American Psychological Association

Journal of Experimental Psychology:
Learning, Memory, and Cognition

ISSN: 0278-7393

2021, Vol. 47, No. 8, 1288–1298
https://doi.org/10.1037/xlm0000972

Diagnostic Feature Training Improves Face Matching Accuracy

Alice Towler, Michelle Keshwa, Bianca Ton, Richard I. Kemp, and David White
School of Psychology, University of New South Wales, Sydney

Identifying unfamiliar faces is surprisingly error-prone, even for experienced professionals who perform this task regularly. Previous attempts to train this ability have been largely unsuccessful, leading many to conclude that face identity processing is hard-wired and not amenable to further perceptual learning. Here, we take a novel expert knowledge elicitation approach to training, based on the feature-based comparison strategy used by high-performing professional facial examiners. We show that instructing novices to focus on the facial features that are most diagnostic of identity for these experts—the ears and facial marks (e.g., scars, freckles and blemishes)—improves accuracy on unfamiliar face matching tasks by 6%. This training takes just 6 min to complete and yet accounts for approximately half of experts' superiority on the task. Benefits of training are strongest when diagnostic features are clearly visible and absent when participants are trained to rely on nondiagnostic features. Our data-driven approach contrasts with theory-driven training that is designed to improve holistic face processing mechanisms associated with familiar face recognition. This suggests that protocols which bypass the core face recognition system—and instead reorient attention to features that are undervalued by novices—offer a more promising route to training for unfamiliar face matching.

Keywords: face recognition, facial image comparison, knowledge elicitation, perceptual expertise, perceptual learning

Supplemental materials: https://doi.org/10.1037/xlm0000972.supp

It typically takes many years of training, feedback and deliberate practice to develop expertise in a domain (see Ericsson et al., 2006). In some cases, however, it has been possible to accelerate the acquisition of expertise by eliciting the knowledge, cognitive strategies, and behaviors underlying experts' superior ability and using this information to develop training. This data-driven expert knowledge elicitation approach to training has successfully improved performance in many domains, including memory (Chase & Ericsson, 1982), mathematics (Staszewski, 1988), landmine detection (Staszewski & Davison, 2000), and tennis (Williams et al., 2002).

Biederman & Shiffrar (1987) provide the clearest example of the benefits of this approach. They examined the perceptual basis of expertise in chicken sexers—a profession that requires fine discrimination of minute features. After observing and interviewing an expert with 50 years' experience sexing 55 million chicks, Biederman and Shiffar learned that the sex of day-old chicks could be determined by a single diagnostic feature: whether the chicks' genital "bead" was convex (male) or concave (female). In a training procedure that took just 1 min, Biederman and Shiffar instructed novices to rely on this diagnostic feature and boosted their accuracy by nearly 40%. In fact, training was so effective that novices became just as accurate as five professional chicken sexers with 18 to 36 years of experience.

Here, we apply this knowledge elicitation training approach to face identification. Knowing whether face identification ability can be improved by training provides important insight into the flexibility and limits of human perceptual learning. We encounter and recognize faces every day from birth and because we are an intensely social species, this ability has been subject to strong selection pressure. It is therefore possible that human accuracy in face identification tasks is asymptotic, with no potential for further learning. Consistent with this view, previous attempts to improve face identification ability in the general population, prosopagnosia patients, and forensic practitioners have been largely unsuccessful (see Bate & Bennetts, 2014; DeGutis et al., 2015; Towler et al., 2019; Towler et al., 2021). Notwithstanding, accuracy on *unfamiliar* face identification tasks is typically much poorer than on *familiar* face identification tasks (see Bruce et al., 2001), suggesting there may be scope for learning in unfamiliar face identification. Further, large individual differences in performance show that some people have more effective perceptual strategies for identification than others (see Wilmer, 2017), and so training to equip people with better strategies could improve performance.

This article was published Online First April 29, 2021.

Alice Towler https://orcid.org/0000-0003-4092-8703
Michelle Keshwa https://orcid.org/0000-0002-9054-193X
Bianca Ton https://orcid.org/0000-0003-0285-0554
Richard I. Kemp https://orcid.org/0000-0003-2819-265X
David White https://orcid.org/0000-0002-6366-2699

This research was supported by Australian Research Council Linkage grants to David White and Richard I. Kemp (LP130100702; LP160101523), in partnership with the Department of Foreign Affairs and Trade, Australian Passport Office.

Correspondence concerning this article should be addressed to Alice Towler, School of Psychology, University of New South Wales, Sydney, High Street, Kensington, NSW 2052, Australia. Email: a.towler@unsw.edu.au

Because face processing is thought to rely on holistic representations more than other types of object processing (see Tanaka & Farah, 1993; Tanaka & Simonyi, 2016; Young et al., 1987), previous training attempts have typically focused on improving holistic processing (see Towler et al., 2021 for a review). For example, remedial training for prosopagnosia patients has aimed to increase their sensitivity to the configuration of internal facial features (e.g., DeGutis et al., 2007). However, holistic training approaches have had very little success (see Towler et al., 2021 for a review).

A more promising approach to face identification training is to encourage *featural* face processing (see Towler et al., 2021 for a review). For example, prosopagnosia patients' ability to recognize familiar faces is improved by memorizing each face's distinctive features (Brunsdon et al., 2006; Schmalzl et al., 2008). Consistent with this approach, professional training courses encourage practitioners working at border crossings and in police investigations to adopt a feature-by-feature comparison strategy (Towler et al., 2019). Surprisingly, however, professional training courses do not improve face identification accuracy (Towler et al., 2019), possibly because they do not specify *which* facial features trainees should prioritize.

Previous research has investigated which facial features are the most important for face identification. Early work by Ellis et al. (1979) suggested the internal facial features (eyes, nose and mouth) were most important after they found familiar faces were recognized more accurately from internal than external features (see also Kramer et al., 2018; Logan et al., 2017). Using the "bubbles" technique, Schyns et al. (2002) found that participants tended to rely on the eyes, mouth and chin when determining which of 10 identities were presented (see also Tardif et al., 2019). Sadr et al. (2003) suggested that eyebrows are particularly important for face recognition after finding that familiar faces are difficult to recognize without them. More recently, Abudarham and Yovel (2016) concluded that lip thickness, hair color and eye color are the most important features by estimating their contributions to face similarity in a multidimensional feature space derived from computer-generated faces. Critically, these studies assume that the important facial features are those which people typically use to support identification decisions. However, the features people use to identify unfamiliar faces are probably not the features they *should* use, given that people are, in general, poor at identifying unfamiliar faces (e.g., Bruce et al., 2001).

We recently developed a novel method of calculating the diagnostic value of facial features, by quantifying the amount of identity information contained in each (see Towler et al., 2017). We did this using an unfamiliar face matching task, which is a surprisingly challenging task that involves deciding whether simultaneously presented unfamiliar faces show the same person or different people (see Burton et al., 2010). Participants rated the similarity of 11 facial features on face pairs from 1 (*very dissimilar appearance*) to 5 (*very similar appearance*), before making a same/different person identity decision. To determine the diagnosticity of each facial feature, we calculated the extent to which participants' feature similarity ratings predicted whether the faces showed the same person or different people (see Towler et al., 2017 for more details).

