performance (see, e.g., Fysh & Bindemann, 2018).

Of course, another important difference between previous studies and the experiments reported here is the application of eye-tracking technology to examine how faces are viewed during facial comparison. However, whereas Experiment 1 and 2 measured how the features of faces are *viewed* during facial image comparison, these experiments do not speak to how facial features are *perceived*, in terms of the judgements and the decisions that were made. In Experiment 3, we therefore examined whether an accuracy advantage is found for facial examiners under viewing conditions that specifically require the evaluation of facial features. For this purpose, we tested all observer groups under conditions that more closely resemble the working practices of facial examiners, by performing a listwise analytical comparison of facial features to determine similarities and differences between faces before an identification decision is made (FISWG, 2018). Analysis of eye movements and feature decisions indicates that this manipulation was successful in providing a similar perceptual evaluation of faces across the observer groups. In turn, identification accuracy also appeared to converge, with FFE1 (70%), FFE2 (65%) and FFE3 (70%) performing close to the control means of the fingerprint analyst (68%) and novice groups (66%). Only FFE4 performed significantly above the level of the comparison in groups in Experiment 3, at 95% accuracy.

The question arises of whether the feature evaluation method of Experiment 3 genuinely increased accuracy in the comparison groups or whether facial examiner performance had declined. An exploratory analysis to investigate this question (see Discussion of Experiment 3) suggests that the accuracy of fingerprint analysts and novices increased across experiments. Forcing untrained observers to view faces for longer in Experiment 2 (compared to Experiment 1; see response time means) may lead to some small (but non-significant) changes in the evaluation of facial similarities. This aligns with evidence that face-matching accuracy improves in untrained observers as more viewing time becomes available (e.g., Bindemann et al., 2016; Fysh & Bindemann, 2017; Özbek & Bindemann, 2011). But asking observers to compare facial features *explicitly* during face viewing via similarity decisions in Experiment 3 appeared to lead to a more in-depth evaluation of the looked-at content for making an identity-matching decision. This finding is consistent with previous research in which novice face-matching accuracy improved when observers were required to rate the similarity of each feature in a pair of faces (Towler et al., 2017), or were instructed to focus on features of greater diagnostic value such as facial marks or ears (Towler, Keshwa, et al., 2021). In turn, this suggests that the accuracy advantage of individual facial examiners does not arise from how faces are viewed per se, but from the systematic evaluation of feature information during viewing.

We offer this explanation tentatively considering current sample sizes and data limits. At the time of testing, our group of examiners comprised *all* facial examiner staff of a UK organisation. Correspondingly, we have treated this group as a population here and, in light of the documented individual differences in examiner performance (Phillips et al., 2018), focused our analysis on individual examiner performance. An advantage of this approach is that the examiners that were tested here would have undergone comparable in-house training and mentoring in forensic facial image comparison, providing uniformity to contextualise the individual performance differences that were observed here. However, replication and extension of the current approach with other examiners will be an important endeavour. In such future work, it is also possible that more concrete differences might emerge between facial examiners and comparison groups in the level of detail that is focused on. For example, there is evidence from psychological experiments (Towler et al., 2017) as well as professional guidance (see Moreton, 2021) that fine feature detail, such as scars or facial marks can be particularly diagnostic for identification (see also Biswas et al., 2011; Wirth & Carbon, 2017). In an eye-tracking paradigm, investigation of this detail will require stimuli that are varied more systematically along these dimensions than the face pairs employed here, in which the presence of scars and marks can be difficult to distinguish from image artefacts and the covering of such marks through make-up is unknown.

This issue also points to another important caveat of the current study. The everyday working conditions of facial examiners still differ substantially from the experiments reported here and across the literature, which reduces correspondence between examiner performance under experimental conditions with their ability to identify people in occupational settings. In applied settings, for example, facial identifications might follow more detailed protocols, are likely to rely on a greater set of source images, and are undertaken across much more extended time frames. Facial examiners also do not conduct facial comparisons at the same volume that was required by this study, with a single examination typically taking hours or days (Steyn et al., 2018). Understanding these differences between scientific experimentation and professional practices, and examining these systematically, holds the key for further progress in this field.

