Forensic Expertise in Facial Image Comparison

A thesis submitted for the Degree of Master of Science by Research in the Faculty of Social Sciences at the University of Kent

Jacqueline Claydon

School of Psychology
University of Kent
September 2019

Word Count 15370

Abstract

Deciding whether two images of unfamiliar faces are the same person or two different people is a difficult task, but one in which forensic facial examiners generally outperform untrained observers, although not with perfect accuracy. Here, the ways in which they perform face matching were compared with forensically-trained and non-expert controls whilst eye movements were recorded. In Experiments 1 and 2, examiners were the most accurate group. In Experiment 3, rating the features prior to the same or different matching decision improved the controls' performance which reduced the examiners' accuracy advantage. Across the experiments, all groups showed similar patterns of responses to the face pairs and similar attention to the features, including a bias towards faces on the left of the screen. The higher overall accuracy of examiners was not accounted for by differences in viewing times, or by a more conservative response to feature rating. Further, examining the performance of individual examiners showed how group accuracy was driven by some high performers, although the same examiners were not consistently the most accurate in all experiments. Overall, this study did not find any differences in the way professionals viewed faces which might explain their high performance as a group. However, as the adoption of a feature comparison strategy improved accuracy for both control groups, this suggests high accuracy for facial experts may be due to their methodological approach to face matching rather than any qualitative differences in their viewing behaviours.

Keywords: Unfamiliar face matching, face perception, individual differences, facial image comparison, forensic science

Introduction

Personnel employed by police services and border agencies are frequently required to identify unfamiliar faces. Despite the fact that professional mistakes can deprive a person of their liberty, leave an organisation vulnerable to litigation and deeply undermine public confidence in the criminal justice system, the likelihood of erroneous identifications is currently unknown (PCAST, 2016). Given the judicial weight afforded to expert testimony (Edmond, Valentine & Davis, 2015), recent empirical work suggests that understanding qualitative differences in the performance of forensic professionals may be the key to assessing and improving their accuracy (Towler et al., 2017). However, details surrounding the recruitment of staff for these roles are not in the public domain, and a recent review found the majority of facial comparison training courses are based on only limited scientific evidence (Towler et al., 2019). The inclusion of forensic organisations in research, and the development of experiments which reflect working practices, are therefore vital for the effective recruitment and training of facial comparison experts (Ramon, Bobak & White, 2019; Robertson & Bindemann, 2019), a view supported by the Forensic Science Regulator within the UK (Tully, 2019).

In unfamiliar face matching tasks, the observer needs to determine whether a pair of simultaneously presented faces depict the same person or two different people. Within forensic settings, this may arise when verifying a person's identity using a photographic representation such as a passport, or when comparing crime scene CCTV footage with the image of a suspect. Even in the absence of any operational demands, unfamiliar face matching is typically found to be a difficult task. For example, an experiment to assess the accuracy of person-to-photo matching saw a group of supermarket cashiers wrongly accept more than 50% of fraudulent ID cards as genuine (Kemp, Towell & Pike, 1997). In other research, matching a person to a photograph or deciding if two images were of the same

person reported accuracy of only around 66% (Megreya & Burton, 2008). A widely cited resource to measure the accuracy of unfamiliar face matching is The Glasgow Face Matching Test (GFMT; Burton, White & McNeill, 2010). This uses high-quality pairs of photographs, taken on the same day and showing a clear frontal view of the faces. Even under these optimised conditions, mean accuracy is only around 81%. More recently, the Kent Face Matching Test (KFMT; Fysh & Bindemann, 2018) has been developed to reflect some of the identification difficulties encountered in applied settings by using more realistic images. The image pairs in this test were taken at least three months apart and comprise an unconstrained image from a student ID card and an image taken under controlled conditions. Performance on this more demanding test is lower than normative GFMT scores, with mean accuracy of 66%. A range of experimental methods have therefore shown unfamiliar face matching to be a difficult and error-prone task.

Individual differences in the ability to match unfamiliar faces are revealed when the performance of each observer is compared. In operational settings, this has implications for the recruitment of suitably skilled staff (Balsdon, Summersby, Kemp & White, 2018). Studies have revealed a wide range of face matching ability which is not reflected in the average scores of a group. For example, GFMT accuracy ranges from 51% to 100% and KFMT accuracy is between 40% and 88%. Other studies have shown accuracy varying from 50% to 96% when matching a person to a photograph (Megreya & Burton, 2006) and from 44% to 94% when matching a live target to an image (Megreya et al, 2008). In addition, research has also found considerable variation in the consistency and accuracy of each individual. When undertaking the same matching task across consecutive days, performance on one day did not predict performance on a subsequent day (Bindemann, Avetisyan & Rakow, 2012). The study of individual difference therefore reveals a wide range of face

matching ability within the general population, and a level of accuracy which varies from one day to the next.

Research has considered why unfamiliar face matching is such a difficult task. One source of known errors relates to data limits in which identification is hindered by the lack of information within the to-be-compared images (for review see Fysh & Bindemann, 2017). For example, any changes in the appearance of a face over time through ageing, illness, weight changes or the wearing of glasses reflect within-person variation (Jenkins, White, Van Montfort & Burton, 2011). However, these changes are not necessarily captured in identity documents such as passports which are valid for up to ten years, and matching accuracy is found to deteriorate as the age of the comparison image increases (Megreya, Sandford & Burton, 2013). Factors such as illumination can also affect matching accuracy. Although pairs of faces are better identified under equal, rather than different, lighting conditions (Hill & Bruce, 1996), this situation is rarely encountered in operational settings as one of the to-be compared faces may be illuminated by several sources of natural or artificial light.

Identification accuracy may be further compounded by the device used to capture the image. High resolution images contain the most detail and this supports accurate face matching. However, resolution can be distorted when images are stored digitally and is affected by the quality of the lens used and the condition of the recording equipment (Edmond, Biber, Kemp & Porter, 2009). Identification can also be hindered by the camera-to-subject distance. Faces closer to the camera appear more convex whereas those further away appear to be flatter, and differences of only one to two metres can affect identification accuracy (Noyes & Jenkins, 2017). This becomes problematic with ID documentation as the exact camera-to-person distance for the supporting photograph is seldom stipulated (Noyes et al., 2017). Variations in appearance due to data limits mean that identification is difficult even when matching a person to different versions of their own ID photographs, with

accuracy ranging from 46% to 67% in this task (Bindemann & Sandford, 2011). Efforts to improve accuracy have shown potential benefits (e.g. Megreya & Bindemann, 2018; Towler, White & Kemp, 2017), however, unfamiliar face matching remains a challenging task.

Although laboratory studies have highlighted the difficult nature of unfamiliar face matching, it may be expected that accuracy would be higher within forensic settings where such tasks are undertaken on a daily basis. For security reasons, access to this group of professionals is limited and opportunities for research are rare. Studies therefore tend to incorporate findings from a range of facial comparison roles as a measure of unfamiliar face matching ability. For example, in a series of matching tasks using the GFMT, passport officers were not more accurate than students and their performance was not significantly different to normative scores (White, Kemp, Jenkins, Matheson & Burton, 2014). Similarly, tests to measure the face matching accuracy of police officers have demonstrated that experience with unfamiliar faces does not necessarily benefit performance (Burton, Wilson, Cowan & Bruce, 1999; Wirth & Carbon, 2017). The identification of "super-recognizers" (SR), people with superior face perception abilities (Russell, Duchaine & Nakayama, 2009), has led to their deployment within the Metropolitan Police in tasks such as identifying perpetrators from CCTV footage (Davis, Lander, Evans & Jansari, 2016). Despite high levels of face recognition performance by SR (e.g. Robertson, Noyes, Dowsett, Jenkins & Burton, 2016), this ability does not necessarily translate into unfamiliar face matching accuracy (Bate et al., 2018; Bobak, Hancock & Bate, 2016). Therefore, some of the difficulties encountered during unfamiliar face matching in experimental settings have been reflected in the performance of personnel who routinely undertake these tasks.

A further group of professionals, Forensic Facial Examiners (FFEs), typically work within police settings and provide expert testimony as to whether images depict the same person or different people (FISWG, 2012). They use morphological analysis for facial

comparison, in which the features are assessed, described and compared (FISWG, 2012). As a group, FFEs have demonstrated performance superior to controls in experimental face matching tasks (e.g. Norrell et al., 2015; White, Dunn, Schmid & Kemp, 2015; White, Phillips, Hahn, Hill & O'Toole, 2015a) with the performance of a single examiner equivalent to the combined accuracy of seven or more students (Towler et al., 2017), and a group of examiners attaining almost perfect accuracy when their scores were combined (White et al., 2015a). However, individual FFE do not perform this task perfectly (e.g. Norrell et al., 2015; White et al., 2015; White et al., 2015a) and group means conceal a range of individual differences in ability (Phillips et al., 2018). Therefore, if face matching accuracy is to be improved it will require a more detailed understanding of the processes and viewing behaviours of facial comparison professionals.

Emerging evidence suggests the face matching ability of FFEs may be due to qualitative differences in the way they perform the task. Faces are usually processed holistically, as an integrated whole rather than by individual features, and this is disrupted when faces are inverted (Goffaux & Rossion, 2006; Tanaka & Farah, 1993). Because FFEs show less impairment when matching upside-down faces, this suggests they have less reliance on holistic face processing than the general population (Towler et al., 2017; Towler et al., 2019; White et al., 2015a). In addition, a recent study found that although FFEs and students both rated the ears as being the most diagnostic of identity, the groups rated every other feature differently in terms of their diagnosticity and usefulness (Towler et al., 2017). These findings suggest facial experts may be viewing faces differently to non-experts. However, as access to forensic experts is necessarily restricted for security reasons there have been few studies which have examined their viewing behaviours when comparing faces.

Tracking the eye movements of facial experts during face matching tasks may reveal different viewing strategies to those of non-experts which could account for their superior

accuracy. Eye movements are believed to relate to underlying cognitive processes (Henderson, 2003), with differences between movement (saccades) and attention to stimuli (fixations) being task-specific (Rayner, 1998) and functional (Henderson, Williams & Falk, 2005). Visual information is acquired for processing when the eye gaze is stable (Henderson, 2007). Therefore, fixations indicate the amount of attention being allocated to a particular stimulus such as a facial feature, as well as the timeline and pattern of related eye movements (Barton, Radcliffe, Cherkasova, Edelman & Intriligator, 2006).

When presented with the image of a face, the observer's initial fixation tends to land at the geometric centre of the stimuli, rather than being feature-specific (Bindemann, Scheepers, Ferguson & Burton, 2010). Subsequent fixations mostly land in the central regions of the face encompassing the eyes, nose and mouth (Arizpe, Walsh, Yovel & Baker, 2017; Or, Petersen & Eckstein, 2015; Özbek & Bindemann, 2011). A left visual field bias has also been observed in which faces to the left of a screen receive more fixations than faces to the right (Butler et al., 2005; Hsiao & Cottrell, 2008; Peterson & Eckstein, 2013). These findings have been observed within the non-expert population. The existence of a visual field bias or increased attention to central features during professional face matching is currently unknown but may be revealed by tracking their eye movements during face matching tasks.

The aim of the current study was to examine the viewing behaviours of professionals and non-experts during unfamiliar face matching. The identification of qualitative differences in the way the task is performed may offer an explanation for the high accuracy of facial examiners. The performance of FFEs from the London Metropolitan police was compared with two control groups. Fingerprint Analysts (FPAs) routinely undertake detailed comparisons of unfamiliar images in the form of latent fingerprints and exemplars. Comparing FFEs with forensically-trained controls examined whether accuracy was due to perceptual expertise (Ackerman, 1987) gained through the analytical examination of

fingerprints and faces, or whether FFE performance was related to the face-specific nature of their work. The inclusion of a group of university students (Controls) as non-expert controls allowed comparison of face matching strategies and related eye movements within the general population.

The accuracy and eye movements of FFEs, FPAs and Controls were compared across three face matching experiments using images from the KFMT. This test uses realistic face stimuli, designed to be representative of images encountered in applied settings. In all experiments, the participants were presented with matching and mismatching pairs of faces against which they needed to make a same or different person decision. For the first experiment, the performance and viewing behaviours of the three groups were compared when viewing and response times were self-paced. FFEs' guidelines emphasise an analytical and measured response to facial comparison (FISWG, 2012). Therefore, the absence of time pressure may afford FFEs an accuracy advantage over the other groups. In the second experiment, viewing of stimuli was therefore restricted to thirty seconds to allow the direct comparison of FFE's performance with that of the control groups under equal time pressure. The third experiment incorporated a list of twelve facial features which observers rated as being the "same", "different" or "can't compare" prior to their face matching decision. Across the three experiments, the consistency of group and individual FFE accuracy was also assessed.

Experiment 1

Participants undertook twenty trials in which pairs of either same or different faces were presented onscreen. For these tasks, viewing and response times were unlimited. Although previous research has observed an accuracy advantage for FFEs after only a brief two-second exposure to faces, their accuracy increased following a longer exposure duration

(White et al., 2015a). Removing time pressure from face matching tasks also reflects working practices in that forensic facial comparison requires a measured and analytical response rather than a fast and instinctive matching decision (FISWG, 2012).

Previous research using a similar sample of participants measured face matching accuracy using the GFMT and reported that all groups exceed the normative score for this test (White et al., 2015a). Therefore, the use of a more difficult face matching test in the form of the KFMT not only reflected some of the difficulties encountered within applied settings but served as a more robust measure of FFEs' performance with which to examine qualitative differences in their processing strategies.

Method

Participants

Five Forensic Facial Examiners (FFEs) from a UK police service 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* = 31.58, range 6–84 months). Eight Fingerprint Analysts (FPAs) also from a UK police service (mean age = 41.3 years, *SD* = 7.1, range 32–50 years, 3 males) with mean experience in fingerprint comparison of 162.13 months (*SD* 35.55, range 108–204 months), were a forensically-trained control group. Both police groups undertook the experiment in a quiet office at their usual place of work. A further control group of thirty university students (Controls) (mean age = 21.7 years, *SD* = 7.9, range 18–54 years, 4 males), participated in return for course credit. All participants reported normal, or corrected-to-normal, eyesight and provided informed consent to take part. The research was approved by the University of Kent Ethics Committee (Ethics ID 20181534456681508).

Stimuli

The stimuli in this experiment consisted of twenty face pairs from the KFMT (Fysh & Bindemann, 2018) with equal numbers of identity matches and mismatches. For each face pair, the image on the right of the screen comprised a controlled image of the target with a neutral expression which had been taken against a white background with even illumination. This was scaled to a size of 283 x 332 pixels. The image on the left consisted of an unconstrained image taken from a student ID photograph and was re-scaled to a size of 142 x 192 pixels. Both images were presented at an image resolution of 72ppi and there was a minimum gap of three months between the taking of both images. An example of match and mismatch pairs is shown in Figure 1.

Procedure

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 60cm with a chin rest. The participant's dominant 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. This was validated by successful fixation of a further nine targets. The procedure was repeated 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 drift correction.

Across twenty trials either a match or mismatch pair of faces was presented. Match trials were interspersed with mismatch trials and the order of presentation was maintained for all participants. An on-screen prompt asked the participant to identify whether the face pair was the same or different and the response was recorded by pressing S or D on the keyboard.

Viewing time was not restricted and stimuli remained on screen until a response was provided.

Materials

Photoshop software was used to colour-code fifteen different regions of interest (ROI) onto the image of each pair of faces in this experiment. When eye-tracking data was subsequently mapped onto the ROIs, this allowed identification of the different face regions. An example is shown in Figure 1. The same colour code was used for both match and mismatch faces, and different colours were used to identify whether the face was presented to the left or right of the screen. The ROIs related to twelve items on the feature list used in published guidelines for facial image comparison (FISWG, 2012). The ears, eyebrows and eyes were coded to reflect those displayed on the left and right sides of the face and additional ROIs were created for the hair and neck to encompass all areas that could be viewed by participants. This created a total of fifteen ROIs for each face.

Figure 1. Examples of match (left) and mismatch (right) face pairs from the KFMT (top) and the same images with colour-coded regions of interest (ROI) (bottom).

Results

Accuracy by Group

To analyse face matching performance in this experiment, a mean accuracy score for each group was calculated for match and mismatch trials (Figure 2). This reflected the number of correct trials as a percentage of overall trials. To compare accuracy between groups a 3 (Group: FFE, FPA and Controls) x 2 (Trial Type: Match vs. Mismatch) mixed-model ANOVA was conducted, with Trial Type the within-subjects factor and Group the between-subjects factor.

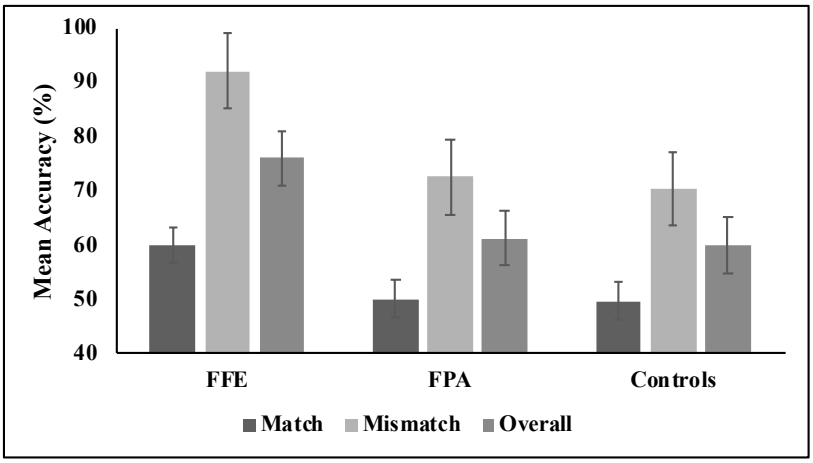

Figure 2. A comparison of mean percentage accuracy by group, trial type and overall performance in Experiment 1. Error bars denote the standard error of the means.

The analysis showed a main effect of Trial Type, $F(1, 40) = 11.26$, $p = .002$, partial $\eta^2 = 0.22$, due to higher accuracy in mismatch trials ($M = 78.28$, $SE = 4.08$) than in match trials ($M = 53.22$, $SE = 4.86$). There was also a main effect of Group, $F(2, 40) = 3.56$, $p = .038$, partial $\eta^2 = .15$ relating to overall accuracy. Post-hoc Tukey HSD tests showed FFEs had more correct responses ($M = 76.00$, SE 5.58) than both FPAs ($M = 61.25$, $SE = 4.40$) and Controls ($M = 60.00$, $SE = 2.27$). No difference between FPAs and Controls was found ($p = .80$). There was no interaction between Trial Type and Group, $F(2, 40) = .20$, $p = .82$, partial $\eta^2 = 0.01$.