In Towler et al. (2017) we collected feature similarity ratings from a group of experts—specialist professionals known as *facial examiners*—who consistently outperform novices on unfamiliar face matching tasks (see White et al., 2021). Facial examiners' identification accuracy was 14% higher than novices' (89% vs. 78%), and their ratings of facial feature similarity were much more diagnostic of identity (Cohen's $d = 1.44$). This finding indicates that examiners are more sensitive to the identity information contained within facial features than novices, which is consistent with the slow, feature-by-feature comparison strategy they use to identify faces (see White et al., 2015). Importantly, examiners' feature similarity ratings for the ears and facial marks[1] were the most diagnostic of identity across trial types. These were also the same features examiners reported finding most useful for comparison (see Materials, Figure 1). By contrast, novices reported these features as only moderately useful, prioritizing the eyes and face shape instead.

In this article we test the hypothesis that orienting novices' attention to the most diagnostic features can improve face identification ability. This approach is similar to the perceptual training approach of Biederman and Shiffrar (1987), except that they trained novices on a perceptual stimulus with which participants had no prior familiarity. Here, we test whether this can also extend to highly familiar stimuli that participants have extensive experience discriminating in daily life. In two experiments, we train novices to use the facial features that were most diagnostic of identity for expert facial examiners in Towler et al. (2017; the ears and facial marks) and assess their face matching accuracy before and after training. In both experiments we compare the effects of this diagnostic feature training to a control group and a nondiagnostic feature training group who are trained to rely on the facial features that were least diagnostic of identity.

Experiment One

In Experiment 1, we test whether diagnostic feature training improves unfamiliar face matching accuracy. Novice participants completed one of three self-paced training courses. The diagnostic feature training instructed participants to focus on diagnostic facial features derived from our Towler et al. (2017) study of facial examiners: ears and facial marks. The nondiagnostic feature training instructed participants to focus on relatively nondiagnostic features derived from the same study: face shape and mouth. Both feature training courses incorporated standard instructions derived from a large-scale international review of professional training courses in face identification (Towler et al., 2019). The control training was unrelated to face identification. Participants completed pre- and posttraining tests so we could track the effects of training on face matching accuracy.

Method

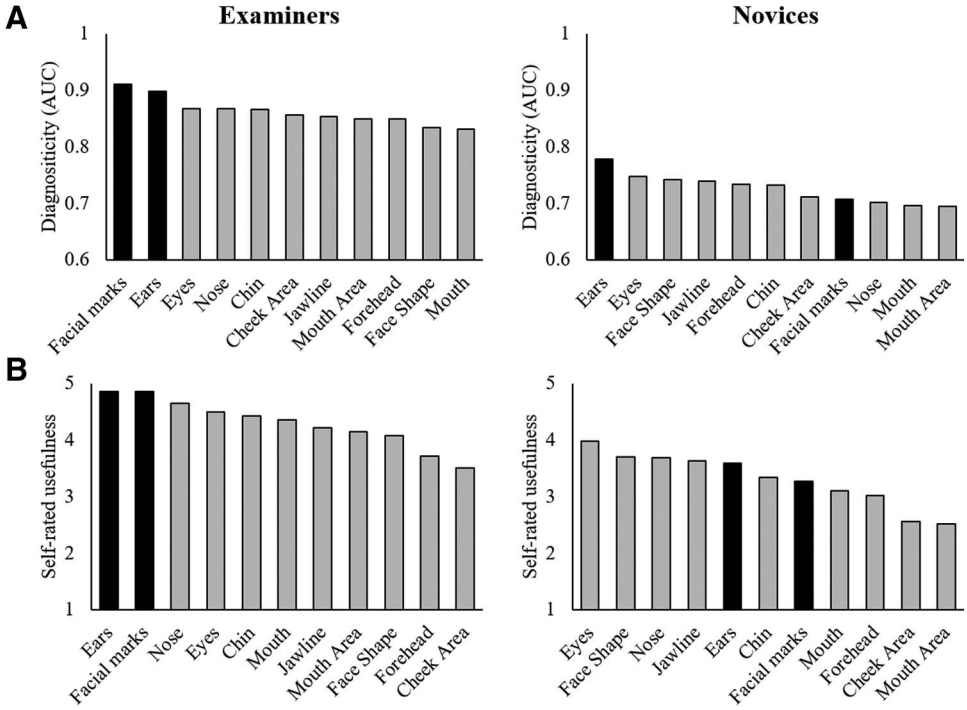
Participants

Sixty undergraduate psychology students participated in return for course credit ($M_{age} = 19$, 22 male, 38 female; see the online supple-

[1] In Towler et al. (2017) we referred to facial marks as "scars and blemishes." This original terminology came from the Facial Identification Scientific Working Group (FISWG)—an international industry standards body—who have since updated their terminology to "facial marks" to better reflect the characteristics originally included in the "scars and blemishes" category (e.g. freckles, moles, acne, birthmarks, vitiligo, dimples etc.; see Facial Identification Scientific Working Group, 2018).

Figure 1

Facial Feature Diagnosticity and Self-Rated Usefulness Data From Towler et al. (2017)



Note. (A) The extent to which facial feature similarity ratings were diagnostic of identity for facial examiners (left) and novices (right). (B) Self-reported facial feature usefulness ratings by facial examiners' (left) and novices (right). AUC = area under the ROC curve. Participants in Towler et al. (2017) rated the degree to which they used each facial feature on a scale from 1 (*never*) to 5 (*all the time*).

mental materials for more details). Twenty participants were randomly allocated to each of three training conditions: diagnostic feature, nondiagnostic feature, or control. A sensitivity power analysis revealed this design can reliably detect an effect size of $\eta_p^2 = .05$ or higher with 95% power, where $\alpha = .05$ and $r = .645$ (Faul et al., 2009). Both experiments were approved by the Human Research Ethics Advisory Committee in the School of Psychology at UNSW Sydney.

Materials

Training Course Development. In Towler et al. (2017), we calculated the extent to which participants' ratings of feature similarity predicted whether face pairs showed the same person or different people (see Towler et al., 2017 for more details). The most diagnostic facial features for facial examiners were the ears and facial marks (see Figure 1A), and these were the same features examiners reported finding most useful for comparison (see Figure 1B). By contrast, novices reported the ears and facial marks as being only moderately useful, prioritizing the eyes and face shape instead (see Figure 1B). We therefore selected ears and facial marks to be the focus of the diagnostic feature training, reasoning that novices typically underestimate the identity information in these features. We selected face shape and mouth to be the focus of the nondiagnostic feature training because these were examiners' two least diagnostic features (see Figure 1A) and because face shape training does not improve face matching accuracy (Towler et al., 2014).

To create the diagnostic feature training, we adapted portions of professional training courses that provide training on the ears and facial marks (see Towler et al., 2019 for details of the professional training courses). We collated these into PowerPoint slides that participants studied at their own pace. To create the nondiagnostic feature training, we repeated the same process by adapting portions of professional training courses that provide training on face shape and the mouth. We created the control training by adapting content on conflict resolution strategies from the Internet, such that the duration of training was roughly equivalent to the diagnostic and nondiagnostic training. The training courses are available from the authors on request.

