Evaluating the feature comparison strategy for forensic face identification

Alice Towler, David White and Richard I. Kemp

School of Psychology, UNSW Australia

To cite this article:

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

Author Note:

This research was supported by an Australian Postgraduate Award to Alice Towler and Australian Research Council Linkage grants to Richard Kemp (LP110100448; LP 130100702), in partnership with the Department of Foreign Affairs and Trade, Australian Passport Office.

Word Count (including Abstract, Figures, Footnotes & References): 10126

Corresponding Author:
Alice Towler
School of Psychology
UNSW Australia
Sydney
Australia NSW 2052
Email: a.towler@unsw.edu.au

Abstract

Face recognition is thought to rely on representations that encode holistic properties. Paradoxically, professional forensic examiners who identify unfamiliar faces by comparing facial images are trained to adopt a feature-by-feature comparison strategy. Here we tested the effectiveness of this strategy by asking participants to rate facial feature similarity prior to making same/different identity decisions to pairs of face images. Experiment 1 provided preliminary evidence that rating feature similarity improves unfamiliar face matching accuracy in novice participants. In Experiment 2, we found benefits of this procedure over and above rating similarity of personality traits and image quality parameters, suggesting that benefits are not solely attributable to general increases in attention. In Experiment 3, we then compared performance of trained forensic facial image examiners to novice participants, and found that examiners displayed: i) superior face matching accuracy; ii) smaller face inversion and *feature* inversion effects; and iii) feature ratings that were more diagnostic of identity. Further, aggregating feature ratings of multiple examiners produced perfect identity discrimination. Based on these quantitative and qualitative differences between experts and novices, we conclude that comparison based on local features confers specific benefits to trained forensic examiners.

Word Count: 190

INTRODUCTION

People are remarkably good at recognising familiar faces, even under challenging conditions. For example, we can effortlessly recognise friends in forgotten photographs taken many years ago (Bahrick, Bahrick, & Wittlinger, 1975). This type of empirical evidence has led to a widely held belief that people are experts at face recognition, and to account for this expertise, current theory suggests that 'holistic' representations of familiar faces enable robust recognition (e.g., Biederman & Kalocsai, 1997; Maurer, Le Grand, & Mondloch, 2002; Rossion, 2008). This is based on the observation that global properties of a stimulus appear to be especially important in face recognition relative to other types of object recognition (e.g., Carey, 1992; Tanaka & Farah, 1993; Yin, 1969; see Tanaka & Gordon, 2011, and Tanaka & Simonyi, 2016 for reviews).

Accounts of expertise in face identification are qualified by the fact that we are not experts at all face identification tasks. In most applied contexts, such as when police officers compare CCTV images to mugshot records, or border control officers compare the face of a traveller to their passport image, the faces being compared are *unfamiliar* to the viewer. When faces are unfamiliar, people show surprisingly poor performance, even under optimal conditions for matching; such as when comparing good quality images taken on the same day, in the same neutral pose, and lighting conditions. Error rates in these tasks, which require participants to compare images presented concurrently and so rely little on memory, typically range between 20 and 30 percent (Bruce, Henderson, Newman, & Burton, 2001; Burton, Wilson, Cowan, & Bruce, 1999; Kemp, Towell, & Pike, 1997).

Given the importance of unfamiliar face matching in applied contexts, and the difficulty people experience in performing this task, recent empirical work has begun to investigate whether training can improve accuracy (e.g., Dowsett & Burton, 2014; Towler, White, & Kemp, 2014; White, Kemp, Jenkins, & Burton, 2014). Further, in professions

where face matching is performed in daily work, training is often provided to staff. Perhaps surprisingly, this training typically encourages *featural* comparisons (Facial Identification Scientific Working Group, 2012; White, Kemp, Jenkins, Matheson, & Burton, 2014; White, Phillips, Hahn, Hill, & O'Toole, 2015). This type of training could be expected to *harm* identification accuracy given that piecemeal feature-by-feature processing usually impairs face recognition accuracy (e.g., Berman & Cutler, 1998; Patterson & Baddeley, 1997; Yin, 1969).

Here we investigate whether a feature-by-feature comparison strategy can improve accuracy in face matching, a task that does not rely on memory. Although contrary to the notion that face recognition relies on holistic representations, evidence suggests that performance in unfamiliar face matching tasks engages qualitatively different processes to those recruited in face memory tasks. Megreya and Burton (2006) found that individual differences in unfamiliar face matching accuracy were not predicted by performance in familiar face matching, but *were* predicted by matching accuracy when faces were presented upside down. Importantly, when faces were familiarised, the relationship between matching accuracy for upright and inverted faces broke down, providing strong evidence that familiar and unfamiliar face processing engage qualitatively different processes.

Further, to the extent that access to holistic information is impeded by image inversion (see e.g., Rossion, 2008; Tanaka & Farah, 1993; Young, Hellawell, & Hay, 1987), Megreya and Burton's finding also suggests that unfamiliar face matching is not as reliant on holistic information as familiar face processing. This is consistent with experimental evidence showing that spatial relations between features are less important for unfamiliar than for familiar face identification (Lobmaier & Mast, 2007; Ramon, 2015), and that forensic experts in unfamiliar face matching are less impaired by face inversion than novices (White, Phillips, Hahn, Hill, & O'Toole, 2015). Given that performance in unfamiliar face matching appears to

be more reliant on one-to-one comparison of local properties, facial image comparison practitioners may be justified in adopting a feature comparison strategy. However, the effectiveness of this approach has not been tested directly. This is a particularly important aim, given that previous studies show no benefit of workplace training (Woodhead, Baddeley, & Simmonds, 1979), and that a common component of modern workplace training – teaching practitioners to classify face shape – does not benefit identification (Towler, White, & Kemp, 2014).

Here we investigate whether feature comparison improves unfamiliar face matching performance. First, we test whether comparing feature similarity improves matching accuracy in university students (Experiments 1 and 2). Second, we compare performance of students to specialist facial examiners from the Australian Passport Office who have received extensive training in feature comparison (Experiment 3). Despite evidence that professional experience alone does not improve face matching performance (Burton, Wilson, Cowan, & Bruce, 1999; White, Kemp, Jenkins, Matheson, & Burton, 2014), some recent studies of forensic facial examiners have shown superior performance in these groups (Norell et al., 2014; White, Phillips, Hahn, Hill, & O'Toole, 2015; White, Dunn, Schmid, & Kemp, 2015; Wilkinson & Evans, 2008), suggesting that forensic training and experience in close examination of facial images may benefit face identification accuracy.

Previous work suggests that superior accuracy of forensic facial examiners is supported by greater reliance on feature comparison (White, Phillips, Hahn, Hill, & O'Toole, 2015). Therefore, Experiments 2 and 3 also test the extent to which similarity ratings are themselves diagnostic of identity – by measuring how well similarity ratings discriminate matching from non-matching image pairs. In addition to theoretical motivations, we hoped this would inform strategic approaches to feature comparison in applied settings by revealing the features that ought to be given highest priority in identification judgements. In

Experiment 3, this approach also enables us to examine differences between students and forensic examiners more closely, by comparing their sensitivity to the identity information available within individual facial features.

EXPERIMENT 1

In Experiment 1, we test whether feature comparison improves unfamiliar face matching performance by asking untrained students to rate the similarity of facial features in image pairs prior to making same/different identity judgments. We also test whether feature comparison carries a training benefit that transfers to subsequent face matching decisions made when participants are not required to compare facial features, by administering the Glasgow Face Matching Test (GFMT; Burton, White, & McNeill, 2010) before and after feature comparison training.

Method

Participants

Eighty-eight undergraduate psychology students (Mean age = 19 years, *SD* = 2 years, 61 females) participated in return for course credit. Participants were randomly allocated to either the ratings or no ratings group, such that each group contained 44 participants.

Materials

Feature Rating Task

A major motivation of this study was to evaluate forensic identification procedures. Therefore, in all experiments we used an unfamiliar face matching task designed to model the challenging conditions encountered in forensic casework. Although forensic face identification decisions are increasingly facilitated by automatic facial recognition (AFR) software, final identification decisions continue to be made by humans (Dessimoz & Champod, 2008; Grother & Ngan, 2014; Jain, Klare, & Park, 2012; White, Phillips, Hahn,

Hill, & O'Toole, 2015). To model this workflow, and to ensure that the test was challenging for expert populations, we selected pairs of face images that produce high proportions of face matching errors in both computer algorithms and humans.

We selected image pairs for the Feature Rating Task from a test that has been used in previous research to assess expertise of forensic facial examiners (*Expertise in Facial Comparison Test*; EFCT; White, Phillips, Hahn, Hill, & O'Toole, 2015). Images in this test were selected from *The Good, The Bad and The Ugly Challenge* dataset (GBU; Phillips et al., 2012), which sorts image pairs into three levels of ascending difficulty according to match score data generated by leading face recognition algorithms. This dataset contains frontal face images captured in challenging, unconstrained environmental conditions (i.e. with minimal control of illumination, expression, and appearance). The EFCT contains images from the Bad and Ugly portions of the GBU dataset, representing pairs that produced moderate and poor levels of algorithm performance respectively. For the purpose of the current study we randomly selected 20 matching and 20 non-matching image pairs from the most challenging portion (i.e. Ugly only).

Glasgow Face Matching Test

The short version of the GFMT is a standardised test of face matching ability and consists of 40 simultaneous pairwise same/different identity decisions (see Figure 1; Burton, White, & McNeill, 2010). In recent work we used item accuracy data to divide this test into two equally difficult sub-tests of 20 items each that can be administered before and after training (see Towler, White, & Kemp, 2014 for details). Increased accuracy from pre-to-post training is indicative of training-based improvements, so we used this test to measure transfer effects from the Feature Rating Task to face matching decisions in which participants did not rate feature similarity.