In summary, accuracy in mismatch trials was higher than in match trials and the overall accuracy of FFEs was superior to both FPAs and Controls. FFEs exceeded the KFMT normative score of 66%, although this level was not reached by the other groups. Differences in performance between FPAs and Controls were not significant.

Accuracy by Item

To compare group responses to each pair of faces presented in the experiment, mean group accuracy was calculated for each of the twenty items to reflect a percentage of correct responses (Figure 3). A 3 (Group: FFE, FPA and Controls) x 2 (Trial Type: Match vs. Mismatch) mixed-model ANOVA was used to compare the mean item accuracy of each group by type of trial. Trial Type was the within-subjects factor and Group the between-subjects factor.

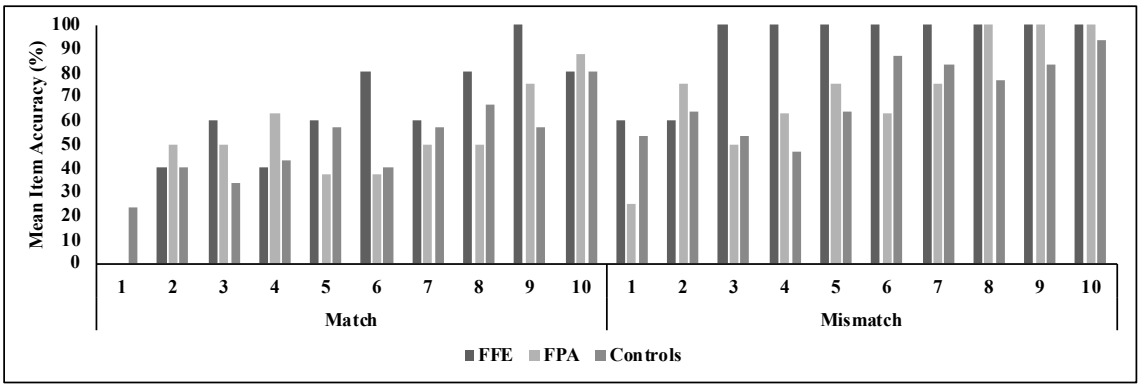

Figure 3. A comparison of mean item accuracy (%) by trial type and group for each item in Experiment 1.

The analysis showed a main effect of Trial Type, $F(1, 27) = 72.24$, $p < .001$, partial $\eta^2 = 0.73$, due to more mismatch face pairs being correctly identified ($M = 78.27$, $SD = 21.22$) than match face pairs ($M = 53.30$, $SD = 23.12$). There was no main effect of Group, $F(2, 27) = 1.98$, $p = .16$, partial $\eta^2 = 0.13$, and no interaction between Trial Type and Group, $F(2, 27) = 1.47$, $p = .25$, partial $\eta^2 = 0.10$.

Thus, item responses showed that accuracy was higher in mismatch trials than in match trials. There were no differences in mean item accuracy between the groups, suggesting they each had a similar pattern of responses to each face pair, although Controls were the only group to provide a correct response in Match Trial 1 (Figure 3).

Correlations

To examine any differences in the pattern of responses to each item, a Pearson product-moment correlation coefficient was computed to assess the relationship between the three groups and their mean accuracy score for each item. The results are shown in Table 1. As all correlations between groups were positive and of similar magnitude, this reflected a similar pattern of responses to each item for each group.

Table 1

Comparison of Between-Group Correlations for Mean Item Accuracy in Experiment 1.

	FFE	FPA	Controls
FFE	1.00	.70*	.70*
FPA		1.00	.77*
Controls			1.00

* Significant at the 0.01 level (2-tailed)

Response Times

Viewing and response times (RT) were unrestricted in this experiment and the mean RT for each group was calculated for correct match and correct mismatch trials (Figure 4).

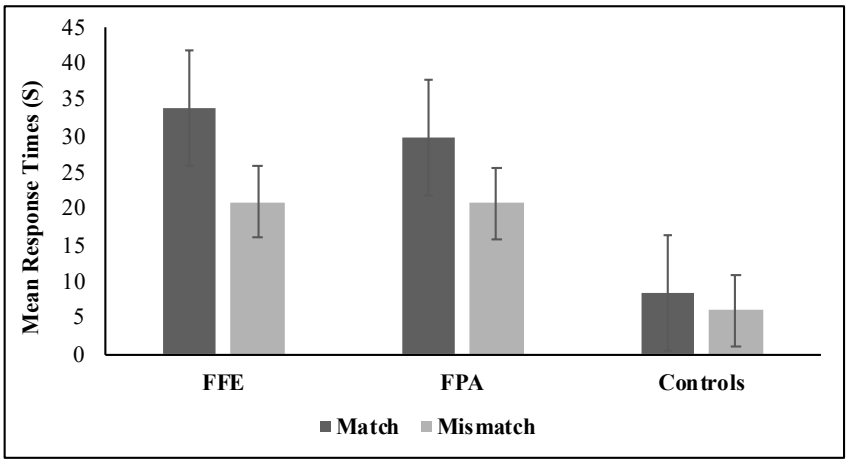

Figure 4. A comparison of mean RT in seconds by group and trial type in Experiment 1. Error bars denote the standard error of the means.

To analyse whether RT differed between groups, a 3 (Group: FFE, FPA and Controls) x 2 (Trial Type: Match vs. Mismatch) mixed-model ANOVA was conducted using the mean response times, with Group as the between-subjects factor and Trial Type as the within-subjects factor. This analysis revealed a main effect of Group, $F(2, 40) = 23.62$, $p < .001$, partial $\eta^2 = 0.54$. Post-hoc Tukey HSD tests showed the mean RT of the Control group (M = 7.27s, SE = 1.43s) was faster than the mean RT of FFEs (M = 27.43s, SE = 3.50s) and FPAs (M = 23.46s, SE = 2.77s). There were no differences in response times between FFEs and FPAs ($p = .65$). The analysis also showed a main effect of Trial Type, $F(1, 40) = 12.95$, p = .001, partial $\eta^2 = 0.25$, due to a quicker RT in mismatch trials (M = 15.97s, SE = 1.61s) than in match trials (M = 22.80s, SE = 2.02s). There was no interaction between Group and Trial Type, $F(2, 40) = 2.68$, $p = .08$ partial $\eta^2 = 0.12$.

Overall, the RT data show that both forensically-trained groups took around three-times longer than the untrained student controls to undertake the face matching tasks, with no difference in response times between FFEs and FPAs.

Individual Performance

As a group, FFEs were more accurate than the comparison groups. To determine whether this accuracy advantage reflected individual performance, the mean accuracy of each expert was compared to the mean accuracy of the FPA and Control groups. Modified *t*-tests for single case comparisons (Crawford & Garthwaite, 2002) were used for this analysis and the results are shown in Table 2.

Table 2

Individual Case Analyses Comparing the Overall Mean Accuracy of Individual FFE With the Mean Accuracy of the FPA and Control Groups in Experiment 1.

	Mean Accuracy (%) (*SD*)	**FFE 1**	**FFE 2**	**FFE 3**	**FFE 4**	**FFE 5**
FFE mean accuracy (%)	-	85.00	85.00	65.00	75.00	70.00
Non-expert controls (N = 30)	60 (13.4)					
***t* (29)**	-	1.84	1.84	0.37	1.10	0.73
***p* (one-tailed)**	-	0.04	0.04	0.36	0.14	0.23
***p* (two-tailed)**	-	0.07	0.07	0.72	0.28	0.47
95% CI	-	[89.67, 99.30]	[89.67, 99.30]	[49.99, 77.05]	[74.37, 94.20]	[63.14, 87.43]
Population below individual's score (%)	-	96.16	96.16	64.19	86.01	76.56
FPA (N = 8)	61 (9.9)					
***t* (7)**	-	2.29	2.29	0.38	1.33	0.86
***p* (one-tailed)**	-	0.03	0.03	0.36	0.11	0.21
***p* (two-tailed)**	-	0.05	0.05	0.72	0.22	0.42
95% CI	-	[83.79, 99.99]	[83.79, 99.99]	[36.98, 86.75]	[65.12, 99.17]	[52.08, 95.76]
Population below individual's score (%)	-	97.19	97.19	64.27	88.79	79.01

For transparency, these results show significance (*p*) values for both one-tailed and two-tailed tests. Given the superior accuracy previously demonstrated by FFEs as a group (e.g. Norrell et al., 2015; White et al., 2015; White et al., 2015a) it could be expected that the mean accuracy of individual examiners would be higher than the mean scores of any non-expert control groups. In this case, a one-tailed test would reflect the likelihood of superior accuracy for individual FFEs. However, the KFMT is designed as a challenging test of ability and the use of a two-tailed test takes into account the possibility that individual FFE may not

perform at a higher level than both control groups. Here, using a one-tailed test showed only FFE 1 and FFE 2 performed better than both control groups. With a two-tailed test, there was no difference in accuracy between any of the individual examiners and the non-expert controls, and differences between FFEs 1 and 2 and the FPA group were only marginally significant.

In summary, single-case analysis showed that not all of the individual examiners performed better than both control groups. High accuracy for facial examiners as a group did not therefore reflect high performance by all individuals within the group.

Screen (visual field) bias

Previous research has observed a left visual field bias in which faces presented to the left-hand side of a screen are fixated more frequently than those presented to the right. To examine whether this bias was present during unfamiliar face matching by professionals, the mean percentage of fixations to features on faces displayed to the left side of the screen was combined to produce a score for each group (Figure 5).

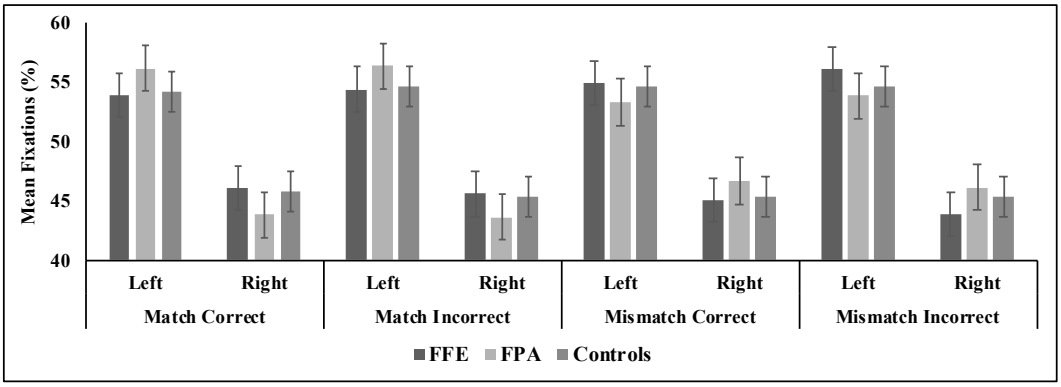

Figure 5. A comparison of mean percentage fixations to faces displayed on the left and right side of the screen by trial type and accuracy for each group in Experiment 1. Error bars denote the standard error of the mean.

As the total number of fixations to faces on the left and right sides of the screen equalled one hundred percent, a one-sample *t*-test was used to compare the percentage of fixations to the faces on the left against a test value of fifty percent in correct and incorrect match and mismatch trials. The results of the analysis are shown in Table 3.

Table 3

Comparison of Scores from One-Sample T-tests Comparing the Mean Percentage of Fixations to Faces Displayed on the Left of the Screen by Trial Type and Accuracy for Groups in Experiment 1.

Trial Type	FFE	FPA	Controls
Correct Match (*t*)	2.87**	4.71**	4.47*
Mean (*SD*)	53.90 (3.04)	56.17 (3.50)	54.17 (5.10)
Incorrect Match (*t*)	3.36**	4.56**	6.37*
Mean (*SD*)	54.41 (2.94)	56.38 (3.96)	54.59 (3.88)
Correct Mismatch (*t*)	6.16**	5.28**	5.06*
Mean (*SD*)	54.90 (1.77)	53.30 (1.77)	54.58(4.96)
Incorrect Mismatch (*t*)	1.71	5.41**	3.58**
Mean (*SD*)	56.14 (6.21)	53.87 (1.89)	54.58 (6.76)

* $p < .001$ ** $p < .05$

Positive *t*-values indicated higher mean percentage fixations to faces displayed on the left of the screen rather than faces on the right. Therefore, all groups displayed a left visual field bias in correct and incorrect match trials and correct mismatch trials. There was no difference in the percentage of fixations to faces on the left and right sides of the screen by FFEs during incorrect mismatch trials ($p = .23$). This reflects a small sample size as two FFEs did not make any errors during mismatch trials and the *t*-test was based on data from only three participants.

In summary, these results provide converging evidence in support of previous research findings of a left visual field bias towards faces presented to the left-hand side of a screen. It was observed in all groups and shows that facial comparison experts did not devote equal attention to the faces displayed on both sides of the screen.

Fixations to Eyes, Nose and Mouth

Research has found that eyes, nose and mouth regions typically receive the most attention when viewing a face. To examine whether this viewing pattern was adopted by professionals during unfamiliar face matching, the percentage of eye movements to these regions was calculated for the three groups (Figure 6). 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.

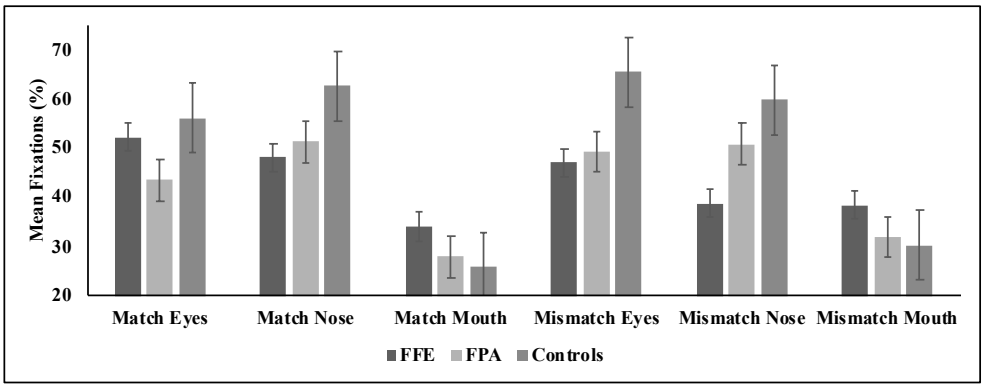

Figure 6. A comparison of mean percentage fixations to eyes, nose and mouth by trial type and groups in Experiment 1. Error bars denote the standard error of the mean.

To compare the percentage of fixations to these features by groups across match and mismatch trials a 3 (Group: FFE, FPA and Controls) x 2 (Trial Type: Match or Mismatch) x 3 (Feature: Eyes, Nose, Mouth) mixed-model ANOVA was used, with Trial Type and Feature as the within-subjects factors and Group as the between-subjects factor. This analysis showed a main effect of Group, $F(2, 40) = 5.64$, $p = .007$, partial $\eta^2 = 0.22$. Post-hoc analysis

with Tukey HSD showed this was due to more fixations to the eyes, nose and mouth by Controls (M = 49.99, SE = 1.20) than by FPAs (M = 42.42, SE = 2.33). There was no difference between FFEs (M = 43.03, SE = 2.95) and either FPAs or Controls. There was no main effect of Trial, $F(1, 40) = .39$, $p = .54$, partial $\eta^2 = 0.01$. There was a main effect of Feature, $F(2, 80) = 8.62$, $p < .001$, partial $\eta^2 = 0.18$, due to fewer fixations to the mouth region (M = 31.38, SE = 3.30) than to the eyes (M = 52.22, SE = 4.32) or the nose (M = 51.85, SE = 2.95). Differences in the percentages of fixations to the eyes and nose were not significant (p = 1.00). There was no interaction between Trial and Feature, $F(2, 80) = 2.80$, $p = .07$, partial $\eta^2 = 0.07$, or between Trial, Feature and Group, $F(4, 80) = 0.72$, $p = .58$, partial $\eta^2 = 0.04$.

In summary, the analysis of eye movement data for the eyes, nose and mouth revealed more attention was given to these features by the Controls than FPAs, with no differences between FFEs and either of the other groups. This suggests FFEs adopted a similar viewing strategy to both FPAs and Controls during the face matching tasks.

Correlations

To analyse the pattern of fixations to the different face areas, the mean percentage of fixations to each feature was calculated for each group for correct and incorrect match trials (Figure 7) and correct and incorrect mismatch trials (Figure 8).

