Feedback training for facial image comparison

David White[1], Richard Kemp[1], Rob Jenkins[2] & A. Mike Burton[3]

[1]School of Psychology, University of New South Wales
[2] School of Psychology, University of Glasgow
[3]School of Psychology, University of Aberdeen

Corresponding Author:
Dr. David White
School of Psychology
University of New South Wales
Kensington
Sydney
NSW 2052
david.white@unsw.edu.au
(+61) 2 9385 3254

Author Note

ABSTRACT

People are typically poor at matching the identity of unfamiliar faces from photographs. This observation has broad implications for face matching in operational settings (e.g. border control). Here, we report significant improvements in face matching ability following feedback training. In Experiment 1 we show cumulative improvement in performance on a standard test of face matching ability when participants were provided with trial-by-trial feedback. More importantly, Experiment 2 shows that training benefits can generalize to novel, widely-varying, unfamiliar face images for which no feedback is provided. The transfer effect specifically benefited participants who had performed poorly on an initial screening test. These findings are discussed in the context of existing literature on unfamiliar face matching and perceptual training. Given the reliability of the performance enhancement, and its generalization to diverse image sets, we suggest that feedback training may be useful for face matching in occupational settings.

INTRODUCTION

Many important security and forensic procedures rely on people's ability to verify the identity of unfamiliar individuals from photographs. However, research has consistently found that people perform poorly at this task. For example, Bruce et al. (1999) report 30% error-rates for a 1-in-10 identification decision, where participants must identify a target face from a simultaneously-presented array of face images. This poor performance was observed despite the fact that all photos were taken on the same day, under standardized lighting conditions and in standardized full-face pose. Similar results have been reported across a range of different stimulus sets, viewing conditions and experimental procedures (e.g. Burton, Wilson, Cowan, & Bruce, 1999; Megreya & Burton, 2006; Burton, White, & McNeill, 2010; Megreya, White, & Burton, 2011). Performance is just as poor when matching a photo to a live person, rather than matching two photos (Kemp, Towell, & Pike, 1997; Megreya & Burton, 2008).

From the outset, the extent of human error in simultaneous face matching tasks was seen as counterintuitive, and early reports of the phenomenon emphasized the surprising nature of the findings (Bruce et al. 1999; Kemp et al. 1997). Indeed the task difficulty also surprises study participants, who typically predict that they will perform well (see Bruce et al. 1999). It has recently been proposed that overconfidence in face matching ability might stem from the ease with which we recognize familiar faces from photographs, this facility leading to the mistaken belief that expert processing also extends to unfamiliar faces. The misconception is likely to be compounded by a lack of feedback concerning the identities of unfamiliar faces in everyday life (see Jenkins & Burton, 2011). Without such feedback, it is difficult for observers to assess their own level of performance.

In this paper we investigate whether feedback training can improve face matching performance. Previous studies have shown that when subjects perform face matching tasks without feedback, there is no cumulative improvement in performance (see O'Toole et al. 2007). This finding indicates that practice alone does not promote learning in this task. However, feedback has recently been found to support perceptual learning in *sequential* matching of faces (e.g. Hussain, Sekuler, & Bennett, 2009). This makes it all the more surprising that effects of training on *simultaneous* face matching performance have not yet been examined. Simultaneous face matching more closely resembles the important task of checking identity in many applied settings (e.g. comparing a photo-identity document to the document holder), as it involves perceptual comparison of two concurrently presented images, and does not place demands on recognition memory.

As is now well established, performance on face matching tasks is improved by prior exposure to the particular faces that appear in the test (Bruck, Cavanagh, & Ceci, 1991; Clutterbuck & Johnston, 2005; Megreya & Burton, 2006; Hussain et al., 2009). However, to be of practical use, any benefits of training must generalize to new identities that the viewer has not encountered before. Hussain et al (2009) did show that feedback training produced task-general improvement in a sequential face matching task, but the improvement was small compared to the stimulus-specific benefit. Furthermore, the highly standardized images used in that study do not represent the variability encountered in everyday experience (see Jenkins, White, Van Montfort, & Burton, 2011; Burton, 2013), and so learning may have been caused by familiarity with image-level parameters.