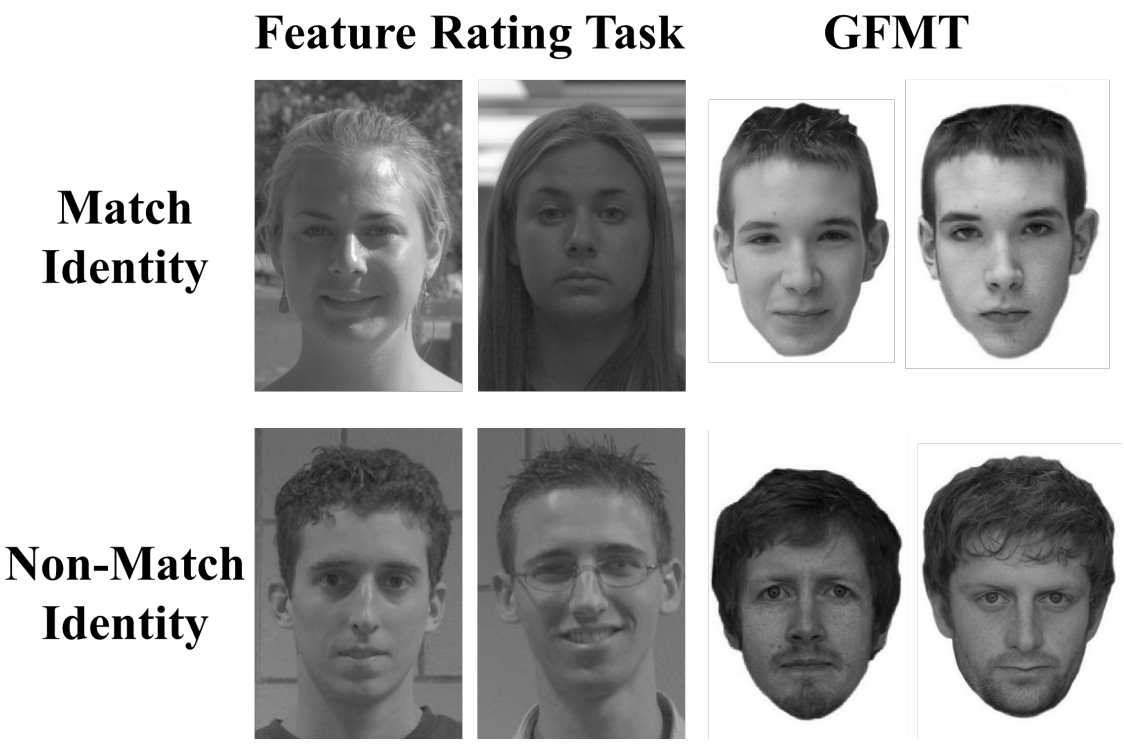

Figure 1. Example image pairs from the GBU dataset (Phillips et al., 2012) used in the Feature Rating Tasks in Experiments 1, 2 and 3 (left), and example items from the Glasgow Face Matching Test (Burton et al., 2010) used in Experiment 1 (right).

Procedure

For the Feature Rating Task, participants in the ratings group were presented with image pairs simultaneously on a computer monitor. Below each image pair was a list of eleven facial features taken from international best practice guidelines for forensic facial examination (Facial Identification Scientific Working Group, 2012; see Figure 4 for the full list of facial features). Participants were asked to '*rate the similarity of the following facial features*' using a 5-point scale from 1 (very dissimilar appearance) to 5 (very similar appearance). After a response had been made for all facial features, participants were able to submit their ratings. The similarity rating scales were replaced by a text prompt '*Are these images of the same person or different people?*' and participants responded using buttons

labelled ‘same’ and ‘different’. The face pairs remained onscreen for the duration of the trial and the entire task was self-paced. Participants in the no ratings group did not make feature similarity ratings and instead made a single same/different identity response on each trial. Both groups of participants completed a randomised sequence of 40 trials in the Feature Rating Task, with half of trials displaying matching identities and half non-matching. All participants completed the two GFMT sub-tests, one before and one after the Feature Rating Task, with sub-test order counterbalanced across participants. The entire session took approximately 1 hour to complete for participants in the ratings group, and 30 minutes for the no ratings group.

Results

Feature Rating Task

In all experiments our primary analysis was performed on response accuracy for match and non-match trials. Expressing accuracy in this way is important to translate performance measures into real-world face identification tasks, and errors in these trial types have distinct implications for different applied settings. Classifying errors in this way may also reveal strategies adopted by forensic facial examiners who are taught to look for ‘unexplainable’ and ‘exclusionary’ differences between face images (e.g., Scientific Working Group Imaging Technology, 2010; c.f. Ulery, 2011). Nevertheless, it is important to confirm whether effects observed in accuracy data are reflective of improved sensitivity to the task relevant information, or instead reflect a shift in response criteria. Therefore, we also report signal detection measures of sensitivity (d-prime) and response bias (criterion) for each experiment.

Accuracy

A 2 x 2 mixed ANOVA was conducted on the accuracy of the same/different identity decisions, with Ratings (ratings, no ratings) as a between-subjects factor and Trial Type (match, non-match) as a within-subjects factor (see Figure 2). Main effects of Ratings [$F(1, 86) = 14.22$, $p < .001$, $\eta_p^2 = .14$] and Trial Type [$F(1, 86) = 43.52$, $p < .001$, $\eta_p^2 = .34$] were significant, as was the interaction between Ratings and Trial Type [$F(1, 86) = 18.48$, $p < .001$, $\eta_p^2 = .18$]. Simple Main Effects analysis revealed that rating the similarity of facial features improved matching accuracy on match [$F(1, 172) = 32.18$, $p < .001$, $\eta_p^2 = .16$, Cohen's $d = 1.12$ (95% CI: 1.09 – 1.18)], but not non-match trials [$F(1, 172) = 2.09$, $p > .05$, $\eta_p^2 = .01$, Cohen's $d = 0.35$ (95% CI: 0.31 – 0.39)].

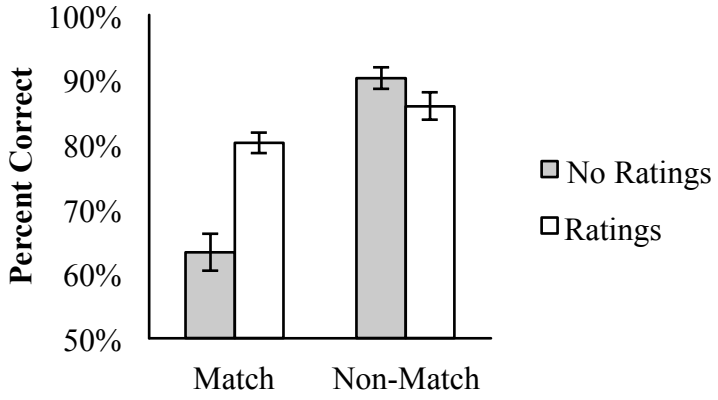

Figure 2. Accuracy scores on match and non-match trials for the no ratings (grey bars) and ratings groups (white bars). Error bars show standard error of the mean.

Signal Detection Analysis

Summary signal detection measures are shown in Table 1. For sensitivity, the difference between the ratings and no ratings groups was non-significant [$t(86) = 1.16$, $p > .05$, Cohen's $d = 0.25$ (95% CI: -0.17 – 0.67)]. For criterion scores, the difference between groups was significant [$t(86) = 2.91$, $p < .05$, Cohen's $d = 0.62$ (95% CI: 0.18 – 1.04)], with

bias reduced in the feature similarity ratings condition, suggesting that the overall improvement in accuracy measures reflects a more appropriately balanced response criteria in the feature rating condition.

	Sensitivity (d')		*Response bias (c)*	
	Ratings	*No Ratings*	*Ratings*	*No Ratings*
Experiment 1	2.38 (0.14)	2.15 (0.13)	0.22 (0.09)	0.62 (0.11)
Experiment 2				
Facial Features	2.74 (0.15)	2.17 (0.14)	-0.32 (0.11)	0.26 (0.08)
Image Quality	2.29 (0.18)	1.88 (0.15)	0.05 (0.11)	0.39 (0.11)
Personality Traits	2.56 (0.21)	2.49 (0.21)	-0.03 (0.15)	0.46 (0.13)
	Upright	*Inverted*	*Upright*	*Inverted*
Experiment 3				
Examiners	4.35 (0.55)	3.03 (0.34)	0.19 (0.26)	0.21 (0.39)
Students	2.87 (0.17)	1.40 (0.17)	0.12 (0.12)	0.03 (0.10)

Table 1. Signal detection measures of sensitivity (d') and response bias (c) for Experiments 1, 2, and 3. Standard Errors are shown in parentheses.

GFMT

Overall accuracy on the GFMT was 83.1% (*SD* = 9.9%) and 80.7% (*SD* = 11.2%) for the ratings and no ratings groups respectively. These scores align with normative performance on this test (81.3%, SD = 9.7%; Burton et al., 2010). A 2 x 2 mixed ANOVA was conducted on the GFMT data with Ratings (ratings, no ratings) as a between-subjects factor and Test (pre- and post-test) as a within-subjects factor. There was a significant main effect of Test [$F(1, 86) = 4.88$, $p < .05$, $\eta_p^2 = .05$], with accuracy across the groups declining from 83.3% to 80.5% between the two sub-tests. The main effect of Ratings [$F(1, 86) = 1.17$, $p > .05$, $\eta_p^2 = .01$], and interaction of Ratings and Test [$F(1, 86) = 1.45$, $p > .05$, $\eta_p^2 = .02$] however, were not significant. Thus, results show that any benefit of feature ratings was

specific to the image pairs for which feature similarity had been compared and did not generalise to new stimuli.