Training Course Content. The diagnostic feature and nondiagnostic feature training both instructed trainees to avoid looking at the face as a whole and to avoid fixating on the "triangle of recognition" (the internal region of the face triangulated by the eyes and mouth). Instead, trainees were encouraged to break faces down into parts and compare each facial feature individually. This instruction was common to all professional training courses reviewed in Towler et al. (2019), and we used it here to encourage a feature-based approach to the task.

Trainees were then told that, according to scientific research, *some features are more useful than others*. In the diagnostic feature training, trainees were told to rely on the ears and facial marks. In the nondiagnostic feature training, trainees were told to rely on the face shape and mouth. Alongside this instruction,

trainees were shown a graph ranking a selection of facial features from most to least useful. In the diagnostic feature training, this graph correctly ranked the features from most to least useful (facial marks, ears, eyes, face shape and mouth). In the nondiagnostic feature training, we reversed the feature labels so that mouth and face shape appeared to be most useful.

Finally, the training described the different subparts (e.g., ear lobe, tragus) and characteristics (e.g., shape, thickness) of each "useful" feature, using information derived from the professional training courses reviewed in Towler et al. (2019). Trainees then saw example face identification comparisons with the respective useful features highlighted to illustrate that similarities between features provides evidence the photos show the same person, and that differences provide evidence the photos show different people.

Pre- and Posttraining Face Matching Task. To test for training effects, we split the Expertise in Facial Comparison Test (EFCT; see White et al., 2015) into two equally difficult 84-item subtests using existing performance data. The EFCT contains 168 challenging color, front-facing face pairs, captured on different days and under varying lighting conditions. Participants completed one subtest before training, and the other after training. The order of subtests was counterbalanced across participants. Participants viewed each face pair for a maximum of 30 s and decided whether the images showed the same person or different people using a 5-point scale from 1 (*sure same person*) to 5 (*sure different people*) before or after the images were removed from the screen.

Procedure

Participants completed the pretraining face matching test, followed by either the diagnostic feature, nondiagnostic feature, or control training, and then completed the posttraining face matching test. Participants were then asked whether training had made face matching easier, harder, or had no effect.

Data Analysis

We assessed the effectiveness of training in both experiments using 3 × 2 mixed ANOVAs, with Training (diagnostic feature, nondiagnostic feature, control) as a between-subjects factor and Test (pretraining, posttraining) as a within-subjects factor. For brevity, we only report the critical interaction between Training and Test, which indicates whether the change in accuracy from pre- to posttraining differs between the groups and follow-up simple main effects. We confirmed that significant interactions between Training and Test remained when the nondiagnostic training group and outliers were excluded from the analyses and verified our conclusions with ANCOVA, using pretraining accuracy as a covariate. Full details of these analyses and complete data sets are provided in the online supplemental materials.

Finally, we used one-sided Bayesian *t* tests indicate the strength of evidence that each training course improved (H_+) or did not improve (H_0) accuracy from pre- to posttraining. Bayes Factors of 1–3, 3–10, and 10–30 indicate anecdotal, moderate, and strong evidence, respectively, for a hypothesis (see Jeffreys, 1961; Lee & Wagenmakers, 2014). Priors are described by the JASP (0.13.1.0) default Cauchy distribution centered on a zero effect size and a width of .707 (JASP Team, 2020).

Results

Face Matching Accuracy

Because the EFCT requires participants to respond using a 5-point scale, the standard measure of accuracy on this task is area under the ROC curve (AUC; White et al., 2015; see Figure 2). AUC scores on the pre- and posttraining tests are shown separately for each training group in Figure 2, where values of 1 indicate perfect performance and 0.5 indicates chance-level performance.

The interaction between Training and Test was significant, $F(2, 57) = 4.91$, $p < .05$, $\eta_p^2 = .15$, and exceeded the minimum effect size that could be reliably detected. Participants who completed the diagnostic feature training showed a significant 6% improvement from pre- to posttraining (pre $M = .83$, $SD = .09$, post $M = .88$, $SD = .05$), $F(1, 57) = 9.90$, $p < .05$, $\eta_p^2 = .15$. Participants who completed the nondiagnostic feature, $F < 1$, $\eta_p^2 = .00$, or control training, $F(1, 57) = 1.42$, $p > .05$, $\eta_p^2 = .02$, showed no change in accuracy from pre- to posttraining (nondiagnostic: pre $M = .82$, $SD = .10$, post $M = .83$, $SD = .08$; control: pre $M = .87$, $SD = .08$, post $M = .85$, $SD = .10$).

Bayesian analysis confirmed the observed data is 10.1 times more likely to occur when the diagnostic feature training *improves* accuracy (H_+) than when it does not (H_0), providing strong evidence for the effectiveness of the diagnostic feature training. Equivalent tests for the control and nondiagnostic feature training groups showed the observed data are 8.7 and 3.7 times more likely to occur, respectively, when these training courses *do not* improve accuracy (H_0) than when they do (H_+), providing moderate evidence the control and nondiagnostic feature training are ineffective.

Visual inspection of individual participant data in Figure 2 indicates that diagnostic feature training is most beneficial for low-performers—training appears to have lifted the tail of the distribution, rather than improving all participants equally. This is consistent with previous research showing effective training in face matching tasks (Dowsett & Burton, 2014; White, Kemp, et al., 2014).

Perceived Effectiveness of Training

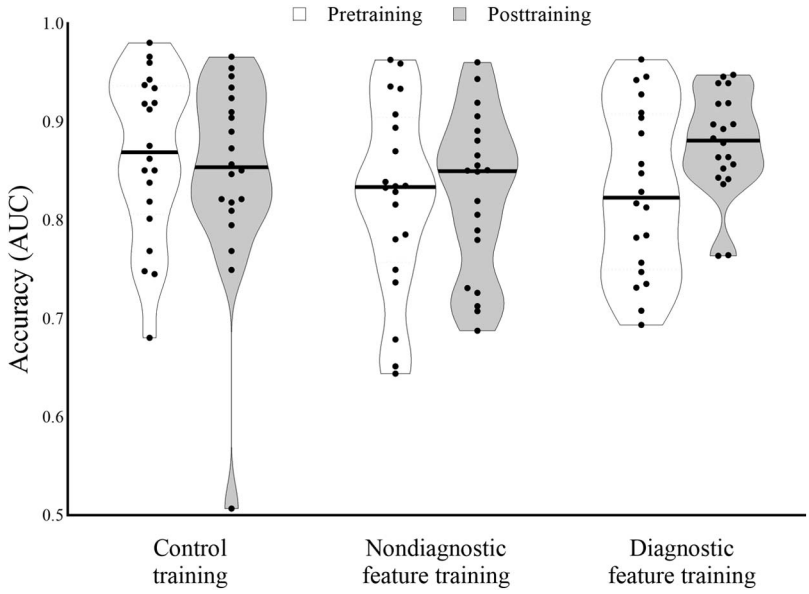
Most participants *thought* training made face matching easier regardless of whether it improved accuracy (see Figure 3). Seventy percent of participants in the diagnostic feature training group and 65% of participants in the nondiagnostic feature training group reported that training made face matching easier, even though only the diagnostic feature training improved accuracy (see the online supplemental materials for full details). This lack of insight into the effectiveness of training is consistent with previous evaluations of professional face identification training (Towler et al., 2019).