In conclusion, the current study shows that forensic facial examiners from the same organisation and with the same training exhibit individual differences in facial comparison accuracy. Moreover, these facial examiners do not consistently outperform other forensic professionals and untrained novices. There is some evidence that facial examiners studied faces more carefully, by viewing face pairs for longer, scanning faces more systematically, and examining the same facial areas in depth over successive fixations. However, these characteristics were not relatable to examiners' identification accuracy in a clear systematic manner. Moreover, the behaviour of facial examiners also showed many similarities to the comparison groups in terms of how faces were viewed and how they were perceived, based on the feature judgements and identification decisions that were made. This indicates that an examiner advantage in facial identification accuracy might reflect the systematic evaluation of facial features rather than qualitatively-different viewing strategies.

FUNDING INFORMATION

This work was funded in part by a Graduate Teaching Assistantship to Jacky Claydon and by a research grant from the Economic and Social Research Council (ESRC Grant no. ES/S010181/1) to Markus Bindemann.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The datasets generated during and/or analysed during the current study are available in the OSF repository https://osf.io/rhyq6/.

ORCID

Matthew C. Fysh https://orcid.org/0000-0002-3812-3749
Markus Bindemann https://orcid.org/0000-0002-9608-4186

ENDNOTES

[1] We note that 'super-recognisers' are sometimes deployed in professional settings and their accuracy has been found to be on-par with forensic examiners (see Towler, Dunn, et al., 2021; White et al., 2021), but questions also remain about their proficiency and legal status (see Bate et al., 2021; Roberts, 2021).

[2] In a pilot study ($N = 101$), we collected accuracy scores for all face pairs employed in the experiments here online. The face pairs were then divided into three stimulus sets of comparable difficulty for the experiments (Experiment 1: $M = 56.3\%$, $SD = 18.6$; Experiment 2: $M = 56.9\%$, $SD = 17.8$; Experiment 3: $M = 56.5\%$, $SD = 18.1$).

[3] The data for all experiments reported here can be accessed at https://osf.io/rhyq6/.

[4] For FFE2, the eye movement data for the first nine trials of Experiment 1 was lost due to a power cut during testing. For this examiner, all analyses of eye movements in this experiment are therefore based on trials 10–20.

[5] At the suggestion of a reviewer, we have also created heatmaps for all participants for all experiments reported here. These are available, along with the datasets for the current study, in the OSF repository https://osf.io/rhyq6/.

[6] Although ears have previously been identified as a key feature in facial image comparison (see Towler et al., 2017; Towler, Keshwa, et al., 2021), ears were fully visible in the KFMT in only 26% of image pairs and were therefore not included in this analysis.

REFERENCES

Althoff, R. R., & Cohen, N. J. (1999). Eye-movement-based memory effect: A reprocessing effect in face perception. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 997–1010. https://doi.org/10.1037//0278-7393.25.4.997

Arizpe, J. M., Noles, D. L., Tsao, J. W., & Chan, A. W.-Y. (2019). Eye movement dynamics differ between encoding and recognition of faces. *Vision*, *3*(1), 9. https://doi.org/10.3390/vision3010009

Attard, J., & Bindemann, M. (2014). Establishing the duration of crimes: An individual differences and eye-tracking investigation into time estimation. *Applied Cognitive Psychology*, *28*, 215–225. https://doi.org/10.1002/acp2986

Bate, S., Mestry, N., & Portch, E. (2021). Individual differences between observers in face matching. In M. Bindemann (Ed.), *Forensic face matching: Research and practice*. Oxford University Press. https://doi.org/10.1093/oso/9780198837749.003.0006

Bindemann, M. (2010). Scene and screen center bias early eye movements in scene viewing. *Vision Research*, *50*, 2577–2587. https://doi.org/10.1016/j.visres.2010.08.016

Bindemann, M. (Ed.). (2021). *Forensic face matching: Research and practice*. Oxford University Press. ISBN 978-0-19-883774-9.