Discussion

Experiment 1 provides some evidence that comparing the similarity of facial features improves the accuracy of face matching decisions, with higher accuracy observed in match trials for the feature rating group. Although no difference was observed in non-match accuracy, the feature rating group were more accurate overall compared to participants in the control group. However, one possible account of this result is that rating feature similarity simply made participants more likely to respond 'match', without improving their ability to discriminate matching from non-matching pairs. This is supported by signal detection analysis, which shows no difference in sensitivity between groups, but rather a difference in response criterion.

It is possible that the stimulus set used in this study contributed to this result. The difficulty of image pairs in the GBU dataset is determined by varying the similarity of matching pairs only, while keeping similarity of non-match pairs constant across the Good, Bad and Ugly partitions of the image set (see Phillips et al., 2012). Randomly selecting pairs from the Ugly portion of this set for use in Experiment 1 appears to have resulted in a more difficult set of match compared to non-match trials. Non-match trial accuracy was close to ceiling accuracy, so it is possible that this masked any differences in sensitivity and that changes instead emerged as differences in response criterion. We address this limitation in Experiment 2. We also extend our investigation to ask whether other types of similarity comparisons that are not directed to individual facial features also improve face matching accuracy.

EXPERIMENT 2

In Experiment 2 we aimed to resolve methodological issues in the previous experiment, and also test whether benefits of similarity ratings are observed for other attributes of facial images. To address this question, participants compared faces by rating the similarity of either facial features, personality traits or image quality. Memory for faces is superior when participants rate personality traits during encoding of face stimuli, relative to when they make feature judgements (for meta-analysis see Coin & Tiberghein, 1997). Further, elaborate encoding of face stimuli in a social setting has been shown to improve face matching accuracy (Bruce, Henderson, Newman, & Burton, 2001). We therefore tested whether rating the similarity of personality traits would also improve matching accuracy. To test the boundaries of the effect observed in Experiment 1 further, we test whether comparing superficial elements of the images that do not pertain to identity of the face also improves matching accuracy.

Method

Participants

Participants were 102 undergraduate psychology students (Mean age = 20 years, *SD* = 4 years, 63 females) who had not participated in Experiment 1. Participants were randomly allocated to the facial feature, personality trait or image quality rating groups, such that each group contained 34 participants. This experiment employed a within-subjects design such that each participant completed face matching trials in both rating and no rating conditions.

Materials

We sampled a new set of stimuli from the EFCT items (White, Phillips, Hahn, Hill, & O'Toole, 2015). To increase the number of test items, image pairs were sampled from both Bad and Ugly portions of the image data described by Phillips et al. (2012). Because we observed very high accuracy on non-match trials in both conditions in Experiment 1, it is

possible that the absence of any benefit of feature rating in non-match trials was caused by ceiling levels of performance. To address this possibility in Experiment 2 we used human accuracy data from O'Toole, An, Dunlop, & Natu (2012) to equate the difficulty of match and non-match trials, by selecting 60 of the most difficult non-matching face pairs, and 60 matching face pairs of equivalent difficulty. Further, to facilitate analysis of feature rating data in this experiment, we only selected image pairs where all rated features were visible.

Procedure

Participants completed two blocks of trials. First, to obtain a baseline measure of matching performance, all participants completed a control block of 60 trials where participants made self-paced same/different identity judgements to simultaneously presented face pairs. Following the control block, participants completed a ratings block in which they made similarity ratings prior to making same/different identity judgments, using the same procedure as in Experiment 1. Each participant rated the similarity of 11 stimulus attributes: participants in the *facial feature* group rated similarity of the same set of facial features described in Experiment 1 (e.g. ears, eyes, nose etc.), participants in the *personality trait* group rated the similarity of the faces with respect to 11 social attributions (e.g. trustworthiness, creativity, honesty etc.) and participants in the *image quality* group rated the similarity of image-level artefacts (e.g. sharpness, contrast, brightness etc.). A full list of attributes is shown in Figure 4.

Results

Accuracy

Accuracy scores for same/different identity decisions are shown in Figure 3. These data were analysed using a 2 x 2 x 3 mixed ANOVA with Ratings (ratings, no ratings) and Trial Type (match, non-match) as within-subjects factors and Group (feature, personality, image) as a between-subjects factor. The main effects of Ratings [$F(1, 99) = 13.34$, $p < .001$,

η_p^2 = .12] and Trial Type [$F(1, 99) = 15.04$, $p < .001$ η_p^2 = .13] were significant, and the main effect of Group was marginally significant [$F(2, 99) = 2.76$, $p = .068$, η_p^2 = .05]. Consistent with Experiment 1, these effects were qualified by a significant interaction between Ratings and Trial Type [$F(1, 99) = 99.99$, $p < .001$, η_p^2= .50], with ratings improving performance on match trials but not non-match trials. The interaction between Trial Type and Group was marginally significant [$F(2, 99) = 2.49$, $p = .088$, η_p^2 = .05], driven by generally superior performance in the feature ratings group in match trials. The interaction between Group and Ratings [$F(2, 99) = 1.04$, $p > .05$, η_p^2 = .02], and the three-way interaction between Group, Ratings and Trial Type were non-significant [$F<1$, η_p^2 = .02].

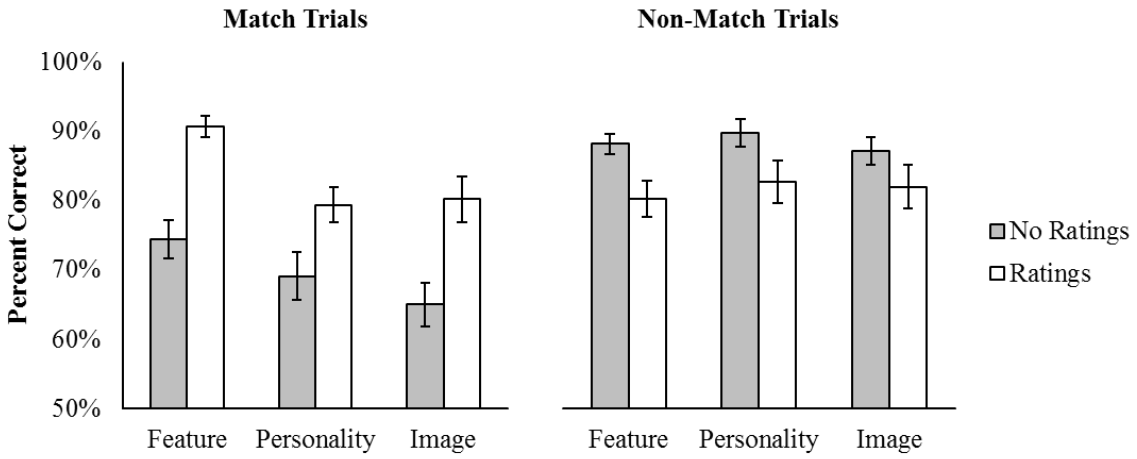

Figure 3. Face matching accuracy data on match and non-match trials for the feature, personality and image groups. Error bars show standard error of the mean.

Signal Detection Analysis

Summary signal detection measures are shown in Table 1. Analysis of sensitivity scores revealed a significant main effect of Ratings [$F(1, 99) = 13.19$, $p < .001$, η_p^2 = .12], but a non-significant main effect of Group [$F(2, 99) = 2.32$, $p > .05$, η_p^2 = .05]. Main effects were qualified by a significant interaction between Ratings and Group [$F(2, 99) = 2.37$, $p > .05$, η_p^2 = .05]. Simple Main Effects analysis revealed that sensitivity was greater in the ratings block

compared to the no ratings block for the facial feature rating group [$F(1, 33) = 11.70$, $p < .001$, Cohen's $d = 0.68$ (95% CI: 0.46 – 0.90)] and image quality group [$F(1, 33) = 6.06$, $p < .05$, Cohen's $d = 0.41$ (95% CI: 0.16 – 0.66)]. Sensitivity in the personality trait group did not differ between the no ratings block to the ratings block [$F < 1$, Cohen's $d = 0.07$ (95% CI: -0.15 – 0.29)].

For criterion scores, the main effect of Ratings was significant [$F(1, 99) = 60.99$, $p < .001$, $\eta_p^2 = .38$]. The main effect of Group was non-significant [$F(2, 99) = 1.88$, $p > .05$, $\eta_p^2 = .04$], as was the interaction between Ratings and Group [$F(2, 99) = 1.46$, $p > .05$, $\eta_p^2 = .03$], reflective of a tendency to respond "match" more often in the ratings condition for all groups [feature: $t(33) = 5.23$, $p < .001$, Cohen's $d = 0.91$ (95% CI: 0.69 – 1.13); image: $t(33) = 3.32$, $p < .05$, Cohen's $d = 0.58$ (95% CI: 0.42 – 0.74); personality: $t(33) = 4.91$, $p < .001$, Cohen's $d = 0.86$ (95% CI: 0.72 – 1.00)].

