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

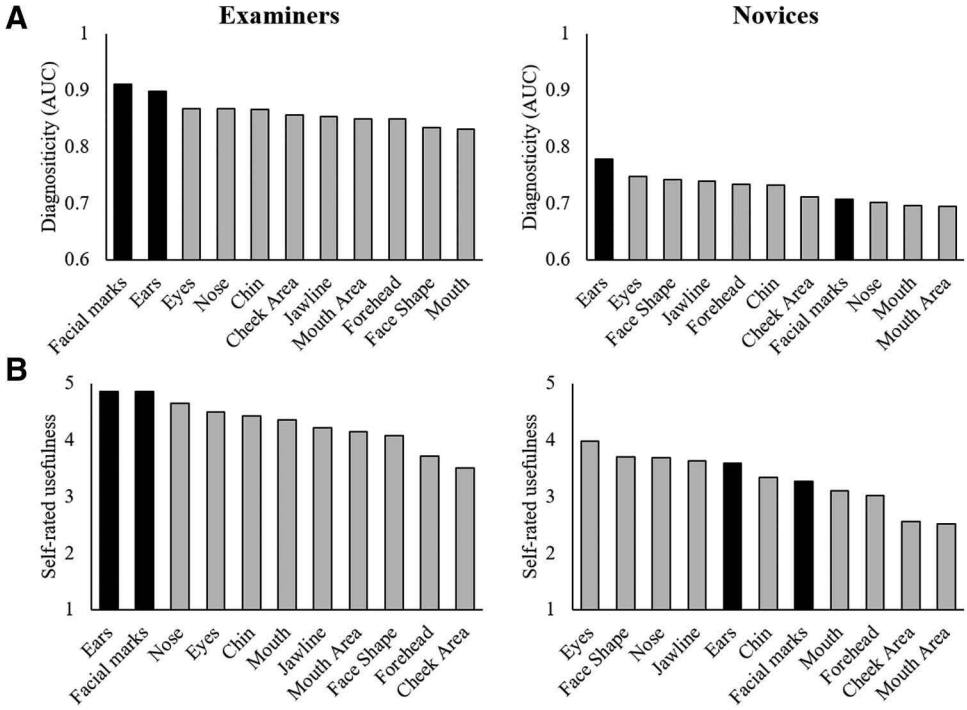
Participants

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

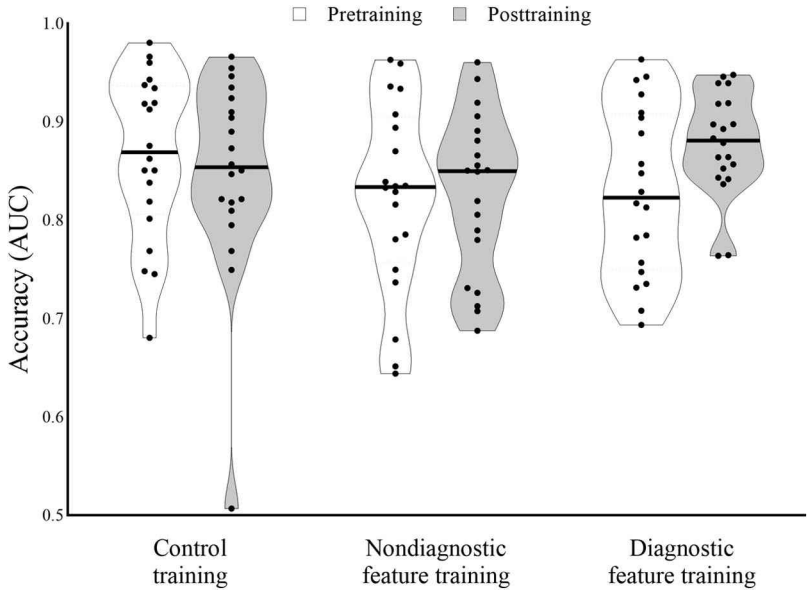
Most participants *thought* training made face matching easier regardless of whether it improved accuracy (see Figure 3). Seventy percent of participants in the diagnostic feature training group and 65% of participants in the nondiagnostic feature training group reported that training made face matching 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

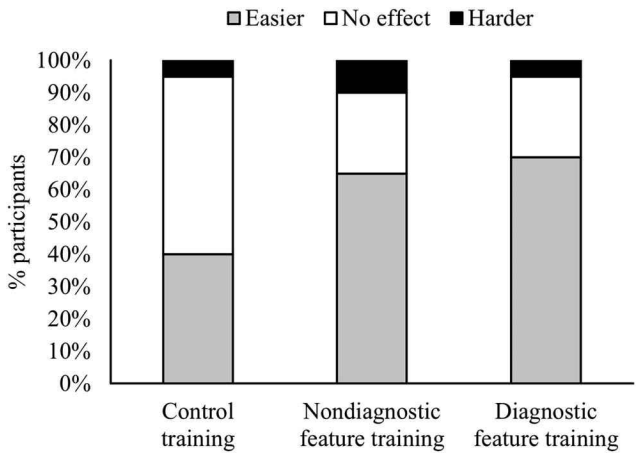


Note. AUC = area under the ROC curve. Participants who received the control training or nondiagnostic feature training did not show any improvement. Markers represent individual data points, horizontal lines represent the medians, and means are reported in the main text.

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).