In the present study we test whether simultaneous face matching performance is improved by

providing participants with trial-by-trial performance feedback on a standardized test of face matching ability, the Glasgow Face Matching Test (GFMT, Burton et al., 2010). In this test, subjects are shown pairs of images that were collected under controlled conditions, using two different cameras. The conditions for matching in this test are close to optimal. All photographs were taken under the same lighting from the same viewpoint, and show the same neutral expression. Photos of the same person were taken only minutes apart. Despite these favorable conditions, performance on this test is rather poor overall and reveals large individual differences. Some observers attain near perfect accuracy, while others perform at near chance (Burton et al. 2010; see also Megreya & Burton, 2006).

We then go on to assess the effect of GFMT feedback training on performance in a completely separate face matching test, to investigate generalization of performance gains. Importantly, the latter test is composed of entirely different identities, and photos that vary widely in terms of lighting, viewpoint, expression, hairstyle, color balance, camera-to-subject distance, and many other image parameters. We have previously referred to such stimuli as *ambient images* to emphasize the fact that they sample natural variability in face images rather than attempting to control it away (Jenkins et al. 2011). Given that perceptual learning is often specific to low-level properties of training set (e.g. Sowden, Rose, & Davies, 2002), it is essential here to determine whether benefits of training propagate to the identity level (Bruce & Young, 1986; Young & Bruce, 2011; Burton, Jenkins, & Schweinberger, 2011). Note also that wide image variability better reflects the challenge of face matching in applied settings, as when comparing photo-identity documents against the document holders.

EXPERIMENT 1

In this experiment, we incorporated trial-by-trial feedback into a standard test of unfamiliar face matching, the GFMT (short version; Burton et al., 2010). In everyday face matching situations, observers seldom receive feedback because it is often difficult to establish whether the correct decision was made (see Jenkins & Burton, 2011). Here we predicted that providing feedback on each decision would improve performance on subsequent trials. To separate effects of feedback and mere practice, we compared performance of Feedback and No Feedback participant groups.

Method

Participants

Eighty-four students at the University of New South Wales took part in the study (52 female, 32 male; mean age 18.8 years, sd 1.3 years). Equal numbers of male and female participants were randomly assigned to Feedback and No Feedback groups.

Stimuli and procedure

The long version of the GFMT comprises 84 Same Identity and 84 Different Identity pairs. Here we used the short version of the test, which comprises the 40 most difficult pairs from the full test (i.e. 20 Same and 20 Different trials; see Burton et al. 2010 for details). Same Identity pairs show two images of the same person taken under similar conditions, but using different digital cameras. Different Identity pairs show images of two similar looking people, also taken under similar conditions using different cameras. For this experiment, we constrained the order in which trials were presented so that responses could be analysed into five trial bins of equal difficulty, each consisting of 4 Same and 4 Different trials. Normative data from Burton et al. (2010) was used to equate difficulty across the bins. Example pairs are shown in Figure 1a.

The GFMT pairs were presented on a 1280 x 1024 pixel computer monitor, with bin order counterbalanced across participants. On each trial, a face pair was presented centrally on the screen. Participants were instructed to respond 'same' or 'different' to each pair, using response buttons that appeared below the face display. They were then asked to rate their confidence in the decision on a scale of 1-100. Immediately following this response, participants in the Feedback condition were presented with a feedback message below the face images ("*You answered correctly/incorrectly: these images are of the same person/different people*"). The control group continued to view the faces with no accompanying feedback. After three seconds, the feedback window was replaced by an advance button allowing participants to progress to the next trial. Participants were asked to be as accurate as possible and were informed that there was no time limit for the task. On average, participants took 10 minutes to complete all 40 trials.

Results and Discussion

So that training effects could be translated into real world contexts, we used overall accuracy as our main dependent variable. However, it is also important for practical and theoretical reasons to determine whether Feedback training improves face matching performance independently of response bias (i.e. tendency to respond either 'same' or 'different'), and so we also analyze performance using signal detection measures for Sensitivity (d') and Response Criterion (C), as shown in Table 1.

