PROCEEDINGS B

rspb.royalsocietypublishing.org

Research

Cite this article: White D, Phillips PJ, Hahn CA, Hill M, O'Toole AJ. 2015 Perceptual expertise in forensic facial image comparison. *Proc. R. Soc. B* **282**: 20151292.
http://dx.doi.org/10.1098/rspb.2015.1292

Received: 31 May 2015
Accepted: 5 August 2015

Subject Areas:
cognition

Keywords:
visual expertise, face recognition, person identification, biometrics, forensic science

Author for correspondence:
David White
e-mail: david.white@unsw.edu.au

Perceptual expertise in forensic facial image comparison

David White[1], P. Jonathon Phillips[2], Carina A. Hahn[3], Matthew Hill[3] and Alice J. O'Toole[3]

[1]School of Psychology, The University of New South Wales, Sydney, New South Wales 2052, Australia
[2]National Institute of Standards and Technology, 100 Bureau Drive, MS 8940, Gaithersburg, MD 20899, USA
[3]The University of Texas at Dallas, Richardson, TX 75080, USA

Forensic facial identification examiners are required to match the identity of faces in images that vary substantially, owing to changes in viewing conditions and in a person's appearance. These identifications affect the course and outcome of criminal investigations and convictions. Despite calls for research on sources of human error in forensic examination, existing scientific knowledge of face matching accuracy is based, almost exclusively, on people without formal training. Here, we administered three challenging face matching tests to a group of forensic examiners with many years' experience of comparing face images for law enforcement and government agencies. Examiners outperformed untrained participants and computer algorithms, thereby providing the first evidence that these examiners are experts at this task. Notably, computationally fusing responses of multiple experts produced near-perfect performance. Results also revealed qualitative differences between expert and non-expert performance. First, examiners' superiority was greatest at longer exposure durations, suggestive of more entailed comparison in forensic examiners. Second, experts were less impaired by image inversion than non-expert students, contrasting with face memory studies that show larger face inversion effects in high performers. We conclude that expertise in matching identity across unfamiliar face images is supported by processes that differ qualitatively from those supporting memory for individual faces.

1. Introduction

Proliferation of CCTV, mobile image capture and face recognition technology entails a critical role for facial images in modern forensic identification. As a result, facial image comparison is a major source of evidence in criminal investigations and trials [1,2], and wide deployment of automatic recognition systems over recent years has been accompanied by substantial gains in reliability [3]. Importantly, forensic applications of this biometric software—as with automatic fingerprint recognition systems—are configured to provide lists of potential matches according to the computed scoring metric. For final identity judgements, such as those provided as evidence in court, trained facial forensic examiners adjudicate suspected matches [1,2,4]. Given this reliance, and evidence of DNA-based exonerations owing to errors in forensic judgements [5], there is a pressing need for research that can benchmark the skills of examiners relative to untrained humans and computer-based face recognition systems [6].

There is striking evidence that untrained individuals perform poorly on the apparently straightforward task of matching the identity of an unfamiliar face across two different images [7–11]. Even under optimal matching conditions in laboratory tests conducted using images that are taken on the same day, in the same neutral pose, and under similar environmental conditions; error rates for untrained individuals are in the range of 20–30% [8,9]. In suboptimal capture conditions when environmental factors are unconstrained, such as when matching between CCTV footage and high-quality mug-shots, performance can approach chance [10]. Moreover, in field tests conducted outside of the

THE ROYAL SOCIETY PUBLISHING

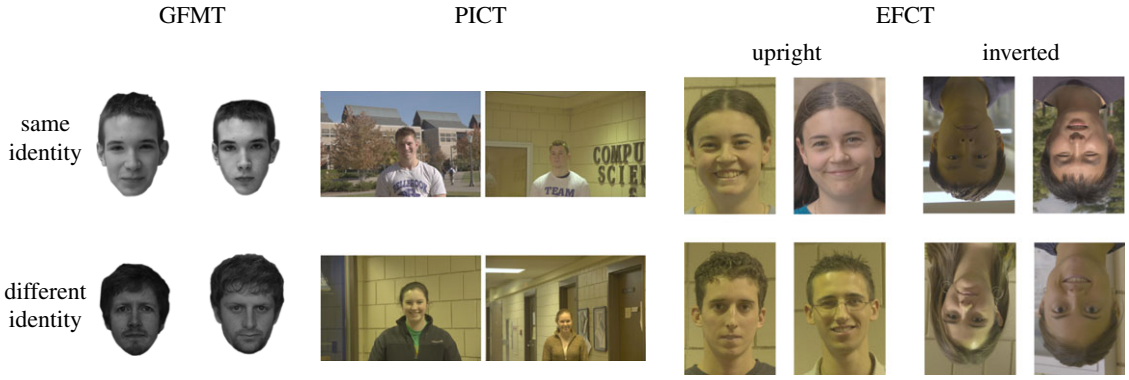

Figure 1. Examples of image pairs from the GFMT, PICT and EFCT. Images in the GFMT were all taken on the same day and under very similar lighting conditions, whereas images from the PICT and EFCT were captured in diverse ambient conditions and over a longer time period. For details of test construction, please refer to Methods. (Online version in colour.)

laboratory, professional police officers and passport officers make the same number of face matching errors as standard groups of student participants [10,11], despite performing face matching as part of their daily work.