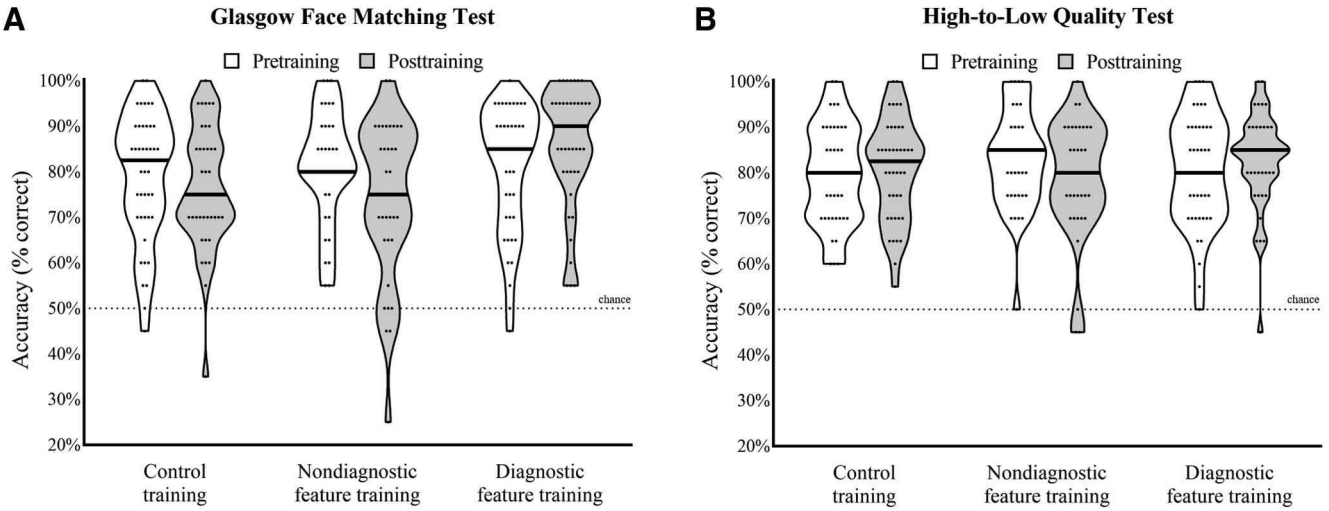
GFMT Accuracy

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 ■

Subject	Training Group	Pre-test AUC (/1)	Post-test AUC (/1)	Pre-test Average Trial RT (ms)	Post-test Average Trial RT (ms)
1	Control	0.8506235828	0.85062358277	2270	3190.5
2	Control	0.8506235828	0.84693877551	6107.5	4414.5
3	Control	0.8381519274	0.76870748299	8369.5	6279
4	Control	0.9600340136	0.96626984127	2527.5	3300
5	Control	0.9662698413	0.80952380952	2488	2410.5
6	Control	0.9192176871	0.90419501134	3455.5	3595.5
7	Control	0.9126984127	0.92403628118	3034.5	2777
8	Diagnostic	0.7568027211	0.76445578231	3775	4969
9	Diagnostic	0.8571428571	0.93934240363	3720.5	6888
10	Non-Diagno	0.6513605442	0.71258503401	5421.5	5117
11	Non-Diagno	0.7806122449	0.81944444444	2660	3136
12	Diagnostic	0.9041950113	0.84183673469	3829.5	3643
13	Diagnostic	0.7823129252	0.85685941043	3268	5405.5
14	Non-Diagno	0.8330498866	0.89087301587	3650	4009.5
15	Diagnostic	0.8171768707	0.89285714286	2402.5	1810
16	Non-Diagno	0.9631519274	0.78968253968	7246	8635
17	Diagnostic	0.9280045351	0.94784580499	3042	4134
18	Non-Diagno	0.8344671202	0.80555555556	4423	6349
19	Control	0.8755668934	0.87301587302	16509.5	16559
20	Non-Diagno	0.8939909297	0.85090702948	5717	4930
21	Diagnostic	0.7080498866	0.83673469388	2894	4968.5
22	Diagnostic	0.7474489796	0.76388888889	4111	4797
23	Non-Diagno	0.6788548753	0.73100907029	4032.5	2964.5
24	Control	0.8015873016	0.82142857143	2004.5	1739.5
25	Non-Diagno	0.9336734694	0.94359410431	9539.5	20194
26	Diagnostic	0.7845804989	0.86422902494	4259	5218.5
27	Diagnostic	0.8883219955	0.93934240363	8198	12347
28	Control	0.7451814059	0.81802721088	1489.5	1349.5
29	Diagnostic	0.735260771	0.86394557823	4851.5	6177.5
30	Non-Diagno	0.6439909297	0.72619047619	2020.5	4438.5
31	Control	0.6805555556	0.50680272109	2285.5	1825.5
32	Diagnostic	0.9634353741	0.84325396825	8408.5	10468
33	Control	0.818877551	0.90986394558	1412	1388
34	Non-Diagno	0.8160430839	0.68764172336	3611.5	2504
35	Non-Diagno	0.8350340136	0.85572562358	1887.5	2238.5
36	Control	0.8625283447	0.82142857143	4438	2824
37	Non-Diagno	0.8390022676	0.86592970522	4196.5	2605
38	Diagnostic	0.8477891156	0.91893424036	3580.5	4040.5
39	Non-Diagno	0.7366780045	0.7074829932	2051.5	2925
40	Control	0.9804421769	0.94642857143	3027	2527.5
41	Control	0.768707483	0.74943310658	3760	6942.5
42	Non-Diagno	0.8287981859	0.88095238095	4828.5	8361.5
43	Non-Diagno	0.9359410431	0.9056122449	3026.5	1926.5
44	Control	0.9183673469	0.89002267574	3136	6052.5
45	Non-Diagno	0.8701814059	0.85034013605	2847	2605.5
46	Diagnostic	0.945861678	0.88321995465	10023	7893.5

47	Non-Diagno	0.7497165533	0.78004535147	1958	3556.5
48	Control	0.9430272109	0.95464852608	3674	3081
49	Diagnostic	0.9092970522	0.91836734694	4258.5	6170
50	Non-Diagno	0.9075963719	0.91950113379	5093.5	7878
51	Control	0.7482993197	0.79506802721	4750.5	3580
52	Diagnostic	0.9424603175	0.945861678	3712.5	4024.5
53	Non-Diagno	0.9591836735	0.96060090703	12340	10522
54	Control	0.9373582766	0.85685941043	2222.5	1560
55	Diagnostic	0.7315759637	0.85260770975	4524.5	6255
56	Diagnostic	0.8287981859	0.87868480726	5600	6224
57	Diagnostic	0.8129251701	0.89739229025	3775	7784.5
58	Non-Diagno	0.785430839	0.84948979592	3183	7043
59	Diagnostic	0.6935941043	0.89767573696	5187	7738
60	Control	0.9342403628	0.93480725624	3838	2691