Bindemann, M., Attard, J., Leach, A., & Johnston, R. A. (2013). The effect of image pixelation on unfamiliar face matching. *Applied Cognitive Psychology*, *27*, 707–717. https://doi.org/10.1002/acp.2970

Bindemann, M., Avetisyan, M., & Rakow, T. (2012). Who can recognize unfamiliar faces? Individual differences and observer consistency in person identification. *Journal of Experimental Psychology: Applied*, *18*, 277–291. https://doi.org/10.1037/a0029635

Bindemann, M., Fysh, M., Cross, K., & Watts, R. (2016). Matching faces against the clock. *i-Perception*, *7*, 2041669516672219. https://doi.org/10.1177/2041669516672219

Bindemann, M., & Sandford, A. (2011). Me, myself, and I: Different recognition rates for three photo-IDs of the same person. *Perception*, *40*, 625–627. https://doi.org/10.1068/p7008

Bindemann, M., Scheepers, C., & Burton, A. M. (2009). Viewpoint and centre of gravity affect eye movements to human faces. *Journal of Vision*, *9*(2), 1–16. https://doi.org/10.1167/9.2.7

Bindemann, M., Scheepers, C., Ferguson, H. J., & Burton, A. M. (2010). Face, body and centre of gravity mediate person detection in natural scenes. *Journal of Experimental Psychology: Human Perception & Performance*, *36*, 1477–1485. https://doi.org/10.1037/a0019057

Birmingham, E., Bischof, W. F., & Kingstone, A. (2008). Social attention and real-world scenes: The roles of action, competition and social content. *Quarterly Journal of Experimental Psychology*, *61*, 986–998. https://doi.org/10.1080/17470210701410375

Biswas, S., Bowyer, K. W., & Flynn, P. J. (2011). A study of face recognition of identical twins by humans. In *2011 IEEE International workshop on information forensics and security, WIFS 2011, (November)*. https://doi.org/10.1109/WIFS.2011.6123126

Blais, C., Jack, R. E., Scheepers, C., Fiset, D., & Caldara, R. (2008). Culture shapes how we look at faces. *PLoS One*, *3*(8), e3022. https://doi.org/10.1371/journal.pone.0003022

Bobak, A. K., Parris, B. A., Gregory, N. J., Bennetts, R. J., & Bate, S. (2017). Eye-movement strategies in developmental prosopagnosia and "super" face recognition. *Quarterly Journal of Experimental Psychology*, *70*, 201–217. https://doi.org/10.1080/17470218.2016.1161059

Burton, A. M., White, D., & McNeill, A. (2010). The Glasgow Face Matching Test. *Behavior Research Methods*, *42*, 286–291. https://doi.org/10.3758/BRM.42.1.286

Crawford, J. R., Garthwaite, P. H., & Porter, S. (2010). Point and interval estimates of effect sizes for the case-controls design in neuropsychology: Rationale, methods, implementations, and proposed reporting standards. *Cognitive Neuropsychology*, *27*, 245–260. https://doi.org/10.1080/02643294.2010.513967

Dowsett, A. J., Sandford, A., & Burton, A. M. (2016). Face learning with multiple images leads to fast acquisition of familiarity for specific individuals. *Quarterly Journal of Experimental Psychology*, *69*, 1–10. https://doi.org/10.1080/17470218.2015.1017513

Facial Identification Scientific Working Group. (2018). *Facial image comparison feature list for morphological analysis*. https://fiswg.org/FISWG_Morph_Analysis_Feature_List_v2.0_20180911.pdf

Fletcher-Watson, S., Findlay, J. M., Leekam, S. R., & Benson, V. (2008). Rapid detection of person information in a naturalistic scene. *Perception*, *37*, 571–583. https://doi.org/10.1068/p5705

Fysh, M. C., & Bindemann, M. (2017). Effects of time pressure and time passage on face matching accuracy. *Royal Society Open Science*, *4*, 170249. https://doi.org/10.1098/rsos.170249