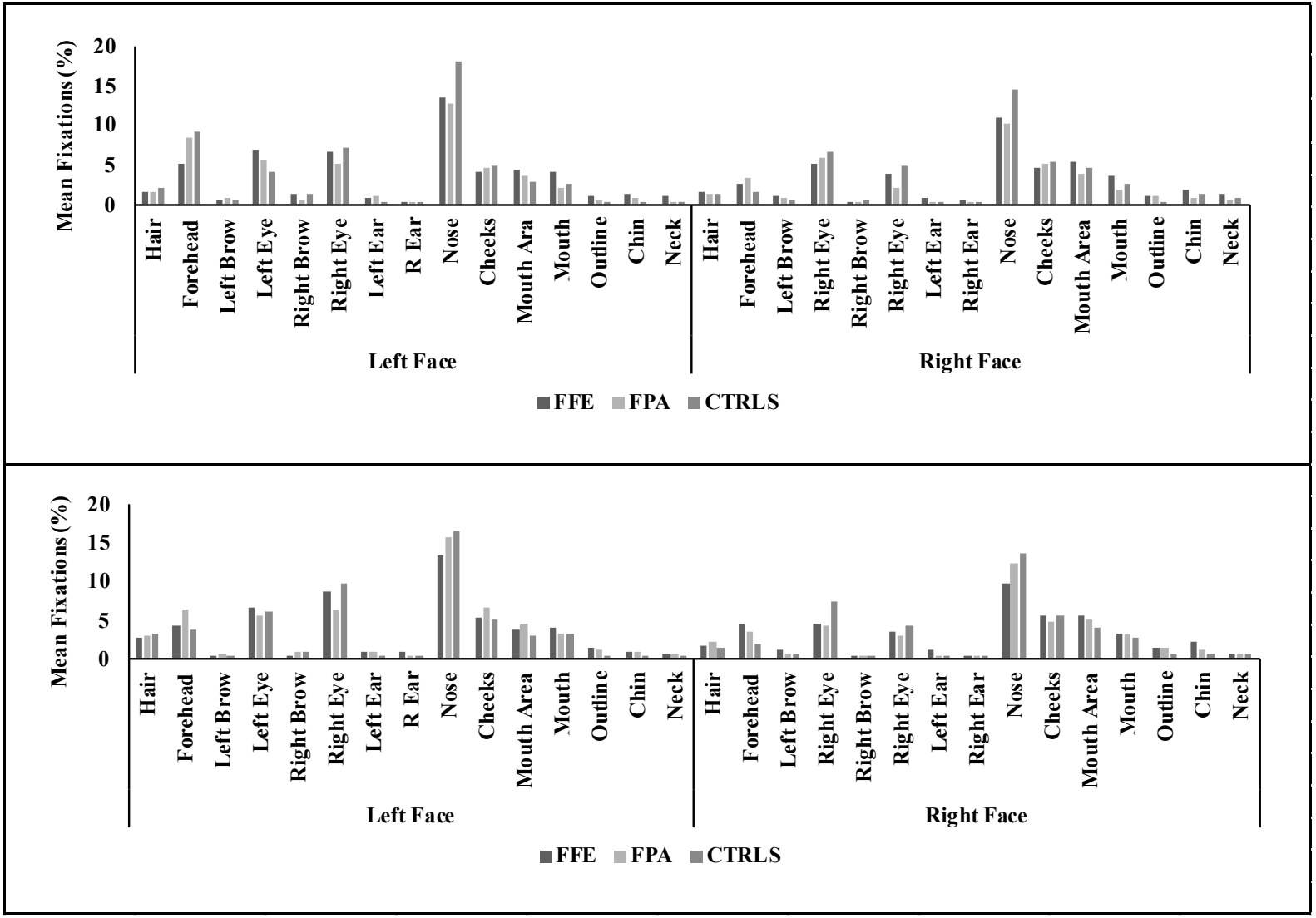

Figure 7. Comparison of mean percentage fixations to each feature by groups in correct match trials (top) and incorrect match trials (bottom).

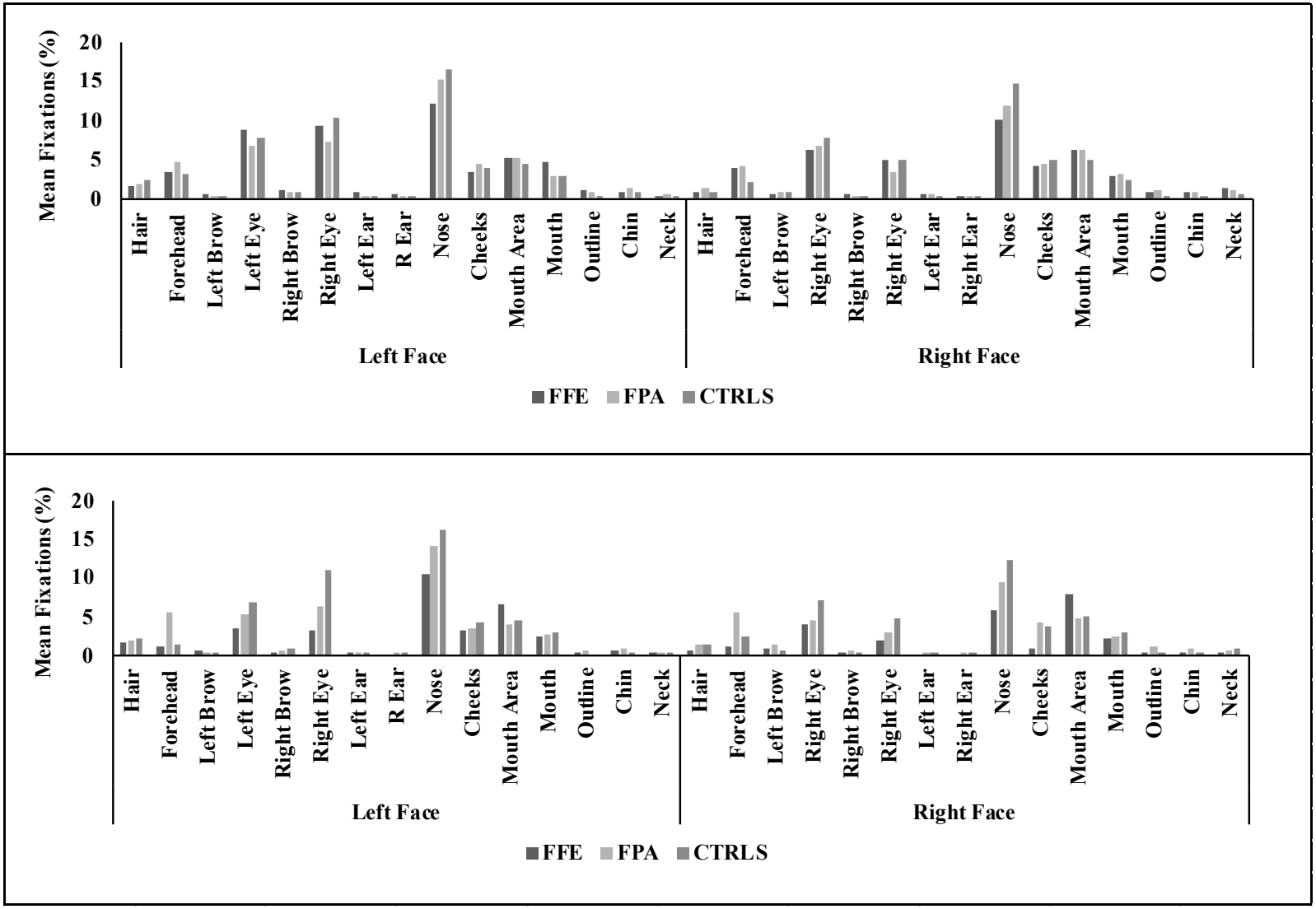

Figure 8. Comparison of mean percentage fixations to each feature by groups in correct mismatch trials (top) and incorrect mismatch trials (bottom).

The mean percentages were used to compute a Pearson product-moment correlation coefficient which assessed the relationship between FFEs, FPAs and Controls and their fixations to each feature. Correlational analyses were conducted separately for correct and incorrect match and mismatch trials and the results are shown in Table 4.

Table 4

A Comparison of Correlations Between Groups by Trial Type and Accuracy for Mean Percentage Fixations to Each Feature in Experiment 1.

Trial Type	FFE vs. Controls	FFE vs. FPA	Controls vs. FPA
Correct Match	.96	.96	.97
Incorrect Match	.96	.96	.96
Correct Mismatch	.96	.96	.97
Incorrect Mismatch	.84	.84	.93

All r significant at the $p < .001$ level (2-tailed)

In summary, high, positive correlations reflected a similar percentage of fixations to each feature for all three groups. This showed that attention to features followed a similar pattern for both facial experts and non-experts, which suggests that the same features had the same diagnostic value for all groups

Discussion

This experiment provides evidence of an accuracy advantage for FFEs in unfamiliar face matching tasks when comparing their performance to that of forensically trained controls and untrained participants. This supports previous research findings of superior performance by facial comparison experts (e.g. Norrell et al., 2015, Towler et al., 2017, White et al., 2015a). As a group, the FFEs' mean accuracy exceeded the normative KFMT score of 66%, although the comparison groups did not reach this level which suggests this was a particularly challenging test of ability. Therefore, the high mean score for FFEs across these

trials reflects superior face matching accuracy rather than accuracy that is merely better than the poorer performance of the other groups. However, this high level of accuracy was not consistent for FFEs across all trials as comparing responses to each item showed they were not always the most accurate group. In one trial all of the examiners incorrectly identified a matching face pair, although correct identification by some of the non-experts would suggest that FFE errors were not due to item difficulty. As this occurred in the first trial, and item responses showed FFEs' accuracy tended to improve as the experiments progressed, this may reflect a more cautious response by professionals than non-experts in the earlier trials.

Whereas group means reflected superior accuracy by FFEs in this experiment, single case analyses revealed individual differences in performance, with only two of the five examiners performing significantly better than the mean scores for each control group. Although this finding is in line with previous studies which have also observed a wide range in individual face-matching ability among facial comparison professionals (White et al, 2014; White et al, 2015), it does reflect an accuracy advantage that was driven by some high performing FFEs rather than the group as a whole.

The results from the behavioural data suggest the superior accuracy of FFEs in face matching tasks is not due to different response strategies. Positive correlations between groups for by-item accuracy reflect a similar pattern of responses across trials. This is supported by eye movement data in which strong, positive correlations between the number of fixations to each feature suggest all groups deployed comparable, rather than idiosyncratic, eye movements during face matching. In addition, there were no differences between groups for the percentage of fixations to the eyes, nose and mouth regions and all groups demonstrated a bias towards the faces presented to the left of the screen.

Although the results reflect similar viewing behaviours by FFEs and the controls, some differences were observed. The FFE and FPA groups took almost three times as long as

the student group to complete the face matching tasks. Both forensic groups may have equated slower processing speed with greater accuracy, in line with the analytical approach emphasised by their working practices (FISWG, 2012; FSR, 2017) which could account for these differences. Given the comparatively longer response times for FFEs than Controls, coupled with their superior performance as a group, it seems feasible that FFE accuracy may be due to prolonged viewing of the faces prior to making a matching decision. Therefore, equating the available viewing time for all observers will examine the relationship between FFE accuracy and exposure duration in Experiment 2.

Experiment 2

In this experiment participants undertook the same face matching tasks as Experiment 1 using novel face pairs. Eye movements were tracked as previously. The effect of time constraints on accuracy was measured by restricting stimulus presentation time to thirty seconds. FFEs' working practices emphasise a measured approach to facial comparison (FISWG, 2012), although previous research found that peak matching accuracy for non-experts was achieved following 2000ms exposure to faces with no performance gain for unlimited viewing time (Özbek et al., 2011). Therefore, restricting the exposure duration in Experiment 2 was unlikely to compromise face matching accuracy for these controls. Comparing the performance of all groups under equal time constraints would determine whether the accuracy advantage for FFEs in Experiment 1 was dependent on unlimited exposure to the face pairs. A decline in FFE performance in Experiment 2 could suggest a speed-accuracy trade-off.

As eye movement data had identified a left-visual field bias in Experiment 1 this was also examined in Experiment 2. Fixations to the eyes, nose and mouth regions, and the percentage of fixations to each feature, were again compared between groups to identify

whether any differences emerged when viewing time was restricted that were not revealed when viewing time was unlimited.

Method

Participants, Stimuli and Procedure

Experiments 1 and 2 were conducted on the same day and all participants from Experiment 1 participated in Experiment 2 following a short break. The stimuli consisted of twenty 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 thirty seconds rather than being unlimited. As previously, participants were prompted to make a same or different matching decision immediately after viewing the face pairs although the response time was not limited. Eye movements were tracked, as in Experiment 1.

Results

Accuracy by Group

To analyse face matching performance, a mean accuracy score was calculated for each group for match and mismatch trials. This reflected the number of correct trials as a percentage of overall trials (Figure 9).

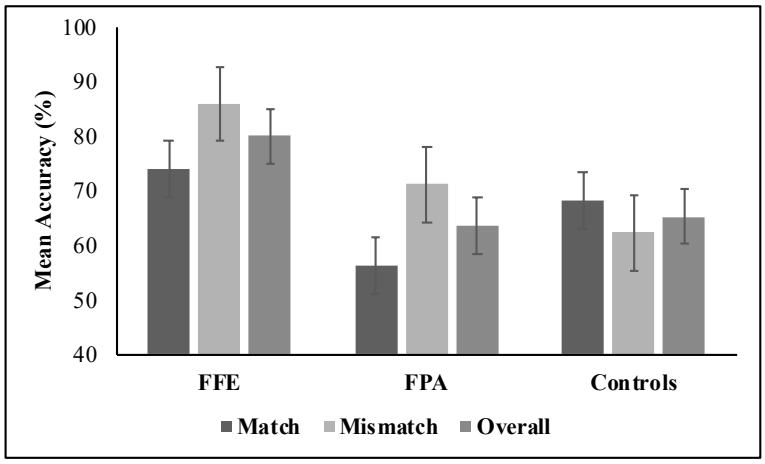

Figure 9. A comparison of mean percentage accuracy by group and trial and overall accuracy in Experiment 2. Error bars denote the standard error of the means.

A 2 (Trial Type: Match vs. Mismatch) x 3 (Group: FFE, FPA and Controls) mixed-model ANOVA was conducted for mean percentage accuracy, with Trial Type as the within-subjects factor and Group the between-subjects factor. The analysis showed a main effect of Group, $F(2, 40) = 5.41$, $p = .008$, partial $\eta^2 = 0.21$. Post-hoc Tukey HSD tests showed that, in terms of overall mean accuracy, FFEs' performance was better ($M = 80.00$, $SE = 4.32$) than the FPAs ($M = 63.75$, $SE = 3.42$) and Controls ($M = 65.33$, $SE = 1.76$), with no differences in accuracy between FPAs and Controls ($p = .91$). There was no main effect of Trial Type, $F(1, 40) = 1.31$, $p = .26$, partial $\eta^2 = 0.03$, and no interaction between Group and Trial, $F(2, 40) = 1.92$, $p = .16$, partial $\eta^2 = 0.09$.

As observed in Experiment 1, the overall accuracy of FFEs was superior to the FPAs and Controls type. They again exceeded the KFMT normative score of 66%, and this level was almost attained by the FPA and Control groups.

Accuracy by Item

To compare group responses to each pair of faces presented in the experiment, mean group accuracy was calculated for each item as a percentage of correct responses (Figure 10).

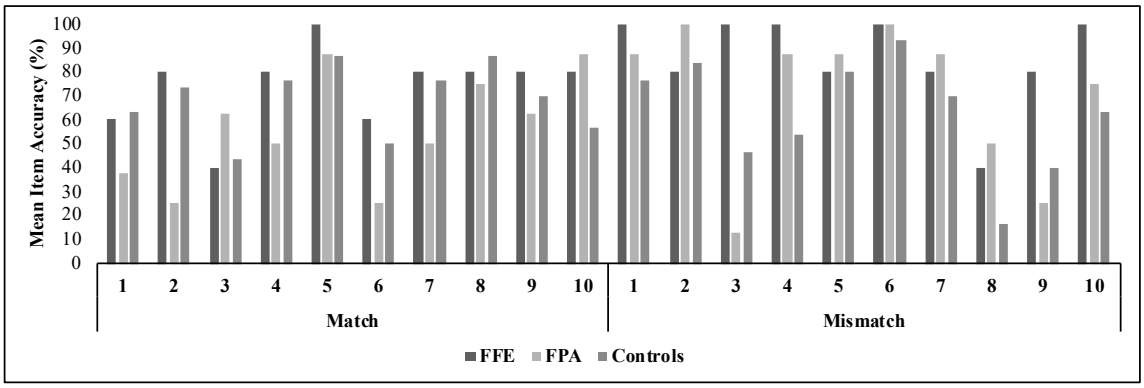

Figure 10. A comparison of mean item accuracy (%) by trial type and group for each face pair in Experiment 2.

A 3 (Group: FFE, FPA and Controls) x 2 (Trial Type: Match vs. Mismatch) mixed-model ANOVA was used to compare the mean item accuracy of each group by type of trial. Trial Type was the within-subjects factor and Group the between-subjects factor. The analysis showed a main effect of Group, $F(2, 27) = 5.11$, $p = .013$, partial $\eta^2 = 0.29$. Post-hoc comparisons using Tukey HSD showed the overall mean item accuracy for FFEs ($M = 80.00$, $SE = 3.93$) was superior to FPAs ($M = 64.00$, $SE = 3.93$) and Controls ($M = 65.35$, $SE = 3.93$). No difference in item accuracy between FPAs and Controls was found ($p = .81$). There was no main effect of Trial Type, $F(1, 27) = 1.11$, $p = .30$, partial $\eta^2 = 0.04$, and no interaction between Group and Trial Type, $F(2, 27) = .989$, $p = .39$, partial $\eta^2 = 0.07$.

In contrast to the results of Experiment 1 which saw no differences in item accuracy between groups, in Experiment 2 the mean item accuracy for FFEs was higher than that of FPAs and Controls. There were no differences in accuracy between FPAs and Controls.

Correlations

To examine any differences in the pattern of responses to each item, a Pearson product-moment correlation coefficient was computed to assess the relationship between the groups and their mean accuracy score for each item. The results are shown in Table 5.

Table 5

Comparison of Between-group Correlations for Mean Item Accuracy in Experiment 2.

	FFE	FPA	Controls
FFE	1.00	.34	.56*
FPA		1.00	.52*
Controls			1.00

* Significant at the 0.01 level (2-tailed)

Positive correlations between FFEs and Controls suggest a similar pattern of responses to each item pair in this experiment, although these correlations were weaker than those observed in Experiment 1. Correlations between FFEs and FPAs were not significant, which reflects lower mean item scores for FPAs in match trials rather than a different pattern of responses for FFEs.

Response Times

For this experiment face pairs were displayed for thirty seconds prior to participants being prompted to make a same or different matching decision, although the actual time allowed to make the response was unlimited. The mean RT for each group was calculated for correct match and mismatch trials and reflected the time taken to make a match decision after the face pair for comparison had been cleared from the screen. (Figure 11)

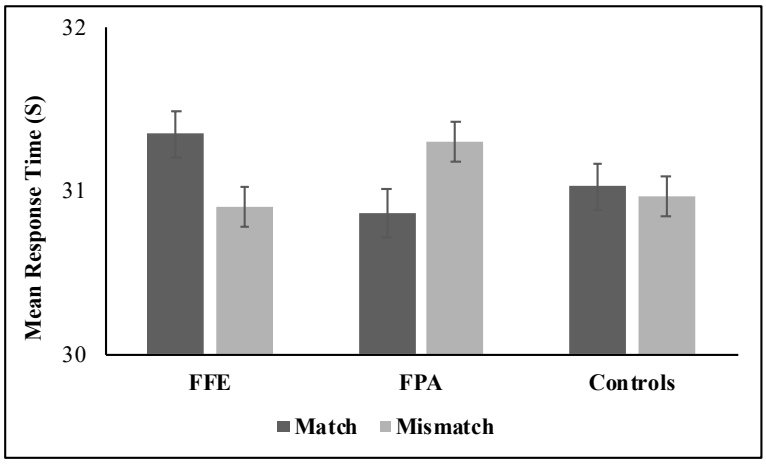

Figure 11. A comparison of mean RT by group and trial type in Experiment 2. Error bars denote the standard error of the mean.