Pre-test Duration (mins)	Post-test Duration (mins)	Effect of training
5.045	7.059	Easier
11.001	8.448	No effect
14.383	11.042	No effect
5.998	6.912	No effect
5.161	5.266	No effect
6.326	6.682	Easier
6.594	5.689	No effect
9.066	9.324	Easier
7.355	11.186	Easier
9.304	11.361	No effect
5.786	7.392	Easier
7.365	7.726	Easier
6.882	10.062	Easier
7.857	8.691	No effect
5.502	5.026	No effect
12.847	14.791	Easier
6.455	8.393	Easier
8.130	10.868	Easier
25.560	24.932	Easier
9.876	9.029	Easier
6.485	9.548	Easier
7.088	8.149	Easier
7.785	6.357	Easier
4.694	4.639	No effect
16.997	28.639	Easier
8.489	9.873	No effect
16.092	21.765	No effect
3.793	3.530	No effect
8.789	11.060	Easier
4.761	8.971	Easier
5.701	4.645	No effect
15.965	19.816	Easier
3.384	3.643	Harder
6.819	5.697	Easier
5.008	5.421	Easier
8.541	5.717	Easier
7.734	5.712	Easier
8.386	8.511	Easier
4.372	6.042	Easier
6.436	5.351	No effect
7.626	11.959	Easier
9.811	14.752	No effect
6.168	4.988	No effect
6.890	10.843	Easier
6.715	5.527	No effect
18.402	16.441	Easier

4.192	7.191	Easier
7.021	5.800	No effect
8.270	10.141	Easier
8.992	13.561	Harder
9.561	7.410	Easier
7.417	8.304	Easier
20.887	18.077	Harder
4.955	3.823	Easier
11.252	12.404	No effect
13.072	12.807	Easier
7.387	14.725	Harder
6.151	12.053	Easier
10.621	12.931	No effect
7.381	5.508	No effect

		GFMT			
		Pre-test			
Subject	Training Group	Overall accuracy (/100)	Match trial accuracy (/100)	Non-match trial accuracy (/100)	Overall accuracy (/100)
1	Diagnostic	60	100	20	55
2	Diagnostic	95	90	100	85
3	Diagnostic	95	90	100	95
4	Diagnostic	50	0	100	55
5	Diagnostic	85	70	100	75
6	Diagnostic	85	80	90	85
7	Diagnostic	90	100	80	95
8	Diagnostic	95	100	90	95
9	Diagnostic	85	100	70	90
10	Diagnostic	85	70	100	80
11	Diagnostic	95	90	100	95
12	Diagnostic	80	60	100	70
13	Diagnostic	85	90	80	95
14	Diagnostic	65	50	80	65
15	Diagnostic	85	90	80	80
16	Diagnostic	60	60	60	55
17	Diagnostic	70	50	90	60
18	Diagnostic	90	90	90	95
19	Diagnostic	90	80	100	95
20	Diagnostic	95	90	100	90
21	Diagnostic	75	100	50	95
22	Diagnostic	65	90	40	85
23	Control	70	50	90	70
24	Control	90	100	80	80
25	Control	95	90	100	75
26	Control	60	50	70	75
27	Control	100	100	100	100
28	Control	70	50	90	90
29	Control	90	100	80	95
30	Control	95	100	90	75
31	Control	90	80	100	95
32	Control	85	80	90	70
33	Control	45	0	90	35
34	Control	60	80	40	60
35	Control	95	100	90	85
36	Control	75	50	100	90
37	Control	95	100	90	85
38	Control	70	50	90	70
39	Control	90	80	100	70
40	Control	65	40	90	70
41	Control	85	70	100	90
42	Control	70	40	100	75
43	Control	80	80	80	75
44	Non-Diagr	80	100	60	45

45	Non-Diagn	80	100	60	75
46	Non-Diagn	85	90	80	50
47	Non-Diagn	95	100	90	85
48	Non-Diagn	80	60	100	65
49	Non-Diagn	60	20	100	75
50	Non-Diagn	90	100	80	70
51	Non-Diagn	75	100	50	70
52	Non-Diagn	100	100	100	85
53	Non-Diagn	85	100	70	70
54	Non-Diagn	100	100	100	90
55	Non-Diagn	100	100	100	90
56	Non-Diagn	70	70	70	70
57	Non-Diagn	80	90	70	80
58	Non-Diagn	95	90	100	70
59	Non-Diagn	90	80	100	85
60	Non-Diagn	80	60	100	65
61	Non-Diagn	85	90	80	90
62	Non-Diagn	95	90	100	80
63	Diagnostic	95	100	90	100
64	Diagnostic	70	80	60	85
65	Diagnostic	70	100	40	95
66	Diagnostic	90	90	90	100
67	Diagnostic	65	80	50	85
68	Diagnostic	65	100	30	80
69	Diagnostic	75	60	90	85
70	Diagnostic	90	80	100	100
71	Diagnostic	75	70	80	85
72	Diagnostic	100	100	100	100
73	Diagnostic	95	90	100	95
74	Diagnostic	90	80	100	80
75	Diagnostic	45	70	20	70
76	Diagnostic	95	100	90	95
77	Diagnostic	80	80	80	90
78	Diagnostic	95	90	100	85
79	Diagnostic	55	80	30	80
80	Diagnostic	90	90	90	90
81	Diagnostic	85	70	100	100
82	Diagnostic	90	90	90	100
83	Diagnostic	80	90	70	100
84	Control	85	90	80	70
85	Control	95	90	100	85
86	Control	85	80	90	80
87	Control	70	100	40	70
88	Control	80	100	60	70
89	Control	75	70	80	85
90	Control	75	100	50	70
91	Control	50	40	60	65
92	Control	60	60	60	65

