Psychon Bull Rev
DOI 10.3758/s13423-013-0475-3

BRIEF REPORT

Feedback training for facial image comparison

David White · Richard I. Kemp · Rob Jenkins · A. Mike Burton

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 important, 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.

Keywords Face recognition · Unfamiliar face matching · Identity verification · Perceptual learning

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) reported 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, White, & McNeill, 2010; Burton, Wilson, Cowan, & Bruce, 1999; Megreya & Burton, 2006; 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 article, 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

D. White (✉) · R. I. Kemp
School of Psychology, University of New South Wales, Kensington, Sydney, NSW 2052, Australia
e-mail: david.white@unsw.edu.au

R. Jenkins
Department of Psychology, University of York, York, UK

A. M. Burton
School of Psychology, University of Aberdeen, Aberdeen, UK

Published online: 09 July 2013

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 with the document holder), since 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; Hussain et al., 2009; Megreya & Burton, 2006). 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, as compared with the stimulus-specific benefit. Furthermore, the highly standardized images used in that study do not represent the variability encountered in everyday experience (see Burton, 2013; Jenkins, White, Van Montfort, & Burton, 2011), 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; Burton, Jenkins, & Schweinberger, 2011; Young & Bruce, 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 analyzed into five trial bins of equal difficulty, each consisting of four *same* and four *different* trials. Normative data from Burton et al. (2010) were used to equate difficulty across the bins. Example pairs are shown in Fig. 1a.

The GFMT pairs were presented on a 1,280 × 1,024 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

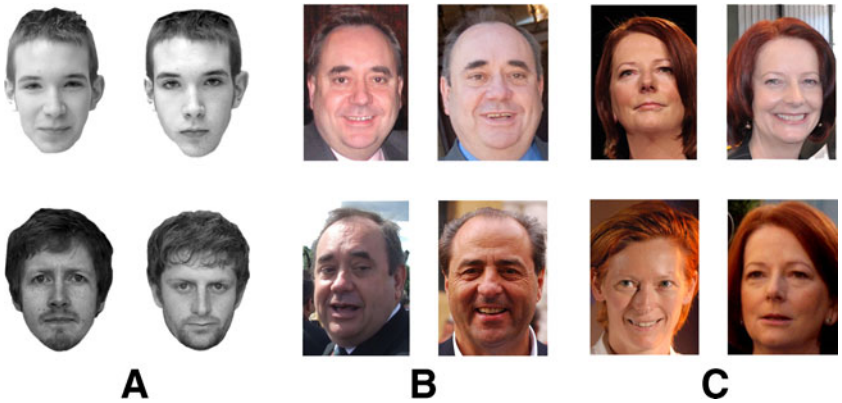

Fig. 1 Example image pairs from the GFMT feedback training test (**a**), the transfer unfamiliar test (**b**) and the transfer familiar test (**c**) in Experiment 2. In each case, the task is to decide whether the two images are of the same person or of different people. *Same* pairs are shown in the top row, and *different* pairs are shown in the bottom row. For copyright reasons, we cannot reproduce the precise images of the celebrities that were used in the experiment, and so transfer test images (panels b and c) are representative of the image variability in the experiment

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 3 s, 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 min 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.

Table 1 Performance data for Experiment 1 (with standard deviations in parentheses)

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	7,452 (*3,757*)	7,392 (*4,113*)	6,934 (*3,558*)	6,394 (*3,788*)	6,111 (*2,785*)	6,857 (*3,109*)
	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	8,141 (*5,243*)	8,030 (*4,430*)	7,011 (*3,593*)	6,728 (*3,980*)	5,841 (*2,738*)	7,150 (*3,327*)
	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*)
	Confidence (incorrect)	69.8 (*16.8*)	72.0 (*16.4*)	71.1 (*13.9*)	73.5 (*24.7*)	71.3 (*18.9*)	69.3 (*13.6*)

