The effect of feedback on face matching accuracy

Hamood M. Alenezi[1,2] & Markus Bindemann[1]

[1] School of Psychology, University of Kent, UK

[2] Department of Education and Psychology, Northern Borders University, KSA

Correspondence to:

Hamood Alenezi, School of Psychology, University of Kent, CT2 7NP, UK.

Email: ha229@kent.ac.uk

Tel: +44(0)1227 823087

Fax: +44(0)1227 827030

Word count (excluding abstract, references and figure captions): 13748

Abstract

In face matching, observers have to decide if two photographs depict the same person or different people. This is a remarkably difficult task so the current study investigated whether it can be improved when observers receive feedback for their performance. In five experiments, observers' initial matching performance was recorded before feedback for their accuracy was administered across three blocks. Improvements were then assessed with faces that had been seen previously with or without feedback and with completely new, previously unseen faces. In all experiments, feedback failed to improve face-matching accuracy. However, trial-by-trial feedback helped to maintain accuracy at baseline level after feedback was withdrawn again, even with new faces (Experiments 1 to 3). By contrast, when no feedback was given throughout the experiment (Experiments 1 to 3) or when outcome feedback was administered at the end of blocks (Experiments 4 and 5), a continuous decline in matching accuracy was found, whereby observers found it increasingly difficult to tell different facial identities apart. A sixth experiment showed that this decline in accuracy continues throughout when the matching task is prolonged substantially. Together, these findings indicate that observers find it increasingly difficult to differentiate faces in matching tasks over time, but trial-by-trial feedback can help to maintain accuracy. The theoretical and practical implications of these findings are discussed.

Introduction

Face matching refers to the process by which an observer has to decide if two simultaneous presentations of a face, such as a pair of photographs, belong to the same person or different people. This task can be performed with high accuracy when observers have to match two different images of a familiar face, such as a colleague or teacher that they know well (Burton, Wilson, Cowan, & Bruce, 1999; Harmon, 1973), or when two identical images of the same face are compared (Jenkins & Burton, 2011). By contrast, this task can be surprisingly difficult when different pictures of unfamiliar faces, which are unknown to an observer prior to the task, are shown (see, e.g., Bindemann, Avetisyan, & Blackwell, 2010; Henderson, Bruce, & Burton, 2001; Megreya & Burton, 2006).

Many of the factors that give rise to this difficulty in unfamiliar face matching are now well understood. Variation in face photographs in, for example, lighting direction (e.g., Hill & Bruce, 1996; Johnston, Hill, & Carman, 1992; Longmore, Liu, & Young, 2008), viewpoint (e.g., Bruce, 1982; Bruce et al., 1999; Hill, Schyns, & Akamatsu, 1997; Longmore et al., 2008), facial expression (e.g., Bruce, 1982; Bruce, Valentine, & Baddeley, 1987), and age (Jenkins & Burton, 2011) can induce many changes in the appearance of a face (for some striking examples, see Burton, Jenkins, & Schweinberger, 2011; Jenkins, White, Van Montfort, & Burton, 2011). These changes can produce many differences between two to-be-compared faces and, as a result, matching accuracy declines. Such changes can incur over 40% errors under challenging task demands (see, e.g., Bindemann & Sandford, 2011; Henderson et al., 2001), but performance also remains error-prone under seemingly optimized viewing conditions. For example, observers continue to average 10-30% errors when they have to compare two high-quality photographs that were only taken a few moments apart and show faces in the same lighting, expression and view

(see, e.g., Burton, White, & McNeill, 2010, Megreya, Bindemann, & Havard, 2011; Özbek & Bindemann, 2011).

While most studies in this field have assessed face matching with pairs of photographs under controlled laboratory conditions, it is notable that these difficulties persist when photographs have to be matched to high-quality video footage of a face (Bruce et al., 1999), or to a live person (Kemp, Towell, & Pike, 1997; Megreya & Burton, 2008), and also when observers try to match a live person to surveillance video footage (Davis & Valentine, 2009). For this reason, face matching is seen as a problem of considerable applied importance that is relevant to key forensic identification tasks, such as photo-identity verification at airports and national borders and criminal identification from CCTV (see, e.g., Costigan, 2007; Jenkins & Burton, 2008, 2011).