Here, we assessed perceptual expertise in facial image comparison by developing and administering a battery of tests to an international group of forensic facial examiners. Small-scale studies have recently reported higher identification accuracy in court practicing facial image examiners [12,13], but are limited by use of unstandardized tests containing small numbers of image comparisons. Inconsistency across experiments also raises the possibility that group performance is specific to the workplaces tested. To ensure that experts represented the highest global standards in forensic image comparison, we approached organizers of the Facial Identification Scientific Working Group (FISWG; www.fiswg.org). The May 2014 meeting of this group was held in the FBI academy in Quantico, Virginia, and presented a unique opportunity to test an international group of facial forensic examiners with rigorous training and many years of professional experience identifying unfamiliar faces (henceforth *examiners*). As comparison groups, we tested the FISWG meeting attendees who do not perform forensic facial examination as part of their daily work, but were attending the meeting as managers, technical experts or administrators in biometric systems (*controls*). Because people in this group were knowledgeable of the types of training used for forensic examiners and difficulties associated with unfamiliar face matching [14], we also tested untrained university students, representing the most commonly tested cohort in previous research (*students*).

In all tests, participants decided if pairs of images were of the same identity or of two different identities. We mapped expert performance to established human accuracy using a standard psychometric test of unfamiliar face matching ability—the Glasgow Face Matching Test (GFMT) [9]—and created two new tests to benchmark forensic examiners against both human and state-of-the-art algorithm-based matching performance. Tests were designed to be sufficiently challenging for examiners, and representative of the types of decisions encountered in daily work (see figure 1 for example images). To create the Expertise in Facial Comparison Test (EFCT), we selected pairs of images for identity comparisons that were challenging for computers and untrained humans based on data from pilot work and previous evaluations of human and computer face matching performance [15,16]. Image pairs in the Person Identification Challenge Test (PICT) included body cues and were selected to have no computationally useful identity information in the face, as indicated by the fact that leading algorithms make 100% errors on this set [17] (for details of test construction, please refer to Methods).

The FISWG meeting was a unique opportunity to address a key theoretical question in the study of face identification. Decades of research have shown that, relative to other classes of objects, face recognition is a skill for which humans are experts. Because face recognition in the general population is disproportionately impaired by inverted presentation compared with other objects [18], the face inversion effect (FIE) has been taken as an index of this expertise [18–21]. This view is bolstered by evidence that FIE (i) increases as face processing abilities improve with development [19,20], (ii) is weaker in people with impairments in face identification ability [22,23] and (iii) is stronger in those with exceptionally good face processing ability [23]. However, this evidence is almost entirely based on face memory tasks. In facial image comparison tasks performed by forensic examiners, it has been proposed that different mechanisms are recruited [24] and so it is not clear whether expertise of facial examiners will be indexed by FIEs. To test this, we included an inverted face matching test in the EFCT.

To probe the nature of expertise in face matching further, we manipulated exposure duration of image pairs in identity comparisons. Previous research suggests that 2 s is optimal for face matching decisions in untrained participants [25], but training in forensic facial examination emphasizes careful comparison and effortful analysis of facial images prior to identification decisions [14]. Thus, we predicted that the accuracy of the forensic experts would be superior to the controls and students on the longer (30 s), but not the shorter exposure times (2 s).

2. Results

Aggregated results across tests confirmed expertise of the facial examiners (figure 2 summarizes GFMT and PICT results and figure 3 summarizes EFCT). The rank order of identification performance for examiners, controls and students was stable for all results. In all experimental conditions, across the three tests (GFMT; EFCT: 2 s upright, 2 s inverted, 30 s upright, 30 s inverted; and PICT), accuracy was ranked as follows: examiners > controls > students. The consistency of this order proved statistically reliable (examiners > controls, Wilcoxon

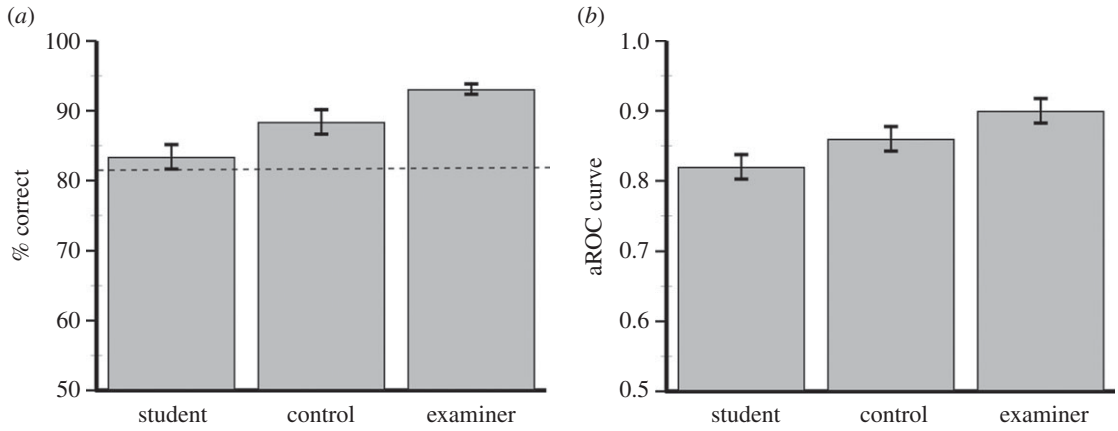

Figure 2. (*a*) Mean percentage correct (±1 s.e.) for experimental groups in the GFMT. (*b*) PICT performance is plotted as average area under the receiver operating characteristic (ROC) curve. Mean normative accuracy in the GFMT is indicated by the dashed line.