To analyse whether RT differed between groups a 2 (Trial Type: Match vs. Mismatch) x 3 (Group: FFE, FPA and Controls) mixed-model ANOVA was conducted on the mean RT for correct trials, with Trial Type as the within-subjects factor and Group as the between-subjects factor. The analysis showed no main effect of Trial Type, $F(1, 40) = .06$, $p = .80$, partial $\eta^2 = 0.002$, and no main effect of Group, $F(2, 40) = .46$, $p = .63$, partial $\eta^2 = 0.02$, although there was an interaction between Trial Type and Group, $F(2, 40) = 5.63$, $p = .007$, partial $\eta^2 = 0.22$. Simple effects analysis with Tukey HSD showed FFEs responded faster in mismatch trials ($M = 30.90$s, $SE = 0.20$s) than in match trials ($M = 31.35$s, $SE = 0.17$s), whereas FPAs responded faster in match trials ($M = 30.86$s, $SE = 0.13$s) than mismatch trials ($M = 31.30$s, $SE = 0.16$s). For Controls, there was no difference in RT between match and mismatch trials ($p = .49$).

In summary, there were no differences between the groups in the time they took to make a match decision after presentation of the face pairs. Differences in RT by trial type for FFEs and FPAs were of only 0.5s duration. This seems unlikely to represent a meaningful difference in RT in the context of unlimited responses times and may reflect differences in the speed of reading onscreen instructions.

Individual Performance

As in Experiment 1, FFEs as a group were more accurate than FPA and Controls. To determine if this accuracy advantage reflected individual performance the mean accuracy of each FFE was compared to the mean accuracy of the FPA and Control groups. Modified *t*-tests for single case comparisons (Crawford et al., 2002) were used for this analysis and the results are shown in Table 6.

Table 6

Individual Case Analyses Comparing Accuracy of FFE with Mean Accuracy of FPA and Controls in Experiment 2.

	Mean Accuracy (%) (*SD*)	**FFE 1**	**FFE 2**	**FFE 3**	**FFE 4**	**FFE 5**
FFE mean accuracy (%)	-	85.00	85.00	70.00	85.00	75.00
Non-expert controls (N = 30)	65 (10.5)					
***t* (29)**	-	1.88	1.88	0.47	1.88	0.94
***p* (one-tailed)**	-	0.04	0.04	0.32	0.04	0.18
***p* (two-tailed)**	-	0.07	0.07	0.64	0.07	0.36
95% CI	-	[90.21, 99.39]	[90.21, 99.39]	[53.76, 80.25]	[90.21, 99.39]	[69.62, 91.62]
Population below individual's score (%)	-	96.45	96.45	67.85	96.45	82.17
FPA (N = 8)	64 (6.9)					
***t* (7)**	-	2.87	2.87	0.82	2.87	1.50
***p* (one-tailed)**	-	0.01	0.01	0.22	0.01	0.08
***p* (two-tailed)**	-	0.02	0.02	0.44	0.02	0.18
95% CI	-	[90.77, 100]	[90.77, 100]	[50.96, 95.28]	[90.77, 100]	[69.17, 99.59]
Population below individual's score (%)	-	98.80	98.80	78.03	98.80	91.17

As in Experiment 1, these results show significance (*p*) values for both one-tailed and two-tailed tests. In Experiment 2, a one-tailed test showed that FFE 1, FFE2 and FFE 4 all performed better than both control groups. With a two-tailed test, FFE1, FFE 2 and FFE 4 were more accurate than the FPA group, although there were no differences in accuracy between these examiners and the non-expert controls.

In summary, comparing the performance of individual FFEs showed that FFE 1 and FFE 2 both retained their accuracy advantage from Experiment 1. Accuracy of FFE 4

improved from the previous experiment and was now higher than the mean scores of both control groups. The single-case analysis showed there was no difference between the mean scores of two examiners and the control groups. Therefore, superior group accuracy for FFEs was driven by some high performers rather than the group as a whole.

Screen (visual field) bias

To examine whether the left visual field bias identified in Experiment 1 was still present during unfamiliar face matching when viewing time was limited, the mean percentage of fixations to the features on faces displayed to the left side of the screen was combined to produce a score for each group (Figure 12).

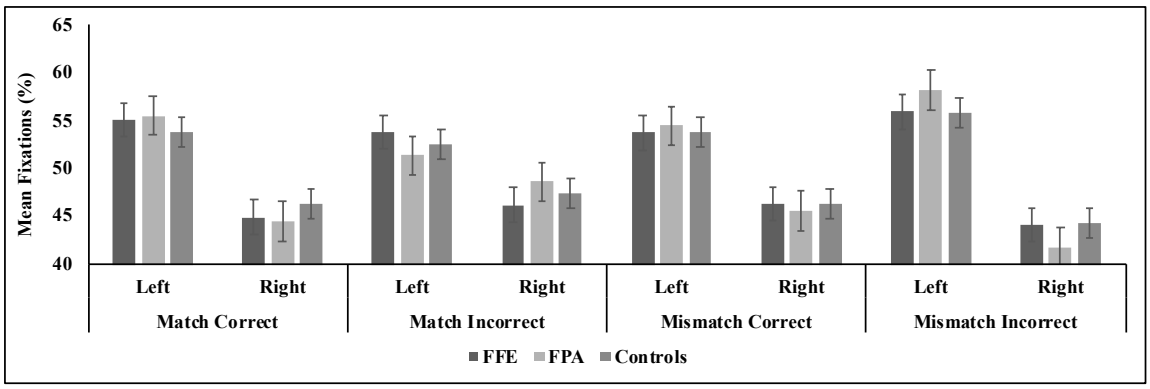

Figure 12. A comparison of mean percentage fixations to faces displayed on the left or right side of the screen by trial type and accuracy for groups in Experiment 2. Error bars denote the standard error of the mean.

As the total number of fixations across faces on the left and right of the screen equalled one hundred percent, a one-sample *t*-test was used to compare the mean percentage of fixations to faces on the left against a test value of fifty percent. This was calculated for all groups across correct and incorrect match and mismatch trials and the results are shown in Table 7

Table 7

Comparison of Scores from One-Sample T-tests Comparing the Mean Percentage of Fixations to Faces Displayed on the Left of the Screen by Trial Type and Accuracy for Groups in Experiment 2.

Trial Type	FFE	FPA	Controls
Correct Match (*t*)	4.53**	6.96*	6.25*
Mean (*SD*)	55.08 (2.51)	55.49 (2.23)	53.78 (3.32)
Incorrect Match (*t*)	1.72	1.27	3.38**
Mean (*SD*)	53.79 (4.92)	51.35 (3.00)	52.58 (3.90)
Correct Mismatch *(t)*	4.08**	6.35*	7.74*
Mean (*SD*)	53.70 (2.02)	54.43 (1.97)	53.70 (2.60)
Incorrect Mismatch *(t)*	3.02**	6.65*	7.62*
Mean (*SD*)	55.90 (4.36)	58.18 (3.25)	55.71 (4.04)

* $p = < .001$ ** $p < .05$

Positive *t*-values reflect higher mean percentage fixations to faces displayed on the left of the screen rather than faces on the right (Figure 11). This suggests a left visual field bias by all groups across all trials, although there was no statistical difference in the percentage of fixations to the left and right faces for FPAs and FFEs during incorrect match trials. This is likely due to smaller sample sizes for these groups, rather than different viewing strategies, as the fixations to the left of the screen were numerically higher than fifty percent for both groups.

Fixations to Eyes, Nose and Mouth

In Experiment 1 there were fewer overall fixations to the mouth region than to the eyes and nose, but there were no differences between the groups in relation to the percentage of fixations to each of these features. To examine whether the same viewing strategy was adopted when exposure to the face pairs was limited to thirty seconds, the mean percentage

of fixations to the eyes, nose and mouth regions was compared across the three groups (Figure 13).

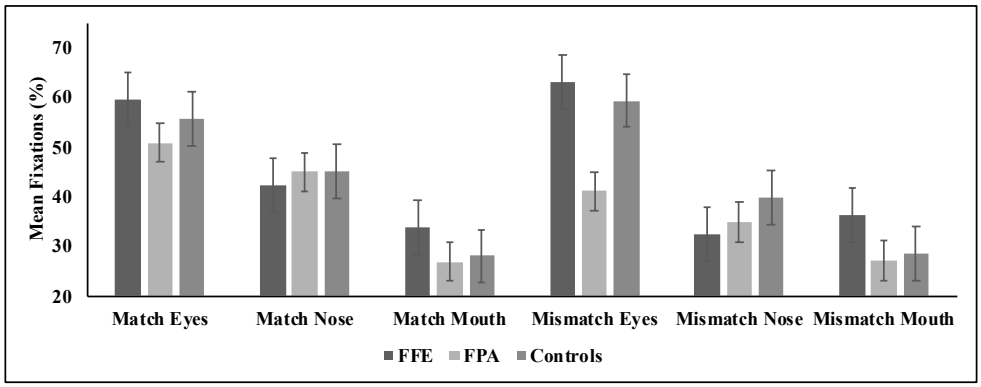

Figure 13. A comparison of the percentages of fixations to the eye, nose and mouth by groups across match and mismatch trials in Experiment 2. Error bars denote the standard error of the mean.

As in Experiment 1, data from the eyes and eyebrows were combined into a single score, as were data from the mouth and mouth region. To compare the mean percentage of fixations to these regions by groups across match and mismatch trials a 3 (Group: Controls, FPA, FFE) x 2 (Trial Type: Match or Mismatch) x 3 (Feature: Eyes, Nose, Mouth) mixed-model ANOVA was used with Trial Type and Feature as the within-subjects factors and Group as the between-subjects factor.

There was no main effect of Group, $F(2, 40) = 3.01$, $p = .06$, partial $\eta^2 = 0.13$, or Trial Type, $F(1, 40) = 1.91$, $p = .17$, partial $\eta^2 = 0.05$, although there was a main effect of Feature, $F(1, 40) = 20.39$, $p = < .001$, partial $\eta^2 = 0.34$. Post hoc comparisons with Tukey HSD showed this was due to the eyes receiving more fixations ($M = 55.03$, $SE = 3.40$) than the mouth ($M = 39.95$, $SE = 1.80$) or nose ($M = 30.15$, SE = 2.17), with the mouth receiving more fixations than the nose. There was no interaction between Trial Type and Group, $F(2, 40) = 1.20$, $p = .31$, partial $\eta^2 = 0.06$, or between Feature and Group, $F(4, 80) = 1.00$, $p = .41$, partial $\eta^2 = 0.05$. There was an interaction between Trial Type and Feature, $F(2, 80) = 9.60$, p

<.001, partial $\eta^2 = 0.19$, and simple effects analysis identified this was due to the nose receiving more fixations during match trials (*M* = 44.16, *SE* = 2.60) than mismatch trials (*M* = 35.72, *SE* = 1.77).

The interaction between Trial Type, Feature and Group was marginally significant, $F(4, 80) = 2.53$, $p = .05$, partial $\eta^2 = 0.11$. This was analysed with a series of 3 x 3 ANOVAs to compare the percentage of fixations by each group to each feature. This identified differences between groups only for the percentage of fixations to the nose within mismatch trials, $F(2, 42) = 4.27$, $p = .02$. Post hoc analysis with Tukey HSD showed this was due to more fixations to the nose by Controls (*M* = 39.83, *SD* = 12.35) than by FPAs (*M* = 34.96, *SD* = 12.35).

In summary, the eyes received the most fixations, followed by the mouth and then the nose. The only difference between the groups was during mismatch trials in which the Controls fixated the nose more frequently than FPAs. Overall, these results show a tendency for all groups to pay similar attention to the eyes, nose and mouth regions during face matching.

Correlations

In Experiment 1, strong positive correlation between all groups reflected a similar pattern of fixations to features by each group. To analyse whether this pattern remained when viewing time was limited, the mean percentage of fixations to each feature was calculated for each group for correct and incorrect match trials (Figure 14) and correct and incorrect mismatch trials (Figure 15).

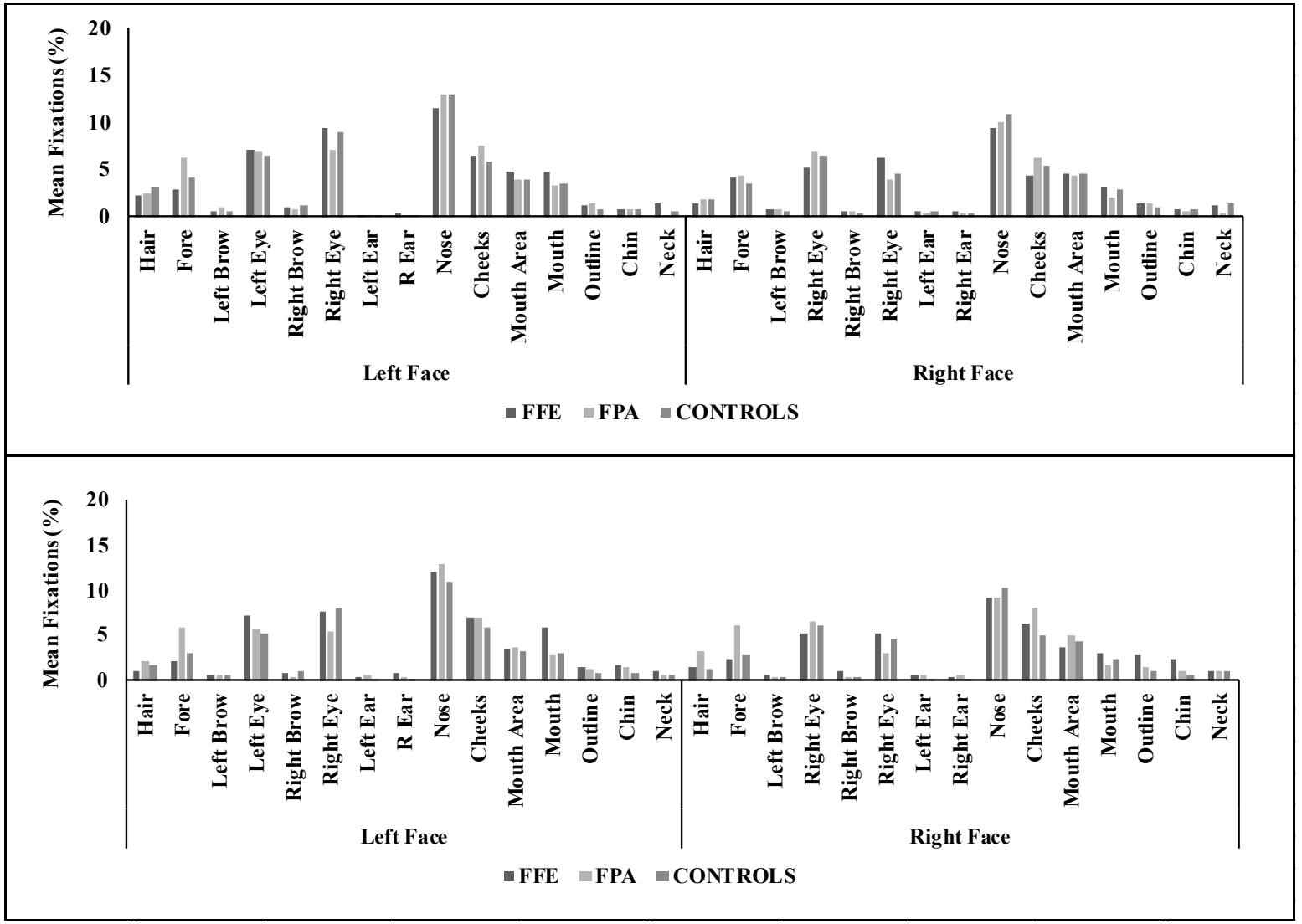

Figure 14. Comparison of mean percentage fixations to each feature by groups in correct match trials (top) and incorrect match trials (bottom) in Experiment 2.