Note. Analysis of signal detection data shows a nonreliable main effect of feedback for sensitivity, $F(1, 82) = 2.10$, $p > .05$, a significant main effect of trial bin, $F(1, 82) = 4.33$, $p < .05$, and a significant interaction between feedback and trial bin, $F(4, 328) = 3.04$, $p < .05$. For criterion scores, positive values indicate a tendency to respond *different*, while negative scores indicate a bias toward *same* responses. Response criterion data show a reliable main effect of trial bin, $F(1, 82) = 6.29$, $p < .05$, but no main effect of feedback ($F < 1$) and no interaction, $F(4, 328) = 2.00$, $p > .05$. This suggests that the improvement in overall accuracy observed in the feedback group (see the 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)

Performance data for Experiment 1 are 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 analyzed using a 2 (no feedback/feedback) × 5 (trial bins 1–5) mixed ANOVA. This analysis revealed a marginal main effect of feedback, $F(1, 82) = 3.25$, $p < .05$, and a significant main effect of trial bin, $F(1, 82) = 3.90$, $p < .05$. More important, we found a significant interaction between trial bin and feedback, $F(4, 328) = 2.92$, $p < .05$. Planned comparisons showed that the effect of training was not reliable for early trial bins one, $t(82) = 1.13$, $p > .05$, two, $t(82) = 1.21$, $p > .05$, or three, $t(82) = 1.13$, $p > .05$, but was reliable for later trial bins four, $t(82) = 2.35$, $p < .05$, $d = 0.512$, and five, $t(82) = 2.68$, $p < .05$, $d = 0.587$, 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—that is, 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, 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 posttraining 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 (Burton et al., 2011; Jenkins & Burton, 2011; Megreya & Burton, 2006). For this reason, we expected that benefits of trainingon *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

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 (U.K. public figures, such as Alex Salmond). Since these images were downloaded from the Internet, they covered a much wider diversity of image characteristics than those making up the GFMT (see Fig. 2). All of the images in the transfer test showed a full-color face in roughly frontal pose, with no occlusions and an interocular 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

Using the 10 most difficult items (5 same and 5 different) from the GFMT (Burton et al., 2010), 1,440 volunteer Australian students were screened for face matching aptitude. Median accuracy across all participants was 8 ($M = 8.0$, $SD = 1.6$). From the initial cohort of 1,440, 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 and procedure

The 112 participants (56 high aptitude, 56 low aptitude) took part in the experiment between 3 and 12 weeks after the initial screening. All participants completed the GFMT (long version; 168 items) as described in Experiment 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 2 × 2 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

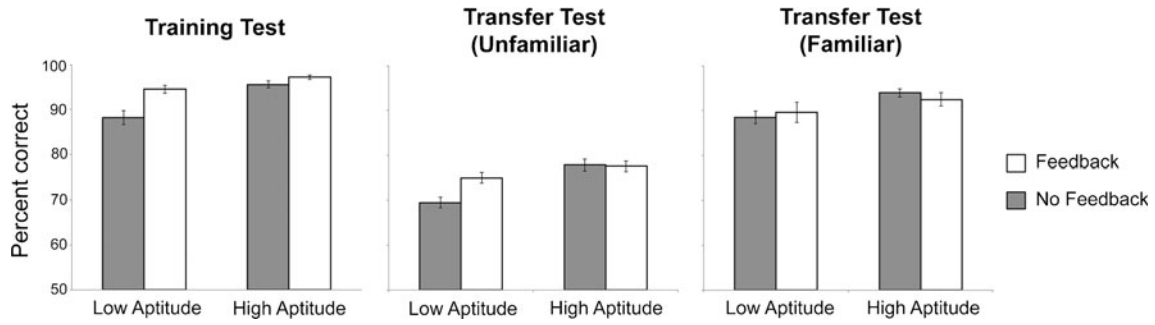

Fig. 2 Accuracy data from Experiment 2. Statistical tests are reported in the text. Error bars denote standard errors