Discussion

We applied a data-driven expert knowledge elicitation approach to unfamiliar face matching training, by instructing novices to rely on facial features that were most diagnostic of identity for facial examiners in Towler et al. (2017). Instructing novices to focus on the ears and facial marks improved participants' face matching accuracy from pre- to posttraining by 6%. However, the EFCT images used in this experiment were sourced from the Good, Bad, and Ugly image set (see Phillips et al., 2011), which is the same image set we used to identify the diagnostic facial features in Towler et al. (2017). The

Figure 2

Diagnostic Feature Training—to Focus on the Ears and Facial Marks—Significantly Improved Face Matching Accuracy From Pre- to Posttraining

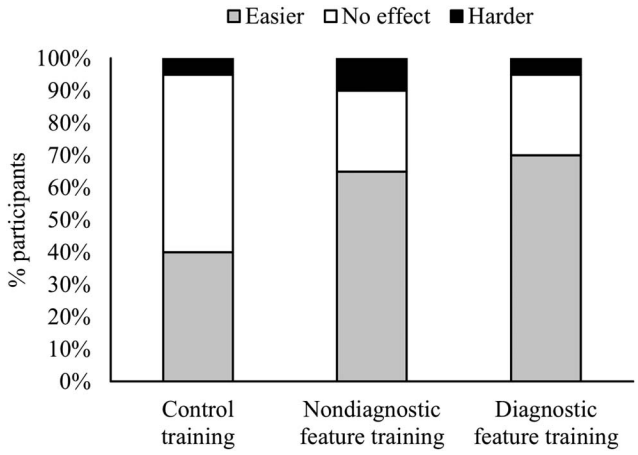


Note. AUC = area under the ROC curve. Participants who received the control training or nondiagnostic feature training did not show any improvement. Markers represent individual data points, horizontal lines represent the medians, and means are reported in the main text.

diagnostic features—and consequently, the training effects observed in Experiment 1—may therefore be specific to the idiosyncratic imaging conditions in this dataset. An essential requirement of face identification training is that it produces generalizable improvements in accuracy, so in Experiment 2 we test whether diagnostic feature training improves accuracy on tests created using image sets of different people, captured in different imaging conditions.

Figure 3

A Similar Proportion of Participants in the Diagnostic Feature (70%) and Nondiagnostic Feature (65%) Training Groups Reported That Training Made Face Matching Easier, Despite Only the Diagnostic Training Improving Accuracy



Experiment Two

In Experiment 2 we test the effectiveness of diagnostic feature training on two new image sets. These tests model the range of imagery encountered in applied settings, from high-quality imagery encountered at border control (e.g., passport photos), to low-quality images encountered by law enforcement (e.g., CCTV). Because fine facial feature detail is not necessarily visible in low resolution imagery, this provides a strict test of the generalizability of the diagnostic feature training. These tests use a binary response scale, so they allow us to examine training effects on match and nonmatch trials separately. This is an important theoretical question because dissociable cognitive skills are thought to underlie accuracy on these two trial types (see Megreya & Burton, 2007). We also track how long participants spend on the training to check whether the improvement observed in Experiment 1 can be explained by longer training duration.

Method

Participants

A power analysis indicated we required 27 participants to have a 95% chance of detecting the effect size observed in Experiment 1 (η_p^2 = .15), where α = .05 and r = .5 (Faul et al., 2009). However, because we ran the study online, we decided to collect data from approximately 40 participants per group to improve the reliability of our data. One hundred and twenty-one participants recruited via Amazon's Mechanical Turk were paid US$2 to participate ($M_{age}$ = 38, 52 male, 69 female; see the online supplemental materials for more details). Random allocation to each

training course meant that 42 participants received the control training, 36 received the nondiagnostic feature training, and 43 received the diagnostic feature training.

Materials

Participants completed the Glasgow Face Matching Test (GFMT) to model applied casework involving high-quality imagery (Burton et al., 2010). The GFMT is a standardized face matching test consisting of high-quality, greyscale and front-facing face pairs captured on the same day in studio conditions with a neutral expression. To model casework involving comparison between high-quality (e.g., mugshot) and low-quality images (e.g., CCTV), we included the high-to-low image quality test (see Towler et al., 2019). This test consists of one high-quality front-facing face photograph and one low-quality front-facing face photograph, presented in color and with neutral expressions.

Both tests consist of 40 simultaneous face pairs, which we divided into two equally difficult versions of 20 items each (10 match, 10 nonmatch) using itemized accuracy data. Allocation of each test version to pre- and posttraining was counterbalanced across participants. On each trial of the GFMT and high-to-low image quality tests, participants saw a face pair for up to 30 s and decided if the faces showed the same person or different people. Participants made binary same/different identity decisions before or after the images were removed.

Procedure

Participants completed the two pretraining face matching tests in a random order before being randomly allocated to the diagnostic feature, nondiagnostic feature or control training course. Participants took a median of 5.5 min to complete the diagnostic feature training (SD = 13.5), 5.6 min to complete the nondiagnostic feature training (SD = 11.3), and 5.5 min to complete the control training (SD = 7.3). A one-way ANOVA confirmed there were no significant differences in training duration between the three groups, $F(2, 115) = 1.28$, $p > .05$ (see the online supplemental materials for more details). Participants were then asked to make a binary yes/no decision about whether training had improved their face identification accuracy. Finally, participants completed the posttraining face matching tests in a random order.

Results

Participants responded using a binary same person/different people scale, so we measured accuracy on each test using percent correct (see Figure 4).