*** *INSERT TABLE 1 HERE PLEASE* ***

Performance data for Experiment 1 is summarized in Table 1. Overall accuracy for the Feedback group improved over trial bins (from 82% to 92%), whereas accuracy for the No Feedback group remained comparatively flat (from 85% to 85%). Accuracy data were analysed using a 2 (No

Feedback/Feedback) X 5 (Trial Bins 1-5) mixed ANOVA. This analysis revealed a marginal main effect of Feedback, $F(1,82) = 3.25$, $p<0.05$, and a significant main effect of Trial Bin $F(1,82) = 3.90$, $p<0.05$. More importantly, we found a significant interaction between Trial Bin and Feedback, $F(4,328) = 2.92$, $p<0.05$. Planned comparisons showed that the effect of training was not reliable for early trial bins one, $t(82) = 1.13$, $p>0.05$, two, $t(82) = 1.21$, $p>0.05$, or three, $t(82) = 1.13$, $p>0.05$, but was reliable for later trial bins four, $t(82) = 2.35$, $p<0.05$, $d = 0.512$, and five, $t(82) = 2.68$, $p<0.05$, $d= 0.587$, confirming that the benefit of training accumulated as the task progressed. Supplementary performance data are also in accordance with this observation (see Table 1 for details).

We propose that feedback may help participants by drawing their attention to the fact that the task is harder than they expected, perhaps causing them to devote more attention to task-relevant information. If this is true, i.e. if the feedback is affecting participants' strategic approach to the task, then we might expect the improvement to generalize to a different stimulus set. On the other hand, if the effect of the feedback is to sensitize viewers to particularly diagnostic aspects of the stimulus set, then we might expect poor generalization. It is important to discriminate between these two possibilities, for both theoretical and practical reasons.

EXPERIMENT 2

In this experiment, we trained viewers using the same procedure, and the same faces as in Experiment 1. However, we now assessed face matching performance post-training using a completely separate Transfer test. This Transfer test comprised ambient images of both unfamiliar and familiar faces, none of which had been presented during the training phase. Many previous studies have shown that unfamiliar and familiar faces are processed very differently (Megreya & Burton, 2006; Burton et al. 2011; Jenkins & Burton, 2011). For this reason we

expected that benefits of training on *unfamiliar* faces would not transfer to a *familiar* face matching test.

We also had the opportunity in this study to ask how well training generalizes across participants. Given the well-established individual differences in face matching ability, we specifically recruited equal numbers of High Aptitude and Low Aptitude face matchers in order to assess the effect of training on groups of contrasting ability.

Method

Stimuli

*** INSERT FIGURE 1 HERE PLEASE ***

In addition to the GFMT (stimuli as described in Experiment 1), we constructed a Transfer Matching Test. This Transfer test was constructed from a set of 40 faces that were *Familiar* to our Australian participants (Australian public figures, such as Julia Gillard) plus a further 40 that were *Unfamiliar* to these participants (UK public figures, such as Alex Salmond). As these images were downloaded from the internet, they covered a much wider diversity of image characteristics than those making up the GFMT (see Figure 2). All of the images in the Transfer test showed a full color face in roughly frontal pose, with no occlusions, and an inter-ocular distance of at least 100 pixels. However, these were the only selection criteria. The images were unconstrained with respect to facial variables (e.g. emotional expression), environmental variables (e.g. lighting conditions), and image variables (e.g. camera characteristics) that affect the appearance of a face photograph. Using these new images, we created one *Same Identity* and one *Different Identity* pair for each face. *Same* pairs were made by pairing two randomly chosen photos of one individual. *Different* pairs were made by pairing randomly chosen photos of two

individuals who matched the same basic verbal description (e.g. middle aged male with black hair).

Screening and participants

1440 volunteer Australian students were screened for face matching aptitude using the ten most difficult items (5 Same & 5 Different) from the GFMT (Burton et al., 2010). Median accuracy across all participants was 8 [M = 8.0, SD = 1.6]. From the initial cohort of 1440, we recruited 112 volunteers who were at least one standard deviation above or below this mean score (i.e. above 9.6 or below 6.4, respectively). This resulted in a group of 56 High Aptitude matchers (who all scored 10) and a group of 56 Low Aptitude matchers [M = 5.2; SD = 1.1].