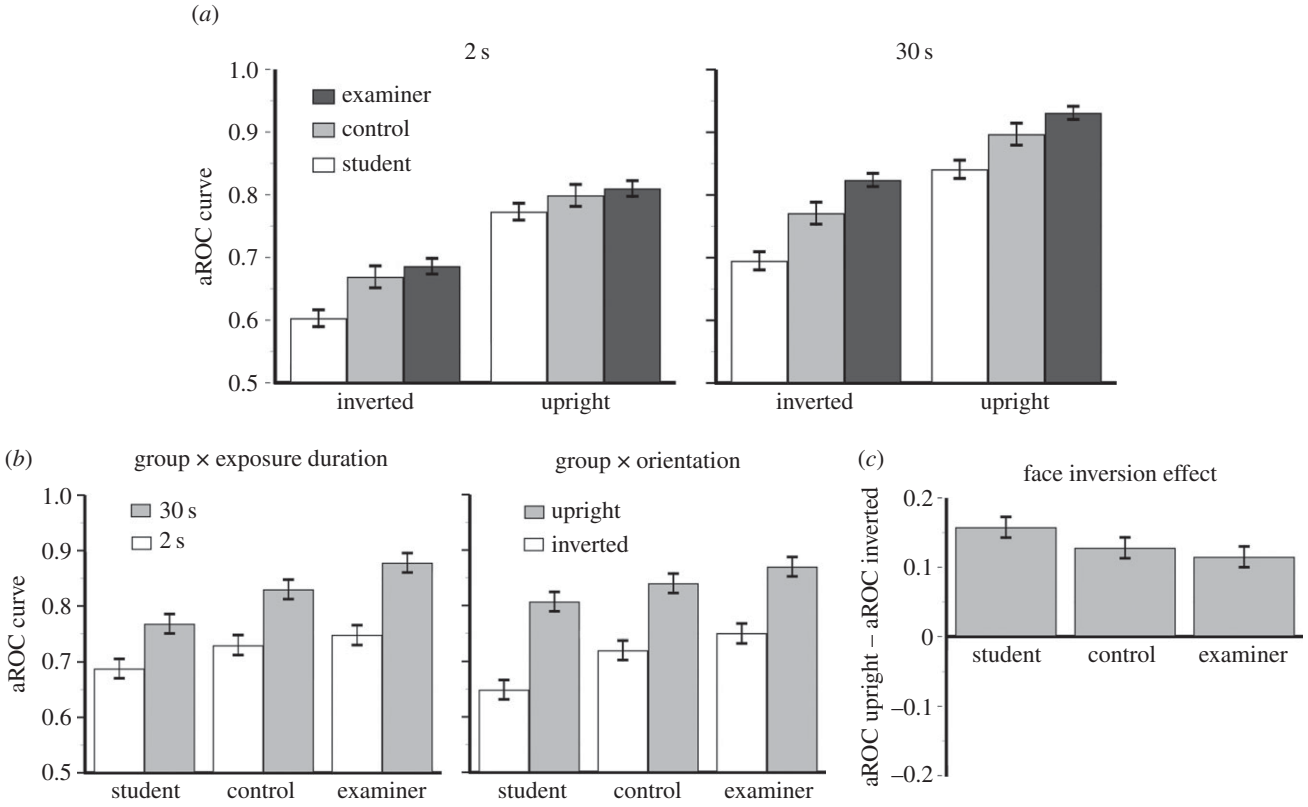

Figure 3. Analysis of EFCT results. (*a*) Mean ROC scores for three groups on the EFCT (±1 s.e.). (*b*) Significant interactions in EFCT data. Simple main effects revealed stronger effects of group in 30 s compared with 2 s exposure durations (left). In addition, there was a significant interaction between group and orientation (right). (*c*) Contrary to hypotheses based on FIEs in memory tasks, inversion effects are larger for students than for experts. Details of analysis are provided in the text.

sign test, $t_5 = 4.85$, $p = 0.0313$; controls > students, $t_5 = 4.85$, $p = 0.0313$), showing a general superiority of forensic examiners across tests.

(a) Glasgow Face Matching Test

To map performance of experimental groups to the population at large, we compared performance on the GFMT with normative data for the test [9]. Mean GFMT scores for the three groups are shown in figure 2. Accuracy of examiners exceeded normative, control and student accuracy ($t_{219} = 6.35$, $p < 0.0001$, Cohen's $d = 0.858$; $t_{39} = 2.34$, $p < 0.05$, Cohen's $d = 0.749$; $t_{57} = 4.09$, $p < 0.05$, Cohen's $d = 1.08$). Performance of control participants also exceeded normative accuracy ($t_{206} = 2.77$, $p = 0.006$, Cohen's $d = 0.385$), but performance did not differ significantly between controls and students ($t_{44} = 1.51$, $p > 0.05$, Cohen's $d = 0.455$). Student performance did not differ significantly from normative accuracy scores ($t_{224} = 1.14$, $p > 0.05$, Cohen's $d = 0.1$). As far as we are aware, experts and controls are the only groups reported to have exceeded normative accuracy on the GFMT.

(b) Person Identification Challenge Test

To analyse scores on the PICT, for each participant we computed the area under the receiver operator characteristic

(aROC). Summary aROC scores are shown in figure 2. Scores were analysed by ANOVA with group (student, control, examiner) as a between-subjects factor. There was a significant main effect of group ($F_{2,70} = 4.89$, $p = 0.01$, $\eta_p^2 = 0.122$). Contrast analyses indicated that examiners performed more accurately than students ($F_{1,70} = 9.66$, $p = 0.003$, $\eta_p^2 = 0.121$). Differences between examiners and controls ($F_{(1,70)} = 1.06$, $p = 0.307$, $\eta_p^2 = 0.015$), and between controls and students ($F_{1,70)} = 2.18$, $p = 0.144$, $\eta_p^2 = 0.030$), were non-significant.

(c) Expertise in facial comparison test

The EFCT was designed to test three key predictions. First, examiners would be more accurate than both other groups. Second, this advantage would only be observed in conditions that enabled careful examination of image pairs (i.e. 30 s exposure). Third, owing to perceptual expertise comparing upright facial images, the examiner group would show larger inversion effects compared with controls and students.