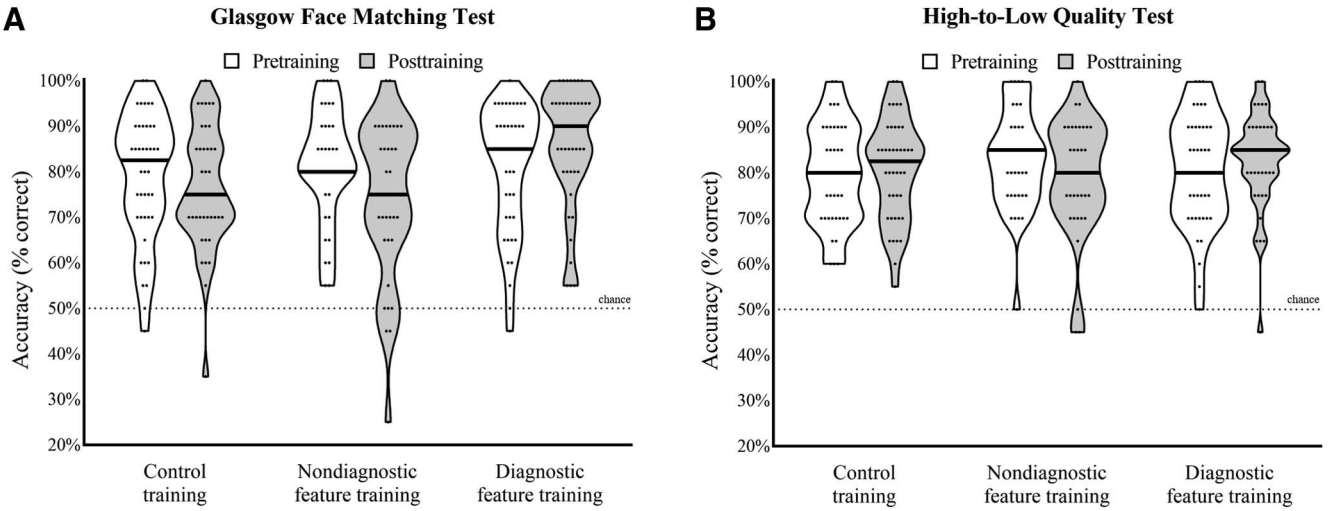
GFMT Accuracy

Overall Accuracy. The interaction between Training and Test was significant, $F(2, 118) = 9.81$, $p < .05$, $\eta_p^2 = .14$. Consistent with Experiment 1, participants who completed the diagnostic feature training showed a significant 6% improvement from pre- to posttraining (pre M = 81%, SD = 14%, post M = 86%, SD = 13%), $F(1, 118) = 7.13$, $p < .05$, $\eta_p^2 = .06$. Interestingly, participants who completed the nondiagnostic feature training showed a significant 9% *decrease* in accuracy (pre M = 81%, SD = 12%, post M = 74%, SD = 17%), $F(1, 118) = 11.99$, $p < .05$, $\eta_p^2 = .10$. Participants in the control group showed no change in accuracy from pre- to posttraining (pre M = 79%, SD = 14%, post M = 76%, SD = 13%), $F(1, 118) = 1.58$, $p > .05$, $\eta_p^2 = .01$.

Bayesian analysis confirmed the observed data are 20.7 times more likely to occur when the diagnostic feature training improves accuracy (H_+) than when it does not (H_0), providing strong evidence for the effectiveness of the diagnostic feature training. Equivalent tests for the control and nondiagnostic feature training groups showed the observed data are 13.6 and 18.9 times more

Figure 4

Accuracy on the GFMT (A) and High-to-Low Quality Test (B) Before (Pretraining) and After (Posttraining) Completing the Control, Nondiagnostic Feature, or Diagnostic Feature Training



Note. GFMT = Glasgow Face Matching Test. Markers represent individual data points, horizontal lines represent the medians, and means are reported in the main text.

likely to occur, respectively, when these training courses *do not* improve accuracy (H_0) than when they do (H_+), providing strong evidence the control and nondiagnostic feature training are ineffective.

Match and Nonmatch Trial Accuracy. We repeated the 3 (Training) × 2 (Test) ANOVA analysis above for match and nonmatch trials separately. The critical interaction between Training and Test was nonsignificant for match trials, $F(2, 118) = 2.60$, $p > .05$, $\eta_p^2 = .04$, but significant for nonmatch trials, $F(2, 118) = 7.17$, $p < .05$, $\eta_p^2 = .11$, suggesting that diagnostic feature training specifically improved participants' ability to tell pairs of different faces apart.

Bayesian analyses showed the observed data are 6.0 times more likely to occur when the diagnostic feature training *improves* nonmatch trial accuracy (H_+) than when it does not (H_0) and 3.8 times more likely to occur when it does not improve match trial accuracy (H_0) than when it does (H_+). These analyses provide moderate evidence that diagnostic feature training improves nonmatch trial accuracy only.

Equivalent Bayesian analyses for the control and nondiagnostic feature training groups showed the observed data are more likely to occur when they *do not* improve match (control $BF_{0+} = 2.7$, nondiagnostic $BF_{0+} = 15.3$) or nonmatch (control $BF_{0+} = 19.3$, nondiagnostic $BF_{0+} = 13.8$) trial accuracy (H_0) than when they do (H_+).

Follow-up signal detection analyses are reported in the online supplemental materials and indicate the benefit of diagnostic feature training is driven by a change in sensitivity not response criterion.

High-to-Low Quality Test Accuracy

Overall Accuracy. The interaction between Training and Test was nonsignificant, $F(2, 118) = 2.80$, $p > .05$, $\eta_p^2 = .05$. Bayesian analyses for the control, nondiagnostic and diagnostic feature training groups showed the observed data are 4, 14, and 1 times more likely to occur, respectively, when these training courses *do not* improve accuracy (H_0) than when they do (H_+).

Match and Nonmatch Trial Accuracy. We repeated the 3 (Training) × 2 (Test) analysis above for match and nonmatch trials separately and found the interaction between Training and Test was nonsignificant for match trials, $F < 1$, $\eta_p^2 = .01$, but significant for nonmatch trials, $F(2, 118) = 5.35$, $p < .05$, $\eta_p^2 = .08$, again suggesting that diagnostic feature training specifically improved participants' ability to tell faces apart.

Bayesian analyses showed the observed data are 28.9 times more likely to occur when the diagnostic feature training *improves* nonmatch trial accuracy (H_+) than when it does not (H_0), and 13.5 times more likely to occur when it does not improve match trial accuracy (H_0) than when it does (H_+). These analyses provide strong evidence that diagnostic feature training improves nonmatch trial accuracy only.

Equivalent Bayesian analyses for the control and nondiagnostic feature training groups showed the observed data are more likely to occur when they *do not* improve match (control $BF_{0+} = 8.7$, nondiagnostic $BF_{0+} = 10.4$) or nonmatch trial accuracy (control $BF_{0+} = 1.5$, nondiagnostic $BF_{0+} = 11.4$; H_0) than when they do (H_+).

Together, our findings indicate that the benefits of diagnostic feature training are specific to nonmatch trials, but these benefits are somewhat attenuated in imagery where fine feature detail is not clearly visible.

Perceived Effectiveness of Training

Strikingly, but consistent with the results of Experiment 1, 86% of participants in the nondiagnostic feature training condition reported the training *improved* their overall accuracy despite evidence that it substantially *impaired* their accuracy on the GFMT. Ninety-one percent of participants in the diagnostic feature training condition and 31% of participants in the control condition reported that training improved their accuracy. This is an interesting result and may relate to a general tendency to underestimate the difficulty of unfamiliar face matching tasks (Ritchie et al., 2015) and the limited insight people have into their face identification ability (Bindemann et al., 2014; Bobak et al., 2018; Palermo et al., 2017).

Meta-Analysis of Cumulative Bayesian Support for the Effectiveness of Training

To assess the accumulated evidence that each training course improved face matching accuracy from pre- to posttraining we pooled the data across all three tests in Experiment 1 and 2 and conducted one-sided Bayesian *t* tests. These analyses indicate the observed data are 15.1 times more likely to occur when diagnostic feature training *improves* accuracy (H_+) than when it does not (H_0), providing strong evidence that diagnostic feature training improves face matching accuracy. Equivalent analyses for the control and nondiagnostic feature training groups indicated the observed data are 13.8 and 34.7 times more likely to occur, respectively, when these training courses *do not* improve accuracy (H_0) than when they do (H_+), providing strong and very strong evidence the control and nondiagnostic feature training are ineffective.

General Discussion

We found that training participants to focus on the ears and facial marks, features that were most diagnostic of identity for expert facial examiners (see Towler et al., 2017), produced generalizable improvements in people's ability to identify unfamiliar faces. These improvements were strongest when facial features were clearly visible and absent when participants were trained to rely on nondiagnostic features, confirming that the benefit of diagnostic feature training is due to increased attention to *diagnostic* facial features. Our findings make important contributions to face identification theory and practice, which we outline below.