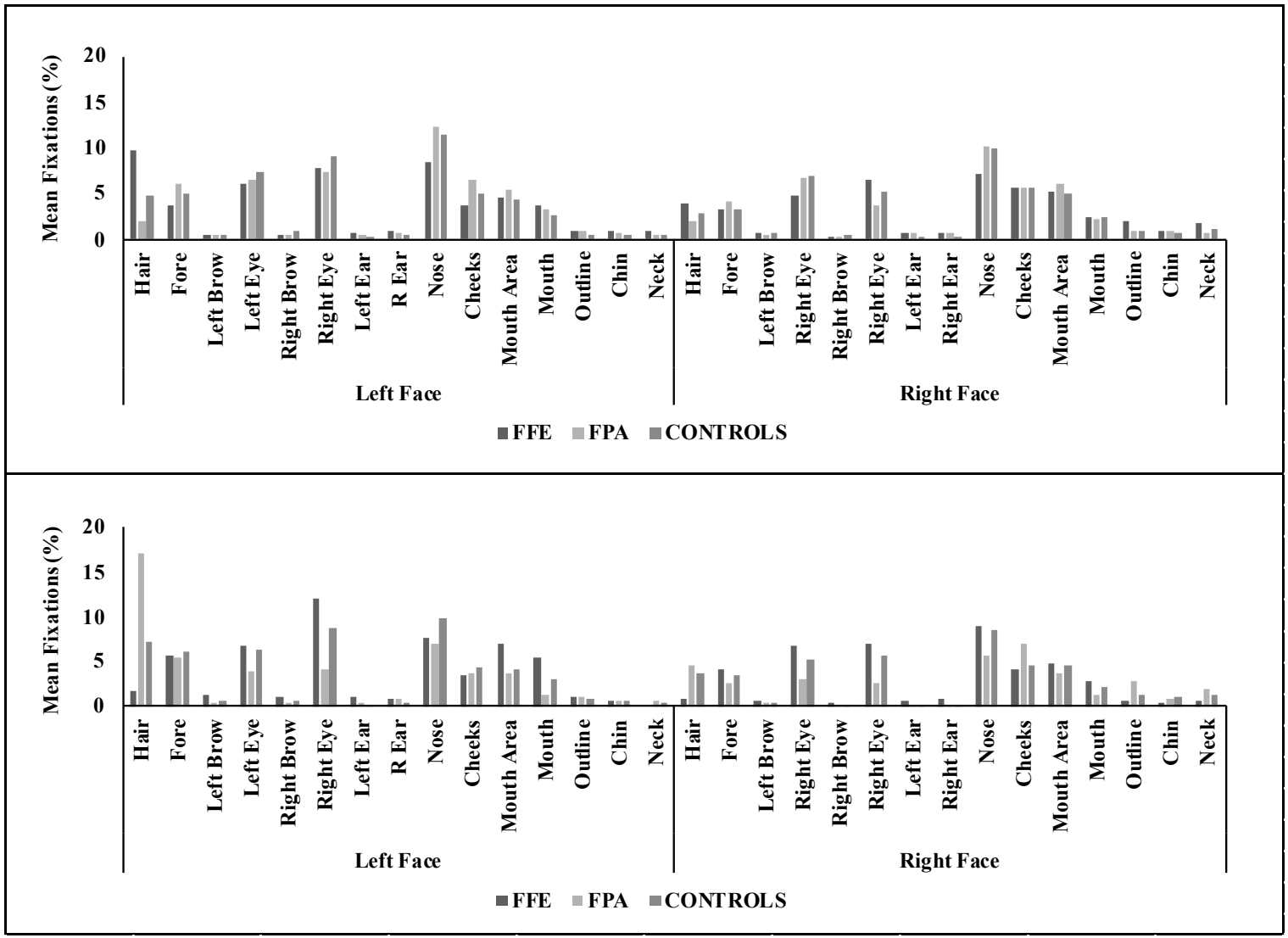

Figure 15. Comparison of mean percentage fixations to each feature by groups in correct mismatch trials (top) and incorrect mismatch trials (bottom) in Experiment 2.

The mean percentage of fixations was used to compute a Pearson product-moment correlation coefficient which assessed the relationship between FFEs, FPAs and Controls and their fixations to each feature. Correlational analyses were conducted separately for trial type and accuracy and the results are shown in Table 8.

Table 8

A Comparison of Correlations Between Groups for Mean Percentage Fixations to Each ROI by Trial Type and Accuracy in Experiment 2.

Trial Type	FFE vs. Controls	FFE vs. FPA	Controls vs. FPA
Correct Match	.97*	.94*	.98*
Incorrect Match	.96*	.88*	.92*
Correct Mismatch	.90*	.79*	.96*
Incorrect Mismatch	.86*	.35	.72*

* r significant at the 0.01 level (2-tailed)

Strong, positive correlations between groups reflected a similar pattern of fixations to each feature in most of the trials. The correlation between FFEs and FPAs was not significant during incorrect mismatch trials. Here, examining mean fixations showed FPAs focused less on the eyes, nose and mouth regions, and more on the hair, than FFEs.

Discussion

In this experiment, FFEs continued to demonstrate a group accuracy advantage over FPAs and Controls in terms of their overall performance. Although the KFMT normative score of 66% was not reached by the comparison groups it was again exceeded by FFEs, thereby demonstrating their superior group performance in unfamiliar face matching. There were individual differences in the accuracy of facial examiners, with only three of five examiners performing better than the control groups. As in Experiment 1, this suggests FFEs'

accuracy is driven by high performers within the group rather than reflecting the performance of the group as a whole.

Although mean item accuracy was higher for FFEs than the other groups, correlational analysis of the data revealed associations between FFEs and Controls which reflected a similar pattern of responses. Converging evidence from eye movement data suggests there were no differences in the viewing behaviours of FFEs during these trials. All groups demonstrated a bias towards faces on the left of the screen and all showed a similar pattern of fixations, and therefore attention, to individual features. However, Experiments 1 and 2 do not necessarily capture the working practices of FFEs as depicted in guidelines for facial image comparison (FISWG, 2012). Therefore, incorporating a feature list from this guidance into the face matching tasks should reflect the real-life routines of FFEs and this may reveal differences in the viewing behaviours of professionals. In turn, requiring FPAs and Controls to adopt similar feature-rating strategies to FFEs may also diminish any accuracy differences between the groups.

Experiment 3

Guidance for forensic facial image professionals (FISWG, 2012) incorporates a list of twelve facial features for comparison across the face pairs. In Experiment 3 this list was displayed on screen, adjacent to the to-be-compared faces. Prior to deciding whether the pair of faces depicted the same person or two different people, participants needed to rate whether each facial feature was the same or different in both faces. The option to rate the feature as "can't compare" was also given to eliminate the risk of guessing, should a feature be obscured or unclear in the presented images.

The aim of this experiment was to more closely reflect the working practices of facial examiners, in so far as they are required to methodically work through and evaluate a list of

facial features prior to making a same or different matching decision. Eye movements were tracked so that fixations to features and any viewing biases could be examined when feature rating was incorporated into unfamiliar face matching. *The analysis of the eye tracking data for this experiment is not included in this thesis but will be completed at a later stage. *

Method

Participants

Experiment 3 was conducted on the same day as Experiments 1 and 2. All previous participants took part in Experiment 3 following a short break.

Stimuli and Procedure

The stimuli in this experiment consisted of twenty new face pairs from the KFMT with equal numbers of identity matches and mismatches. These were displayed onscreen to the right of a feature list (Figure 16), along with instructions to classify each feature as Similar (S), Dissimilar (D) or Can't Compare (C) using the keyboard. Having rated each feature, participants were then required to identify the faces as the Same Identity (S) or Different Identity (D), again using the keyboard. Faces remained onscreen for the duration of the trial, responses were self-paced, and participants could only progress to the next trial when all features had been rated.

As in Experiments 1 and 2, eye movements of the participants were tracked whilst they undertook face matching tasks. The fifteen ROIs relating to the facial features were retained from Experiments 1 and 2 and a further twelve colour codes were used to create additional ROIs for the items on the feature list displayed to the left side of the screen.

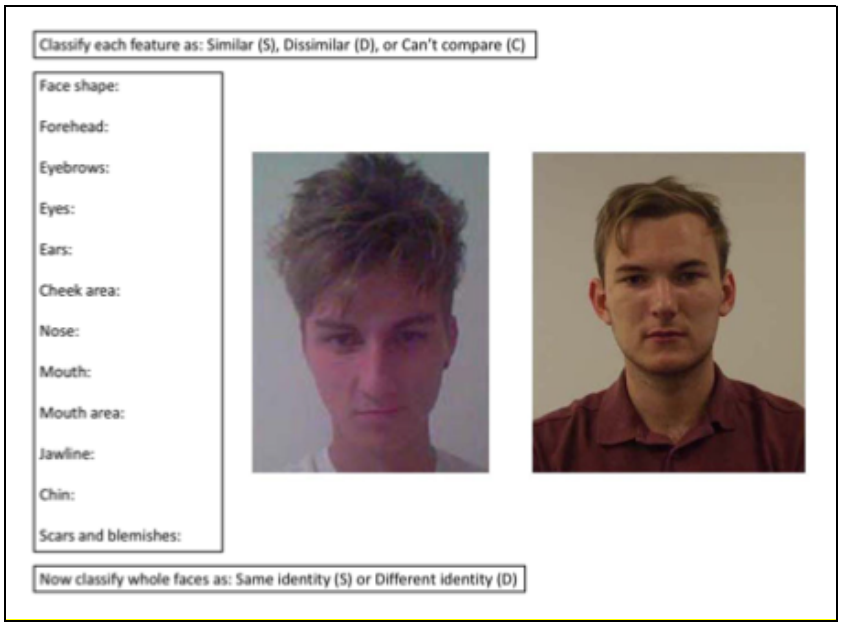

Figure 16. Example of a mismatch face pair from the KFMT with feature list (left) and instructions to participants (top & bottom).

Results

Accuracy by Group

To analyse face matching performance with feature rating, a mean accuracy score was calculated for each participant for match and mismatch trials. This reflected the number of correct trials as a percentage of overall trials (Figure 17).

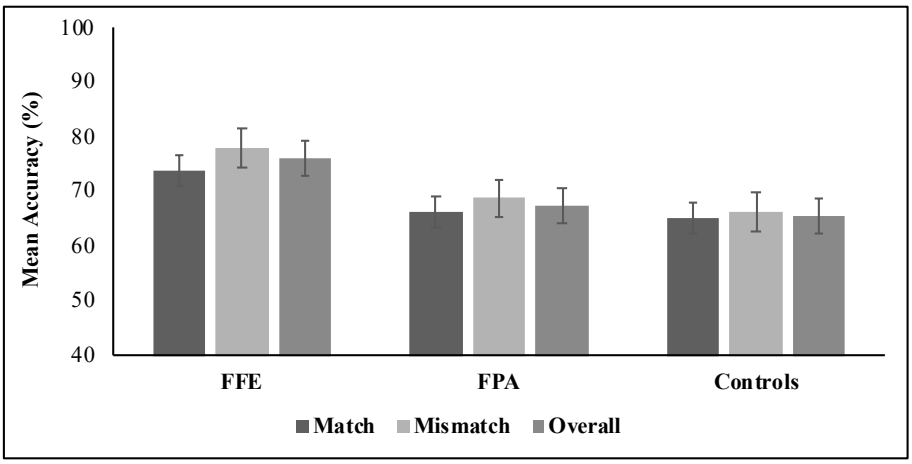

Figure 17. A comparison of mean percentage accuracy by group and trial type in Experiment 3. Error bars denote the standard error of the means.

A 3 (Group: FFE, FPA and Controls) x 2 (Trial Type: Match vs. Mismatch) mixed-model ANOVA was conducted for mean percentage accuracy, with Trial Type as the within-subjects factor and Group the between-subjects factor. The analysis showed there was no main effect of Trial Type, $F(1, 40) = .17$, $p = .68$, partial $\eta^2 = 0.004$, or Group, $F(2, 40) = 2.04$, $p = .14$, partial $\eta^2 = 0.09$, and no interaction between Trial Type and Group, $F(2, 40) = .017$, $p = .98$, partial $\eta^2 = 0.001$.

Therefore, these data show that when required to rate features prior to making a match decision, there were no differences in accuracy between the groups and no differences in accuracy by trial type. Therefore, FFEs did not retain the accuracy advantage they had shown in Experiments 1 and 2.

Accuracy by Item

To compare group responses to each pair of faces presented in the experiment, mean group accuracy was calculated for each item as a percentage of correct responses (Figure 18).

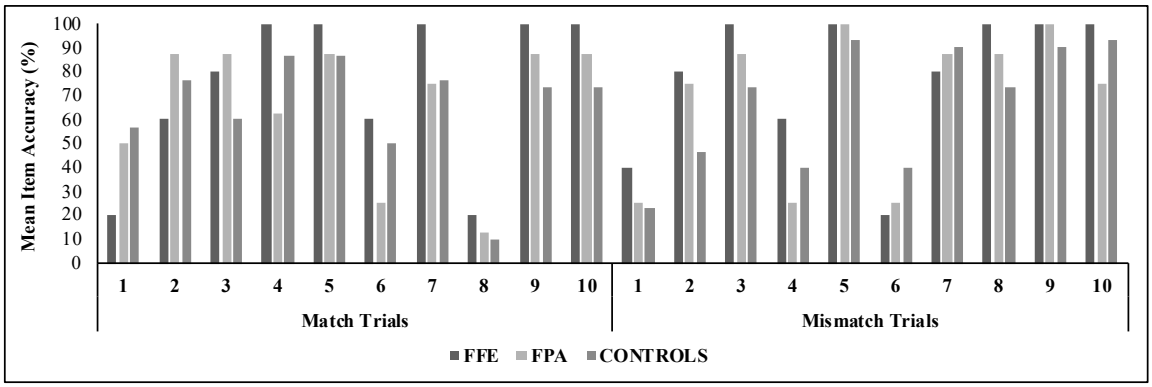

Figure 18. A comparison of mean item accuracy (%) by trial type and group for each face pair in Experiment 3.

A 3 (Group: FFE, FPA and Controls) x 2 (Trial Type: Match vs. Mismatch) mixed-model ANOVA was used to compare the mean item accuracy of each group by type of trial. Trial Type was the within-subjects factor and Group the between-subjects factor. The analysis

showed there was no main effect of Trial Type, $F(1, 27) = .17$, $p = .68$, partial $\eta^2 = 0.006$ and no main effect of Group, $F(2, 27) = .54$, $p = .59$, partial $\eta^2 = 0.04$. There was no interaction between Trial Type and Group, $F(2, 27) = .02$, $p = .98$, partial $\eta^2 = 0.001$.

In contrast to the results from Experiments 1 and 2, FFEs did not demonstrate an accuracy advantage over FPAs and Controls when required to rate the similarity of features prior to the match decision. However, the overall mean accuracy for FFEs was high and still exceeded the KFMT normative score of 66%. In contrast to Experiments 1 and 2, FPAs also performed better than the KFMT normative score and it was equalled by the Controls.

Correlations

To examine any differences in the pattern of responses to each item, a Pearson product-moment correlation coefficient was computed to assess the relationship between groups and mean item accuracy. All correlations were positive and reached statistical significance (Table 9). These strong correlations between all groups reflected similar mean item accuracy scores, and therefore a similar pattern of responses.

Table 9

Comparison of Between-group Correlations for Mean Item Accuracy in Experiment 3

	FFE	FPA	Controls
FFE	1.00	.80*	.79*
FPA		1.00	.84*
Controls			1.00

* Significant at the 0.01 level (2-tailed)

Response Times

In this experiment, participants were required to rate features prior to making a same or different matching decision, with self-paced viewing and response times. The mean RT for each group was calculated for correct match and mismatch trials (Figure 19).

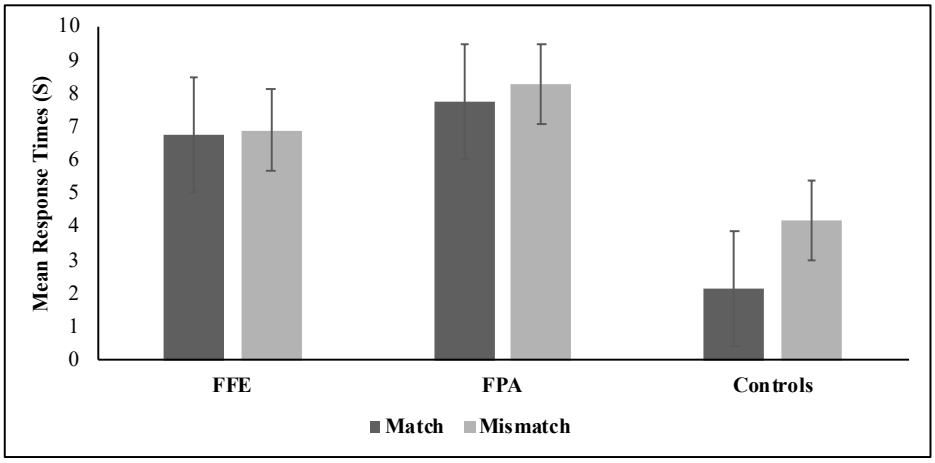

Figure 19. A comparison of mean RT in seconds by group and trial type in Experiment 3. Error bars reflect the standard error of the mean.

To analyse whether RT differed between groups a 3 (Group: FFE, FPA and Controls) x 2 (Trial Type: Match vs. Mismatch) mixed-model ANOVA was conducted on mean RT for correct trials, with Trial Type as the within-subjects factor and Group as the between-subjects factor.

The analysis showed a main effect of Group, $F(2, 40) = 7.64$, $p = .002$, partial $\eta^2 = 0.03$. Post-hoc tests with Tukey HSD showed this was due to Controls ($M = 3.16$s, $SE = 0.62$s) being faster than FPAs ($M = 7.99$s, $SE = 1.21$s). There was no difference between the RT of FFEs ($M = 6.80$s, $SE = 1.53$s) and the RT of FPAs or Controls. There was no main effect of Trial Type, $F(1, 40) = .70$, $p = .41$, partial $\eta^2 = 0.02$, and no interaction between Group and Trial Type, $F(2, 40) = .41$, $p = .67$, partial $\eta^2 = 0.02$.

Therefore, the inclusion of feature rating in the matching task did not reveal any difference in the response times of facial experts and non-experts. The only difference between groups was accounted for by Controls responding more quickly in trials than FPAs.

Individual Performance

There were no differences in group accuracy between FFEs, FPAs and Controls in this experiment. To examine how this reflected the performance of individual FFEs, modified *t*-tests for single case comparisons (Crawford et al., 2010) were used to compare their mean accuracy with that of FPAs and Controls in the same block (Table 10).

Table 10

Individual Case Analyses Comparing Accuracy of FFE with Mean Accuracy of FPAs and Controls in Experiment 3.

	Mean Accuracy (%) (*SD*)	FFE 1	FFE 2	FFE 3	FFE 4	FFE 5
FFE mean accuracy (%)	-	70.00	65.00	80.00	70.00	95.00
Non-expert controls (N = 30)	66 (10.8)					
***t* (29)**	-	0.36	-0.09	1.28	0.36	2.64
***p* (one-tailed)**	-	0.36	0.46	0.11	0.36	0.01
***p* (two-tailed)**	-	0.72	0.93	0.21	0.72	0.01
95% CI	-	[49.89, 76.96]	[32.62, 60.52]	[78.87, 96.23]	[49.89, 76.96]	[97.16, 99.97]
Population below individual's score (%)	-	64.09	46.40	89.38	64.09	99.34
FPA (N = 8)	68 (8.9)					
***t* (7)**	-	0.21	-0.32	1.27	0.21	2.86
***p* (one-tailed)**	-	0.42	0.38	0.12	0.42	0.01
***p* (two-tailed)**	-	0.84	0.76	0.24	0.84	0.02
95% CI	-	[31.37, 82.11]	[14.91, 65.11]	[63.55, 98.94]	[31.37, 82.11]	[52.08, 95.76]
Population below individual's score (%)	-	58.09	38.00	87.79	58.09	98.78

As in previous experiments, the results of both one-tailed and two-tailed tests are shown for transparency. The analysis shows that in this experiment, only FFE 5 was more accurate than the mean scores of the control groups, although this examiner had not outperformed either group previously. FFE 1 and FFE 2 did not retain their accuracy

advantage from Experiments 1 and 2, and FFE 4 did not retain theirs from Experiment 2. Although there were no differences in accuracy between the groups, mean accuracy for FFEs as a group was still high at 76%. Single-case analysis shows only FFE 3 and FFE 5 exceeded this score. Therefore, high average scores for the group do not reflect the performance of all FFEs.