Diagnosticity of feature similarity ratings

Results of Experiment 2 suggest that similarity ratings improve the accuracy of same/different face matching decisions. However, we also predicted that individual similarity ratings would themselves carry identifying information that could be useful in determining identity. To examine the diagnostic value of each attribute, we calculated the Area Under the ROC Curve (AUC) for each attribute, separately for each participant. This provides a measure of the extent to which a similarity ratings, to a given attribute, accurately predicted whether image pairs were of matching identities. An AUC score of 1 represents perfect discrimination of match and non-match trials, such that similarity ratings predicted identity for 100% of image pairs, whereas an AUC score of 0.5 indicates chance-level discrimination.

Mean AUC values for each attribute are shown in Figure 4. This figure clearly shows that facial feature ratings provide good levels of discrimination, with all features achieving an AUC of 0.77 or higher. Quite surprisingly, ears ranked as the most identifying feature (AUC

= 0.84), contrasting with *memory*-based identification tasks where eyes have instead been particularly diagnostic of identity (e.g. Schyns, Bonnar & Gosselin, 2002; Vinnette, Goselin & Schyns, 2004). What is also very clear is that similarity ratings of perceived personality traits and image quality were not at all useful for identification. To confirm this, we averaged ratings across attributes for each image pair, before computing AUC scores separately for each participant, using the average scores as the predictor variable (denoted as 'Average' in Figure 4). Comparing these scores to chance-level discrimination (.5) confirms that facial feature ratings were very diagnostic of identity [$M = 0.90$; $t(33) = 34.01$, $p < .001$, Cohen's $d = 5.83$ (95% CI: 5.50 – 6.16)]. In contrast, personality traits did not discriminate identity above chance [$M = 0.50$; $t(33) = 0.01$, $p > .05$, Cohen's $d = 0.00$ (95% CI: -0.33 – 0.33)] and image quality scores were significantly below chance [$M = 0.43$; $t(33) = 4.32$, $p < .001$, Cohen's $d = 0.74$ (95% CI: 0.41 – 1.07)].

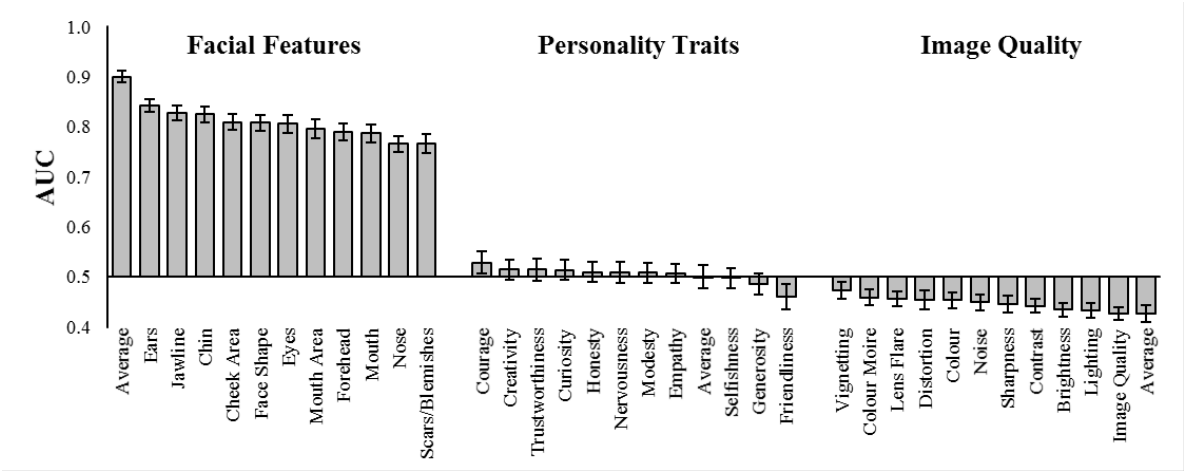

Figure 4. Diagnosticity of similarity ratings in discriminating between match and non-match trials for each rating type, ordered from left to right by mean Area Under the ROC Curve (AUC) scores. Error bars show standard error of the mean.

Discussion

Results of Experiment 2 show that requiring participants to make similarity ratings improved overall accuracy of subsequent same/different face matching decisions. As in Experiment 1, the effect of similarity ratings was specific to match trials, with no improvement to accuracy in non-match trials. In addition to the effects in accuracy, Experiment 2 revealed higher *sensitivity* when making similarity ratings prior to identity judgments. This confirms that improvements in accuracy were not caused by a shift in response bias alone, but instead reflect improved identity discrimination. Importantly, signal detection analysis also show that the effect of similarity ratings on sensitivity varied as a function of rating type. Facial feature ratings produced the strongest benefit to performance, with image quality ratings also benefiting performance[1]. However, the effect of personality trait ratings on sensitivity was non-significant, suggesting that discrimination performance is not only determined by the degree of attention paid to the stimulus, but also by the nature of the comparison performed.

Further benefits of facial feature comparison were observed when we analysed the extent to which similarity ratings were themselves diagnostic of identity. Facial feature ratings reliably discriminated matching from non-matching image pairs, whereas similarity ratings to personality traits and image quality did not[2]. Moreover, the average of facial feature similarity ratings proved to be a very accurate predictor of identity, producing an AUC score of .90. This represents impressive discrimination in the context of mean same/different accuracy in feature rating groups of 85%, reflecting similar benefits of response averaging that have been found when averaging identity judgments made by different individuals

[1] The advantage of image quality ratings reinforces the notion that face matching engages image-based processes to a greater extent than 'holistic' face processes (see Megreya & Burton, 2006; Longmore, Liu, & Young, 2008).

[2] That perceived personality was not predictive of identity is consistent with the observation that variation in perceptions of personality traits are often larger between different images of the same face than between different faces (Jenkins, White, Van Montfort, & Burton, 2011; Todorov & Porter, 2014).

(e.g.White, Burton, Kemp, & Jenkins, 2013). Here, averaging multiple facial feature judgements made by the same individual provided benefits relative to the single most diagnostic feature – and also over explicit same/different judgments

Given these apparent benefits of facial feature comparison for face identification, it is possible that close analysis of facial features also benefits face matching performance in the longer-term, enabling forensic examiners to extract more identifying information from facial features with experience. To address this, in the next experiment we test performance of forensic examiners who have received rigorous training in featural comparison of face images as part of their role in the Identity Resolution Unit at the Australian Passport Office.

EXPERIMENT 3

In this experiment, we compare student performance on the Feature Rating Task to a group of forensic facial examiners who have received rigorous training emphasising deliberate feature-by-feature comparison of facial images (see Facial Identification Scientific Working Group, 2012; White, Phillips, Hahn, Hill, & O'Toole, 2015). To probe the nature of expertise in forensic facial examiners further, we also compare the size of face inversion effects in examiners and untrained students (Carey, 1992; Carey & Diamond, 1977; Diamond & Carey, 1986; Valentine, 1988). In a recent study by White, Phillips et al. (2015), an international group of examiners, with similar training and experience to the examiners tested here, showed smaller face inversion effects relative to control groups, suggesting that expertise in forensic facial examiners differs qualitatively from the type of expertise that drives high levels of performance in untrained participants (e.g. Russell, Duchaine, & Nakayama, 2009). Here we examine this possibility in more detail by comparing the sensitivity of students and examiners to identifying information contained in individual facial features.

Method

Participants

We recruited seven forensic facial examiners who make up the specialist Identity Resolution Unit at the Australian Passport Office (Mean age = 41 years, *SD* = 11 years, 4 females). These staff perform detailed comparison of images referred to them in cases of suspected identity fraud and have been shown to perform better than students and non-specialist passport officers in previous work (see White, Dunn, Schmid, & Kemp, 2015: Experiment 2, *Facial Examiner* group). To make identification decisions, facial examiners are trained to compare the similarity of individual facial features across images. In addition, when identity fraud is suspected, feature-based analysis is used as the basis of formal reports that can be submitted as evidence in legal proceedings. We compared performance of these facial examiners to a control group of forty undergraduate psychology students (Mean age = 19 years, *SD* = 1 year, 21 females)[3], who participated in the experiment for course credit. We made a post-hoc exclusion of an outlying student participant because their matching accuracy for upright faces fell three standard deviations below the group mean.

Materials and procedure

Face pairs were sourced from the same image set used in the previous experiments. However, to ensure the task was sufficiently challenging for experts, we included images from the Ugly partition of the GBU dataset only. We then used human accuracy data (O'Toole et al., 2012) to select the 32 most difficult matching and 32 most difficult non-matching image pairs. These were then split into two stimulus sets of equal difficulty, each containing 16 match and 16 non-match pairs. One of these sets was assigned to the upright test, and the other to the inverted. Participants completed a block of upright and a block of

[3] Although there is a clear age difference between the examiners and students we did not expect this factor to systematically influence their ability to perform the task (see Burton, White, & McNeill, 2010; Grady, McIntosh, Horwitz, & Rapoport, 2000; Konar, Bennett, & Sekuler, 2013).

inverted stimulus comparisons with order counterbalanced across participants. Participants performed feature ratings on every trial using the same procedure as in previous experiments. Immediately following the Feature Rating Task participants reported the degree to which each facial feature influenced their decisions, using a scale from 1 (never) to 5 (all the time).