First, our diagnostic feature training provides a new and efficient method of improving unfamiliar face matching ability. After decades of research and practice seeking to develop training to improve this ability, only two other methods have shown generalizable improvements (see Towler et al., 2021 for a review). One method is feedback training—where participants receive extensive trial-by-trial feedback on face matching decisions (White, Kemp, et al., 2014). However, evidence for its effectiveness is mixed (see Alenezi & Bindemann, 2013). The other is paired decision-making—where two people work together on a set of face matching decisions, improving the ability of the low-performer in the pair

(Dowsett & Burton, 2014). Here, we show that simply directing trainees' attention to the diagnostic features used by experts leads to generalized improvements in face matching ability, using far less time and resources than other methods.

Second, our results provide empirical evidence for two distinct cognitive routes to expertise in face identification. Seminal work shows that face identification involves two separable cognitive routes (Bartlett et al., 2003; Bruce & Young, 1986; Farah, 1991). One is a quick, holistic route that we use to recognize familiar faces with near perfect accuracy. The other is a slow, featural route that exploits domain-general directed visual processing (see Bruce & Young, 1986). This featural route is considered abnormal, presumably because it is typically associated with impaired performance (see Coin & Tiberghien, 1997; McKone & Yovel, 2009; Tanaka & Farah, 1993; Yin, 1969), and the strategies used by prosopagnosia patients (Adams et al., 2020). Unsurprisingly, previous attempts to train prosopagnosia patients and the general population have therefore tended to adopt procedures inspired by the *holistic* processes supporting familiar face recognition (see Towler et al., 2021 for a review). These have been largely unsuccessful, leading many researchers to conclude that face identification ability is static and not amenable to training (e.g., Ramon et al., 2016; Wilmer, 2017).

Recent evidence demonstrates this conclusion is incorrect. Facial examiners achieve very high levels of accuracy using a feature-based comparison strategy (Towler et al., 2017), and their skills are qualitatively different to those with naturally occurring expertise ('superrecognizers'; see Noyes et al., 2017; Russell et al., 2009). This indicates that facial examiners have *learned* to identify faces in a feature-based way. Further, the most promising training for prosopagnosia patients is in fact to adopt feature-based strategies (see Bate & Bennetts, 2014; DeGutis et al., 2015).

Elsewhere, we have argued that this evidence indicates that the core holistic face recognition route is not trainable, but that the featural route that bypasses this system *is* trainable, at least for unfamiliar face matching tasks (see Towler et al., 2021). There, we also argued that this evidence indicates that *both* separable cognitive routes involved in face identification provide legitimate routes to *expertise* in this task—a significant departure from the notion that the featural route is abnormal (see Towler et al., 2021). Here, we find empirical evidence to support both proposals—that the featural route is trainable and a legitimate route to expertise—by demonstrating that training people to use a feature-based comparison strategy improves face matching accuracy.

Third, our results shed light on the nature of expertise in facial examiners (see Phillips et al., 2018; White et al., 2015). In Towler et al. (2017), we calculated facial examiners' diagnostic facial features and found that examiners outperformed novices by 14%. Here, we found that training novices to focus on these diagnostic features conferred a 6% improvement in face matching accuracy—accounting for roughly *half* of facial examiners' expertise. We interpret this as further evidence that the expertise of facial examiners stems from selective attention to facial features (see Towler et al., 2017; White et al., 2015). Critically, it also suggests that at least part of the perceptual learning underpinning their expertise is discovering *which* of these features carry useful identity information. Combined with our finding that nondiagnostic feature training conferred no benefits, this finding indicates that training the featural route is not simply about getting people to adopt a feature-based comparison strategy (e.g., Megreya, 2018; Megreya & Bindemann, 2018). Rather, it appears contingent on learning which features contain useful sources of identity information that would otherwise have been overlooked.

Interestingly, facial examiners' trajectory of perceptual learning in face identification—from intuitive, holistic processing to more analytic, featural comparison (see White et al., 2015)—is precisely the *opposite* shift that is typically thought to characterize perceptual learning and the development of expertise more generally (see Chase & Simon, 1973; Kahneman & Klein, 2009; White et al., 2021). The effectiveness of feature-based comparison observed here therefore has broader implications for the study of perceptual expertise. Our findings suggest that analytic, feature-based perceptual strategies can confer important performance benefits, even in overlearned stimuli like faces, by aiding the discovery of useful features that are ordinarily missed when viewing such stimuli (cf. Drew et al., 2013; Wolfe et al., 2017).

Fourth, we found that the benefits of diagnostic feature training were specific to nonmatch trials, adding to growing evidence that dissociable cognitive and perceptual processes underpin accuracy on matching and nonmatching face pairs in unfamiliar face matching tasks (see Megreya & Burton, 2007). For example, multiple images, motivation, anxiety, feature similarity ratings, and sleep deprivation affect accuracy on one trial type but not the other (Attwood et al., 2013; Beattie et al., 2016; Moore & Johnston, 2013; Towler et al., 2017; White, Burton, et al., 2014), and developmental prosopagnosia patients show deficits on match but not nonmatch trials (White et al., 2017). Here, we show that diagnostic feature training improves people's ability to detect nonmatching identities. This finding suggests that the featural route described above is particularly useful for detecting differences between faces, providing the first evidence of mechanistic differences in the cognitive strategies underpinning match and nonmatch trial accuracy in unfamiliar face matching tasks. Anecdotally, our participants often report experiencing an "Aha!" moment when the correct answer to a challenging image pair suddenly becomes obvious after noticing dissimilarities in the ears or facial marks (see Kounios & Beeman, 2014). This might suggest that the featural route is engaged after an initial holistic assessment of facial similarity that does not ordinarily encapsulate these features.

Fifth, the effectiveness of diagnostic feature training validates the Towler et al. (2017) method of determining facial feature diagnosticity. Given that feature diagnosticity plays an important role in theoretical models of face processing (e.g., Valentine, 1991), we propose that this method can help to understand how diagnosticity varies as a function of face and viewer characteristics in future work. For example, the tendency for people to perform worse on identification tasks involving faces from another ethnicity (e.g., Megreya et al., 2011; Meissner & Brigham, 2001) has been explained as a misapplication of diagnostic features derived from one ethnic group to another. The diagnostic feature extraction method described in Towler et al. (2017) can therefore provide a basis for testing these predictions directly and may also be applied more broadly to examine differences in the feature representations supporting expert performance in other pattern-matching domains, such as fingerprint comparison (Tangen et al., 2011) and radiology (Siegle et al., 1998).

Finally, this work makes important applied contributions to real-world forensic practice. Diagnostic feature training significantly improved unfamiliar face matching accuracy in just 6 min. This stands in stark contrast to professional training courses, which typically run over 1 or more *days* and do not improve accuracy despite adhering to international best-practice guidelines (see Towler et al., 2019). Diagnostic feature training therefore provides a more effective and efficient replacement for professional training courses.

Notably, the benefits of diagnostic feature training were specific to nonmatch trials and most pronounced with high-quality imagery. It is therefore likely to be most useful for detecting imposters in situations such as border control and passport issuance, and for eliminating innocent suspects in criminal investigations where relatively high-quality imagery is available. It may be less useful in situations that require the detection of matches in low-quality imagery, such as tracking an offender across CCTV cameras. Given that our diagnostic features were initially elicited from high-quality imagery, future research to establish the extent to which diagnostic features vary as a function of image, viewer, and face characteristics would enable broader benefits to practitioners.