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 U.K. 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 Fig. 2. There was a main effect of feedback, $F(1, 108) = 15.0$, $p < .05$, a main effect of group, $F(1, 108) = 24.4$, $p < .05$, and a significant interaction, $F(1, 108) = 5.25$, $p < .05$. Analysis of simple main effects revealed a reliable effect of feedback for the low-aptitude matchers, $F(1, 54) = 19.0$, $p < .05$, $d = 0.923$, but not the high-aptitude matchers, $F(1, 54) = 1.25$, $p > .05$, $d = 0.468$, and also a significant difference between high- and low-aptitude matchers in the no-feedback condition, $F(1, 54) = 26.10$, $p < .05$, but not in the feedback condition, $F(1, 54) = 3.49$, $p > .05$. 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 U.K. 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 fewer than 25 % (10) of the Australian celebrities were excluded from the familiar analysis.

Accuracy scores for the transfer test are summarized in Fig. 2. On unfamiliar trials, the main effect of feedback was reliable, $F(1, 108) = 4.74$, $p < .05$, as was the main effect of group, $F(1, 108) = 20.58$, $p < .05$. In addition, there was a significant interaction between these factors, $F(1, 108) = 5.93$, $p < .05$. Analysis of simple main effects revealed a significant effect of feedback for the low-aptitude matchers, $F(1, 54) = 10.63$, $p < .05$, $d = 0.861$, 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, $F(1, 54) = 24.29$, $p < .05$, but not in the feedback condition, $F(1, 54) = 2.21$, $p > .05$.

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, $F(1, 95) = 6.66$, $p < .05$. 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. More important, 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), as compared with typical perceptual learning procedures, which often involve thousands of trials (e.g., Hussain et al., 2009; Sowden et al., 2002).

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-

Table 2 Performance data for Experiment 2 (with standard deviations in parenthesss)

		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	3,748 (*1,480*)	2,274 (*835*)	1,991 (*683*)	4,393 (*2,057*)	2,169 (*714*)	1,775 (*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	3,620 (*1,008*)	2,564 (*835*)	1,812 (*381*)	3,664 (*1,378*)	2,440 (*1,084*)	2,088 (*1,267*)

Note. Analysis of signal detection training test data is available on request from the authors. For transfer test sensitivity scores in unfamiliar trials, the main effect of feedback was nonsignificant, $F(1, 108) = 1.82$, $p > .05$; however, there was a reliable main effect of group, $F(1, 108) = 17.6$, $p < .05$, and a significant interaction between these factors, $F(1, 108) = 4.33$, $p < .05$. Analysis of simple main effects revealed a significant effect of feedback for the low-aptitude matchers, $F(1, 54) = 5.884$, $p < .05$, $d = 0.725$, but not the high-aptitude matchers, $F < 1$, $d = 0.128$. For criterion scores, positive values indicate a tendency to respond *different*, while negative scores indicate a bias toward *same* responses. The main effect of feedback on response criterion was nonsignificant ($F < 1$), as was the main effect of group, $F(1, 108) = 1.98$, $p < .05$, 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 the training effect was confined to unfamiliar face matching (details available from the authors on request). There were no significant effects in response latency data

aptitude matchers so that 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, since 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 postperceptual 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, Bennett, & Sekuler, 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 nonrepeating items at training (Gonzalez & Madhavan, 2011; Hussain et al., 2012). If so, one might predict that the training effects in the present 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 procedure similar 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.

Author Note This research was supported by an ARC grant to Kemp (LP110100448), a bilaterally funded grant to Kemp (ARC: LX0083067), Burton, and Jenkins (ESRC: RES-000-22-2519), and an ESRC Professorial Fellowship to Burton (ES/J022950/1). We thank Graham Nisbett (B), Filippo Caranti (B), Troy Constable (B), Ian Short (C) and the Edinburgh International Film Festival (C) for making the photographs in Fig. 1 available for publication under Creative Commons licenses (CC BY 2.0).

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. *Proceedings 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.