Results

Accuracy

Overall face matching accuracy is shown in Figure 5. Accuracy data were analysed by a 2 x 2 x 2 mixed ANOVA with Group (examiners, students) as a between-subjects factor and Orientation (upright, inverted) and Trial Type (match, non-match) as within-subjects factors. The main effect of Group was significant [$F(1, 44) = 10.47$, $p < .05$, $\eta_p^2 = .19$], with examiners ($M = 89\%$; $SD = 5\%$) performing more accurately than students ($M = 78\%$; $SD = 9\%$). The main effect of Orientation was also significant [$F(1, 44) = 56.12$, $p < .001$, $\eta_p^2 = .56$], with higher accuracy in upright ($M = 87\%$; $SD = 8\%$) than inverted trials ($M = 72\%$; $SD = 11\%$). The main effect of Trial Type [$F(1, 44) = 1.49$, $p > .05$, $\eta_p^2 = .03$], and the two-way interactions were non-significant [Group x Orientation: $F(1, 44) = 2.45$, $p > .05$, $\eta_p^2 = .05$; Group x Trial Type: $F<1$, $\eta_p^2 = .02$; Orientation x Trial Type: $F<1$, $\eta_p^2 = .01$]. Main effects were qualified by a significant three-way interaction of Group, Orientation and Trial Type [$F(1, 44) = 4.28$, $p < .05$, $\eta_p^2 = .09$]. To investigate this interaction we analysed performance data by 2 x 2 mixed ANOVA, separately for match and non-match trials.

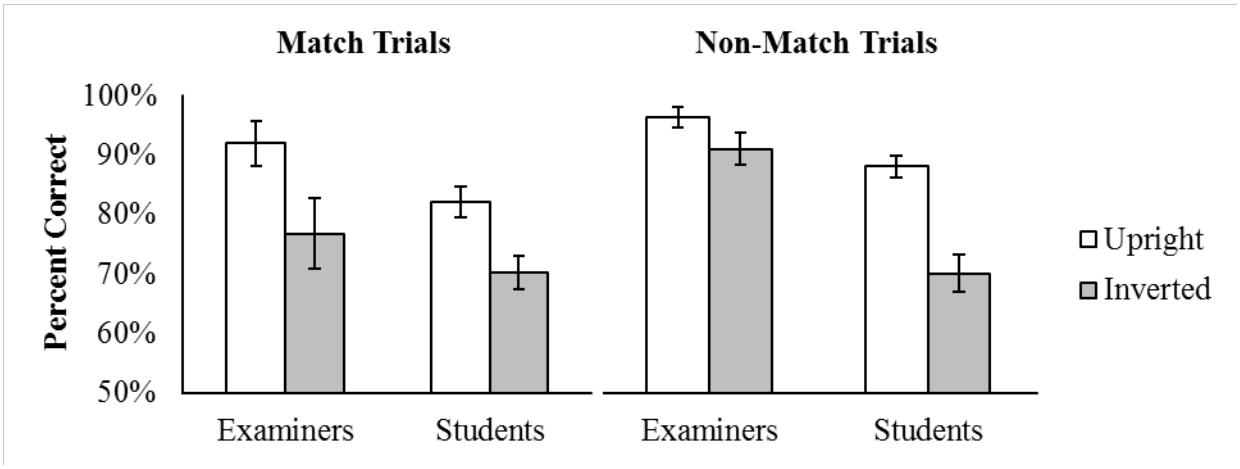

Figure 5. Accuracy on upright (white bars) and inverted (grey bars) face pairs by examiners and students on match (left) and non-match (right) trials. Error bars show standard error of the mean.

For match trials, the main effect of Orientation was significant [$F(1, 44) = 35.16$, $p < .001$, $\eta_p^2 = .44$]. The main effect of Group [$F(1, 44) = 1.41$, $p > .05$, $\eta_p^2 = .03$] and the interaction between Orientation and Group [$F<1$, $\eta_p^2 = .01$] were non-significant. For non-match trials, main effects of Orientation [$F(1, 44) = 18.13$, $p < .001$, $\eta_p^2 = .29$] and Group [$F(1, 44) = 6.98$, $p < .05$, $\eta_p^2 = .14$] were significant. Further, the two-way interaction between Orientation and Group was significant [$F(1, 44) = 5.71$, $p < .05$, $\eta_p^2 = .12$]. Simple Main Effects analysis revealed a significant effect of stimulus orientation for students, [$F(1, 38) = 72.61$, $p < .001$, Cohen's $d = 1.36$ (95% CI: 1.30 – 1.42)], but not for facial examiners, [$F(1, 6) = 1.03$, $p > .05$, Cohen's $d = 0.59$ (95% CI: 0.35 – 0.83)], showing that examiners were not impaired by inversion on non-match trials. This result could be interpreted as showing that examiners' ability to tell faces apart was not impeded by inversion. Interestingly, forensic examiners' training emphasises detection of 'unexplainable' and 'exclusionary' differences in images of faces (e.g., Scientific Working Group Imaging

Technology, 2010), raising the possibility that the strategy examiners employed to determine two images were of different people was resistant to inversion.

Signal Detection Analysis

Summary signal detection measures are shown in Table 1. Sensitivity and criterion scores were analysed by a 2 x 2 mixed ANOVA with Group (examiners, students) as a between-subjects factor and Orientation (upright, inverted) as a within-subjects factor. For sensitivity, there was a significant main effect of Group [$F(1, 44) = 16.03$, $p < .001$, $\eta_p^2 = .27$] and Orientation [$F(1, 44) = 47.54$, $p < .001$, $\eta_p^2 = .52$]. The interaction between Group and Orientation was non-significant [$F<1$, $\eta_p^2 = .00$]. For Criterion scores, main effects of Group [$F<1$, $\eta_p^2 = .01$], and Orientation [$F<1$, $\eta_p^2 = .00$], and the interaction between Group and Orientation [$F<1$, $\eta_p^2 = .00$] were non-significant.

We also computed the effect of face inversion on sensitivity individually for each participant. Given the very substantial superiority of examiners over students in d-prime scores for upright stimuli, and following from previous studies comparing inversions effects in high performers to control participants (Russell, Duchaine, & Nakayama, 2009), we calculated the size of face inversion effects as a proportion of performance with upright image pairs. An independent t-test of this data revealed a significantly smaller inversion effect for examiners ($M = 0.25$; $SD = 0.30$) than for students [$M = 0.53$; $SD = 0.30$; $t(44) = 2.24$, $p < .05$; *Cohen's d* = 0.93 (95% CI: 0.09 – 1.74)].

Diagnosticity of feature ratings

Analysis of accuracy and sensitivity show a reduced inversion effect in forensic examiners, suggesting that examiners are less reliant on holistic properties of face images when making identity judgments. Following from Experiment 2, we next calculated the extent to which feature ratings themselves were diagnostic of identity by computing AUC scores separately for each participant. These data are shown in Figure 6A.

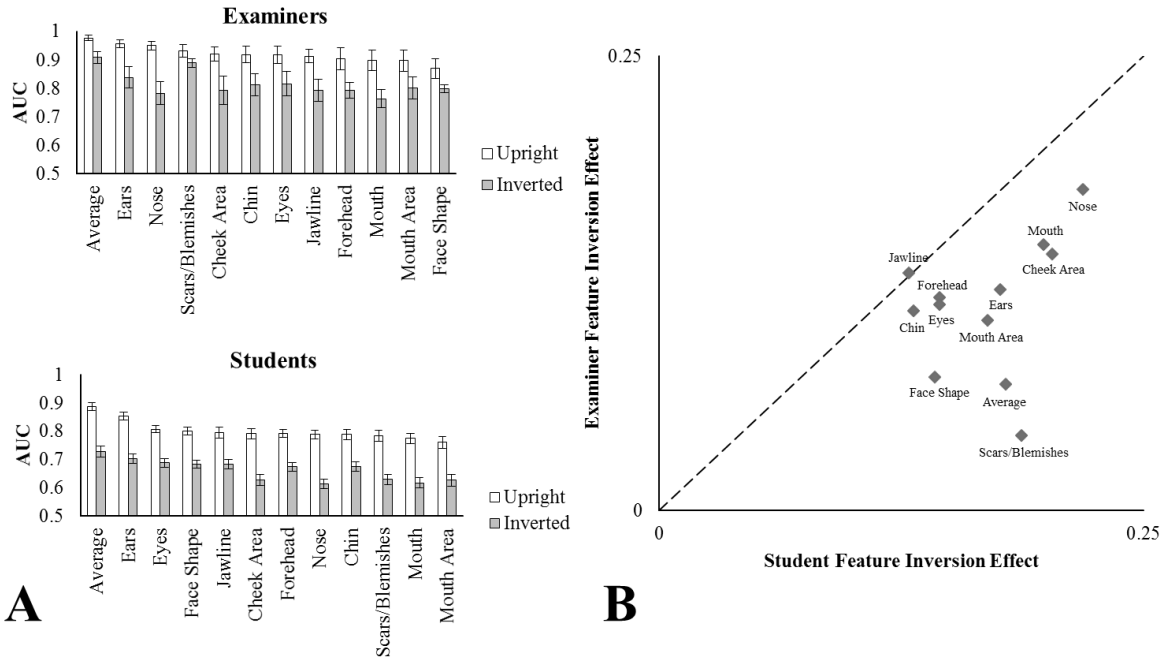

Figure 6. Diagnosticity of feature ratings made by examiners and students in Experiment 3. (A) AUC scores for facial feature similarity ratings for upright (white bars) and inverted face pairs (grey bars). Features are ordered from left to right by AUC scores for upright features separately for each group. Error bars show standard error of the mean. (B) Forensic examiner Feature Inversion Effect scores for each facial feature plotted against student Feature Inversion Effect scores. In this graph distance from the diagonal line indexes the divergence between feature inversion effects in students and examiners, with data points below the line signifying smaller inversion effects for examiners compared to students.