We computed aROC scores individually for each participant (figure 3*a*). Scores were submitted to a $2 \times 2 \times 3$ ANOVA with exposure time (2, 30 s) and orientation (upright, inverted) as within-subject factors and group (student, control, examiner) as a between-subjects factor. Significant main effects were found for exposure time, orientation and group ($F_{1,70} = 176.33$, $p < 0.0001$, $\eta_p^2 = 0.716$; $F_{1,70} = 341.07$, $p < 0.0001$, $\eta_p^2 = 0.83$; $F_{2,70} = 14.54$, $p < 0.0001$, $\eta_p^2 = 0.293$). In line with our second prediction, there was a significant interaction between group and exposure time ($F_{1,70} = 4.82$, $p = 0.0109$, $\eta_p^2 = 0.121$). In addition, we observed a statistically significant interaction between group and orientation ($F_{1,70} = 4.02$, $p = 0.022$, $\eta_p^2 = 0.103$). The three-way interaction was non-significant ($F < 1$). Significant interactions are plotted in figure 3*b*.

For the group and exposure time interaction, simple main effects tests revealed group effects at both shorter and longer exposure times ($F_{2,70} = 14.55$, $p < 0.0001$, $\eta_p^2 = 0.294$; $F_{2,70} = 46.66$, $p < 0.00001$, $\eta_p^2 = 0.571$), but these effects were more pronounced when participants had more time to examine each image pair. In the 2 s condition, examiners performed more accurately than students ($F_{1,70} = 27.02$, $p = 0.00001$, $\eta_p^2 = 0.279$), but did not differ from the controls. In the 30 s condition, the examiners were more accurate than the controls and students ($F_{1,70} = 9.28$, $p = 0.003$, $\eta_p^2 = 0.117$; $F_{1,70} = 91.83$, $p < 0.0001$, $\eta_p^2 = 0.567$), supporting the prediction of greater differentiation of the examiners from the other groups, when they had more time to examine the image pairs.

To examine the group and orientation interaction, we collapsed across study duration (figure 3*b*). Simple main effects tests revealed significant effects of group for both inverted and upright faces ($F_{2,70} = 50.38$, $p < 0.00001$, $\eta_p^2 = 0.59$; $F_{2,70} = 18.18$, $p < 0.0001$, $\eta_p^2 = 0.342$). This interaction is consistent with differences in the size of the FIE across groups. Because the FIE is an established index of perceptual expertise, we predicted a larger inversion effect for the more accurate participant groups (i.e. examiner FIE > control FIE > student FIE). A cursory examination of the data proved inconsistent with that prediction. We examined this further by computing FIE strength (aROC upright − aROC inverted) for each participant, in each exposure duration condition. A 2×3 ANOVA with study duration (2, 30 s) as a within-subject factor and group (student, control, examiner) as a between-subjects factor revealed a significant main effect of group ($F_{2,70} = 4.02$, $p = 0.022$, $\eta_p^2 = 0.103$). Means for these difference scores appear in figure 3*c* and show stronger FIE for students than for examiners ($F_{1,70} = 7.68$, $p = 0.007$, $\eta_p^2 = 0.099$) opposite to predictions based on the perceptual expertise hypothesis. Contrasts between students and controls ($F_{1,70} = 2.53$, $p = 0.116$, $\eta_p^2 = 0.036$), and between examiners and controls ($F_{1,70} = 0.425$, $p = 0.517$, $\eta_p^2 = 0.036$), were non-significant.

(d) Fusion analysis

In forensic practice, to assure consistency and consensus, it is common for multiple examiners to repeat a single comparison judgement. To model the effectiveness of this process, we conducted simulations to 'fuse' or combine judgements at the level of individual image pairs. The simulations followed previous work showing that aggregating the judgements of multiple participants improves identification accuracy [26,27]. We focused on data from the 30 s upright EFCT as this experimental condition most closely resembles working practice of forensic examiners; however, we also carried out these simulations with the PICT and found comparable results.

Effects of aggregation were calculated separately for each group (students, controls, examiners) by resampling participants' identity ratings (i.e. from 1 = sure same to 5 = sure different) for 84 image pairs. We randomly sampled n participants from within a group and averaged their responses for each image pair separately. This sampling procedure was repeated 100 times for each value of n, and accuracy was computed at each iteration by calculating the group aROC. Aggregate accuracy for a given sample size was measured as the average aROC across all iterations. We report results for aggregate sample sizes that vary from 1 to 14 participants, with the upper limit dictated by the smallest group of participants (for controls, by definition, all iterations of sample size 14 include the entire group).

Figure 4 shows the aggregation effect as a function of participant sample size and serves as a practical guide to the performance benefit that can be expected by combining identity judgements across participants. Closer inspection of results with smaller n shows substantial improvements in accuracy as additional judgements were aggregated. Comparisons highlight differences in the relative value of participants according to group. That is, one examiner (0.936) is roughly equal to two controls (0.946) or four students (0.942). With only one subject, examiner performance surpassed control performance in 62% of the 100 iterations, and controls surpassed student performance in 79% of the iterations. As sample size increased, however, aggregated judgements from examiners were more likely to surpass controls and aggregated judgements by controls were more likely to surpass student decisions. At the maximum sample size, examiners surpassed controls in 99% of the iterations and controls surpassed students in 91% of the iterations. The bars in the bottom of figure 4 have this analysis for all participant sample sizes.