93	Control	75	90	60	85
94	Control	85	70	100	70
95	Control	80	80	80	75
96	Control	90	80	100	95
97	Control	55	40	70	60
98	Control	85	80	90	80
99	Control	90	100	80	85
100	Control	55	50	60	55
101	Control	85	90	80	60
102	Control	85	100	70	95
103	Control	100	100	100	95
104	Control	75	70	80	65
105	Non-Diagı	80	90	70	90
106	Non-Diagı	80	90	70	85
107	Non-Diagı	75	50	100	65
108	Non-Diagı	85	80	90	90
109	Non-Diagı	90	90	90	100
110	Non-Diagı	60	70	50	85
111	Non-Diagı	55	40	70	50
112	Non-Diagı	65	100	30	50
113	Non-Diagı	70	50	90	45
114	Non-Diagı	90	90	90	90
115	Non-Diagı	80	60	100	25
116	Non-Diagı	55	70	40	55
117	Non-Diagı	95	90	100	100
118	Non-Diagı	65	80	50	70
119	Non-Diagı	85	70	100	90
120	Non-Diagı	80	70	90	75
121	Non-Diagı	85	70	100	90

		High-to-Lo		
Post-test		Pre-test		
Match trial accuracy (/100)	Non-match trial accuracy (/100)	Overall accuracy (/100)	Match trial accuracy (/100)	Non-match trial accuracy (/100)
10	100	70	100	40
70	100	65	80	50
90	100	85	70	100
30	80	55	60	50
80	70	80	80	80
80	90	85	100	70
100	90	95	100	90
90	100	75	60	90
100	80	75	50	100
70	90	90	80	100
90	100	75	50	100
60	80	70	40	100
90	100	60	60	60
60	70	65	30	100
90	70	70	100	40
50	60	75	100	50
30	90	75	70	80
90	100	80	70	90
100	90	80	60	100
90	90	80	70	90
90	100	100	100	100
100	70	85	70	100
40	100	90	90	90
100	60	90	100	80
70	80	80	60	100
60	90	70	50	90
100	100	95	100	90
90	90	70	40	100
100	90	90	90	90
100	50	80	90	70
100	90	80	90	70
70	70	75	90	60
50	20	65	80	50
60	60	85	80	90
80	90	85	90	80
100	80	70	40	100
90	80	95	100	90
70	70	60	80	40
50	90	75	50	100
40	100	60	20	100
100	80	85	70	100
60	90	70	50	90
90	60	65	80	50
40	50	70	100	40

100	50	80	100	60
40	60	75	90	60
80	90	75	80	70
70	60	95	100	90
70	80	70	50	90
90	50	100	100	100
70	70	85	100	70
100	70	100	100	100
80	60	85	90	80
90	90	90	80	100
100	80	85	100	70
100	40	75	100	50
80	80	70	40	100
90	50	80	100	60
90	80	80	70	90
30	100	75	50	100
100	80	90	100	80
80	80	85	90	80
100	100	85	100	70
80	90	85	70	100
100	90	80	90	70
100	100	95	90	100
90	80	70	80	60
100	60	90	80	100
100	70	70	100	40
100	100	95	90	100
90	80	70	90	50
100	100	100	100	100
90	100	100	100	100
60	100	85	80	90
100	40	50	100	0
90	100	95	90	100
80	100	90	100	80
80	90	75	60	90
90	70	90	90	90
90	90	70	60	80
100	100	90	100	80
100	100	90	90	90
100	100	90	80	100
90	50	100	100	100
80	90	70	40	100
80	80	80	70	90
70	70	75	70	80
100	40	70	100	40
80	90	95	100	90
100	40	70	100	40
100	30	80	90	70
90	40	75	90	60

80	90	70	100	40
60	80	90	100	80
70	80	85	80	90
90	100	100	100	100
40	80	75	80	70
80	80	90	100	80
100	70	90	80	100
50	60	60	70	50
20	100	80	60	100
100	90	85	100	70
100	90	90	80	100
100	30	80	80	80
100	80	85	90	80
80	90	95	100	90
30	100	85	80	90
80	100	85	70	100
100	100	95	100	90
80	90	75	60	90
0	100	80	70	90
60	40	50	70	30
20	70	80	90	70
90	90	75	50	100
30	20	70	50	90
70	40	80	90	70
100	100	90	90	90
50	90	85	90	80
90	90	100	100	100
70	80	90	90	90
80	100	100	100	100

w Quality Test					
Post-test					
Overall accuracy (/100)	**Match trial accuracy (/100)**	**Non-match trial accuracy (/100)**	**Pre-test Duration (sec)**	**Training Duration (sec)**	**Post-test Duration (sec)**
75	50	100	2057	-	779
90	90	90	570	189	704
85	70	100	1952	852	571
45	30	60	810	240	532
85	80	90	1391	270	1405
80	60	100	1405	724	538
90	90	90	3424	2521	1487
90	80	100	2161	1081	922
85	70	100	746	328	866
75	50	100	2139	1114	836
85	70	100	1610	222	567
85	70	100	2242	1434	1232
75	60	90	606	168	667
65	30	100	1849	899	1388
80	90	70	890	209	974
80	100	60	846	2016	536
65	30	100	3716	330	688
95	90	100	1773	356	799
80	60	100	1588	738	543
95	90	100	1954	3808	1229
65	40	90	5413	1403	1185
90	100	80	1413	856	558
85	70	100	1011	514	550
85	100	70	757	273	419
65	30	100	1905	274	674
75	50	100	628	200	506
95	90	100	818	235	353
55	40	70	421	646	397
95	90	100	1028	247	901
95	100	90	769	738	484
100	100	100	833	312	737
70	70	70	606	1120	297
70	90	50	218	470	164
85	90	80	3094	1272	977
80	60	100	1272	489	1054
95	90	100	1993	-	1853
100	100	100	634	209	467
75	80	70	813	-	1285
85	70	100	2400	1388	818
65	30	100	602	468	575
75	60	90	3469	1691	1156
70	40	100	1629	892	455
90	100	80	632	290	671
75	80	70	730	281	574