References

Abudarham, N., & Yovel, G. (2016). Reverse engineering the face space: Discovering the critical features for face identification. *Journal of Vision*, *16*(3), 40–57. https://doi.org/10.1167/16.3.40

Adams, A., Hills, P. J., Bennetts, R. J., & Bate, S. (2020). Coping strategies for developmental prosopagnosia. *Neuropsychological Rehabilitation*, *30*(10), 1995–2015. https://doi.org/10.1080/09602011.2019.1623824

Alenezi, H. M., & Bindemann, M. (2013). The effect of feedback on face-matching accuracy. *Applied Cognitive Psychology*, *27*(6), 735–753. https://doi.org/10.1002/acp.2968

Attwood, A. S., Penton-Voak, I. A., Burton, A. M., & Munafo, M. R. (2013). Acute anxiety impairs accuracy in identifying photographed faces. *Psychological Science*, *24*(8), 1591–1594. https://doi.org/10.1177/0956797612474021

Bartlett, J. C., Searcy, J. H., & Abdi, H. (2003). What are the routes to face recognition? In M. Peterson & G. Rhodes (Eds.), *Perception of faces, objects, and scenes: Analytic and holistic processes* (pp. 21–52). Oxford University Press.

Bate, S., & Bennetts, R. J. (2014). The rehabilitation of face recognition impairments: A critical review and future directions. *Frontiers in Human Neuroscience*, *8*, 1–30. https://doi.org/10.3389/fnhum.2014.00491

Beattie, L., Walsh, D., McLaren, J., Biello, S. M., & White, D. (2016). Perceptual impairment in face identification with poor sleep. *Royal Society Open Science*, *3*(10), 160321. https://doi.org/10.1098/rsos.160321

Biederman, I., & Shiffrar, M. M. (1987). Sexing day-old chicks: A case study and expert systems analysis of a difficult perceptual-learning task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*(4), 640–645. https://doi.org/10.1037/0278-7393.13.4.640

Bindemann, M., Attard, J., & Johnston, R. A. (2014). Perceived ability and actual recognition accuracy for unfamiliar and famous faces. *Cogent Psychology*, *1*(1), 1–15. https://doi.org/10.1080/23311908.2014.986903

Bobak, A. K., Mileva, V. R., & Hancock, P. J. (2018). Facing the facts: Naive participants have only moderate insight into their face recognition and face perception abilities. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, *72*(4), 872–881. https://doi.org/10.1177/1747021818776145

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. https://doi.org/10.1037/1076-898X.7.3.207

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

Brunsdon, R., Coltheart, M., Nickels, L., & Joy, P. (2006). Developmental prosopagnosia: A case analysis and treatment study. *Cognitive Neuropsychology*, *23*(6), 822–840. https://doi.org/10.1080/02643290500441841

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

Chase, W. G., & Ericsson, K. A. (1982). Skill and working memory. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 16, pp. 1–58). Academic Press.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology*, *4*(1), 55–81. https://doi.org/10.1016/0010-0285(73)90004-2

Coin, C., & Tiberghien, G. (1997). Encoding activity and face recognition. *Memory*, *5*(5), 545–568. https://doi.org/10.1080/741941479

DeGutis, J. M., Bentin, S., Robertson, L. C., & D'Esposito, M. (2007). Functional plasticity in ventral temporal cortex following cognitive rehabilitation of a congenital prosopagnosic. *Journal of Cognitive Neuroscience*, *19*(11), 1790–1802. https://doi.org/10.1162/jocn.2007.19.11.1790

DeGutis, J. M., Chiu, C., Grosso, M. E., & Cohan, S. (2015). Face processing improvements in prosopagnosia: Successes and failures over the last 50 years. *Frontiers in Human Neuroscience*, *8*, 561. https://doi.org/10.3389/fnhum.2014.00561

Dowsett, A. J., & Burton, A. M. (2014). Unfamiliar face matching: Pairs out-perform individuals and provide a route to training. *British Journal of Psychology*, *106*(3), 433–445. https://doi.org/10.1111/bjop.12103

Drew, T., Vo, M. L. H., & Wolfe, J. M. (2013). The invisible gorilla strikes again: Sustained inattentional blindness in expert observers. *Psychological Science*, *24*(9), 1848–1853. https://doi.org/10.1177/0956797613479386

Ellis, H. D., Shepherd, J. W., & Davies, G. M. (1979). Identification of familiar and unfamiliar faces from internal and external features: Some implications for theories of face recognition. *Perception*, *8*(4), 431–439. https://doi.org/10.1068/p080431

Ericsson, K. A., Charness, N., Feltovich, P. J., & Hoffman, R. R. (Eds.). (2006). *The Cambridge handbook of expertise and expert performance*. Cambridge University Press. https://doi.org/10.1017/CBO9780511816796

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

Farah, M. J. (1991). Patterns of co-occurence among the associative agnosias: Implications for visual object representation. *Cognitive Neuropsychology*, *8*(1), 1–19. https://doi.org/10.1080/02643299108253364

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

JASP Team. (2020). JASP (Version 0.13.1) [Computer software]. https://jasp-stats.org/faq/how-do-i-cite-jasp/

Jeffreys, H. (1961). *The theory of probability* (3rd ed.). Oxford University Press.

Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologist*, *64*(6), 515–526. https://doi.org/10.1037/a0016755

Kounios, J., & Beeman, M. (2014). The cognitive neuroscience of insight. *Annual Review of Psychology*, *65*, 71–93. https://doi.org/10.1146/annurev-psych-010213-115154

Kramer, R. S. S., Manesi, Z., Towler, A., Reynolds, M. G., & Burton, A. M. (2018). Familiarity and within-person facial variability: The importance of the internal and external features. *Perception*, *47*(1), 3–15. https://doi.org/10.1177/0301006617725242

Lee, M. D., & Wagenmakers, E. J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge University Press.

Logan, A. J., Gordon, G. E., & Loffler, G. (2017). Contributions of individual face features to face discrimination. *Vision Research*, *137*, 29–39. https://doi.org/10.1016/j.visres.2017.05.011

McKone, E., & Yovel, G. (2009). Why does picture-plane inversion sometimes dissociate perception of features and spacing in faces, and sometimes not? Toward a new theory of holistic processing. *Psychonomic Bulletin & Review*, *16*(5), 778–797. https://doi.org/10.3758/PBR.16.5.778

Megreya, A. M. (2018). Feature-by-feature comparison and holistic processing in unfamiliar face matching. *PeerJ*, *6*(e4437), e4437–e4446. https://doi.org/10.7717/peerj.4437

Megreya, A. M., & Bindemann, M. (2018). Feature instructions improve face-matching accuracy. *PLoS ONE*, *13*(3), e0193455. https://doi.org/10.1371/journal.pone.0193455

Megreya, A. M., & Burton, A. M. (2007). Hits and false positives in face matching: A familiarity-based dissociation. *Perception & Psychophysics*, *69*(7), 1175–1184. https://doi.org/10.3758/BF03193954

Megreya, A. M., White, D., & Burton, A. M. (2011). The other-race effect does not rely on memory: Evidence from a matching task. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, *64*(8), 1473–1483. https://doi.org/10.1080/17470218.2011.575228

Meissner, C. A., & Brigham, J. C. (2001). Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychology, Public Policy, and Law*, *7*(1), 3–35. https://doi.org/10.1037/1076-8971.7.1.3

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

Noyes, E., Phillips, P. J., & O'Toole, A. J. (2017). What is a super-recogniser? In M. Bindemann & A. M. Megreya (Eds.), *Face processing: Systems, disorders and cultural differences* (pp. 173–201). Nova Science.