So, although examiner performance did not reach ceiling levels at 30 s exposures when computing performance measures at the individual level, these limits were largely overcome by response aggregation—which produced near-perfect accuracy and revealed a highly stable performance advantage for professional examiner groups. Given the highly challenging nature of the images used in the EFCT, this suggests that a fusion approach can help support identification decisions in forensic practice.

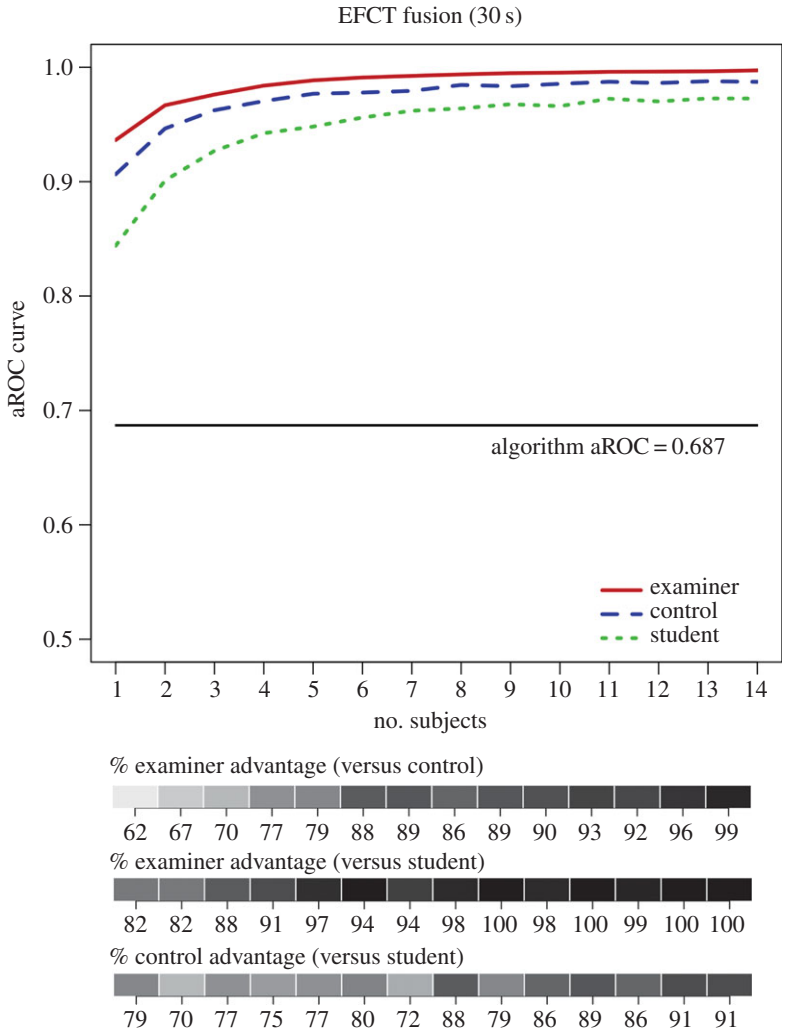

Figure 4. Accuracy according to number of judgements fused in the EFCT simulations. Accuracy increases as more judgements are fused. As indicated in the bars below the chart, as the number of subjects fused increases, the reliability of the pattern examiners > controls > students also increases. Clear improvements for aggregating judgements are seen within all three groups of subjects. For all groups, these improvements plateau well before the maximum sample size: at asymptote, average aROCs for groups of roughly eight raters were close to perfect for all three groups (examiners = 0.997; controls = 0.987; students = 0.973). (Online version in colour.)

3. Discussion

In this study, we report the first systematic assessment of face matching performance by a diverse group of international forensic examiners. The rank order of identification performance in all six experimental tests placed examiners over controls and students, and examiner performance exceeded normative levels established in previous studies [9]. Although examiners' performance was not statistically superior to controls in all experiments, it was always statistically superior to the student performance. To the best of our knowledge, this is the first convincing demonstration of a professional group showing higher accuracy on face matching tasks.

Closer analysis also revealed two qualitative differences between examiners and non-expert groups. First, examiners' superiority was not specific to longer exposures but was also observed when permitted just 2 s per comparison in the EFCT, suggesting that expertise improved intuitive as well as considered judgements. Differences between groups were most pronounced with 30 s exposure, however, pointing to a more entailed and effective identity examination process by examiners than by less experienced participants. This contrasts with accounts of perceptual expertise in radiographers [28] and fingerprint examiners [29], where expertise appears primarily driven by a shift in perceptual strategy towards fast and global image analysis [30]. Consistent with the training forensic examiners receive [14], our results suggest an opposing trajectory of expertise in forensic facial image comparison characterized by a transition towards controlled and effortful analysis.

Second, inverting images produced less impairment in examiners compared with students. This is perhaps surprising, as it is not entirely consistent with the expertise hypothesis of face processing, which predicts increasing impairment as perceptual expertise develops [19,20]. This finding can be reconciled, however, with existing evidence that unfamiliar face matching may rely on separate processes to those supporting face memory. For example, individual differences in unfamiliar face matching accuracy are not predicted by accuracy in face memory tasks [9,24]. Our results extend this work, suggesting that processes underlying expertise in unfamiliar face matching may be dissociable from those driving expertise in face memory [20,31]. Because the expertise of forensic examiners extended to images presented upside down, our results are also consistent with the proposal that mechanisms supporting unfamiliar face matching performance may not be entirely face-specific, but may instead reflect general image comparison abilities [24].