75	100	50	1795	425	704
45	30	60	4512	2913	2183
95	100	90	797	276	796
70	90	50	587	2714	636
85	90	80	1765	796	1016
70	70	70	1571	498	1612
85	90	80	1329	803	561
100	100	100	560	207	648
70	80	60	3070	474	1479
85	80	90	776	324	627
70	100	40	642	246	621
45	90	0	1327	654	290
90	90	90	1052	243	700
90	100	80	705	219	816
90	90	90	1894	1287	922
65	30	100	2476	329	1211
80	100	60	1151	654	440
75	60	90	1582	243	681
95	90	100	2344	1124	909
75	60	90	500	204	394
85	90	80	1215	306	1121
90	80	100	885	261	749
95	100	90	923	224	707
90	90	90	680	266	617
70	80	60	3016	1752	1264
85	70	100	821	226	832
85	90	80	462	272	589
100	100	100	618	182	710
100	100	100	2473	202	1056
80	60	100	1067	206	851
80	80	80	653	182	461
85	70	100	2205	1257	972
85	80	90	1618	982	817
85	70	100	2514	1320	990
80	100	60	946	160	574
90	100	80	3235	1684	1459
85	90	80	965	186	704
85	80	90	5098	2485	2151
80	60	100	823	331	976
90	100	80	6653	222	4695
75	50	100	1392	264	1031
70	70	70	661	227	495
80	60	100	824	304	696
90	100	80	341	221	276
85	80	90	2364	1622	1358
65	100	30	2299	786	540
80	80	80	474	223	377
80	100	60	7145	-	1660

85	100	70	465	370	498
90	90	90	2341	1232	734
65	40	90	861	224	919
90	90	90	1473	786	593
70	80	60	786	204	842
85	90	80	861	258	736
80	90	70	1354	346	1497
85	100	70	952	1107	681
85	80	90	2913	-	749
80	100	60	3488	837	1124
85	70	100	455	173	402
60	50	70	768	202	575
85	100	70	6469	326	1220
90	100	80	2828	1347	998
80	70	90	1454	317	1531
85	70	100	1112	197	927
75	50	100	968	232	570
75	50	100	619	478	730
50	0	100	6654	2034	912
75	100	50	1039	185	340
75	60	90	2488	1233	817
90	90	90	819	210	516
80	90	70	1830	1234	502
80	70	90	878	272	1055
70	90	50	1119	340	1056
90	80	100	1061	724	468
90	100	80	1852	1055	731
90	80	100	824	206	606
95	100	90	1321	210	308

Do you think that the training made you more accurate in face matching? (Y/N)

Yes
Yes
Yes
No
Yes
Yes
Yes
Yes
Yes
No
No
Yes
Yes
Yes
Yes
Yes
Yes
Yes
Yes
Yes
Yes
Yes
No
Yes
No
No
No
No
No
No
Yes
No
No
Yes
No
No
Yes
Yes
Yes
No
No
No
Yes

Yes
Yes
No
Yes
Yes
Yes
Yes
No
Yes
Yes
Yes
Yes
No
Yes
Yes
Yes
Yes
Yes
Yes
Yes
Yes
Yes
Yes
Yes
Yes
Yes
Yes
Yes
Yes
Yes
Yes
No
Yes
Yes
Yes
Yes
Yes
Yes
Yes
Yes
No
No
No
No
No
No
Yes
No
Yes

No
No
No
No
Yes
No
No
No
Yes
Yes
No
Yes
Yes
Yes
Yes
Yes
Yes
Yes
Yes
Yes
No
No
Yes
Yes
Yes
Yes
Yes
Yes
Yes

SUPPLEMENTARY MATERIALS

Diagnostic feature training improves face matching accuracy

Alice Towler*, Michelle Keshwa, Bianca Ton, Richard I. Kemp & David White
University of New South Wales, Australia
* Corresponding author: a.towler@unsw.edu.au

EXPERIMENT 1

Main analysis

We analysed the AUC data using a 3 x 2 mixed ANOVA, with Training (diagnostic feature, non-diagnostic feature, control) as a between-subjects factor and Test (pre-training, post-training) as a within-subjects factor. The main effects of Training [$F<1$, $\eta_p^2 = .03$] and Test [$F(1, 57) = 1.53$, $p > .05$, $\eta_p^2 = .03$] were non-significant. Importantly however, the critical interaction between Training and Test was significant [$F(2, 57) = 4.91$, $p < .05$, $\eta_p^2 = .15$].

Simple main effects analyses revealed that participants who completed the diagnostic feature training significantly improved by 6% pre- to post-training [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 non-diagnostic feature [$F <1$, $\eta_p^2 = .00$] or control [$F(1, 57) = 1.42$, $p > .05$, $\eta_p^2 = .02$] training showed no change in accuracy pre- to post-training (non-diagnostic: pre $M = .82$, $SD = .10$, post $M = .83$, $SD = .08$; control: pre $M = .87$, $SD = .08$, post $M = .85$, $SD = .10$).

Supplementary analyses

2 x 2 ANOVAs

In addition to the main analysis, we analysed the AUC data using a 2 x 2 mixed ANOVA, with Training (control, diagnostic feature) as a between-subjects factor and Test (pre-training, post-training) as a within-subjects factor. The main effects of Training [$F<1$, $\eta_p^2 = .00$] and Test [$F(1, 38) = 1.81$, $p > .05$, $\eta_p^2 = .05$] were non-significant. Importantly however, the critical interaction between Training and Test was significant [$F(1, 38) = 8.89$, $p < .05$, $\eta_p^2 = .19$].

Simple main effects analysis revealed that participants who completed the diagnostic feature training showed a significant 6% improvement in performance from pre-training ($M = .83$, $SD = .09$) to post-training ($M = .88$, $SD = .05$), $F(1, 38) = 9.36$, $p < .05$, $\eta_p^2 = .20$. Participants in the control condition showed no significant change in performance from pre-training ($M = .87$, $SD = .08$) to post-training ($M = .85$, $SD = .10$), $F(1, 38) = 1.34$, $p > .05$, $\eta_p^2 = .03$.