First, we computed AUC scores using average feature similarity ratings for each participant as the predictor variable. AUC data were analysed by 2 x 2 mixed ANOVA with Group (examiners, students) as a between-subjects factor and Orientation (upright, inverted) as a within-subjects factor. The main effects of Group [$F(1, 44) = 14.98$, $p < .001$, $\eta_p^2 = .25$] and Orientation were significant [$F(1, 44) = 29.91$, $p < .001$, $\eta_p^2 = .41$]. Consistent with the accuracy of same/different decisions, the interaction between Group and Orientation was

significant [$F(1, 44) = 4.86$, $p < .05$, $\eta_p^2 = .10$], with diagnosticity of examiners' feature ratings less impaired by inversion than students' ratings.

Consistent with Experiment 2, ears were most diagnostic of identity in both upright and inverted orientations (see Figure 6A). Some differences are also apparent between groups. In particular, examiners were more sensitive to the similarity of scars and blemishes than the students. The usefulness of facial marks for identification is emphasised in professional facial identification training courses (Facial Identification Scientific Working Group, 2011), and so workplace training may explain why students appear less able to use facial marks as a cue to identity than examiners.

Compared to students, examiners clearly exhibited smaller inversion effects both in same/different judgment accuracy and in the overall discrimination of their feature ratings. Figure 6B enables comparison of the effect of inversion on the discriminative value of individual feature ratings for students and examiners. Two things are clear from this figure. First, examiners consistently display smaller inversion effects for individual feature ratings compared to students, as indicated by data points falling below the diagonal line. Second, as indicated by the outlying data point, inversion has almost no effect on the diagnosticity of the scars and blemishes ratings made by examiners, but has a large effect on the diagnosticity of these ratings by students. This suggests that when trained to examine scars and blemishes, comparison of facial marks is largely invariant to inversion, and so may partly account for reduced face inversion effects in forensic examiners reported here and in previous studies (White, Phillips, Hahn, Hill, & O'Toole, 2015).

Feature usefulness ratings

At the end of the rating task, participants were asked to rate the degree to which they had used each facial feature when making same/different identity decisions on a scale from 1 (never) to 5 (all the time). These data are shown in Figure 7. Overall, examiners reported

higher levels of feature usefulness in comparison to students. Moreover, examiners appeared to consciously prioritise different feature sets than students. Most strikingly, examiners reported using the ears and scars and blemishes more than the other features. This same pattern was not observed in the students' ratings, suggesting that examiners are more aware of the features influencing their decisions.

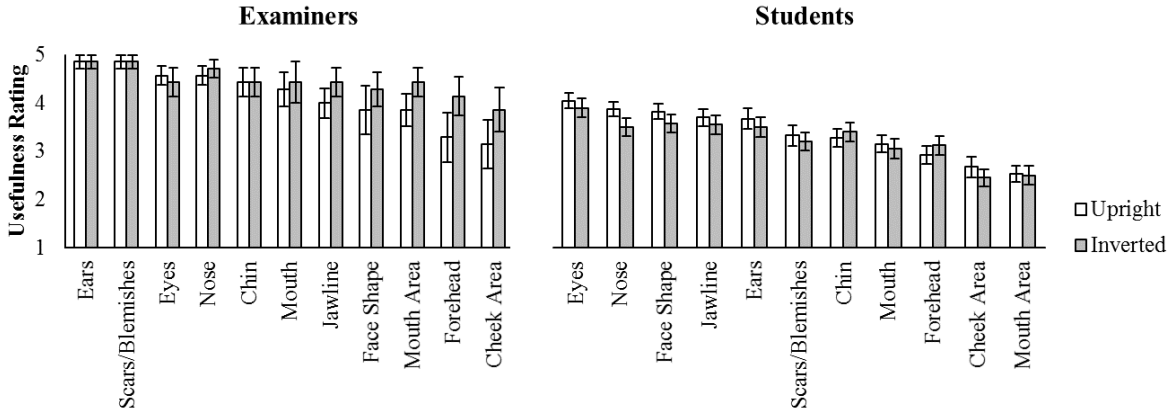

Figure 7. Self-reported facial feature usefulness ratings (1-5) for each facial feature for upright (white bars) and inverted (grey bars) faces made by examiners and students. Error bars show standard error of the mean.

Fusion analysis

When comparing facial images in forensic and security settings, it is common to solicit multiple examiners to make independent identification judgments. In addition, previous studies have shown large gains in identity discrimination can be achieved by aggregating similarity ratings made to image pairs in increasing group sizes (White, Burton, Kemp, & Jenkins, 2013; White, Phillips, Hahn, Hill, & O'Toole, 2015). Here, we conducted simulations to examine the effect of averaging feature similarity ratings on identity discrimination, separately for examiner and student groups and for upright and inverted stimuli. First, we randomly sampled *n* participants and calculated their group average similarity rating for each facial feature, for each image pair in the set. We then calculated the

extent to which this average feature rating vector discriminated between match and non-matching identity pairs, by computing Area Under the ROC Curve (AUC). This sampling procedure was repeated 100 times for each value of *n* (i.e. group sizes 1 to 7) and average diagnosticity of a given feature was measured as the average AUC score across all iterations.

Results of the fusion analysis are shown in Figure 8. Extending previous work showing benefits of aggregating whole face comparisons (White, Burton, Kemp, & Jenkins, 2013; White, Phillips, Hahn, Hill, & O'Toole, 2015), we found large benefits in identification accuracy by aggregating feature similarity ratings across participants. Combining across individual feature ratings to produce a single similarity rating for each participant on each image pair produced very impressive identity discrimination, with examiners producing AUC scores of .99 at group sizes of just 2, and perfect discrimination (AUC = 1.0) at group size of 7. The examiners' expertise is further highlighted by the observation that performance of a single examiner (0.98) was equal to groups of 7 or more students (0.98).

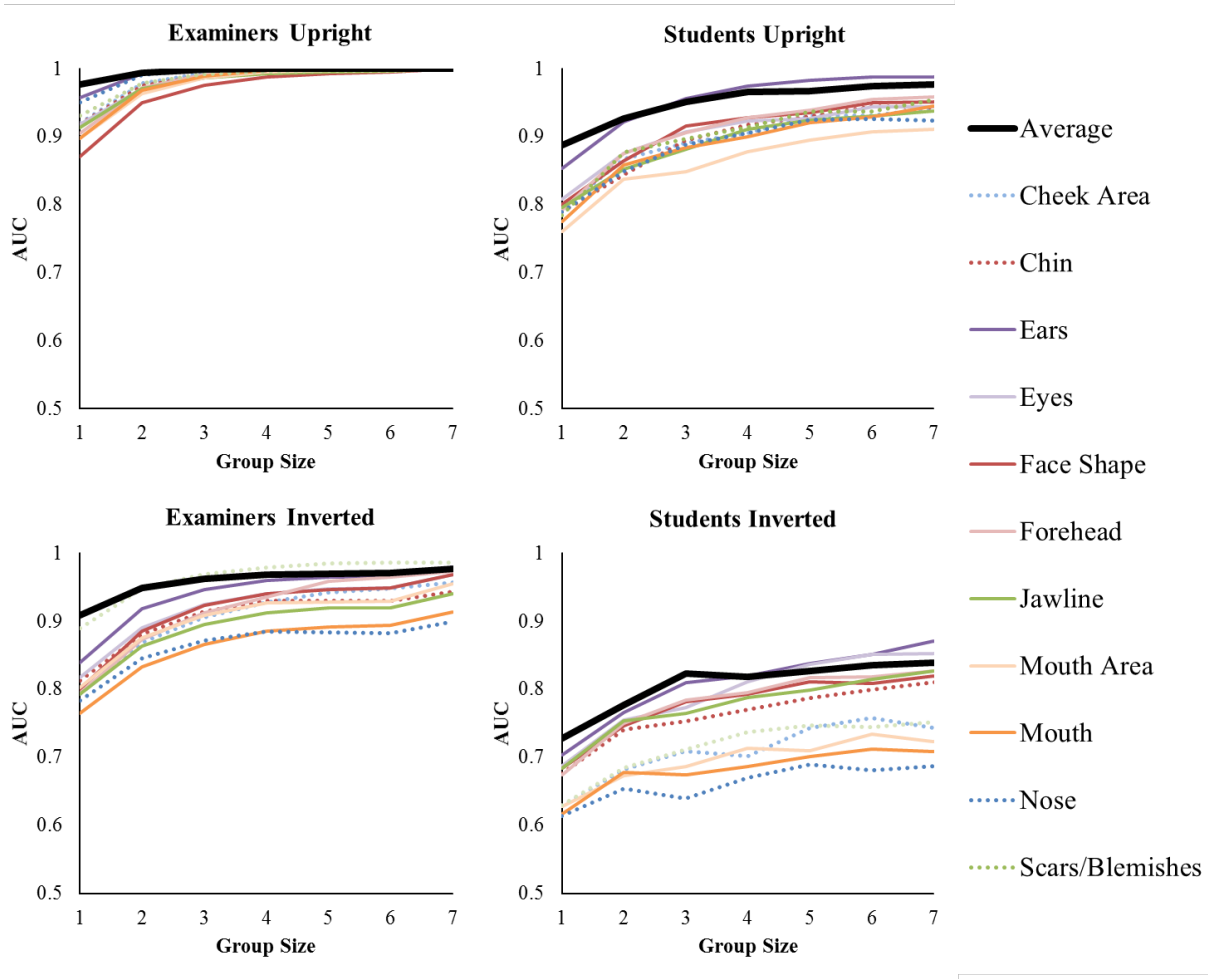

Figure 8. Combinatorial 'Fusion' analysis showing the extent to which facial feature ratings were diagnostic of identity, after combining scores for group sizes between 1 and 7. Diagnosticity is measured as Area Under ROC Curve (AUC), which provides an unbiased measure of the extent that feature rating vectors discriminate match from non-matching trials. Details of this analysis are provided in the main text. A colour version of the figure is presented in Supplementary Materials.