To complicate this issue, deficits in face memory and perception have been associated with smaller FIEs [22,23], and people with exceptional face processing ability show larger FIEs in both memorial and perceptual tasks (Cambridge Face Perception Test [23]). Our data show the opposite pattern. This raises the additional possibility that visual expertise of forensic facial examiners differs qualitatively from that which underpins face processing in the population at large. This type of visual expertise may be dissociable from that shown by high performers with no specific training in facial image comparison. Given the emphasis on feature-by-feature approaches to comparison in professional training [14,32] and the interaction between image inversion and expertise reported here, it is possible that increased selective attention to facial features improves performance. Thus, future studies that examine benefits of part-based comparison strategies and identify visual cues subsisting accuracy in forensic examiners promise to elucidate the foundations of expertise in forensic comparison.

It is also important to note that control participants performed very well despite not performing facial image comparison in their daily work. As a group, control participants' scores exceeded normative levels on the GFMT. Moreover, in two tests, their performance did not differ statistically from examiners (EFCT 2 s exposure, PICT). This raises the possibility that controls were more motivated than students, and that examiners were more motivated than controls. Thus, it is possible that differing levels of motivation might account for performance differences across groups. Indeed, in any test where groups of observers from different backgrounds are compared, motivation may differ across groups and affect performance. Here, we think it possible, and even likely, that students, controls and examiners differed in their 'self-investment' in the results, and thereby in their motivation to perform well. Previous research suggests, however, that benefits of motivation alone are limited. Two different tests of police and passport officers reported equivalent levels of face matching accuracy to untrained students [10,11], despite a clear motivation for these groups to perform well. Moreover, in the present study, controls did not surpass professionals in all tests. For example, examiners performed more accurately in the EFCT test, but only at the longer exposure duration, when the test conditions supported the employment of the special skills and experience of the examiners.

Nevertheless, it will be important in future work to establish the relative contribution of natural ability, motivation, experience and training to expertise in forensic examination. As is typical in studies of expert populations [33], separating their contributions to the emergence of expertise is problematic, and so longitudinal studies of forensic professionals may be necessary to address this important question. Decoupling the influence of these factors will aid development of recruitment and training methods for forensic examination. It is also important when evaluating forensic evidence provided in the courtroom, where accurate assessments of expertise are critical in establishing the appropriate weight to be given to identity judgements [34]. For now, qualitative differences in examiner performance suggest that differences in cognitive processing contribute to their superior accuracy.

Finally, although forensic facial examiners performed more accurately than the control and student groups, perfect performance was not attained on any test, with average misclassification rates of around 7% on both the EFCT (30 s upright) and GFMT. Because these were strictly perceptual tests, examiners were not permitted unlimited time or access to digital tools that would support decisions in daily work. Further, participants made identity judgements on a Likert scale that may not reflect normal reporting of identity judgements in forensic practice. In a recent study of professional fingerprint examiners, participants could skip comparison decisions on the basis that they did not provide sufficient evidence for identification [35]. Although this approach inhibits the measurement of underlying perceptual skill [35,36], these types of decisions are critical in minimizing costly workplace errors, and recent work suggests that forensic examiners are skilled in these types of judgements [11]. Thus, our results provide an estimate of the *perceptual* abilities of facial forensic examiners that can serve as a benchmark for future tests of identification accuracy in standard forensic practice, and for computer-based face recognition systems that support this practice.

4. Methods

(a) Participants

Three groups of participants completed each test; we refer to these groups as examiner, control and student. The examiner and control groups comprised 41 volunteers (19 females; mean age = 42.4, s.d. = 9.9) who attended the FISWG meeting in May 2014. The role of FISWG is to create policy for best practices in facial image comparison and training. Each participant in these two groups completed a questionnaire designed to assess their professional experience and training in forensic and facial examination. The examiner group consisted of 27 FISWG attendees who stated that they regularly performed facial examination as a part of their employment (average years experience = 7.3, s.d. = 5.8; average hours per week = 11.8, s.d. = 12.1). The remaining 14 attendees were also government employees. Although this group was relatively small, they provided a valuable control group and so we tested them in the same experimental session. A detailed comparison of the demographics of the examiner and control groups was not possible because of requirements to protect the anonymity of participants. Thus, it was not possible to match groups on age. Although we have no reason to suspect that these groups differed in average age characteristics, previous large-scale studies have found that age of participants is not correlated with accuracy on perceptual face matching tasks [9].

Students represent the most commonly tested population in unfamiliar face matching experiments. Thus, we expected that student performance should approximate levels of performance reported in the literature. We also expected students to be less cognizant of task demands when compared with controls. The controls were attending the FISWG meeting to create policy documents outlining 'best practice' in facial identification, and so we anticipated somewhat better performance from the controls than from the students based on their inherent interest in the task. Students were undergraduates at the University of New South Wales ($n = 32$; 19 females, average age 21.4, s.d. = 5.76).

Participants completed the tests in the following order: EFCT with 2 s exposures (upright block, then inverted block), PICT, GFMT and then the EFCT again with 30 s exposures (upright block, then inverted block). All tests were administered on laptop computers. Example image pairs from each test are shown in figure 1.

(b) Glasgow Face Matching Test

The GFMT is a psychometric test designed to evaluate an individual's ability to match identity across images of unfamiliar

faces [9]. Stimuli for the short version GFMT consisted of 20 same- and 20 different-identity image pairs. Same-identity pairs show two images of the same person taken under similar lighting conditions, on the same day, but using different digital cameras. For different-identity pairs, one of these images was paired with a similar-looking person from the database, so that each identity appears once in a same-identity pair and once in a different-identity pair. For each image pair, participants responded 'same' or 'different' identity. The task was self-paced, and image pairs remained on the computer monitor until participants made their response, at which point the next image pair was presented.