We repeated the 2 x 2 ANOVA analysis above, comparing the control training to the non-diagnostic feature training. The main effects of Training [$F(1, 38) = 1.33$, $p > .05$, $\eta_p^2 = .03$] and Test [$F < 1$, $\eta_p^2 = .02$] were non-significant. Importantly, the critical interaction between Training and Test was also non-significant [$F(1, 38) = 1.06$, $p > .05$, $\eta_p^2 = .03$], demonstrating no difference in performance between the control and non-diagnostic feature training groups.

1-way ANCOVA

We conducted a one-way ANCOVA to determine if post-training accuracy differed between the three training groups (control, diagnostic, non-diagnostic), after controlling for pre-training accuracy. We found a significant effect of Training on post-training accuracy, after controlling for pre-training accuracy [$F(2, 56) = 4.60$, $p < .05$, $\eta_p^2 = .14$].

Post-hoc analysis showed that post-training accuracy was significantly higher in the diagnostic feature training group compared to the control group [mean difference = .053 (95% CI .014 to .091), $p < .05$] and the non-diagnostic training group [mean difference = .047 (95% CI .009 to .085), $p < .05$]. There was no statistical difference between the control and non-diagnostic training groups [mean difference = .005 (95% CI -.034 to .044), $p > .05$].

Outliers

To check the robustness of our findings, we removed outliers and re-ran the main 3 (Training) x 2 (Test) mixed ANOVA analysis above. We classified participants as outliers if they scored at or below AUC = 0.55 on the pre-training or post-training test, where chance is 0.50. One participant from the control training qualified as an outlier against this criteria. After removing this participant, the main effects of Training [$F(2, 56) = 2.07$, $p > .05$, $\eta_p^2 = .07$] and Test [$F(1, 56) = 2.55$, $p > .05$, $\eta_p^2 = .04$] remained non-significant, and the critical interaction between Training and Test remained significant [$F(2, 56) = 4.27$, $p < .05$, $\eta_p^2 = .13$]. The results of Experiment 1 are therefore not driven by an outlier.

Participant demographics

A breakdown of participant demographics for each group in Experiment 1 is presented below. There were no differences in age, but the control condition contained more females than males. We collapsed accuracy across all three conditions and both timepoints to compare male vs. female accuracy and found no significant difference ($p = .320$). We therefore do not believe the results of Experiment 1 can be explained by participant demographics.

	Age (years)	Male	Female
Control	19.1	2	18
Non-diagnostic feature	19.5	11	9
Diagnostic feature	19.8	9	11

Perceived effectiveness of training

	Easier	No effect	Harder
Control	40%	55%	5%
Non-diagnostic feature	65%	25%	10%
Diagnostic feature	70%	25%	5%

EXPERIMENT 2

GFMT

Main analysis

Accuracy data were analysed using a 3 x 2 mixed ANOVA with Training (diagnostic feature, non-diagnostic feature, control) as a between-subjects factor and Test (pre-training, post-training) as a within-subjects factor. The main effects of Training [$F(2, 118) = 2.84$, $p > .05$, $\eta_p^2 = .05$] and Test [$F(1, 118) = 1.77$, $p > .05$, $\eta_p^2 = .02$] were non-significant, but the critical interaction between Training and Test was significant [$F(2, 118) = 9.81$, $p < .05$, $\eta_p^2 = .14$].

Simple main effects analyses revealed a significant 6% improvement from pre- to post-training for participants who received the diagnostic feature training [pre $M = 81\%$, $SD = 14\%$, post $M = 86\%$, $SD = 13\%$; $F(1, 118) = 7.13$, $p < .05$, $\eta_p^2 = .06$], and a significant 9% *decrease* in accuracy for those who received the non-diagnostic feature training [pre $M = 81\%$, $SD = 12\%$, post $M = 74\%$, $SD = 17\%$; $F(1, 118) = 11.99$, $p < .05$, $\eta_p^2 = .10$]. The control group's accuracy did not change from pre- to post-training [pre $M = 79\%$, $SD = 14\%$, post $M = 76\%$, $SD = 13\%$; $F(1, 118) = 1.58$, $p > .05$, $\eta_p^2 = .01$].

Match vs. non-match trial accuracy

We analysed match and non-match trial accuracy data using separate 3 x 2 mixed ANOVAs, with Training (diagnostic feature, non-diagnostic feature, control) as a between-subjects factor and Test (pre-training, post-training) as a within-subjects factor.

For match trials, the main effects of Training [$F(2, 118) = 1.35$, $p > .05$, $\eta_p^2 = .02$] and Test [$F<1$, $\eta_p^2 = .00$], and the interaction between Training and Test were non-significant [$F(2, 118) = 2.60$, $p > .05$, $\eta_p^2 = .04$].

For non-match trials, the main effects of Training [$F(2, 118) = 1.45$, $p > .05$, $\eta_p^2 = .02$] and Test [$F(1, 118) = 1.06$, $p > .05$, $\eta_p^2 = .01$] were non-significant. The interaction between Training and Test was significant [$F(2, 118) = 7.17$, $p < .05$, $\eta_p^2 = .11$].

Simple main effects analysis revealed a significant 11% improvement on non-match trials from pre- to post-training for participants who received the diagnostic feature training [pre $M = 80\%$, $SD = 24\%$, post $M = 88\%$, $SD = 15\%$; $F(1, 118) = 6.32$, $p < .05$, $\eta_p^2 = .05$]. Participants in the control training group showed a significant 10% decrease in accuracy from pre- to post-training [pre $M = 82\%$, $SD = 17\%$, post $M = 74\%$, $SD = 22\%$; $F(1, 118) = 5.44$, $p < .05$, $\eta_p^2 = .04$], and participants in the non-diagnostic feature training group did not change from pre- to post-training [pre $M = 82\%$, $SD = 20\%$, post $M = 75\%$, $SD = 21\%$; $F(1, 118) = 3.35$, $p > .05$, $\eta_p^2 = .03$].