Design & Procedure

The 112 subjects (56 High Aptitude; 56 Low Aptitude) took part in the experiment between three and twelve weeks after the initial screening. All participants completed the GFMT (long-version; 168 items) as described in Study 1, with trials presented in a different random order for each participant. Participants were randomly allocated to Feedback and No Feedback groups, resulting in a 2x2 between subjects design (High Aptitude vs Low Aptitude matchers; Feedback vs No Feedback).

Following the GFMT, participants completed the Transfer face matching test. The Transfer test comprised 80 *Same* and 80 *Different* pairs that were presented in a different random order for each participant. No feedback was given during the Transfer test phase. To verify participants' familiarity with the familiar faces, participants then viewed printed names of the Australian and UK celebrities, and classified these as familiar or unfamiliar. This confirmed high familiarity with Familiar items, and low familiarity with Unfamiliar items.

Results

Training Phase

Accuracy scores are summarized in Figure 2. There was a main effect of feedback, $F(1,108) = 15.03$, $p<0.05$, a main effect of group, $F(1,108) = 24.35$, $p<0.05$, and a significant interaction, $F(1,108) = 5.25$, $p<0.05$. Analysis of Simple Main Effects revealed a reliable effect of feedback for the Low Aptitude matchers, $F(1,54) = 18.99$, $p<0.05$, $d = 0.923$, but not the High Aptitude matchers, $F(1,54) = 1.25$, $p>0.05$, $d = 0.468$, and also a significant difference between High and Low Aptitude matchers in the No Feedback condition, $F(1,54) = 26.10$, $p<0.05$, but not in the Feedback condition, $F(1,54)= 3.49$, $p>0.05$. Statistical tests for signal detection and response latency data are consistent with the accuracy data, and are presented in Table 2.

Transfer face matching test

Analyses of performance on the Transfer matching test were conducted separately for unfamiliar and familiar faces, because of unequal exclusion of data across these tests. For each participant, unfamiliar faces were defined as UK celebrities who were categorized as unfamiliar in the name familiarity task (38 faces on average), and familiar faces as Australian celebrities who were categorized as familiar (23 faces on average). All other trials were excluded from the analysis. Data from 13 participants who were familiar with less than 25% (10) of the Australian celebrities were excluded from the familiar analysis.

Accuracy scores for the Transfer test are summarized in Figure 2. In unfamiliar trials, the main effect of feedback was reliable, $F(1,108) = 4.74$, $p<0.05$, as was the main effect of group, $F(1,108) = 20.58$, $p<0.05$. In addition, there was a significant interaction between these factors, $F(1,108) = 5.93$, $p<0.05$. Analysis of Simple Main Effects revealed a significant effect of feedback for the Low Aptitude matchers, $F(1,54) = 10.63$, $p<0.05$, $d = 0.861$, but not the High Aptitude matchers, $F<1$, $d = 0.061$, and also a significant difference between High and Low Aptitude matchers in the No Feedback condition, $F(1,54) = 24.29$, $p<0.05$, but not in the Feedback condition, $F(1,54)=2.21$, $p>0.05$.

In matching trials with images of familiar faces, we found no effects of training on accuracy ($F<1$) however we did find a significant effect of group, $F(1,95) = 6.66$, $p<0.05$. The interaction between these factors was not reliable ($F<1$). Thus training improved unfamiliar face matching, but not familiar face matching.

General Discussion

Our results show that baseline measures of unfamiliar face matching performance can be improved by feedback training. In both experiments, we found that providing trial-by-trial feedback significantly improved performance on the GFMT, and in experiment 2 this improvement was observed in both High Aptitude and Low Aptitude groups. More importantly, feedback training on the GFMT also improved accuracy on a subsequent unfamiliar face matching test based on naturally varying images. This is the first time that training has been shown to benefit a simultaneous face matching task, and also the first demonstration that training benefits generalize to a task involving a realistic range of image variability. Moreover, the feedback phase that led to these performance benefits was very brief (40 trials in Experiment 1; 168 trials in Experiment 2) compared with typical perceptual learning procedures, which often

involve thousands of trials (e.g. Sowden et al. 2002; Hussain et al. 2009).