(c) Person Identification Challenge Test

The stimuli for the PICT were sampled from those used in a recent study that compared human and computer algorithm performance on a special set of image pairs for which machine performance in the face recognition vendor test [15] was 100% incorrect [17]. Specifically, similarity scores computed between same-identity faces were uniformly lower than those computed for the different-identity image pairs, suggesting that they contain no computationally useful identity information in the face. Interestingly, in a recent study, untrained observers achieved above-chance identification accuracy for these image pairs owing to non-face identity cues from the body [17]. We included this as a test of person identification ability for a set of image pairs for which face recognition software fails.

We sampled 40 pairs of images (20 same-identity pairs) from this dataset for the PICT. Participants were presented pairs in a random order and the image pairs remained on the screen until the participant's response was registered. Response options were as follows: (i) sure they are the same person; (ii) think they are the same person; (iii) do not know; (iv) think they are different people; and (v) sure they are different people. After participants made a response, the next image pair was presented.

(d) Expertise in Facial Comparison Test

In designing the EFCT, our goal was to measure performance of examiners with image pairs that challenge both computer face recognition systems and untrained observers. We selected images from The Good, the Bad and the Ugly Challenge [15], an image dataset containing images from diverse and unconstrained ambient conditions. This dataset was specifically designed to test state-of-the-face recognition algorithms under challenging environmental conditions, and contains frontal views of faces, taken with minimal control of illumination, expression and appearance.

First, image pairs ($n = 1\,177\,225$) were ranked according to machine performance using a fusion of top-performing algorithms in the Face Recognition Vendor Test, an international benchmarking test for leading face recognition algorithms [15]. Images were then stratified according to match score data into three subsets based on item accuracy: easy (the Good), moderate (the Bad) or poor (the Ugly). For the ECFT, we choose only image pairs from the Bad and Ugly portions of this dataset. Second, performance of untrained human observers on a sample ($n = 480$) of the challenging items [16] was used as the basis for a second selection. We combined these data with unpublished student performance on the same test, and recalculated item performance for only the highest-performing 25% of participants. Items on which 8% or more of high-performing participants made errors were selected for the EFCT.

Participants completed the EFCT with upright and inverted image pairs in two conditions that varied by exposure time. The same image pairs were tested in 2 s and 30 s exposure time conditions. Because PICT and GFMT were administered between the 2 s and 30 s conditions, there was a gap of roughly 1 h between the 2 s and 30 s EFCT tests. For each trial, images remained visible for the prescribed exposure duration (2 or 30 s), and then disappeared. Response options were as follows: (i) sure they are the same person; (ii) think they are the same person; (iii) do not know; (iv) think they are different people; and (v) sure they are different people. Participants could enter a response at any time during the image display or after the image pair disappeared. The next trial followed immediately. In total, the EFCT consisted of 168 image pairs (half same-identity, half different-identity), with half of these allocated to the upright test and half to the inverted test.

Ethics. This study was approved by the Institutional Review Board at the University of Texas at Dallas. All participants provided written informed consent and appropriate photographic release.

Data accessibility. Participant performance data: Dryad (http://dx.doi.org/10.5061/dryad.ng720). Fusion analysis data: Dryad (http://dx.doi.org/10.5061/dryad.ng720).

Authors' contributions. D.W., P.J.P., C.A.H., M.H. and A.J.O. designed the study; D.W., C.A.H., M.H. and A.J.O. collected the data; D.W., P.J.P., C.A.H., M.H. and A.J.O. analysed the data. All authors discussed the results and commented on the manuscript.

Competing interests. The authors have no competing interests.

Funding. This research was supported by US Department of Defense funding to A.J.O., and funding from the Australian Research Council and the Australian Passport Office to D.W. and Richard Kemp (LP110100448, LP130100702).

References

1. Jain AK, Klare B, Park U. 2012 Face matching and retrieval in forensics applications. *IEEE MultiMedia* **19**, 20–28. (doi:10.1109/MMUL.2012.4)
2. Dessimoz D, Champod C. 2008 Linkages between biometrics and forensic science. In *Handbook of biometrics* (eds AKJP Flynn, AA Ross), pp. 425–459. New York, NY: Springer.
3. Philips PJ, O'Toole AJ. 2014 Comparison of human and computer performance across face recognition experiments. *Image Vision Comput.* **32**, 74–85. (doi:10.1016/j.imavis.2013.12.002)
4. Jain AK, Ross A. 2015 Bridging the gap: from biometrics to forensics. *Phil. Trans. R. Soc. B* **370**, 20140254. (doi:10.1098/rstb.2014.0254)
5. Sacks MJ, Koehler JJ. 2005 The coming paradigm shift in forensic identification science. *Science* **309**, 892–895. (doi:10.1126/science.1111565)
6. National Research Council. 2009 *Strengthening forensic science in the United States: A path forward*. Washington, DC: The National Academies Press.
7. Kemp R, Towell N, Pike G. 1997 When seeing should not be believing: photographs, credit cards and fraud. *Appl. Cognit. Psychol.* **14**, 211–222. (doi:10.1002/(SICI)1099-0720(199706)11:3<211::AID-ACP430>3.0.CO;2-0)
8. Bruce V, Henderson Z, Greenwood K, Hancock PJB, Burton AM, Miller P. 1999 Verification of face identities from images captured on video. *J. Exp. Psychol. Appl.* **5**, 339–360. (doi:10.1037/1076-898X.5.4.339)
9. Burton AM, White D, McNeill A. 2010 The Glasgow face matching test. *Behav. Res. Methods* **42**, 286–291. (doi:10.3758/BRM.42.1.286)
10. Burton AM, Wilson S, Cowan M, Bruce V. 1999 Face recognition in poor-quality video: evidence from security surveillance. *Psychol. Sci.* **10**, 243–248. (doi:10.1111/1467-9280.00144)
11. White D, Kemp RI, Jenkins R, Matheson M, Burton AM. 2014 Passport officers' errors in face matching. *PLoS ONE* **9**, e103510. (doi:10.1371/journal.pone.0103510)