Palermo, R., Rossion, B., Rhodes, G., Laguesse, R., Tez, T., Hall, B., Albonico, A., Malaspina, M., Daini, R., Irons, J., Al-Janabi, S., Taylor, L. C., Rivolta, D., & McKone, E. (2017). Do people have insight into their face recognition abilities? *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, *70*(2), 218–233. https://doi.org/10.1080/17470218.2016.1161058

Phillips, P. J., Beveridge, J. R., Draper, B. A., Givens, G., O'Toole, A. J., Bolme, D. S., Dunlop, J., Lui, Y. M., Sahibzada, H., & Weimer, S. (2011). *An introduction to the good, the bad, & the ugly face recognition challenge problem* [Paper presentation]. IEEE International Conference on Automatic Face & Gesture Recognition.

Phillips, P. J., Yates, A. N., Hu, Y., Hahn, C. A., Noyes, E., Jackson, K., Cavazos, J. G., Jeckeln, G., Ranjan, R., Sankaranarayanan, S., Chen, J., Castillo, C. D., Chellappa, R., White, D., & O'Toole, A. J. (2018). Face recognition accuracy in forensic examiners, super-recognisers and algorithms. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(24), 6171–6176. https://doi.org/10.1073/pnas.1721355115

Ramon, M., Miellet, S., Dzieciol, A. M., Konrad, B. N., Dresler, M., & Caldara, R. (2016). Super-memorisers are not super-recognisers. *PLoS ONE*, *11*(3), e0150972. https://doi.org/10.1371/journal.pone.0150972

Ritchie, K. L., Smith, F. G., Jenkins, R., Bindemann, M., White, D., & Burton, A. M. (2015). Viewers base estimates of face matching accuracy on their own familiarity: Explaining the photo-ID paradox. *Cognition*, *141*(0), 161–169. https://doi.org/10.1016/j.cognition.2015.05.002

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

Sadr, J., Jarudi, I., & Sinha, P. (2003). The role of eyebrows in face recognition. *Perception*, *32*(3), 285–293. https://doi.org/10.1068/p5027

Schmalzl, L., Palermo, R., Green, M., Brunsdon, R., & Coltheart, M. (2008). Training of familiar face recognition and visual scan paths for faces in a child with congenital prosopagnosia. *Cognitive Neuropsychology*, *25*(5), 704–729. https://doi.org/10.1080/02643290802299350

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. https://doi.org/10.1111/1467-9280.00472

Siegle, R. L., Baram, E. M., Reuter, S. R., Clarke, E. A., Lancaster, J. L., & McMahan, C. A. (1998). Rates of disagreement in imaging interpretation in a group of community hospitals. *Academic Radiology*, *5*(3), 148–154. https://doi.org/10.1016/S1076-6332(98)80277-8

Staszewski, J. J. (1988). Skilled memory and expert mental calculation. In M. T. H. Chi, R. Glaser, & M. J. Farr (Eds.), *The nature of expertise* (pp. 71–28). Erlbaum.

Staszewski, J. J., & Davison, A. (2000). *Mine detection training based on expert skill* [Paper presentation]. Proceedings of SPIE, Orlando, United States.

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. https://doi.org/10.1080/14640749308401045

Tanaka, J. W., & Simonyi, D. (2016). The "parts and wholes" of face recognition: A review of the literature. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, *69*(10), 1876–1889. https://doi.org/10.1080/17470218.2016.1146780

Tangen, J. M., Thompson, M. B., & McCarthy, D. J. (2011). Identifying fingerprint expertise. *Psychological Science*, *22*(8), 995–997. https://doi.org/10.1177/0956797611414729

Tardif, J., Morin Duchesne, X., Cohan, S., Royer, J., Blais, C., Fiset, D., Duchaine, B., & Gosselin, F. (2019). Use of face information varies systematically from developmental prosopagnosics to super-recognizers. *Psychological Science*, *30*(2), 300–308. https://doi.org/10.1177/0956797618811338

Towler, A., Kemp, R. I., Burton, A. M., Dunn, J. D., Wayne, T., Moreton, R., & White, D. (2019). Do professional facial image comparison training courses work? *PLoS ONE*, *14*(2), e0211037. https://doi.org/10.1371/journal.pone.0211037

Towler, A., Kemp, R. I., & White, D. (2021). Can face identification ability be trained? Evidence for two routes to expertise. In M. Bindemann (Ed.), *Forensic face matching: Research and practice*. Oxford University Press.

Towler, A., White, D., & Kemp, R. I. (2014). Evaluating training methods for facial image comparison: The face shape strategy does not work. *Perception*, *43*(2–3), 214–218. https://doi.org/10.1068/p7676

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

Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion, and race in face recognition. *Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology*, *43*(2), 161–204. https://doi.org/10.1080/14640749108400966

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*, *20*(2), 166–173. https://doi.org/10.1037/xap0000009

White, D., Kemp, R. I., Jenkins, R., & Burton, A. M. (2014). Feedback training for facial image comparison. *Psychonomic Bulletin & Review*, *21*(1), 100–106. https://doi.org/10.3758/s13423-013-0475-3

White, D., Phillips, P. J., Hahn, C. A., Hill, M., & O'Toole, A. J. (2015). Perceptual expertise in forensic facial image comparison. *Proceedings of the Royal Society of London, Series B: Biological Sciences*, *282*(1814), 20151292. https://doi.org/10.1098/rspb.2015.1292

White, D., Rivolta, D., Burton, A. M., Al-Janabi, S., & Palermo, R. (2017). Face matching impairment in developmental prosopagnosia. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, *70*(2), 287–297. https://doi.org/10.1080/17470218.2016.1173076

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

Williams, A. M., Ward, P., Knowles, J. M., & Smeeton, N. J. (2002). Anticipation skill in a real-world task: Measurement, training, and transfer in tennis. *Journal of Experimental Psychology: Applied*, *8*(4), 259–270. https://doi.org/10.1037/1076-898X.8.4.259

Wilmer, J. B. (2017). Individual differences in face recognition: A decade of discovery. *Current Directions in Psychological Science*, *26*(3), 225–230. https://doi.org/10.1177/0963721417710693

Wolfe, J. M., Alaoui Soce, A., & Schill, H. M. (2017). How did I miss that? Developing mixed hybrid visual search as a 'model system' for incidental finding errors in radiology. *Cognitive Research: Principles and Implications*, *2*(1), 35. https://doi.org/10.1186/s41235-017-0072-5

Yin, R. K. (1969). Looking at upside-down faces. *Journal of Experimental Psychology*, *81*(1), 141–145. https://doi.org/10.1037/h0027474

Young, A. W., Hellawell, D. J., & Hay, D. C. (1987). Configurational information in face perception. *Perception*, *16*(6), 747–759. https://doi.org/10.1068/p160747

Received October 24, 2019
Revision received August 31, 2020
Accepted September 1, 2020 ■