Despite the observed improvements in performance, it is important to note that none of the groups attained perfect accuracy. Indeed, two aspects of our results suggest that the effective ceiling for these tasks may be somewhat less than 100%. First, although training raised the accuracy of Low Aptitude matchers so it was equivalent to that of High Aptitude matchers, it did not raise the accuracy of the High Aptitude matchers above its initial level. One possible reason is that matching unfamiliar face photos is limited by the information available in the images, as well as by cognitive processes (Jenkins & Burton, 2011). Second, training had no effect on matching familiar faces, presumably because accuracy on this task was already as high as could be achieved for the minor celebrities that we presented. In any case, the absence of a training effect for familiar faces makes it unlikely that increased general motivation caused the observed improvement, as increased motivation would presumably improve performance on familiar and unfamiliar tasks alike.

Although our present design did not allow us to examine cognitive changes responsible for the training effect, it is already apparent that the training effect is sufficiently robust to survive large changes in low-level image characteristics. In addition, signal detection data confirmed that training did not improve accuracy by optimizing participants' response strategy, but by increasing sensitivity to task-relevant information (see Table 2). Our suggestion therefore is that feedback training caused participants to attend to features of the face that most reliably predict identity. One possibility is that participants learned to attend more to the internal features of the face, which are more stable over time (see Burton, Jenkins, Hancock, & White, 2005). Future research should examine the cognitive changes that support the task-general learning we observed here. In this work, it will be important to establish the extent to which learning operates

on attentional mechanisms, and whether post-perceptual decision processes are also modulated by feedback.

Our results also raise some important theoretical issues relating to perceptual learning research more generally. First, we found that the specificity of training effects interacted with individual differences in task aptitude. This raises the question of whether group level analyses in previous studies of perceptual learning may have masked important patterns in the data. Second, training generalized to a rather different set of test images. This observation is consistent with recent research demonstrating a task general effect of feedback (e.g. Hussain et al. 2012), and with the notion that perceptual learning is not confined to low-level visual processing (e.g. Green & Bavelier, 2003). Interestingly, recent evidence strongly suggests that the generality of perceptual learning can be improved by using diverse and non-repeating items at training (Gonzales & Madhavan, 2011; Hussain et al. 2012). If so, one might predict that the training effects in the current study would be greater if the order of training and test items were reversed, such that participants were trained on a highly variable stimulus set, and tested on less variable images.

In summary, our findings demonstrate that accuracy on an unfamiliar face matching task can be reliably improved using a brief feedback training procedure. This result has direct practical relevance because it offers the possibility for significant improvements in the accuracy of identity vetting processes in real-world settings. For example, it may be beneficial to introduce short feedback sessions into the routine of people who are required to match unfamiliar faces in their daily work (e.g. passport issuance officers). This is a realistic goal, and could be achieved using a similar procedure to that currently deployed in the workflow of airport security screening staff (Cutler & Paddock, 2009). Future research should test the longevity of the training effects established here, and work to optimize the training procedure to maximize improvements in task

performance.

REFERENCES

Bruce, V., Henderson, Z., Greenwood, K., Hancock, P. J. B., Burton, A. M., & Miller, P. (1999). Verification of face identities from images captured on video. *Journal of Experimental Psychology: Applied, 7*, 207-218.

Bruce, V., & Young, A. W. (1986). Understanding face recognition. *British Journal of Psychology*, *77*(3), 305–327.

Bruck, M., Cavanagh, P. & Ceci, S. J. (1991). Fortysomething: Recognizing faces at one's 25th reunion. *Memory and Cognition*, *19*(3), 221–228.

Burton, A. M. (2013). Why has research in face recognition progressed so slowly? The importance of variability. *The Quarterly Journal of Experimental Psychology*. Advance online publication. doi: 10.1080/17470218.2013.800125

Burton, A. M., Jenkins, R., Hancock, P. & White, D. (2005). Robust representations for face recognition: The power of averages. *Cognitive Psychology, 51*, 256-284.

Burton, A. M., Jenkins, R., & Schweinberger, S. R. (2011). Mental representations of familiar faces. *British Journal of Psychology, 102*(4), 943-958.