Feature Rating

In Experiment 3, participants were presented with a feature list alongside each face pair for comparison. For each of the twelve features listed, participants rated whether they were the "same" or "different" across each pair of faces. There was also the option to rate the features as "can't compare" if the feature was unclear or obscured, such as the ears being covered by the hair.

For each group the mean percentage of times they rated the features as "same", "different" or "can't compare" was calculated separately for correct and incorrect match trials and correct and incorrect mismatch trials. (Figure 20). This allowed comparison of the mean percentage of ratings between FFEs, FPAs and Controls to identify any differences between types of trial and accuracy.

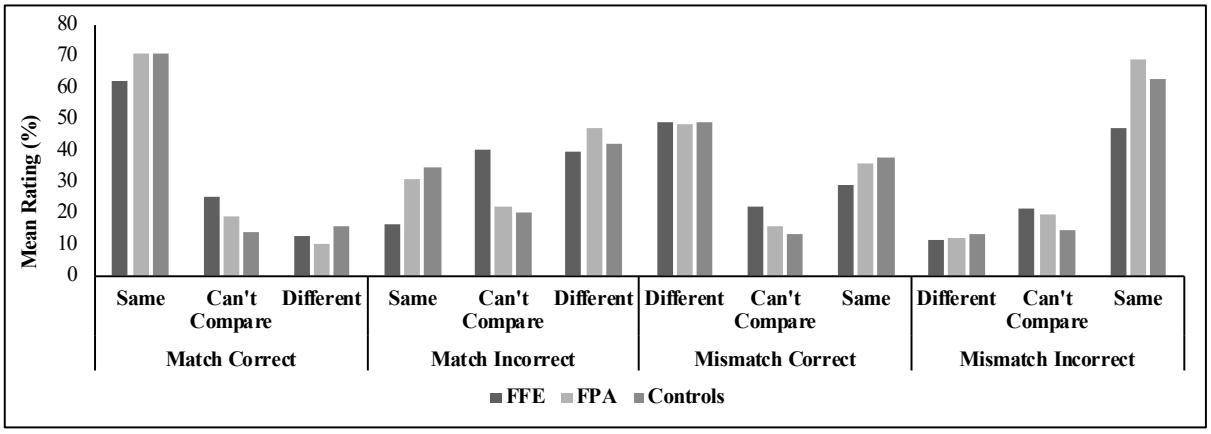

Figure 20. A comparison of the mean percentage ratings of "same", "different" and "can't compare" by trial type and accuracy for groups in Experiment 3.

The first analysis compared “same” and “different” ratings in correct trials using a 3 (Group: FFE, FPA and Controls) x 2 (Trial Type: Match or Mismatch) x 2 (Rating: Same or Different) mixed-model ANOVA. The results are shown in Table 12.

Table 12

Results of Mixed-model ANOVA to Compare “Same” and “Different” Ratings in Correct Trials Across Groups in Experiment 3.

ANOVA	***F* (df)**	***p***	**Partial Eta Sq.**
Group	13.31(2,40)	<.001	0.40
Trial Type	3.96 (1,40)	.05	0.09
Trial Type x Group	1.01 (2,40)	.37	0.05
Rating	48.00 (1,40)	<.001	0.55
Rating x Group	0.69 (2,40)	.51	0.03
Trial Type x Rating	382.98 (1,40)	<.001	0.91
Trial Type x Rating x Group	0.5 (2, 40)	.61	0.02

Post hoc analyses were carried out with, Bonferroni corrections, for main effects and interactions. In correct trials, the “same” and “different” ratings were used less by FFEs (M = 38.24, SE = 0.95) than either FPAs (M = 41.34, SE = 0.75) or Controls (M = 43.39, SE = 0.39), with no differences between FPAs and Controls (p = .08). Across all trials, the “same” rating” was used more (M = 51.09, SE = 1.48) than the rating of “different” (M = 30.82, SE = 1.57). For the interaction between Trial Type and Rating, “same” and “different” ratings were congruent with trial type and accuracy, so a higher percentage of “same” ratings in match trials (M = 67.99, SE = 1.51) than in mismatch trials (M = 34.18, SE = 1.9), and a higher percentage of “different" ratings in mismatch trials (M = 48.84, SE = 2.14) than in match trials (M = 12.81, SE = 1.51).

The second analysis compared "same" and "different" ratings in incorrect trials using a 3 (Group: FFE, FPA and Controls) x 2 (Trial Type: Match or Mismatch) x 2 (Rating: Same or Different) mixed-model ANOVA. The results are shown in Table 13.

Table 13

Results of Mixed-model ANOVA to Compare "Same" and "Different" Ratings in Incorrect Trials Across Groups in Experiment 3.

ANOVA	***F* (df)**	***p***	**Partial Eta Sq.**
Group	4.40 (2, 40)	.02	0.20
Trial Type	.07 (1, 40)	.79	0.00
Trial Type x Group	.07 (2, 40)	.94	0.00
Rating	17.09 (1, 40)	<.001	0.30
Rating x Group	1.26 (2, 40)	.29	0.06
Trial Type x Rating	101.03 (1, 40)	<.001	0.72
Trial Type x Rating x Group	0.95 (2, 40)	.40	0.05

Post hoc analyses were carried out, with Bonferroni corrections, for main effects and interactions. In incorrect trials, the "same" and "different" ratings were used less by FFEs (M = 28.68, SE = 3.14) than either FPAs (M = 39.61, SE = 2.50) or Controls (M = 38.10, SE = 1.30), with no difference between FPAs and Controls (p = 1.00). Across all trials, the rating of "same" (M = 43.37, SE = 2.65) was used more than the rating of "different" (M = 27.56, SE = 2.06). For the interaction between Trial Type and Rating, the use of "same" and "different" were congruent with trial type and accuracy. In incorrect match trials, "same" (M = 27.22, SE = 2.80) was used less than "different" (M = 42.86, SE = 3.33). In incorrect mismatch trials, "same" was used more (M = 59.51, SE = 4.39) than "different" (M = 12.26, SE = 2.04).

The third analysis compared the mean percentage of "can't compare" ratings by groups across all correct and incorrect trials using a 3 (Group: FFE, FPA, Controls) x 2 (Trial

Type: Match or Mismatch) x 2 (Accuracy: Correct or Incorrect) mixed-model ANOVA. The results are shown in Table 14.

Table 14

Results of Mixed-model ANOVA to Compare "Can't Compare" Ratings in Correct and Incorrect Match and Mismatch Trials Across Groups in Experiment 3.

ANOVA	***F* (df)**	***p***	**Partial Eta Sq.**
Group	12.42 (2, 40)	<.001	0.38
Trial Type	22.22 (1, 40)	<.001	0.36
Trial Type x Group	3.74 (2, 40)	.03	0.16
Accuracy	11.21 (1, 40)	.002	0.22
Accuracy x Group	.51 (2, 40)	.61	0.03
Trial Type x Accuracy	7.44 (1, 40)	.009	0.16
Trial Type x Accuracy x Group	2.15 (2, 40)	.13	0.10

Post hoc analyses were carried out, with Bonferroni corrections, for main effects and interactions. The rating of "can't compare" was used more by FFEs (M = 27.06, SE = 2.24) than either FPAs (M = 19.05, SE = 1.77), or Controls (M = 15.29, SE = 0.92), with no difference between FPAs and Controls (p = .20). "Can't compare" was used more in match trials (M = 23.31, SE = 1.33) than in mismatch trials (M = 17.62, SE = 0.98), and used more in incorrect trials (M = 22.83, SE = 1.51) than in correct trials (M = 18.11, SE = 0.85).

For the interaction between Trial Type and Group, in match trials FFEs used "can't compare" more (M = 32.54, SE = 2.97) than either FPAs (M = 20.55, SE = 2.35), or Controls (M = 16.85, SE = 1.21). In mismatch trials, it was used by FFEs (M = 21.58, SE = 2.21) more than Controls (M = 13.73, SE = 0.90) but not more than FPAs (M = 20.55, SE = 2.35). FFEs and Controls used "can't compare" more often in match trials than mismatch trials, whereas there were no differences between trial types for FPAs. For the interaction between Trial Type and Accuracy, "can't compare" was used more in incorrect match trials, (M = 27.42, SE

= 2.10) than correct match trials (*M* = 19.21, *SE* = 1.04), with no difference between correct and incorrect mismatch trials. In incorrect match trials (*M* = 27.42, *SE* = 2.10) the use of "can't compare" was higher than in incorrect mismatch trials (*M* = 18.24, *SE* = 1.62).

In summary, the rating of faces as either "same" or "different" was congruent with trial type and accuracy. Faces were more often rated as "same" in correct match trials and "different" in correct mismatch trials, and more often rated "different" in incorrect match trials and "same" in incorrect mismatch trials. The "can't compare" rating was used more in incorrect match trials than correct match trials, with no difference between correct and incorrect mismatch trials. Differences emerged between the groups in the use of ratings, with FFEs using "same" and "different" less than the other groups across all trials. This was accounted for FFEs using the "can't compare" rating more than FPAs and Controls in match trials and more than Controls in mismatch trials.

Rating by Features

The following twelve facial features were rated in this experiment: 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).

For each face pair, participants rated the twelve facial features as being the "same", "different" or "can't compare". It was predicted that feature ratings would be congruent with trial type, so a greater percentage of features would be rated "same" in match trials and a greater percentage of features would be rated "different" in mismatch trials. The number of times each feature was rated "same" and "different" by trial type and accuracy was calculated to reflect a mean percentage score for each group (Figure 21). The rating of "can't compare" was also calculated for match and mismatch trials as a mean percentage score.

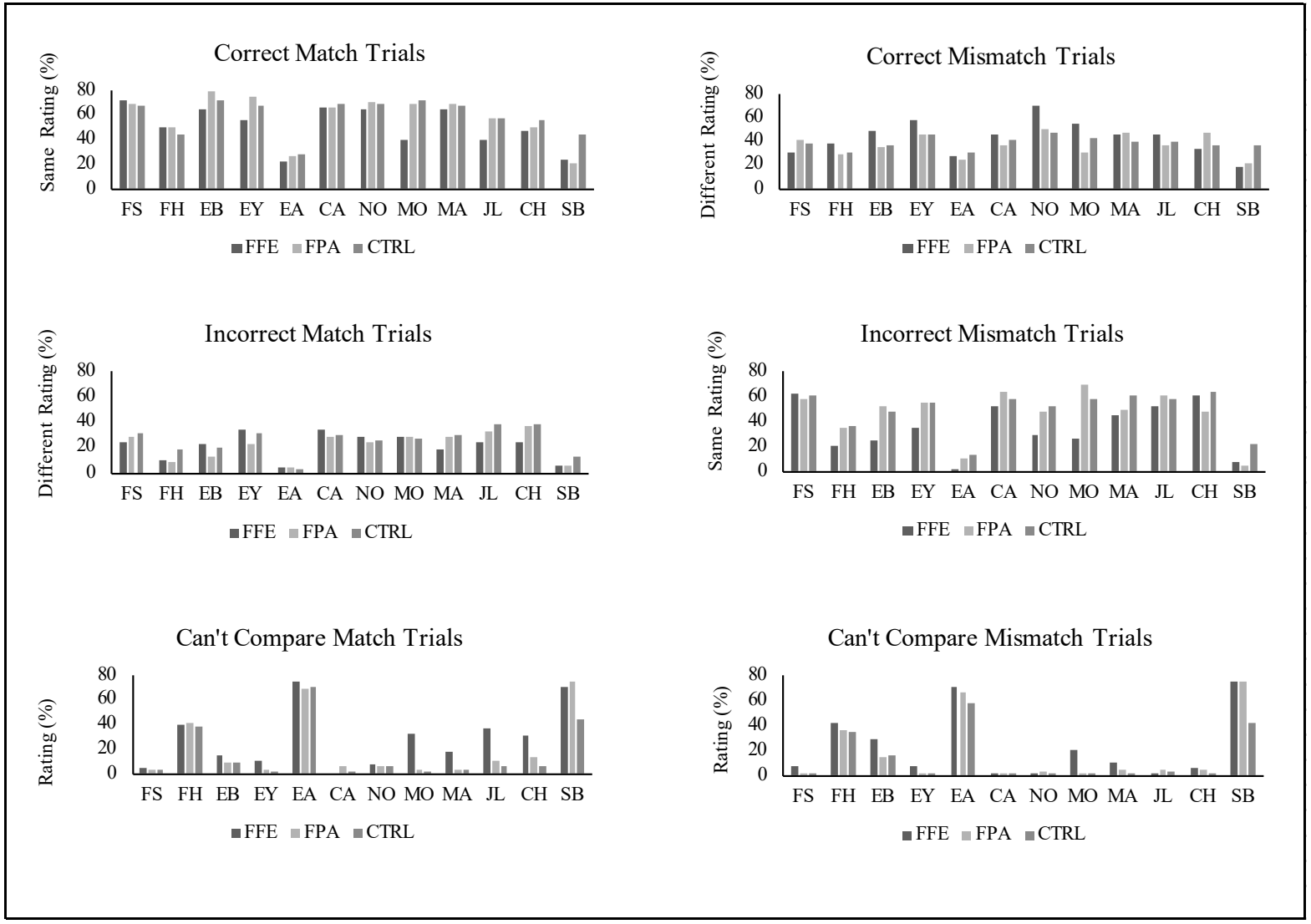

Figure 21. Comparison of group mean percentage ratings of features as "same" and "different" by trial type and accuracy, and the "can't compare" rating by trial type in Experiment 3.

A series of 3 x 12 ANOVAs were used to compare the mean percentage ratings of the twelve features by each group across trials. Separate analyses compared "same" responses in correct match trials, "different" responses in correct mismatch trials, "different" responses in incorrect match trials and "same" responses in incorrect mismatch trials. The results are shown in Table 15. Post-hoc comparisons (with Bonferroni adjustment) are shown where the ANOVAs identified significant differences in the rating of a feature.

Table 15

The Results of 3 x 12 ANOVAs to Compare Mean Percentage Ratings of Features as the "Same" or "Different" by Groups. Post-hoc Comparisons Shown for Significant ANOVAs.

Trial Type & Accuracy	One-Way ANOVA	Feature	Post-Hoc Comparisons (*M, SD*)	
Correct Match	F (2, 27) = 5.33, p = *	Mouth	FFE (40.00, 10.24) FFE (40.00, 10.24) Ctrls (70.67, 4.52)	FPA (68.75, 6.25) ** Ctrls (70.67, 4.52) ** FPA (68.75, 6.25)
Correct Mismatch	Not significant			
Incorrect Match	Not significant			
Incorrect Mismatch	F (2, 27) = 6.20, p = **	Mouth	FFE (26.00, 8.46) FFE (26.00, 8.46) Ctrls (68.75, 10.58)	FPA (68.75, 10.58) * Ctrls (56.98, 7.24) FPA (56.98, 7.24)
	F (2, 27) = 5.44, p = **	Scars/Marks	FFE (8.00, 4.42) FFE (8.00, 4.42) Ctrls (21.68, 14.74)	FPA (3.75, 2.67) Ctrls (21.68, 14.74) FPA (3.75, 2.67) *

* $p < .001$, ** $p < .05$

The analyses showed there were no differences between the groups in correct mismatch and incorrect match trials. In correct match trials, FFEs used the "same" rating for the mouth less than the other groups, with no difference in ratings between FPAs and Controls. In incorrect mismatch trials, FFEs made fewer errors than FPAs in rating the mouth as the "same", with no difference in ratings between FFEs and Controls or between FPAs and Controls. In rating the marks and scars, Controls made more errors than FPAs, with no difference in ratings between FFEs and Controls or between FFEs and FPAs.

Further 3 x 12 ANOVAs were used to compare the mean percentage ratings of the twelve features as "can't compare" by each group across trials. This rating was compared for match and mismatch trials and the results are shown in Table 16. Post-hoc comparisons (with Bonferroni correction) are shown where the ANOVA identified significant differences in the rating of a feature.

Table 16

The Results of 3 x 12 ANOVAs to Compare Mean Percentage Ratings of Features as "Can't Compare" by Groups. Post-hoc Comparisons are Shown for Significant ANOVAs.

Trial Type & Rating	One-Way ANOVA	Feature	Post-Hoc Comparisons (*M, SD*)	
Match Can't Compare	F (2, 27) = 3.66, p = **	Cheek Area	FFE (0.00, 0.00)	FPA (6.25, 2.80) **
			FFE (0.00, 0.00)	Ctrls (1.99, 0.74)
			Ctrls (1.99, 0.74)	FPA (6.25, 2.80)
	F (2, 27) = 14.35, p = *	Mouth	FFE (32.00, 7.42)	FPA (3.75, 1.91) *
			FFE (32.00, 7.42)	Ctrls (1.99, 0.74) *
			Ctrls (1.99, 0.74)	FPA (3.75, 1.91)
	F (2, 27) = 5.13, p = **	Mouth Area	FFE (18.00, 5.54)	FPA (3.75, 2.67) **
			FFE (18.00, 5.54)	Ctrls (3.65, 1.44) **
			Ctrls (3.65, 1.44)	FPA (3.75, 2.67)
	F (2, 27) = 7.91, p = **	Jawline	FFE (36.00, 9.33)	FPA (10.00, 3.63) *
			FFE (36.00, 9.33)	Ctrls (5.33, 1.81) *
			Ctrls (5.33, 1.81)	FPA (10.00, 3.63)
Mismatch Can't Compare	F (2, 27) = 5.08, p = **	Mouth	FFE (20.00, 8.43)	FPA (1.25, 1.25) **
			FFE (20.00, 8.433)	Ctrls (0.33, 0.33) **
			Ctrls(0.33, 0.33)	FPA (1.25, 1.25)
	F (2, 27) = 6.72, p = **	Scars/Marks	FFE (74.00, 8.97)	FPA (75.00, 7.91)
			FFE (74.00, 8.97)	Ctrls (41.33, 4.56) *
			Ctrls (41.33, 4.56)	FPA (75.00, 7.91) *

* $p < .001$, ** $p < .05$

In match trials, the analyses show FFE used "can't compare" more than the other groups when rating the mouth, mouth area and jawline. For these features, there were no differences in ratings between FPA and Controls. FFE did not use "can't compare" at all for the cheek area, although it was used by the other groups. In mismatch trials, FFE rated the

mouth as “can’t compare” more than the other groups and rated scars/marks as “can’t compare” more than the Controls. There were no differences between FFE and FPA in the rating of scars/marks, but FPA used this rating more than Controls.