12. Norell K, Läthén KB, Bergström P, Rice A, Natu V, O'Toole A. 2015 The effect of image quality and forensic expertise in facial image comparisons. *J. Forensic Sci*. **60**, 331–340. (doi:10.1111/1556-4029.12660)
13. Wilkinson C, Evans R. 2011 Are facial image analysis experts any better than the general public at identifying individuals from CCTV images? *Sci. Justice* **49**, 191–196. (doi:10.1016/j.scijus.2008.10.011)
14. Facial Identification Scientific Working Group. 2011 Guidelines and recommendations for facial comparison training to competency. See www.fiswg.org/document/.
15. Philips PJ *et al.* 2011 An introduction to The Good, the Bad, and the Ugly face recognition challenge problem. In *IEEE Int. Conf. on Automatic Face and Gesture Recognition, 21–25 March 2011, Santa Barbara, CA*, pp. 346–353. Piscataway, NJ: IEEE.
16. O'Toole AJ, An X, Dunlop J, Natu V. 2012 Comparing face recognition algorithms to humans on challenging tasks. *ACM. Appl. Percept.* **9**, 16.
17. Rice A, Philips PJ, Natu V, An X, O'Toole AJ. 2013 Unaware person recognition from the body when face identification fails. *Psychol. Sci.* **24**, 2235–2243. (doi:10.1177/0956797613492986)
18. Valentine T. 1988 Upside-down faces: a review of the effect of inversion upon face recognition. *Brit. J. Psychol*. **79**, 471–491. (doi:10.1111/j.2044-8295.1988.tb02747.x)
19. Carey S, Diamond R. 1977 From piecemeal to configurational representation of faces. *Science* **195**, 312–314. (doi:10.1126/science.831281)
20. Carey S. 1992 Becoming a face expert. *Phil. Trans. R. Soc. B*. **335**, 95–103. (doi:10.1098/rstb.1992.0012)
21. Diamond R, Carey S. 1986 Why faces are and are not special: an effect of expertise. *J. Exp. Psychol. Gen*. **115**, 107–117. (doi:10.1037/0096-3445.115.2.107)
22. Avidan G, Tanzer M, Behrmann M. 2011 Impaired holistic processing in congenital prosopagnosia. *Neuropsychologia* **49**, 2541–2552. (doi:10.1016/j.neuropsychologia.2011.05.002)
23. Russell R, Duchaine B, Nakayama K. 2009 Super-recognizers: people with extraordinary face recognition ability. *Psychon. B, Rev*. **16**, 252–257. (doi:10.3758/PBR.16.2.252)
24. Megreya AM, Burton AM. 2006 Unfamiliar faces are not faces: evidence from a matching task. *Mem. Cogn.* **34**, 865–876. (doi:10.3758/BF03193433)
25. Özbek M, Bindemann M. 2011 Exploring the time course of face matching: temporal constraints impair unfamiliar face identification under temporally unconstrained viewing. *Vision Res*. **51**, 2145–2155. (doi:10.1016/j.visres.2011.08.009)
26. O'Toole AJ, Abdi H, Jiang F, Phillips PJ. 2007 Fusing face-verification algorithms and humans. *IEEE. Trans. Syst. Man. Cybern. B* **37**, 1149–1155. (doi:10.1109/TSMCB.2007.907034).
27. White D, Burton AM, Kemp RI, Jenkins R. 2013 Crowd effects in unfamiliar face matching. *Appl. Cognitive Psychol*. **27**, 769–777. (doi:10.1002/acp.2971)
28. Kundel HL, Nodine CF, Conant EF, Weinstein SP. 2007 Holistic component of image perception in mammogram interpretation: gaze-tracking study 1. *Radiology* **242**, 396–402. (doi:10.1148/radiol.2422051997)
29. Thompson MB, Tangen JM. 2014 The nature of expertise in fingerprint matching: experts can do a lot with a little. *PLoS ONE* **9**, e114759. (doi:10.1371/journal.pone.0114759)
30. Wolfe JM, Võ MLH, Evans KK, Greene MR. 2011 Visual search in scenes involves selective and nonselective pathways. *Trends Cogn. Sci*. **15**, 77–84. (doi:10.1016/j.tics.2010.12.001)
31. Maurer D, Le Grand R, Mondloch CJ. 2002 The many faces of configural processing. *Trends Cogn. Sci*. **6**, 255–260. (doi:10.1016/S1364-6613(02)01903-4)
32. Woodhead MM, Baddeley AD, Simmonds DCV. 1979 On training people to recognize faces. *Ergonomics* **22**, 333–343. (doi:10.1080/00140137908924617)
33. Sternberg RJ. 1996 Costs of expertise. In *The road to excellence: the acquisition of expert performance in the arts and sciences, sports, and games* (ed. KA Ericsson), pp. 347–354. Hillsdale, NJ: Erlbaum.
34. Edmond G *et al.* 2014 How to cross-examine forensic scientists: a guide for lawyers. *Aust. Bar Rev*. **39**, 174–196.
35. Ulery BT, Hicklin RA, Buscaglia J, Roberts MA. 2011 Accuracy and reliability of forensic latent fingerprint decisions. *Proc. Natl Acad. Sci. USA* **108**, 7733–7738. (doi:10.1073/pnas.1018707108)
36. Tangen JM, Thompson MB, McCarthy DJ. 2011 Identifying fingerprint expertise. *Psychol. Sci*. **22**, 995–997. (doi:10.1177/0956797611414729)