Burton, A. M., White, D., & McNeill, A. (2010). The Glasgow Face Matching Test. *Behavior Research Methods, 42*, 286–291.

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

Clutterbuck, R., & Johnston, R.A. (2005). Demonstrating how unfamiliar faces become familiar using a face matching task. *European Journal of Cognitive Psychology*, *17*(1), 97–116.

Cutler, V., & Paddock, S. (2009). Use of threat image projection (TIP) to enhance security performance. In *Security Technology, 2009. 43rd Annual 2009 International Carnahan*

Conference (pp. 46–51).

Green, C. S., & Bavelier, D. (2003). Action video game modifies visual selective attention. *Nature*, *423*(6939), 534–537.

Gonzalez, C., & Madhavan, P. (2011). Diversity during training enhances detection of novel stimuli. *Journal of Cognitive Psychology*, *23* (3), 342-350.

Hussain, Z., Sekuler, A. B., & Bennett, P J. (2009). Perceptual learning modifies inversion effects for faces and textures. *Vision Research, 49*, 2273-2284.

Hussain, Z., Bennett, P. J., & Sekuler, A. B. (2012). Versatile perceptual learning of textures after variable exposures. *Vision Research*, *61*, 89-94.

Jenkins, R., & Burton, A.M. (2011). Stable face representations. *Proceeding of the Royal Society B, 366*, 1671-1683.

Jenkins, R., White, D., Van Montfort, X., & Burton, A. M. (2011). Variability in photos of the same face. *Cognition*, *121*, 313-323.

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

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

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

Megreya, A., 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, 64,* 1473-1483.

O'Toole, A.J., Phillips, J.P., Jiang, F., Ayyad, J., Penard, N. & Abdi, H. (2007). Face Recognition Algorithms Surpass Humans Matching Faces Over Changes in Illumination. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29* (9), 1642-1646.

Sowden, P. T., Rose, D., & Davies, I. R. L. (2002). Perceptual learning of luminance contrast detection: Specific for spatial frequency and retinal location but not orientation. *Vision Research*, *42*(10), 1249–1258.

Young, A. W., & Bruce, V. (2011). Understanding person perception. *British Journal of Psychology*, *102*(4), 959–974.