Supplementary analyses

2 x 2 ANOVAs

In addition to the main analysis, we analysed the accuracy data using a 2 x 2 mixed ANOVA, with Training (control, diagnostic feature) as a between-subjects factor and Test (pre-training, post-training) as a within-subjects factor. The main effect of Training was significant [$F(1, 83) = 4.42$, $p < .05$, $\eta_p^2 = .05$]. The main effect of Test was non-significant [$F(1, 83) = 1.29$, $p > .05$, $\eta_p^2 = .02$]. Critically, the interaction between Training and Test was significant [$F(1, 83) = 10.26$, $p < .05$, $\eta_p^2 = .11$].

Simple main effects analysis revealed that participants who completed the diagnostic feature training showed a significant 6% improvement from pre-training ($M = 81\%$, $SD = 14\%$) to post-training ($M = 86\%$, $SD = 13\%$), $F(1, 83) = 9.53$, $p < .05$, $\eta_p^2 = .10$). Participants who completed the control training showed no significant change in accuracy from pre-training ($M = 79\%$, $SD = 14\%$) to post-training ($M = 76\%$, $SD = 13\%$), $F(1, 83) = 2.11$, $p > .05$, $\eta_p^2 = .03$.

We repeated the 2 x 2 ANOVA analysis above, comparing the control training to the non-diagnostic feature training. The main effect of Training was non-significant [$F < 1$, $\eta_p^2 = .00$]. The main effect of Test was significant [$F(1, 76) = 10.05$, $p < .05$, $\eta_p^2 = .12$]. The critical interaction between Training and Test was non-significant [$F(1, 76) = 2.48$, $p > .05$, $\eta_p^2 = .03$].

1-way ANCOVA

We conducted a one-way ANCOVA to determine if post-training accuracy differed between the three training groups (control, diagnostic, non-diagnostic), after controlling for pre-training

accuracy. We found a significant effect of Training on post-training accuracy, after controlling for pre-training accuracy [$F(2, 117) = 11.44$, $p < .001$, $\eta_p^2 = .16$].

Post-hoc analysis showed that post-training accuracy was significantly higher in the diagnostic feature training group compared to the control group [mean difference = 8.06 (95% CI 3.13 to 12.98), $p < .05$] and the non-diagnostic training group [mean difference = 11.98 (95% CI 6.86 to 17.10), $p < .001$]. There was no statistical difference between the control and non-diagnostic training groups [mean difference = -3.93 (95% CI -9.09 to 1.23), $p > .05$].

Outliers

To check the robustness of our findings, we removed outliers and re-ran the main 3 (Training) x 2 (Test) mixed ANOVA analysis above. We classified participants as outliers if they scored at or below 55% accuracy on the pre-training or post-training test, where chance is 50%. Four participants from the control training group, seven participants from the non-diagnostic feature training group, and five participants from the diagnostic feature training group qualified as outliers against this criteria. After removing these participants, the main effect of Test [$F<1$, $\eta_p^2 = .00$] remained non-significant and the main effect of Training became significant [$F(2, 102) = 4.62$, $p < .05$, $\eta_p^2 = .08$]. Importantly, the critical interaction between Training and Test remained significant [$F(2, 102) = 5.84$, $p < .05$, $\eta_p^2 = .10$]. The GFMT results in Experiment 2 are therefore not driven by outliers.

HIGH-TO-LOW QUALITY TEST

Main analysis

Accuracy data were analysed using a 3 x 2 mixed ANOVA with Training (diagnostic feature, non-diagnostic feature, control) as a between-subjects factor and Test (pre-training, post-training) as a within-subjects factor. The main effects of Training [$F<1$, $\eta_p^2 = .01$], and Test [$F<1$, $\eta_p^2 = .00$], and the interaction between Training and Test [$F(2, 118) = 2.80$, $p > .05$, $\eta_p^2 = .05$] were non-significant.

Match vs. non-match trial accuracy

We analysed match and non-match trial accuracy data using separate 3 x 2 mixed ANOVAs, with Training (diagnostic feature, non-diagnostic feature, control) as a between-subjects factor and Test (pre-training, post-training) as a within-subjects factor.

For match trials, the main effects of Training [$F<1$, $\eta_p^2 = .01$] and Test [$F(1,118) = 3.14$, $p > .05$], and the interaction between Training and Test [$F<1$, $\eta_p^2 = .01$] were non-significant.

For non-match trials, the main effects of Training [$F(2, 118) = 1.27$, $p > .05$, $\eta_p^2 = .02$] and Test [$F(1, 118) = 3.32$, $p > .05$, $\eta_p^2 = .03$] were non-significant. The interaction between Training and Test [$F(2, 118) = 5.35$, $p < .05$, $\eta_p^2 = .08$] was significant.

Simple main effects analysis revealed a significant 12% improvement on non-match trials from pre- to post-training for participants who received the diagnostic feature training [pre $M = 81\%$, $SD = 24\%$, post $M = 90\%$, $SD = 13\%$; $F(1, 118) = 12.02$, $p < .05$, $\eta_p^2 = .09$]. Participants in the control [pre $M = 80\%$, $SD = 20\%$, post $M = 83\%$, $SD = 17\%$; $F(1, 118) = 1.37$, $p > .05$, $\eta_p^2 = .01$] and non-diagnostic feature [pre $M = 82\%$, $SD = 18\%$, post $M = 78\%$, $SD = 22\%$; $F(1, 118) = 1.60$, $p > .05$, $\eta_p^2 = .01$] training groups did not improve.

Supplementary analysis

2 x 2 ANOVAs

In addition to the main analysis, we analysed the accuracy data using a 2 x 2 mixed ANOVA, with Training (control, diagnostic feature) as a between-subjects factor and Test (pre-training, post-training) as a within-subjects factor. The main effects of Training [$F < 1$, $\eta_p^2 = .01$] and Test [$F(1, 83) = 1.77$, $p > .05$, $\eta_p^2 = .02$], and the interaction between Training and Test [$F < 1$, $\eta_p^2 = .00$] were non-significant. This finding suggests that the benefits of diagnostic feature training are specific to imagery where fine feature detail is visible.