Discussion

In this experiment FFEs did not retain their accuracy advantage from Experiments 1 and 2, although they again exceeded the KFMT normative score of 66% which suggests a high level of ability. The accuracy of both control groups improved from previous experiments, with FPAs’ accuracy above the KFMT normative score and Controls matching it for the first time in this study. Comparing item accuracy revealed no differences between the groups and high correlations reflected a similar pattern of responses to each item by each group. There were group differences in RT, which were due to faster responses by the Controls rather than different RTs for FFEs. Analysis of accuracy data therefore revealed no quantitative differences in the performance of FFE as a group when compared to the non-expert groups.

Single-case analyses showed the performance of individual FFEs declined from the previous experiments, with only one of the five examiners more accurate than the mean scores of the other groups. The previous accuracy advantage for two of the FFEs diminished in this experiment, and the only FFE to outperform the control groups had not done so previously. Although the mean group score for FFEs was high at 76%, only two of the five FFEs scored above this level. Therefore, as in Experiments 1 and 2, high group accuracy was driven by some high performers and did not reflect the performance of the whole group.

In relation to the feature rating task, the use of “same” or “different” ratings were congruent with trial type and accuracy. There were differences between the groups in the use of ratings, with FFEs using “same” and “different” less than the control groups. This was

accounted for by FFEs using the "can't compare" rating more often in match trials, and more often than Controls in mismatch trials.

Of the twelve features to be rated, differences between the groups were confined to ratings for the mouth, mouth area, cheek area, scars and blemishes and jawline. In correct match trials, FFE rated the mouth of each face pair as the same less than the other groups. There were no differences between the groups in the rating of features in incorrect match trials and correct mismatch trials. In incorrect trials, differences emerged in the rating of marks and scars, due to more errors by Controls than FPAs, and in the rating of the mouth there were fewer errors by FFEs than FPAs. FFEs showed a more cautious approach to rating the mouth and mouth area, noticeably in mismatch trials, by using the "can't compare" rating more than the other groups. In match trials, FFEs used "can't compare" more than the other groups when rating the jawline, although they did not use this rating at all for the cheek area.

Greater use of the "can't compare" rating by FFEs in this experiment may reflect a more cautious response as a result of the highly accountable nature of forensic facial image comparison rather than a difference in their processing strategies. As FFEs' accuracy was not superior to the other groups in this experiment, it is difficult to conclude that this reflects a qualitatively different approach to face matching by professionals.

Accuracy Across Experiments

Group Accuracy

To provide a clearer picture of the face matching accuracy of FFEs, their performance across all experiments was compared to that of FPAs and Controls. The mean accuracy of each group was calculated as a percentage of correct responses in match and mismatch trials in each experiment (Figure 22).

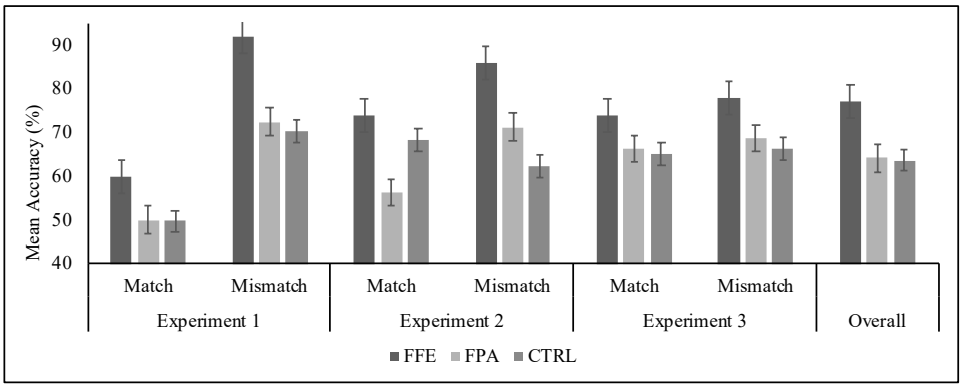

Figure 22. Comparison of group mean accuracy (%) by trial type and experiment and overall group mean accuracy (%) in all experiments.

A 3 (Group: FFE, FPA, Controls) x 2 (Trial Type: Match and Mismatch) x 3 (Experiment:1, 2 and 3) mixed-model ANOVA was used to compare the mean percentage accuracy across the experiments. The results showed a main effect of Group, $F(2, 40) = 6.09$, $p = .005$, partial $\eta^2 = 0.23$. Post-hoc comparisons with Bonferroni corrections showed the overall accuracy of FFEs ($M = 77.33$, $SE = 3.66$) was higher than FPAs ($M = 64.17$, $SE = 2.89$) and Controls ($M = 63.67$, $SE = 1.49$), with no differences in accuracy between FPAs and Controls.

The analysis found no main effect of Trial Type, $F(1, 40) = 3.86$, $p = .06$, partial $\eta^2 = 0.09$, and no interaction between Trial Type and Group, $F(2, 40) = .44$, $p = .65$, partial $\eta^2 = 0.02$. There was also no main effect of Experiment, $F(2, 80) = 1.63$, $p = .20$, partial $\eta^2 = 0.04$, and no interaction between Experiment and Group, $F(4, 80) = .40$, $p = .81$, partial $\eta^2 = 0.02$. There was an interaction between Trial Type and Experiment, $F(2, 80) = 1.27$, $p = <.001$, partial $\eta^2 = .19$. Simple effects analysis with Bonferroni correction showed this was due to lower mean accuracy in match trials in ($M = 53.22$, $SE = 4.86$) than mismatch trials ($M = 78.28$, $SE = 4.08$) in Experiment 1, although there were no differences in accuracy between match and mismatch trials in Experiments 2 and 3. In match trials, accuracy in Experiment 1 ($M = 53.22$, $SE = 4.86$) was lower than in Experiment 2 ($M = 66.19$, $SE = 3.39$) and

Experiment 3 (M = 68.42, SE = 3.87), with no differences in accuracy between Experiments 2 and 3. In mismatch trials, there were no differences in accuracy between Experiment 1 (M = 78.29, SE = 4.08), Experiment 2 (M = 73.19, SE = 3.83) and Experiment 3 (M = 71.03, SE = 3.77).

The results showed the overall face matching accuracy of FFEs in these experiments was superior to that of FPAs and Controls. Other differences in accuracy were specific to trial types and experiments rather than reflecting any differences between the groups.

Consistency of Group Accuracy

To examine whether group accuracy was consistent across the experiments, the accuracy of each group was compared (Figure 23). This used the overall mean accuracy data from each of the three experiments.

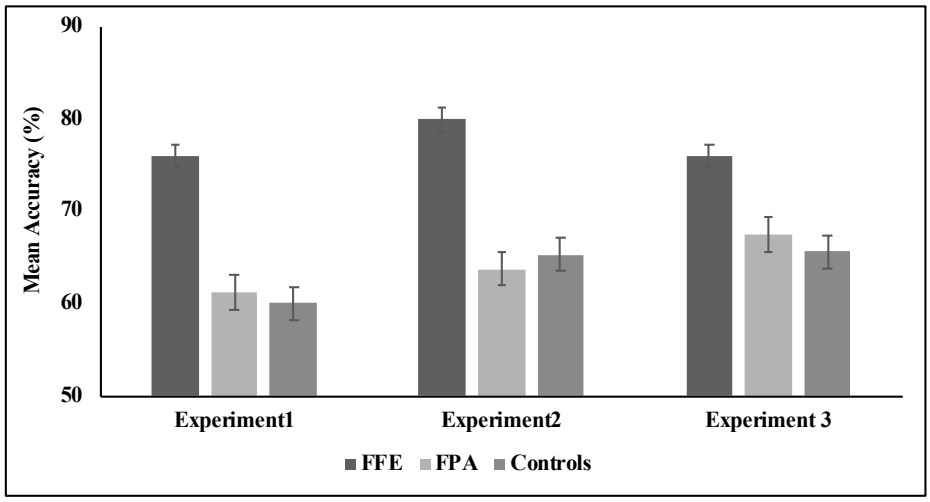

Figure 23. A comparison of overall mean accuracy of each group in each experiment.

This comparison shows that FFEs demonstrated a consistently high level of group accuracy, with peak performance when viewing time was restricted during Experiment 2. For FPAs and Controls, their performance improved with each experiment but did not exceed the

accuracy of FFEs. Although FPAs and Controls were both less accurate than FFEs overall, their accuracy improved in Experiment 3 and reduced the FFEs' accuracy advantage.

Individual Performance

Mean accuracy scores allow comparison of group performances by FFEs, FPAs and Controls but do not show the range of scores for individuals within each group. Figure 24 compares the highest and lowest mean accuracy score within each group in each experiment, together with mean group accuracy.

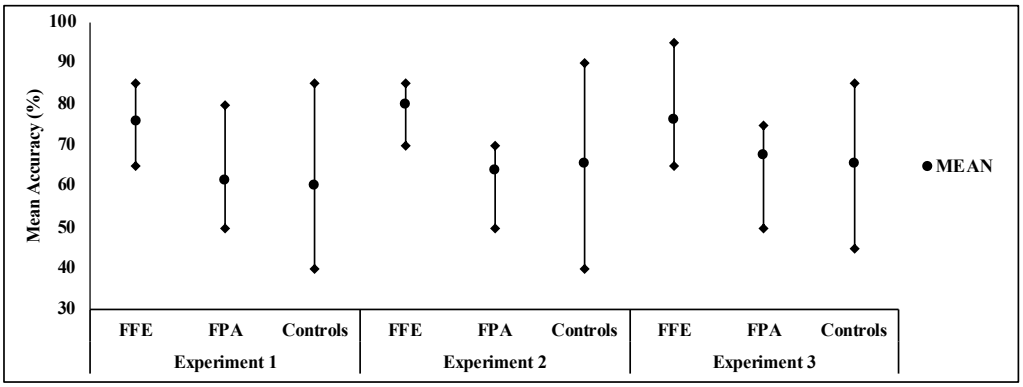

Figure 24. A comparison of the mean group accuracy scores for each experiment, with lowest and highest values to reflect the range of accuracy scores within each group.

This comparison shows a wide range of individual differences in accuracy within each group. In Experiment 1, the accuracy of the best performing Control is equal to the performance of the most accurate FFE. In Experiment 2, the highest performing Control is more accurate than the best performing FFE. Although the overall range of FFEs' accuracy increased in Experiment 3, their highest accuracy scores are not matched by either comparison group. Across all experiments the lowest scores for FFE are considerably higher than the lowest scores of both FPAs and Controls.

In each experiment, single case comparisons have revealed individual differences in the performance of facial examiners. When plotted across all experiments, the mean percentage accuracy scores for individual FFE reveal an inconsistent pattern of accuracy (Figure 25), despite a consistent level of accuracy for the group as a whole.

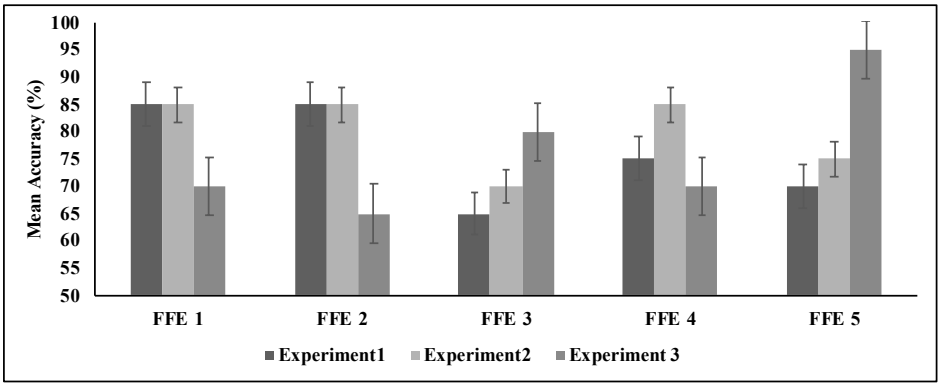

Figure 25. A comparison of the mean percentage accuracy scores for individual FFE in each experiment. Error bars reflect the standard error of the mean.

Professional Experience

Within the FFE group, professional experience in facial comparison ranged from 6 months to 84 months. A Pearson product-moment correlation coefficient was computed to examine whether experience was predictive of accuracy in the face matching tasks. This compared the overall mean accuracy score for each individual with their experience in forensic facial comparison. There was no relationship between experience and accuracy, ($N = 5$, $r = -.06$, $p > .05$). Therefore, professional experience did not predict unfamiliar face matching accuracy as FFEs with more experience did not necessarily perform better than colleagues with less.

In summary, the group performance of FFEs was consistently high across the experiments, although the performance of individual FFEs varied and the same FFE was not the most accurate observer within the group in each experiment. The accuracy of FPAs and

Controls improved with each experiment, with performance gains in Experiment 3 reducing the FFEs' accuracy advantage from Experiments 1 and 2.

General Discussion

Quantitative Differences in Performance

The aim of this research was to examine whether the viewing behaviours of facial comparison professionals were different to those of untrained observers. Face matching ability was first compared to ensure that any qualitative differences in the performance of experts were associated with differences in accuracy. In all three experiments, forensic facial examiners (FFEs) exceeded the normative accuracy score of 66% for the KFMT. This is a lower accuracy score than a comparable test such as the GFMT, where normative data is between 80% to 90%, and therefore reflects a challenging test of ability. Normative data for the KFMT was obtained in self-paced trials with novice participants. In the current study, FFEs' scores were above 66% with restricted viewing times (Experiment 2), smaller images and an additional feature rating task (Experiment 3), and therefore reflected a high level of face matching ability.

FFEs demonstrated a clear accuracy advantage over fingerprint analysts (FPAs) and the student participants (Controls) in the first two experiments. This supports previous research which found that this group of specialists frequently outperform non-specialists in laboratory face matching tests (Norrell et al., 2015; Phillips et al., 2018; Towler et al., 2017; White et al., 2015a; Wilkinson & Evans, 2008). Experiment 3 was designed to reflect some of the working practices of FFEs with the incorporation of feature rating prior to the same or different matching decisions (FISWG, 2102). Here there were no differences in accuracy between the groups, which provides converging evidence that professionals do not consistently perform better than non-professionals in every face matching task (White et al.,

2015a). Although FFEs did not retain their accuracy advantage in Experiment 3, the performance of FPAs and Controls improved. Forcing observers to use a feature-comparison strategy prior to making a same or different matching decision (Towler et al., 2017), or directing participants' attention to specific features (Megreya et al., 2018), have both been found to improve accuracy. It therefore seems feasible that FPAs and Controls benefitted from the feature rating task and thus improved their performance. Because feature rating is a requisite element of forensic facial comparison (FISWG, 2012), its incorporation into the trials did not represent a change to the everyday working practices of FFEs. As they were already demonstrating a high level of accuracy, they may have derived no additional benefit from its inclusion in the task

As FFEs were generally better at matching unfamiliar faces than non-experts, and were not outperformed in any experiment, the accuracy with which FFEs matched each pair of faces was compared to FPAs and Controls. The aim was to identify any differences in their pattern of responses to account for their high performance. However, the analyses revealed that low or high item accuracy tended to reflect the responses of all groups rather than depicting a distinct pattern of responses by FFEs. Their high accuracy was therefore not due to an ability to match pairs of faces that other groups found it difficult to identify.

Although FFEs, FPAs and Controls all demonstrated similar response strategies for each item in the experiment, some differences between the groups did emerge during the feature-rating tasks in Experiment 3. Here, FFEs used the "can't compare" rating more than the other groups. This may reflect, a more conservative approach to face matching than non-experts, in line with the subjective but highly accountable nature of forensic facial comparison. This also supports previous research which found that facial examiners were more likely to provide an inconclusive "did not know" rating than non-experts when identifying an ambiguous face image (Norrell et al., 2015).

In relation to specific facial features, FFEs used "can't compare" more than the other groups when rating the mouth, mouth area and the jawline. As these features are mobile during speech or facial expression, they may provide a less reliable basis for comparison than features with a limited range of movement such as the nose or ears. The only other difference to emerge was in correct match trials. Here, FFEs used the "same" rating for the mouth less than FPAs and Controls, which may also reflect less dependence on a feature which is readily altered by changes in expression. Although FFEs' accuracy was high in Experiment 3, they did not outperform comparison groups who used the "can't compare" rating less often or who used the "same" rating for the mouth more often. It therefore seems unlikely that the high accuracy of FFEs is the result of a more cautious approach to feature rating.

The high performance of FFEs was identified by comparing their mean scores with those of FPAs and Controls. However, a recent observation within the face perception literature is that individual differences in ability can be concealed within average group performances (Balsdon et al, 2018; Lander, Bruce & Bindemann, 2018). This view was supported by the current study which found considerable variation in accuracy between individual FFEs. At best, three examiners were more accurate than both control groups (Experiment 2) and at worst, only one FFE outperformed the other groups (Experiment 3). This provides converging evidence of a range of accuracy for individual facial examiners within superior group performances (Phillips et al., 2018; Towler et al., 2017; White et al., 2015). In addition, individual FFEs were not consistently accurate across all experiments. Superior accuracy in Experiments 1 and 2 did not predict performance in Experiment 3, and the only FFE to outperform the comparison groups in the final experiment had not done so previously. Similar within-person variation in performance has been observed in novice participants across different days (Bindemann et al., 2012) and across different blocks of trials (Alenezi, Bindemann, Fysh & Johnston, 2015). Therefore, variation in the accuracy of

individual FFEs, and inconsistencies in their performance, reflects the range of differences that are also found within the general population (Noyes, Hill & O'Toole, 2018).