GENERAL DISCUSSION

We conducted three experiments to assess the effectiveness of feature comparison strategies for improving unfamiliar face matching accuracy. In Experiments 1 and 2, untrained student participants showed improved face matching accuracy after rating the similarity of facial features. Further, in Experiment 2, benefits of facial feature ratings were

greater than benefits of rating faces on other attributes, and judgments of personality trait similarities did not improve perceptual sensitivity on subsequent identity judgments.

This finding is contrary to findings in studies of face *memory,* where recognition performance is superior after rating abstract qualities of the face, such as personality traits, compared to when judgments are made to facial features (see Coin & Tiberghien, 1997 for a review)[4]. The benefit of personality judgments is thought occur because it encourages encoding of 'holistic' representations that are supportive of recognition memory. Results of Experiment 2 suggest that where there is no requirement to memorise a face – as in the case of unfamiliar face matching – holistic comparisons are not as beneficial. Rather, in these tasks, careful and analytic comparison of facial features appears to be more beneficial to identification accuracy. Thus, the benefits of feature comparison reported here support the proposal that face memory and face matching are reliant on distinct cognitive operations (Megreya & Burton, 2006; Longmore, Liu, & Young, 2008), and suggest that unfamiliar face matching can be improved by enhancing feature encoding.

Benefits of featural comparison were not confined to improvements in accuracy of same/different judgments. Rather, feature ratings were themselves useful in predicting whether pairs of images were of the same person, with AUC analysis showing that certain facial features were particularly diagnostic of identity. Somewhat counterintuitively, across all experiments ears were the most diagnostic feature, for both untrained students and trained examiners. This is also in contrast to studies of memory-based identity classification, where the eyes have been most diagnostic of identity (Vinette, Gosselin, & Schyns, 2004). However, the diagnosticity of ears for face identification in simultaneous matching is not entirely unforeseen. Indeed, best practise manuals and training materials for forensic facial identification stress the importance of ears (FISWG, 2011, 2012). Moreover, ears contain

[4] Because ratings are made during inspection/encoding phases, this literature is distinct from 'overshadowing' studies in which participants describe or attribute labels to faces *after* stimulus encoding (e.g. Schooler and Engstler-Schooler, 1990; Alogna et al. 2014; cf. Brown & Lloyd-Jones, 2005).

distinctive structure and have been shown to be very useful features for biometric identification (Abaza, Ross, Hebert, Harrison, & Nixon, 2013).

We propose that the relative diagnostic value of facial features can form the basis of improvements in feature-based approaches by highlighting *which* features are most useful for identification. For example, consistent with recent work (Towler, White, & Kemp, 2014) we found that face shape was a very poor predictor of identity, and so it appears that this feature should be deemphasised in training courses. In contrast, it may be particularly beneficial for forensic training courses to emphasise careful analysis of ears, especially as it is not intutively obvious that this feature should be useful for the purpose of identification. In future work, it will be important to examine the stability of the relative diagnosticity of facial features found in this study, across different datasets and imaging conditions. This promises to advance approaches to training by optimising strategies for weighting individual features relative to their value in identification decisions.

Results of Experiment 3 uncovered additional benefits of feature comparison that were specific to facial comparison experts. In this experiment, we compared student performance to a specialist group of forensic facial examiners with many years' experience and training in analytic comparison of unfamiliar face images. Consistent with recent tests of people with similar professional experience (White, Phillips, Hahn, Hill, & O'Toole, 2015), we found that examiners outperformed student participants in same/different matching decisions. We also found qualitative differences between examiner and student performance that further emphasise the specialist nature of their expertise in this task.

First of all, examiners showed smaller face inversion effects relative to student participants. This is perhaps surprising given a large body of research in face *memory* showing that face recognition performance is more impaired by inversion, relative to other classes of objects. As memory for non-face 'objects of expertise' (e.g. photos of dogs to dog

breeders) is also more affected by inversion, the size of inversion effects are often thought to index a person's expertise with a class of object (Maurer, Le Grand, & Mondloch, 2002; Tanaka & Farah, 1993; Young, Hellawell, & Hay, 1987). Reduced inversion effects in examiners also appears to be at odds with the finding that 'super-recognisers' – people who naturally excel at face identification – show larger face inversion effects in both memorial and perceptual tasks (Bobak, Bennetts, Parris, Jansari, & Bate, 2016; Russell, Duchaine & Nakayama, 2009). However, these results can be reconciled by the fact that face inversion effects are reduced when task demands require piecemeal, rather than holistic, processing (see McKone & Yovel 2009; Lobmaier & Mast, 2007; Le Grand, Mondloch, Maurer, & Brent, 2001; Barton, Keenan, & Bass, 2001; Hole, 1994).

Thus, we interpret reduced face inversion effects in facial image comparison experts as evidence that the nature of expertise in this task is qualitatively different to the holistic processing thought to underpin recognition *memory*, because it is driven by an increased reliance on individual facial features. This conclusion is further supported by AUC analysis in Experiment 3 that shows examiner's feature ratings were far more diagnostic of identity than feature ratings made by untrained students, and were less affected by face inversion. This analysis also revealed differences in sensitivity to identity information in specific features, highlighting the special role that skin markings play in expert identification. Indeed, differences in self-report ratings between examiners and students also suggest that examiners had conscious knowledge of the diagnostic value of both skin marks and ears (see Figure 7).

We observed a further benefit of feature comparison when averaging the ratings to individual features. Across all experiments, this average similarity score provided impressive discrimination between match and non-match image pairs. Indeed, the average AUC scores for both students and examiners were higher than same/different accuracy scores. The benefit of combining these multiple ratings reflects similar benefits to accuracy when averaging

multiple responses from a single individual in general knowledge judgements (Vul & Pashler, 2008), medical diagnosis (Dawes, 1979) and facial composites (Valentine, Davis, Thorner, Solomon, & Gibson, 2010). Even more impressive was the identification accuracy attained by aggregating these average scores *across* individuals. Consistent with recent work (White, Burton, Kemp, & Jenkins, 2013; White, Phillips, Hahn, Hill, & O'Toole, 2015) the fusion analysis in Experiment 3 elicited near-perfect performance when aggregating responses of seven novice participants, and perfect performance when combining ratings of seven examiners (see Figure 8). Given the very challenging nature of the tests used here, aggregating the independent responses made by group members appears to be a promising method for improving accuracy of forensic facial identification decisions.

Overall, our results suggest that training in forensic examination engenders less reliance on the face-specific processes that govern performance in the population at large, by revealing alternative routes to identification. However, as is often the case when testing expert populations, it is difficult to separate the effects of training, workplace experience and prior ability on performance (Sternberg, 1996). As a result, we cannot rule out the influence of factors such as self-selection on the superiority of examiners in this study. As is clear from recent research (Megreya & Burton, 2006; Russell, Duchaine, & Nakayama, 2009; White, Kemp, Jenkins, Matheson, & Burton, 2014; Robertson, Noyes, Dowsett, Jenkins, & Burton, 2016) there are large individual differences in face matching ability, raising the possibility that people become successful forensic examiners because they are naturally good at the task, rather than as a result of training or experience.

It will be important to address the relationship between natural ability and training in future work. For now, the qualitative differences we observe in this study point towards important differences between experts and novices in unfamiliar face matching that distinguish their abilities from those of high performers identified in recent work (i.e. 'super-

recognisers': see Bobak, Bennetts, Parris, Jansari, & Bate, 2016; Robertson, Noyes, Dowsett, Jenkins, & Burton, 2016; Russell, Duchaine, & Nakayama, 2009). Our results appear to show that forensic facial examiners are expert in the processes involved in unfamiliar face matching, as distinct from face recognition. Given the superior performance of these examiners in this task, such differences offer potential routes to improving accuracy of unfamiliar face matching in novices, and should inform the development of evidence-based training in the future.

References

Abaza, A., Ross, A., Hebert, C., Harrison, M. A. F., & Nixon, M. S. (2013). A survey on ear biometrics. *ACM Computing Surveys (CSUR)*, *45*(2), 22.

Alogna, V. K., Attaya, M. K., Aucoin, P., Bahník, Š., Birch, S., & Birt, A. R., …Zwaan, R. A. (2014). Registered replication report: Schooler and Engstler-Schooler (1990). *Perspectives on Psychological Science*, *9*(5), 556-578.

Andrews, S., Jenkins, R., Cursiter, H., & Burton, A. M. (2015). Telling faces together: Learning new faces through exposure to multiple instances. *The Quarterly Journal of Experimental Psychology*. doi: 10.1080/17470218.2014.1003949

Bahrick, H. P., Bahrick, P. O., & Wittlinger, R. P. (1975). Fifty years of memory for names and faces: A cross-sectional approach. *Journal of Experimental Psychology: General, 104*(1), 54-75. doi: 10.1037/0096-3445.104.1.54

Barton, J. J. S., Keenan, J. P., & Bass, T. (2001). Discrimination of spatial relations and features in faces: Effects of inversion and viewing duration. *British Journal of Psychology, 92*, 527-549.

Berman, G. L., & Cutler, B. L. (1998). The influence of processing instructions at encoding and retrieval on face recognition accuracy. *Psychology, Crime & Law, 4*(2), 89-106.