Considering the documented difficulty of face matching in Psychology and its applied relevance, a question that arises is why observers continue to make errors in this task. One possible explanation is that we rarely receive feedback on our accuracy in the identification of unfamiliar people in real life (Burton & Jenkins, 2011; Jenkins et al., 2011). Consider, for example, a first encounter with a person that we have not met before. If we fail to recognize this individual at a subsequent sighting, then we might simply assume that it is a different person. Similarly, if one fails to notice the correspondence in identity between two face images in a matching task, then there is also no reason to challenge this perception. Indeed, it is difficult to see how such a challenge could be possible without evidence to the contrary, such as some form of explicit feedback on identification errors. This, then, raises the question of whether feedback can be used to improve face-matching accuracy.

There is certainly good reason to expect that feedback will improve performance in face-matching tasks. It has recently been shown, for example, that observers can vary in how they respond to the same face pairs on different days (Bindemann, Avetisyan, & Rakow, 2012). However, it was also found that observers would rarely misclassify the same face pairs repeatedly. This shows that many errors in face matching do not reflect a data-limited problem, whereby some face pairs simply contain insufficient information for observers to solve the task (see, e.g., Norman & Bobrow, 1975). Rather, this finding indicates that many errors in face matching arise because observers occasionally use the available identity information incorrectly. This is an important finding because it suggests that this task is in principle solvable. Therefore, observers might also be able to improve their face-matching accuracy if they are provided with feedback on their performance.

So far, there have only been limited attempts to examine the role of feedback in unfamiliar face identification. A few studies have shown that feedback can improve unfamiliar face recognition in immediate-memory paradigms, in which observers have to decide if two successive face images depict the same person or two different people (Meinhardt-Injac, Persike, & Meinhardt, 2010, 2011). In these studies, task difficulty was increased by asking observers to focus on a few specific target features, such as the eyes, nose and mouth, while other features, such as hairstyle and face outline, were exchanged between successive presentations. The influence of feedback on face recognition has also been examined with another immediate-memory paradigm, in which observers had to select a target, which was degraded with visual noise, from a subsequent array of ten faces (Hussain, Sekuler, & Bennett, 2009). In all of these cases, trial-by-trial feedback for recognition accuracy, provided by a brief tone, helped to improve observers' performance substantially.

However, it is notable that these paradigms assessed recognition accuracy across variants of the same original face *image* (e.g., by substituting select features or by adding visual noise). Such pictorial recognition can be trivial for human observers when non-degraded facial images are used (see, e.g., Brown, Deffenbacher, & Sturgill, 1977; Hochberg & Galper, 1967; Nickerson, 1965; Yin, 1969). Indeed, such pictorial recognition can be achieved also by prosopagnosic observers, who are clinically impaired at face recognition (see Duchaine & Nakayama, 2004; Duchaine & Weidenfeld, 2003), and can be performed with great accuracy with non-face stimuli, such as words, objects and scenes (e.g., Shepard, 1967; Standing, Conezio, & Haber, 1970; Standing, 1973). This suggests that previous experiments that manipulated feedback in unfamiliar face identification may have relied on the encoding of specific pictorial properties of a face image for person identification, rather than typical face processing mechanisms (see, e.g., Bruce, 1982; Jenkins & Burton, 2011; Longmore et al., 2008). Such pictorial recognition tasks are therefore largely irrelevant to the applied issue of face matching, which requires observers to compare two *different* images of the same face (see, e.g., Jenkins & Burton, 2011; Megreya & Burton, 2008).

In addition, there is also some evidence from the study of eyewitness identification to suggest that feedback might improve face-matching performance. In this domain, it has long been known that positive feedback for eyewitnesses who have made a mistaken person identification can produce a range of distortions in their recollection of an event (see, e.g., Wells & Bradfield, 1998, 1999). For example, such false feedback can inflate an eyewitness's confidence in a previously made identification (Bradfield, Wells, & Olson, 2002), or their judgement of the attention that they had originally paid to a perpetrator's face (Douglass & Steblay, 2006). Such findings are useful generally for showing that person recognition is

susceptible to feedback, but do not actually demonstrate improvements in identification performance. However, more recently such an improvement with feedback has also been found in an eyewitness identification paradigm (Palmer, Brewer, & Weber, 2010). In this study, observers were shown a video of a staged crime, which was followed by two photographic identity lineups. Feedback on identification accuracy for the first lineup influenced responses to the second lineup, whereby confirming feedback following a correct initial response improved subsequent identification performance. This finding therefore suggests that feedback can enhance observers' face identification performance in an eyewitness recognition paradigm.