Qualitative Differences in Viewing Faces

The analysis of accuracy and feature rating data did not reveal any differences in the viewing behaviours of FFEs which could account for their high level of performance. It was predicted, however, that FFEs' face matching accuracy may be related to viewing duration. As forensic facial comparison is a relatively slow and analytical process (FISWG, 2012) professionals may require longer to effectively process faces than non-experts. Although both forensic groups (FFEs and FPAs) were slower than the student controls in Experiment 1, there were no meaningful differences between the response times of FFEs and the other groups across the three experiments. As FFEs demonstrated superior accuracy when viewing time was limited to thirty seconds, this suggests FFEs' accuracy is not due to a prolonged comparison strategy when making a matching decision.

It was expected that differences in viewing behaviours may emerge when the eye movements of FFEs, FPAs and Controls were compared. Guidelines for forensic facial comparison emphasis the use of a feature comparison strategy (FISWG, 2012), whereas empirical evidence suggests features are processed holistically, all at the same time, for non-experts during face perception (e.g., Goffaux et al., 2006; Tanaka et al., 1993). Therefore, comparing each group's attention to faces and their features may reveal differences in viewing strategies. If FFEs were using a feature-by-feature viewing strategy, they may be expected to focus attention equally on both sides of the screen to compare each face. However, all groups fixated the faces to the left of the screen more than the faces to the right, which supports previous research findings of a left visual field bias in non-expert observers (Butler et al., 2005; Hsiao, 2005; Peterson et al., 2013). This shows that the high face

matching accuracy of FFEs is not due to devoting equal attention to both faces under comparison.

Research has identified differences in the diagnosticity and usefulness of features to inform the face matching decisions of facial examiners (Towler et al., 2017). The analysis of fixations to facial features may therefore reveal differences in the viewing behaviours of professionals. Within the current study, there were no differences between groups in the percentage of fixations to each of these key areas, with the eyes, nose and mouth receiving equal attention by all groups. Therefore, regardless of professional experience, this provides converging evidence that fixations tend to land in these central regions when viewing faces (Arizpe et al., 2017; Or, et al., 2015; Özbek et al., 2011. Certain features were fixated more than others, although this reflected general viewing behaviours by all of the groups rather than distinct strategies by FFEs. Eye movements did not therefore suggest that some features had more diagnostic value than others for FFEs. This contrasts with research by Towler et al. (2017), in which the ears had the highest diagnostic value for facial examiners. However, direct comparison of results cannot be made as the ears were not visible in all of the images in the current study. The analysis of eye movement data in the current study therefore showed a similar pattern of fixations for experts and non-experts during facial comparison.

Factors Affecting Accuracy

As the face matching accuracy of FFEs could not be accounted for by differences in their response strategies and viewing behaviours, perhaps there were other factors which could explain their level of performance. The FFEs had received training in forensic facial comparison, and a recent review observed some benefits in the training course undertaken by this cohort (Towler et al., 2019). Their high accuracy also reflects the existence of a wide range of face matching skills within the general population (Noyes et al., 2018), and natural

talent may have influenced the choice of career for these professionals. However, professional experience did not predict face matching accuracy, which supports the findings of previous research with other facial experts (White et al., 2015). As these experiments were conducted in the workplace, the effects of motivation on performance may also have improved professional accuracy (Moore & Johnston, 2013; White et al., 2015a). However, as researchers have previously observed, the influence of any of these factors on face matching ability is difficult to extract from the accuracy data (White et al., 2015; White et al., 2015a; Towler et al., 2017).

Limitations

It is noted that there were some limitations within the design of the current research in as much as trial order was not randomized for each participant and the experiment order was not counter-balanced. This was done in an effort to minimise differences between individuals that may have arisen from viewing the images in a different sequence or undertaking the experiments in a different order. For example, Experiment 3 was the most demanding in terms of the attention and time required to complete it. If this had been allocated to some participants as their first experiment, their subsequent performance may have been impaired. Maintaining the same order of experiments for all participants therefore reduced the likelihood of carry-over effects. It would also have been impossible to effectively counter-balance the order of experiments with a small group of participants.

Another factor that warrants consideration is the pairing of the unconstrained and optimized target images during the development of the KFMT. For mismatch trials, face pairs were assigned by the experimenters on the basis of similar hair colour, face and eyebrow shape. As mismatch accuracy tended to be higher in all of the current experiments, perhaps visual differences between the face pairs were too distinct and this inadvertently advantaged

observers' mismatch decisions. For this reason, further evaluation of some of the images used in the KFMT may therefore be necessary prior to its use in future research.

Implications and Future Research

Face matching decisions within forensic settings are dependent on subjective assessment, although verification by a second examiner is recommended (FISWG, 2012). Combining the results of face matching decisions has been found to improve accuracy (Towler et al., 2017; White, Burton, Kemp & Jenkins, 2013; White et al., 2015a), but within a high-pressure working environment this may not be practical. In addition, verification assumes a high level of competence in the second examiner and a surprising finding of the current research is that none of the facial examiners correctly identified one of the face pairs in Experiment 1. A similar "wrong" decision by facial examiners in the workplace may result in an erroneous conviction or allow a perpetrator to remain at large. Therefore, accurate forensic identification depends on the performance of each individual facial examiner within a team rather than the performance of the team as a whole.

To remain relevant to applied settings, future experimental designs should continue to reflect working practice wherever possible, albeit within the confines of experimental controls. This will be best achieved through continued collaboration between forensic experts and scientists, both working towards achieving greater understanding of the processes used during facial comparison (Ramon et al, 2019; Robertson et al, 2019). Future studies should focus on the strategies of individual facial examiners rather than those of the group. The current research revealed a range of accuracy and inconsistent performance by individual FFEs. Therefore, identifying differences between facial examiners may reveal comparison strategies that facilitate or hinder face matching. It may prove beneficial to compare the

processing strategies of the most accurate examiners with those of the worst performing examiners to identify differences which could account for superior performance.

Conclusion

In summary, the main findings of this study are that as a group, FFEs performed at a consistently high level of face matching accuracy during three challenging tests of ability. This was achieved regardless of time constraints or task demands such as feature rating. However, high accuracy was not driven by the performance of all FFEs within the group, and none of the examiners displayed consistently superior performance across all experiments. Although FFEs were more conservative in their rating of features than the non-experts, there were no differences in their responses to items or viewing strategies that could account for their high performance. However, when FPAs and Controls were required to use the same feature comparison strategy as FFEs, their accuracy improved. This suggests the high performance of FFEs may be attributed to their methodological comparison of pairs of faces, in line with their working practices (FISWG, 2012), as opposed to there being any qualitative differences in their viewing behaviours.

References

Ackerman, P. L. (1987). Individual differences in skill learning: An integration of psychometric and information processing perspectives. *Psychological Bulletin, 102* (1), 3-27. doi: 10.1037/0033-2909.102.1.3

Alenezi, H. M., Bindemann, M., Fysh, M. C., & Johnston, R. A. (2015). Face matching in a long task: Enforced rest and desk-switching cannot maintain identification accuracy. *PeerJ, 3*, e1184. doi: 10.7717/peerj.1184

Arizpe, J., Walsh, V., Yovel, G., & Baker, C. I. (2017). The categories, frequencies, and stability of idiosyncratic eye-movement patterns to faces. *Vision Research, 141*, 191-203. doi: 10.1016/j.visres.2016.10.013

Balsdon, T., Summersby, S., Kemp, R. I., & White, D. (2018.). Improving face identification with specialist teams. *Cognitive Research: Principles and Implications, 3*(1), 25. doi: 10.1186/s41235-018-0114-7

Barton, J. J. S., Radcliffe, N., Cherkasova, M. V., Edelman, J., & Intriligator, J. M. (2006). Information processing during face recognition: The effects of familiarity, inversion, and morphing on scanning fixations. *Perception, 35*(8), 1089-1105. doi: 10.1068/p5547

Bate, S., Frowd, C., Bennetts, R., Hasshim, N., Murray, E., Bobak, A. K., … Richards, S. (2018). Applied screening tests for the detection of superior face recognition. *Cognitive Research: Principles and Implications, 3*(1), 22. doi: 10.1186/s41235-018-0116-5

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*(3), 277-291. doi: 10.1037/a0029635

Bindemann, M., & Sandford, A. (2011). Me, myself, and I: Different recognition rates for three photo-IDs of the same person. *Perception, 40*(5), 625-627. doi: 10.1068/p7008

Bindemann, M., Scheepers, C., Ferguson, H. J., & Burton, A. M. (2010). Face, body, and center of gravity mediate person detection in natural scenes. *Journal of Experimental Psychology: Human Perception and Performance, 36*(6), 1477-1485. doi: 10.1037/a0019057

Bobak, A. K., Hancock, P. J., & Bate, S. (2016). Super-recognisers in action: Evidence from face-matching and face memory tasks. *Applied Cognitive Psychology, 30*(1), 81-91. doi: 10.1002/acp.3170

Burton, A.M., White, D., & McNeill, A. (2010). The Glasgow Face Matching Test. *Behavior Research Methods, 42*(1), 286-291. doi:10.3758/BRM.42.1.286

Burton, A. M., Wilson, S., Cowan, M., & Bruce, V. (1999). Face recognition in poor-quality video: Evidence from security surveillance. *Psychological Science, 10*(3), 243-248. doi: 10.1.1.841.7871

Butler, S., Gilchrist, I. D., Burt, D. M., Perrett, D. I., Jones, E., & Harvey, M. (2005). Are the perceptual biases found in chimeric face processing reflected in eye-movement patterns? *Neuropsychologia, 43*, 52-59. doi: 10.1016/j.neuropsychologia.2004.06.005

Crawford, J. R., & Garthwaite, P. H. (2002). Investigation of the single case in neuropsychology: confidence limits on the abnormality of test scores and test score differences. *Neuropsychologia, 40*(8), 1196-1208. doi: 10.1016/S0028-3932(01)00224-X

Davis, J. P., Lander, K., Evans, R., & Jansari, A. (2016). Investigating predictors of superior face recognition ability in police super-recognisers. *Applied Cognitive Psychology, 30*(6), 827-840. doi: 10.1002/acp.3260

Edmond, G., Biber, K., Kemp, R., & Porter, G. (2009). Law's looking glass: Expert identification evidence derived from photographic and video images. *Current Issues in Criminal Justice, 20(*3), 337-377. doi: 10.1080/10345329.2009.12035817

Edmond, G., Valentine, T., & Davis, J. P. (2015). *Expert analysis: Facial image comparison*. In T. Valentine & J. P. Davis (Eds), Forensic facial identification: Theory and practice of identification from eyewitnesses, composites and CCTV (pp. 239-262). Chichester, West Sussex: John Wiley & Sons

Facial Identification Scientific Working Group (FISWG) (2012). *Guidelines for facial comparison methods.* Retrieved from https://www.fiswg.org/document/view Document?id=25

Forensic Science Regulator (FSR). (2017). *Codes of Practice and Conduct. Fingerprint Comparison,* (2), 44. Retrieved from https://assets.publishing.service.gov.uk /government /uploads/ system/uploads/attachment_data/file/638254/128_FSR _fingerprint_appendix__Issue2.pdf

Fysh, M. C., & Bindemann, M. (2017). *Forensic face matching: A Review*. In M. Bindemann & A. M. Megreya (Eds.), Face processing: Systems, Disorders and Cultural Differences (pp. 1-20). New York: Nova Science Publishing, Inc.

Fysh, M. C., & Bindemann, M. (2018). The Kent Face Matching Test. *British Journal of Psychology, 109*(2), 219-231. doi: 10.1111/bjop.12260

Goffaux, V., & Rossion, B. (2006). Faces are 'spatial' - holistic face perception is supported by low spatial frequencies. *Journal of Experimental Psychology: Human Perception and Performance, 32*(4), 1023-1039. doi: 10.1037/0096-1523.32.4.1023

Henderson, J. M. (2003). Human gaze control during real-world scene perception. *Trends in Cognitive Sciences, 7*(11). doi: 10.1016/j.tics.2003.09.006

Henderson, J. M. (2007). Regarding scenes. *Current Directions in Psychological Science, 16*(4), 219-222. doi: 10.1111/j.1467-8721.2007.00507.x

Henderson, J. M., Williams, C. C., & Falk, R. J. (2005). Eye movements are functional during face learning. *Memory and Cognition, 33*(1), 98-106. doi: 10.3758/BF03195300

Hill, H., & Bruce, V. (1996). Effects of lighting on the perception of facial surfaces. *Journal of Experimental Psychology: Human Perception and Performance, 22*(4), 986-1004. doi: 10.1037/0096-1523.22.4.986

Hsiao, J. H. W., & Cottrell, G. (2008). Two fixations suffice in face recognition. *Psychological Science, 19*(10), 998-1006. doi: 10.1111/j.1467-9280.2008.02191.x

Jenkins, R., White, D., Van Montfort, X., & Burton, A. M. (2011). Variability in photos of the same face. *Cognition, 121*(3), 313-323. doi: 10.1016/j.cognition.2011.08.001

Kemp, R., Towell, N., & Pike, G. (1997). When seeing should not be believing: Photographs, credit cards and fraud. *Applied Cognitive Psychology, 11*(3), 211-222. doi: 10.1002/(SICI)1099-0720

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*(1), 26. doi: 10.1186/s41235-018-0115-6

Megreya, A. M., & Bindemann, M. (2018). Feature instructions improve face-matching accuracy. *PLoS ONE, 13*(3), 1-16. doi: 10.1371/journal.pone.0193455

Megreya, A. M., & Burton, A. M. (2006). Unfamiliar faces are not faces: Evidence from a matching task. *Memory and Cognition, 34*(4), 865-876. doi: 10.3758/BF03193433

Megreya, A. M., & Burton, A. M. (2008). Matching faces to photographs: Poor performance in eyewitness memory (without the memory). *Journal of Experimental Psychology: Applied, 14*(4), 364-372. doi: 10.1037/a0013464

Megreya, A. M., Sandford, A., & Burton, A. M. (2013). Matching face images taken on the same day or months apart: The limitations of photo ID. *Applied Cognitive Psychology, 27*(6), 700-706. doi: 10.1002/acp.2965

Moore, R. M., & Johnston, R. A. (2013). Motivational incentives improve unfamiliar face matching accuracy. *Applied Cognitive Psychology, 27*(6), 754-760. doi: 10.1002/acp.2964

Norell, K., Läthén, K. B., Bergström, P., Rice, A., Natu, V., & O'Toole, A. (2015). The effect of image quality and forensic expertise in facial image comparisons. *Journal of Forensic Sciences, 60*(2), 331–340. doi: 10.1111/1556-4029.12660

Noyes, E., Hill, M. Q., & O'Toole, A. J. (2018). Face recognition ability does not predict person identification performance: using individual data in the interpretation of group results. *Cognitive Research: Principles and Implications, 3*(1), 23. doi: 10.1186/s41235-018-0117-4

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

Or, C. C. F., Peterson, M. F., & Eckstein, M. P. (2015). Initial eye movements during face identification are optimal and similar across cultures. *Journal of Vision, 15*(13), 12-12. doi: 10.1167/15.13.12

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*(19), 2145-2155. doi: 10.1016/j.visres.2011.08.009

President's Council of Advisors on Science and Technology (PCAST). (2016). *Report to the President, Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-comparison Methods.* Executive Office of the President of the United States, President's Council of Advisors on Science and Technology.

Peterson, M. F., & Eckstein, M. P. (2013). Individual differences in eye movements during face identification reflect observer-specific optimal points of fixation. *Psychological Science, 24*(7), 1216-1225. doi: 10.1177/0956797612471684

Phillips, P. J., Yates, A. N., Hu, Y., Hahn, C. A., Noyes, E., Jackson, K., … O'Toole, A. J. (2018). Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Proceedings of the National Academy of Sciences of the United States of America, 115*(24), 6171–6176. doi: 10.1073/pnas.1721355115

Ramon, M., Bobak, A. K., & White, D. (2019). Super-recognizers: From the lab to the world and back again. *British Journal of Psychology*. doi: 10.1111/bjop.12368

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124*(3), 372. doi: 10.1037/0033-2909.124.3.372

Robertson, D. J., & Bindemann, M. (2019). Consolidation, wider reflection, and policy: Response to 'Super-recognisers: From the lab to the world and back again'. *British Journal of Psychology*, 8-10. doi: /10.1111/bjop.12393

Robertson, D. J., Noyes, E., Dowsett, A. J., Jenkins, R., & Burton, A. M. (2016). Face recognition by metropolitan police super-recognisers. *PLoS ONE, 11*(2), e0150036. doi: 10.1371/journal.pone.0150036

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

Tanaka, J. W., & Farah, M. J. (1993). Parts and Wholes in Face Recognition. *The Quarterly Journal of Experimental Psychology, 46*(2), 225-245. doi: 10.1080/14640749308401045

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. doi: 10.1371/journal.pone.0211037

Towler, A., White, D., & Kemp, R. I. (2017.). Evaluating the feature comparison strategy for forensic face identification. *Journal of Experimental Psychology: Applied, 23*(1), 47-58. doi: 10.1037/xap0000108

Tully, G. (2019). *Annual Report: November 2017–November 2018*. Birmingham: The Forensic Science Regulator. Retrieved from https://assets publishing.service.gov. uk/government/uploads/system/uploads/attachment_data/file/786137/FSRAnnual_Report_2018_v1.0.pdf

White, D., Burton, A. M., Kemp, R. I., & Jenkins, R. (2013). Crowd effects in unfamiliar face matching. *Applied Cognitive Psychology, 27*(6), 769-777. doi: 10.1002/acp.2971

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. doi: 10.1371/journal.pone.0139827

White, D., Kemp, R. I., Jenkins, R., Matheson, M., & Burton, A. M. (2014). Passport officers' errors in face matching. *PLoS ONE, 9*(8), e103510. doi: 10.1371/journal.pone.0103510

White, D., Phillips, P., Hahn, C. A., Hill, M., & O'Toole, A. J. (2015a). Perceptual expertise in forensic facial image comparison. *Proceedings of the Royal Society B: Biological Sciences, 282(*1814). doi: 10.1098/rspb.2015.1292

Wilkinson, C., & Evans, R. (2008). Are facial image analysis experts any better than the general public at identifying individuals from CCTV images? *Science & Justice, 49*(3), 191-196. https://doi.org/10.1016/j.scijus.2008.10.011

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*(2), 138. doi: 10.1037/xap0000114