Fysh, M. C., & Bindemann, M. (2018). The Kent Face Matching Test. *British Journal of Psychology*, *109*, 219–231. https://doi.org/10.1111/bjop.12260

Hautus, M. J. (1995). Corrections for extreme proportions and their biasing effects on estimated values of d′. *Behavior Research Methods*, *27*, 46–51. https://doi.org/10.3758/BF03203619

Henderson, J. M., Williams, C. C., & Falk, R. J. (2005). Eye movements are functional during face learning. *Memory & Cognition*, *33*, 98–106. https://doi.org/10.3758/bf03195300

Heyer, R., MacLeod, V., Carter, L., Semmler, C., & Ma-Wyatt, A. (2011). Profiling the facial identification practitioner in Australia: Report on the human operator capability project survey. *Report for the Department of Prime Minister and Cabinet: Grant PR09-0078*. https://www.researchgate.net/publication/318206064_Profiling_the_Facial_Comparison_Practitioner_in_Australia

Kelly, D. J., Duarte, D., Meary, D., Bindemann, M., Jaggie, C., & Pascalis, O. (2019). Infants rapidly detect human faces in complex visual scenes. *Developmental Science*, *22*(6), e12829. https://doi.org/10.1111/desc.12829

Kramer, R. S. S., & Ritchie, K. L. (2016). Disguising superman: How glasses affect unfamiliar face matching. *Applied Cognitive Psychology*, *30*, 841–845. https://doi.org/10.1002/acp.3261

Lander, K., Bruce, V., & Bindemann, M. (2018). Use-inspired basic research on individual differences in face identification: Implications for criminal investigation and security. *Cognitive Research: Principles and Implications*, *3*(26), 1–13. https://doi.org/10.1186/s41235-018-0115-6

Maurer, D., Le Grand, R., & Mondloch, C. J. (2002). The many faces of configural processing. *Trends in Cognitive Sciences*, *6*, 255–260. https://doi.org/10.1016/S1364-6613(02)01903-4

McCaffery, J. J. M., Robertson, D. J., Young, A. W., & Burton, A. M. (2018). Individual differences in face identity processing. *Cognitive Research: Principles and Implications*, *3*(21), 1–15. https://doi.org/10.1186/s41235-018-0112-9

Megreya, A. M., Bindemann, M., Havard, C., & Burton, A. M. (2012). Identity-lineup location influences target selection: Evidence from eye movements. *Journal of Police and Criminal Psychology*, *27*, 167–178. https://doi.org/10.1007/s11896-011-9098-7

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*, 700–706. https://doi.org/10.1002/acp.2965

Moreton, R. (2021). Forensic face matching: Procedures and application. In M. Bindemann (Ed.), *Forensic face matching: Research and practice*. Oxford University Press. https://doi.org/10.1093/oso/9780198837749.003.0007

Moreton, R., Havard, C., Strathie, A., & Pike, G. (2021). An international survey of applied face-matching training courses. *Forensic Science International*, *327*, 110947. https://doi.org/10.1016/j.forsciint.2021.110947

Moreton, R., Pike, G., & Havard, C. (2019). A task- and role-based perspective on super-recognizers: Commentary on 'super-recognizers: From the laboratory to the world and back again'. *British Journal of Psychology*, *110*, 486–488. https://doi.org/10.1111/bjop.12394

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*, 331–340. https://doi.org/10.1111/1556-4029.12660

Noyes, E., & Jenkins, R. (2017). Camera-to-subject distance affects face configuration and perceived identity. *Cognition*, *165*, 97–104. https://doi.org/10.1016/j.cognition.2017.05.012

Özbek, M., & Bindemann, M. (2011). Exploring the time course of face matching: Temporal constraints impair unfamiliar face identification under temporally unconstrained viewing. *Vision Research*, *51*, 2145–2155. https://doi.org/10.1016/j.visres.2011.08.009