Figure 1
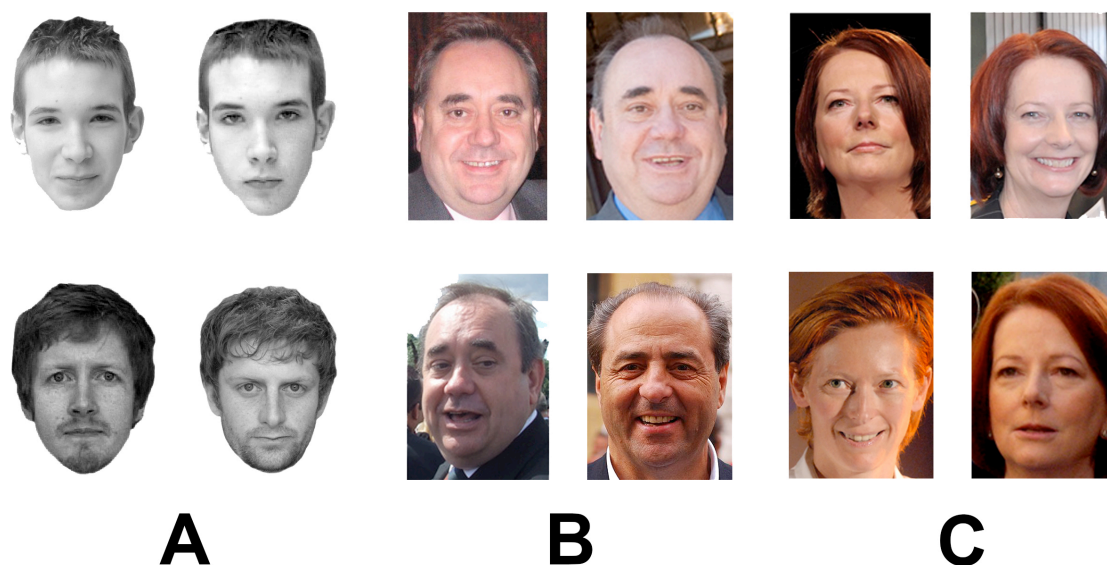
Figure 1: Example image pairs from the GFMT Feedback Training test (A), the Transfer unfamiliar test (B) and the Transfer familiar test (C) in Experiment 2. In each case, the task is to decide if the two images are of the same person or of different people. Same pairs are shown on the top row, and Different pairs are shown on the bottom row. For copyright reasons, we can not reproduce the precise images of the celebrities that were used in the experiment and so Transfer test images (B&C) are representative of the image variability in the experiment.
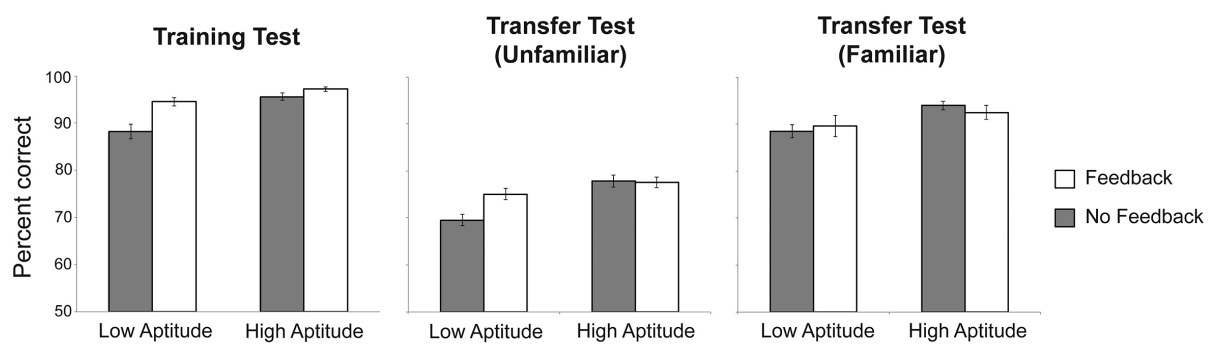
Figure 2

Figure 2: Accuracy data from Experiment 2. Statistical tests are reported in the text. Error bars denote standard error.

Table 1

Performance data for Experiment 1 (standard deviations in parenthesis).

Condition	Measure	Trial Bin					Overall
		1	2	3	4	5	
No Feedback	*Accuracy*	85.1 (*13.6*)	86.3 (*11.3*)	87.2 (*12.8*)	86.3 (*13.8*)	85.4 (*12.9*)	86.1 (*8.10*)
	Sensitivity (d')	2.99 (*1.43*)	3.08 (*1.15*)	3.24 (*1.27*)	3.17 (*1.33*)	3.00 (*1.35*)	3.10 (*1.30*)
	Response Criterion (C)	0.15 (*0.64*)	0.02 (*0.76*)	-0.16 (*0.71*)	-0.06 (*0.74*)	0.22 (*0.65*)	0.03 (*0.71*)
	Response Latency	7452 (*3757*)	7392 (*4113*)	6934 (*3558*)	6394 (*3788*)	6111 (*2785*)	6857 (*3109*)
	Confidence (correct)	79.0 (*13.0*)	77.2 (*13.7*)	77.4 (*13.9*)	75.1 (*15.7*)	75.1 (*14.0*)	76.8 (*12.5*)
	Confidence (incorrect)	71.6 (*21.7*)	71.2 (*19.4*)	69.0 (*19.7*)	67.9 (*21.5*)	66.0 (*21.1*)	67.0 (*17.7*)
Feedback	*Accuracy*	81.5 (*15.4*)	89.6 (*10.7*)	89.0 (*11.1*)	92.6 (*10.4*)	92.0 (*10.4*)	88.9 (*6.13*)
	Sensitivity (d')	2.56 (*1.52*)	3.35 (*1.30*)	3.34 (*1.08*)	3.77 (*1.15*)	3.68 (*1.19*)	3.34 (*1.34*)
	Response Criterion (C)	0.29 (*0.65*)	-0.07 (*0.54*)	-0.13 (*0.64*)	-0.13 (*0.57*)	-0.09 (*0.54*)	-0.03 (*0.60*)
	Response Latency	8141 (*5243*)	8030 (*4430*)	7011 (*3593*)	6728 (*3980*)	5841 (*2738*)	7150 (*3327*)
	Confidence (correct)	77.3 (*12.1*)	79.4 (*10.6*)	80.7 (*11.9*)	81.3 (*13.6*)	81.7 (*12.6*)	80.1 (*10.0*)