Biederman, I., & Kalocsai, P. (1997). Neurocomputational bases of object and face recognition. *Philosophical Transactions of the Royal Society B: Biological Sciences 352*(1358), 1203-1219.

Bobak, A. K., Bennetts, R. J., Parris, B. A., Jansari, A., & Bate, S. (2016). An in-depth cognitive examination of individuals with superior face recognition skills. *Cortex*. *82*, 48-62.

Brown, C. & Lloyd-Jones, T. J. (2005). Verbal facilitation of face recognition. *Memory & Cognition*, 33(8), 1442-1456.

Bruce, V., Henderson, Z., Newman, C., & Burton, A. M. (2001). Matching identities of familiar and unfamiliar faces caught on CCTV images. *Journal of Experimental Psychology: Applied, 7*(3), 207-218.

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: 10.1111/1467-9280.00144

Carey, S. (1992). Becoming a face expert. *Philosophical Transactions of the Royal Society B: Biological Sciences, 335*(1273), 95-103.

Carey, S., & Diamond, R. (1977). From piecemeal to configurational representation of faces. Science, 195(4275), 312-314.

Coin, C., & Tiberghien, G. (1997). Encoding activity and face recognition. *Memory, 5*(5), 545-568.

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist, 34*, 571-582.

Dessimoz, D., & Champod, C. (2008). Linkages between biometrics and forensic science. In A. K. Jain, P. J. Flynn & A. A. Ross (Eds.), *Handbook of biometrics* (pp. 425-459). New York: Springer.

Diamond, R., & Carey, S. (1986). Why faces are and are not special: An effect of expertise. *Journal of Experimental Psychology: General, 115*(2), 107-117.

Dowsett, A. J., & Burton, A. M. (2014). Unfamiliar face matching: Pairs out-perform individuals and provide a route to training. *British Journal of Psychology*.

Facial Identification Scientific Working Group. (2011). Guidelines and Recommendations for Facial Comparison Training to Competency (Version 1.1). from https://www.fiswg.org/document/viewDocument?id=22

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

Grady, C. L., McIntosh, A. R., Horwitz, B., & Rapoport, S. I. (2000). Age-related changes in the neural correlates of degraded and nondegraded face processing. *Cognitive Neuropsychology, 17*(1-3), 165-186.

Grother, P., & Ngan, M. (2014). Face Recognition Vendor Test (FRVT) *Performance of face identification algorithms*: National Institute of Standards and Technology.

Hole, G. J. (1994). Configurational factors in the perception of unfamiliar faces. *Perception, 23*, 65-74.

Jain, A. K., Klare, B., & Park, U. (2012). Face matching and retrieval in forensics applications. *IEEE MultiMedia, 19*, 20-28.

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. I., Towell, N. A., & Pike, G. E. (1997). When seeing should not be believing: Photographs, credit cards and fraud. *Applied Cognitive Psychology, 11*, 211-222.

Konar, Y., Bennett, P., & Sekuler, A. B. (2010). Holistic processing is not correlated with face-identification accuracy. *Psychological Science, 21*(1), 38-43.

Konar, Y., Bennett, P., & Sekuler, A. B. (2013). Effects of aging on face identification and holistic face processing. *Vision Research, 88*, 38-46.

Le Grand, R., Mondloch, C. J., Maurer, D., & Brent, H. P. (2001). Early visual experience and face processing. *Nature, 412*, 786.

Lobmaier, J. S., & Mast, F. W. (2007). Perception of novel faces: The parts have it! *Perception, 36*, 1660-1673.

Longmore, C. A., Liu, C. H., & Young, A. W. (2008). Learning faces from photographs. *Journal of Experimental Psychology Human Perception and Performance*, *34*(1), 77–100.

Maurer, D., Le Grand, R., & Mondloch, C. J. (2002). The many faces of configural processing. *Trends in Cognitive Sciences, 6*(6).

McKone, E., & Yovel, G. (2009). Why does picture-plan inversion sometimes dissociate perception of features and spacing in faces, and sometimes not? Toward a new theory of holistic processing. *Psychonomic Bulletin and Review, 16*(5), 778-797.

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

Murphy, J., Ipser, A., Gaigg, S., & Cook, R. (2015). Exemplar variance supports robust learning of facial identity. *Journal of Experimental Psychology: Human Perception and Performance, 41*(3), 577-584.

Norell, K., Brorsson Lathen, K., Bergstrom, P., Rice, A., Natu, V., & O'Toole, A. J. (2014). The effect of image quality and forensic expertise in facial image comparisons. *Journal of Forensic Sciences, 60*(2), 331-340.

O'Toole, A. J., An, X., Dunlop, J., & Natu, V. (2012). Comparing face recognition algorithms to humans on challenging tasks. *ACM Transactions on Applied Perception*, 1-15.

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

Patterson, K. E., & Baddeley, A. D. (1977). When face recognition fails. *Journal of Experimental Psychology: Human Learning and Memory, 3*(4), 406-417.

Phillips, P. J., Beveridge, J. R., Draper, B. A., Givens, G., O'Toole, A. J., Bolme, D., . . . Weimer, S. (2012). The Good, the Bad, and the Ugly Face Challenge Problem. *Image and Vision Computing, 30*(3), 177-185. doi: 10.1016/j.imavis.2012.01.004

Ramon, M. (2015). Differential processing of vertical inter-feature relations due to real-life experience with personally familiar faces. *Perception, 44*(4), 368-382.

Rice, A., Phillips, P. J., Natu, V., An, X., & O'Toole, A. J. (2013). Unaware person recognition from the body when face identification fails. *Psychological Science, 24*(11), 2235-2243.

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), 1-8.

Rossion, B. (2008). Picture-plane inversion leads to qualitative changes of face perception. *Acta Psychologica, 128*(2), 274-289.

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

Schyns, P. G., Bonnar, L., & Gosselin, F. (2002). Show me the features! Understanding recognition from the use of visual information. *Psychological Science*, *13*(5), 402-409.

Scientific Working Group Imaging Technology. (2010). SWGIT Guidelines for the forensic imaging practitioner Section 6: Guidelines and recommendations for training in imaging technology in the criminal justice system. from https://www.swgit.org/documents/Current%20Documents

Sternberg, R. J. (1996). Costs of expertise. In K. A. Ericsson (Ed.), *The road to excellence: The acquisition of expert performance in the arts and sciences, sports, and games* (pp. 347-354). Hillsdale, NJ: Erlbaum.

Tanaka, J. W., & Farah, M. J. (1993). Parts and wholes in face recognition. *The Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology, 46*(2), 225-245. doi: 10.1080/14640749308401045

Tanaka, J. W., & Gordon, I. (2011). Features, configuration and holistic face processing. In A. J. Calder, G. Rhodes, M. H. Johnson & J. V. Haxby (Eds.), *The Oxford handbook of face perception* (pp. 177-194).

Tanaka, J. W., & Simonyi, D. (2016). The "parts and wholes" of face recognition: A review of the literature. *The Quarterly Journal of Experimental Psychology*, 1-14.

Todorov, A., & Porter, J.M. (2014). Misleading first impressions: Different for different facial images of the same person. *Psychological Science, 25*(7), 1404-1417. doi: 10.1177/0956797614532474

Towler, A., White, D., & Kemp, R. I. (2014). Evaluating training methods for facial image comparison: The face shape strategy does not work. *Perception, 43*, 214-218. doi: 10.1068/p7676

Ulery, B. T., Hicklin, R. A., Buscaglia, J., & Roberts, M. A. (2011). Accuracy and reliability of forensic latent fingerprint decisions. *PNAS, 108*(19), 7733-7738.

Valentine, T. (1988). Upside-down faces: A review of the effect of inversion upon face recognition. British Journal of Psychology, 79, 471-491.

Valentine, T., Davis, J. P., Thorner, K., Solomon, C., & Gibson, S. (2010). Evolving and combining facial composites: Between-witness and within-witness morphs compared. *Journal of Experimental Psychology: Applied*, 16(1), 72-86.

Vinette, C., Gosselin, F., & Schyns, P. G. (2004). Spatio-temporal dynamics of face recognition in a flash: It's in the eyes. *Cognitive Science*, *28*(2), 289-301.

Vul, E., & Pashler, H. (2008). Measuring the crowd within: Probabilistic representations within individuals. *Psychological Science, 19*, 645-647. doi:10.1111/j.1467-9280.2008.02136.x

White, D., Burton, A. M., Jenkins, R., & Kemp, R. I. (2014). Redesigning photo-ID to improve unfamiliar face matching performance. *Journal of Experimental Psychology: Applied*.

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

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

White, D., Kemp, R. I., Jenkins, R., Matheson, M., & Burton, A. M. (2014). Passport officers' errors in face matching. *PLoS ONE, 9*(8), 1-6.

White, D., Phillips, P. J., Hahn, C. A., Hill, M., & O'Toole, A. J. (2015). Perceptual expertise in forensic facial image comparison. *Proceedings of the Royal Society of London B: Biological Sciences, 282*, 1814-1822.

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.

Wilkinson, C., & Evans, R. (2009). Are facial image analysis experts any better than the general public at identifying individuals from CCTV images? *Science and Justice, 49*, 191-196.

Woodhead, M. M., Baddeley, A. D., & Simmonds, D. C. V. (1979). On training people to recognize faces. *Ergonomics, 22*(3), 333-343.

Yin, R. K. (1969). Looking at upside-down faces. *Journal of Experimental Psychology, 81*, 141-145.

Young, A. W., Hellawell, D. J., & Hay, D. C. (1987). Configurational information in face perception. *Perception, 16*, 747-759.