We repeated the 2 x 2 ANOVA analysis above, comparing the control training to the non-diagnostic feature training. The main effects of Training [$F < 1$, $\eta_p^2 = .00$] and Test [$F(1, 76) = 1.14$, $p > .05$, $\eta_p^2 = .02$], and the interaction between Training and Test [$F(1, 76) = 2.90$, $p > .05$, $\eta_p^2 = .04$] were non-significant.

1-way ANCOVA

We conducted a one-way ANCOVA to determine if post-training accuracy differed between the three training groups (control, diagnostic, non-diagnostic), after controlling for pre-training accuracy. We found a non-significant effect of Training on post-training accuracy, after controlling for pre-training accuracy [$F(2, 117) = 2.15$, $p > .05$, $\eta_p^2 = .04$].

Post-hoc analysis showed that post-training accuracy was significantly higher in the diagnostic feature training group compared to the non-diagnostic training group [mean difference = 5.06 (95% CI .22 to 9.89), $p < .05$]. There was no statistical difference between the control training and the diagnostic feature training groups [mean difference = -2.07 (95% CI -6.69 to 2.56), $p >$

.05], and the non-diagnostic training groups [mean difference = 2.99 (95% CI -1.88 to 7.86), $p >$ .05].

Outliers

To check the robustness of our findings, we removed outliers and re-ran the main 3 (Training) x 2 (Test) mixed ANOVA analysis above. We classified participants as outliers if they scored at or below 55% accuracy on the pre-training or post-training test, where chance is 50%. One participant from the control training group, four participants from the non-diagnostic feature training group, and two participants from the diagnostic feature training group qualified as outliers against this criteria. After removing these participants, the main effects of Test [$F<1$, $\eta_p^2 = .00$] and Training remained non-significant [$F(2, 111) = 1.26$, $p > .05$, $\eta_p^2 = .02$]. Importantly, the critical interaction between Training and Test remained non-significant [$F(2, 111) = 1.64$, $p > .05$, $\eta_p^2 = .03$]. The High-to-Low Quality Test results in Experiment 2 are therefore not driven by outliers.

SIGNAL DETECTION ANALYSES

Sensitivity (d′)

We calculated sensitivity on each of the tests, where hits are considered correct decisions on match trials. Sensitivity (d′) data were analysed separately for each test using a 3 x 2 mixed ANOVA with Training (diagnostic feature, non-diagnostic feature, control) as a between-subjects factor and Test (pre-training, post-training) as a within-subjects factor.

On the GFMT, main effects of Training [$F(2, 118) = 2.59$, $p > .05$, $\eta_p^2 = .04$] and Test [$F<1$, $\eta_p^2 = .01$] were non-significant. The critical interaction between Training and Test was significant [$F(2, 118) = 4.42$, $p < .05$, $\eta_p^2 = .07$]. Simple main effects analysis revealed a significant decrease in sensitivity for those in the non-diagnostic feature training group [$F(1, 118) = 4.48$, $p < .05$, $\eta_p^2 = .04$]. Sensitivity in participants who received the diagnostic feature training increased, but did not reach statistical significance [$F(1, 118) = 3.55$, $p = .062$, $\eta_p^2 = .03$]. Participants who received the control training did not show any change in sensitivity from pre- to post-test $F(1, 118) = 1.30$, $p > .05$, $\eta_p^2 = .01$].

On the high-to-low quality test, the main effects of Training [$F<1$, $\eta_p^2 = .00$] and Test [$F<1$, $\eta_p^2 = ,00$] and the critical interaction between Training and Test were non-significant [$F<1$, $\eta_p^2 = .02$].

Criterion (c)

Criterion (c) data were also analysed separately for each test using a 3 x 2 mixed ANOVA with Training (diagnostic feature, non-diagnostic feature, control) as a between-subjects factor and Test (pre-training, post-training) as a within-subjects factor.

On the GFMT, main effects of Training [$F<1$, $\eta_p^2 = .01$] and Test [$F(1, 118) = 1.05$, $p > .05$, $\eta_p^2 = .01$], and the interaction between Training and Test were non-significant [$F(2, 118) = 1.04$, $p > .05$, $\eta_p^2 = .02$].

On the high-to-low quality test, the main effect of Training was non-significant [$F(2, 118) = 2.55$, $p > .05$, $\eta_p^2 = .04$]. The main effect of Test just reached significance [$F(1, 118) = 3.93$, $p = .05$, $\eta_p^2 = .03$], indicating that, in general, participants were somewhat more likely to respond "non-match" after training compared to before. The interaction between Training and Test was non-significant [$F(2, 118) = 2.34$, $p > .05$, $\eta_p^2 = .04$].

CORRELATION BETWEEN TRAINING DURATION AND POST-TRAINING TEST ACCURACY

We investigated the relationship between training duration and accuracy at post-training for the GFMT and High-to-Low Quality Test. Correlation coefficients and p-values are reported below for each group. There were no significant correlations between the post-training tests and training duration for the control group and the diagnostic feature training group. Training duration was significantly negatively correlated with post-training tests for the non-diagnostic feature training group, i.e. the longer participants spent on the non-diagnostic feature training, the worse they performed on the subsequent post-training tests.

	GFMT		**High-to-Low Quality Test**	
	r	*p*	*r*	*p*
Control	-.074	.658	-.005	.977
Non-diagnostic feature	-.403	.015	-.457	.005
Diagnostic feature	.035	.824	.096	.547

PARTICIPANT DEMOGRAPHICS

A breakdown of participant demographics for each group in Experiment 2 is presented below. The demographics are similar across the groups, so we do not believe the results of Experiment 2 can be explained by participant demographics.

	Age (years)	**Male**	**Female**
Control	37.6	20	22
Non-diagnostic feature	39.0	12	24

Diagnostic feature	37.0	20	23

243

Self-Reported Ethnicity	**Control**	**Non-Diagnostic Feature**	**Diagnostic Feature**
African	4	8	4
Asian	3	4	3
Caucasian	31	22	29
Mediterranean	0	0	1
Hispanic	1	0	5
Southern Asian	1	0	0
Other	2	2	1

244