Table 1. Analysis of signal detection data show a non-reliable Main Effect of Feedback for sensitivity, $F(1,82) = 2.10$, $p > 0.05$, a significant Main Effect of Trial Bin, $F(1,82) = 4.33$, $p < 0.05$, and a significant interaction between Feedback and Trial Bin, $F(4,328) = 3.04$, $p < 0.05$. For criterion scores, positive values indicate a tendency to respond 'different' whilst negative scores indicate a bias toward 'same' responses. Response criterion data show a reliable Main Effect of Trial Bin, $F(1,82) = 6.29$, $p < 0.05$, but no Main Effect of Feedback $(F<1)$ and no interaction, $F(4,328) = 2.00$, $p > 0.05$. This suggests that the improvement in overall accuracy observed in the Feedback group (see text) was not caused via changes in response bias. In addition, Feedback led participants to be more confident in correct decisions (details of response latency and confidence ANOVAs are available from the authors on request).

Table 2

Performance data for Experiment 2 (standard deviations in parenthesis).

		Low Aptitude			High Aptitude		
		Training	Transfer Test		Training	Transfer Test	
		GFMT	*Unfamiliar*	*Familiar*	*GFMT*	*Unfamiliar*	*Familiar*
No Feedback	*Sensitivity (d')*	2.84 (*0.86*)	1.16 (*0.41*)	2.83 (*0.94*)	3.73 (*0.68*)	1.70 (*0.50*)	3.55 (*0.78*)
	Response Criterion (C)	-0.23 (*0.46*)	-0.01 (*0.52*)	0.10 (*0.57*)	-0.11 (*0.28*)	0.04 (*0.42*)	0.13 (*0.36*)
	Response Latency	3748 (*1480*)	2274 (*835*)	1991 (*683*)	4393 (*2057*)	2169 (*714*)	1775 (*437*)
Feedback	*Sensitivity (d')*	3.53 (*0.79*)	1.45 (*0.39*)	3.06 (*1.14*)	4.07 (*0.50*)	1.63 (*0.59*)	3.41 (*1.06*)
	Response Criterion (C)	-0.11 (*0.24*)	-0.05 (*0.38*)	-0.05 (*0.45*)	0.00 (*0.29*)	0.11 (*0.55*)	0.03 (*0.45*)
	Response Latency	3620 (*1008*)	2564 (*835*)	1812 (*381*)	3664 (*1378*)	2440 (*1084*)	2088 (*1267*)

Table 2. Analysis of signal detection Training test data are available on request from the authors. For Transfer test sensitivity scores in unfamiliar trials, the main effect of feedback was non-significant, $F(1,108) = 1.82$, $p > 0.05$, however there was a reliable main effect of Group, $F(1,108) = 17.6$, $p < 0.05$, and a significant interaction between these factors, $F(1,108) = 4.33$, $p < 0.05$. Analysis of Simple Main Effects revealed a significant

effect of Feedback for the Low Aptitude matchers, $F(1,54) = 5.884$, $p < 0.05$, $d = 0.725$, but not the High Aptitude matchers, $F<1$, $d = 0.128$. For criterion scores, positive values indicate a tendency to respond 'different' whilst negative scores indicate a bias toward 'same' responses. The main effect of Feedback on response criterion was non-significant ($F<1$), as was the main effect of Group, $F(1,108) = 1.98$, $p < 0.05$, and the interaction between these factors ($F<1$). This result confirms that improvement in matching accuracy was not due to changes in response strategy (i.e. reducing response criterion to zero, which is optimal given the equal proportion of same and different trials). Signal detection analysis for Transfer test trials with images of familiar faces support the conclusion that training effect was confined to unfamiliar face matching (details available from the authors on request). There were no significant effects in response latency data.