Papesh, M. H. (2018). Photo ID verification remains challenging despite years of practice. *Cognitive Research: Principles and Implications*, *3*(19), 1–9. https://doi.org/10.1186/s41235-018-0110-y

Phillips, P. J., Yates, A. N., Hu, Y., Hahn, C. A., Noyes, E., Jackson, K., Cavazos, J. G., Jeckeln, G., Ranjan, R., Sankaranarayanan, S., Chen, J. C., Castillo, C. D., Chellappa, R., White, D., & O'Toole, A. J. (2018). Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Proceedings of the National Academy of Sciences*, *115*, 6171–6176. https://doi.org/10.1073/pnas.1721355115

Prince, J. P. (2012). *Report on emerging use of facial recognition systems and facial image comparison procedures*. https://www.churchilltrust.com.au/media/fellows/2012_Prince_Jason.pdf

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

Roberts, A. (2021). Forensic face matching: A legal perspective. In M. Bindemann (Ed.), *Forensic face matching: Research and practice*. Oxford University Press. https://doi.org/10.1093/oso/9780198837749.003.0008

Stacchi, L., Ramon, M., Lao, J., & Caldara, R. (2019). Neural representations of faces are tuned to eye movements. *The Journal of Neuroscience*, *39*, 4113–4123. https://doi.org/10.1523/JNEUROSCI.2968-18.2019

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods*, *31*, 137–149. https://doi.org/10.3758/BF03207704

Steyn, M., Pretorius, M., Briers, N., Bacci, N., Johnson, A., & Houlton, T. M. R. (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., Dunn, J. D., Martínez, S. C., Moreton, R., Eklöf, F., Ruifrok, A., et al. (2021). Diverse routes to expertise in facial recognition. PsyArxiv.

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. https://doi.org/10.1371/journal.pone.0211037

Towler, A., Kemp, R. I., & White, D. (2021). Can face identification ability be trained? In M. Bindemann (Ed.), *Forensic face matching: Research and practice*. Oxford University Press. https://doi.org/10.1093/oso/9780198837749.003.0005

Towler, A., Keshwa, M., Ton, B., Kemp, R. I., & White, D. (2021). Diagnostic feature training improves face matching accuracy. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *47*(8), 1288. https://doi.org/10.1037/xlm0000972

Towler, A., White, D., & Kemp, R. I. (2017). Evaluating the feature comparison strategy for forensic face identification. *Journal of Experimental Psychology: Applied*, *23*, 47–58. https://doi.org/10.1037/xap0000108

White, D. (2021). Do professional groups outperform novices in unfamiliar face matching tasks? A meta-analysis. *PsyArXiv*.

White, D., Dunn, J. D., Schmid, A. C., & Kemp, R. I. (2015). Error rates in users of automatic face recognition software. *PLoS One*, *10*(10), e0139827. https://doi.org/10.1371/journal.pone.0139827

White, D., Kemp, R. I., Jenkins, R., & Burton, A. M. (2014). Feedback training for facial image comparison. *Psychonomic Bulletin & Review*, *21*, 100–106. https://doi.org/10.3758/s13423-013-0475-3

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*, 20151292. https://doi.org/10.1098/rspb.2015.1292

White, D., Towler, A., & Kemp, R. I. (2021). Understanding professional expertise in unfamiliar face matching. In M. Bindemann (Ed.), *Forensic face matching: Research and practice*. Oxford University Press. https://doi.org/10.1093/oso/9780198837749.003.0004

Wirth, B. E., & Carbon, C. C. (2017). An easy game for frauds? Effects of professional experience and time pressure on passport-matching performance. *Journal of Experimental Psychology: Applied*, *23*, 138–157. https://doi.org/10.1037/xap0000114

How to cite this article: Claydon, J. R., Fysh, M. C., Prunty, J. E., Cristino, F., Moreton, R., & Bindemann, M. (2022). Facial comparison behaviour of forensic facial examiners. *Applied Cognitive Psychology*, 1–20. https://doi.org/10.1002/acp.4027