In this study, we sought to examine whether feedback can also improve accuracy during face matching. For this purpose, we measured observers' initial ability in this task. We then divided observers into two groups, which were either given feedback during face matching, in the *feedback* condition, or not, in the *no feedback* condition. By comparing these groups, we sought to determine if feedback can improve observers' identification accuracy. If such improvements are found, then it is particularly important to establish whether these effects persist when feedback is withdrawn again. Moreover, it is necessary to determine whether such improvements are confined to faces that were seen with feedback during training, or whether they *generalize* to previously unseen stimuli. For these reasons, we compared face-matching accuracy for face pairs that had previously been seen with and without feedback, and for completely new, previously unseen faces. This was examined over a series of six experiments.

Experiment 1

In the first experiment reported here, we examined whether face-matching accuracy could be improved by providing observers with immediate feedback on their performance during this

task. For this purpose, observers were shown pairs of faces comprising photographs of the same person or two different people and match/mismatch decisions to these facial identities were required (as in, e.g., Bindemann et al., 2010; Burton et al., 2010; Megreya et al., 2011). Different photographs of the same person were provided on identity-match trials to eliminate simple picture-matching strategies (see, e.g., Jenkins & Burton, 2011). To examine the effect of feedback on face matching accuracy in detail, observers were given seven blocks of this matching task. The first block was administered without feedback to serve as a baseline measure of face matching performance. The second, third and fourth block then provided observers with feedback for the accuracy of their responses on a trial-by-trial basis, to provoke improvements in performance.

In the remaining three blocks, these improvements were measured. For this purpose, feedback was withdrawn again, to assess if observers had learned to improve their face matching performance generally. In block 6, observers were given the same stimuli as in block 4, to assess if any benefits of feedback would persist for stimuli for which observers had previously received feedback. Similarly, in block 5 observers were given the same stimuli as in the first experimental (baseline) block, to determine if any benefits of receiving feedback would also hold for faces that had been seen previously *without* feedback. Finally, in block 7, observers were shown completely new faces, that had not been seen in any of the preceding blocks.

By comparing performance across these blocks, we sought to establish if any learning from trial-by-trial feedback (in blocks 2-4) persists when this is subsequently removed (blocks 5-7). Moreover, we wished to examine if any learning is tied to the specific face stimuli for which feedback had previously been received (block 6), or if improvements in performance generalize to face pairs that have previously been seen without feedback (block 5) and to previously unseen

faces (block 7). Finally, to establish the influence of face matching in this design fully, we compared two groups of observers, which were either administered the experimental procedure according to the above description (the *feedback* condition) or were given the same sequence of seven blocks but without any feedback (the *no feedback* condition).

Method

Participants

Fifty students (37 female) from the University of Kent, with a mean age of 19 years (SD = 3.3), volunteered to participate in this experiment for a small fee. All reported normal or corrected-to-normal vision.

Stimuli

The stimuli consisted of 100 match pairs (50 male, 50 female) and 100 mismatch pairs (also 50 male, 50 female) from the Glasgow University Face Database (GUFD) (see Burton et al., 2010). These face pairs were constructed so that faces were shown in grayscale on a plain white background. Each face was depicted in full-face view and with a neutral expression, and measured maximally 350 pixels in width at a screen resolution of 72 ppi. The faces in a pair were positioned in such a way that the horizontal distance between the centre of each face measured 500 pixels. In addition, in each match and mismatch face pair, one face image was taken with a high-quality digital camera, while the other image was a frame of a person's face taken from high-quality video footage (for details, see Burton et al., 2010). For each person in the GUFD database, these two images were taken on the same day but a few minutes apart. The resulting match pairs therefore provide similar but not identical face images of a person, to ensure that the

task cannot be performed using simple picture matching processes (see, e.g., Hancock, Bruce, & Burton, 2000; Megreya & Burton, 2006). An example of these stimuli is depicted in Figure 1.

The 200 face-matching pairs were then divided into five sets of 40 pairs, each of which consisted of 20 match and 20 mismatch pairs. The GUFD provides perceived-similarity ratings on a scale of 0-1 for all identity mismatches (see Burton et al., 2010), which were used to ensure that the face pairs were matched along this dimension in each set. The average similarity ratings for the five sets ranged from 0.40/1 to 0.42/1 (SDs ranged from 0.05 to 0.09). A one-factor ANOVA showed that the ratings were similar for all five sets, $F(4,95) = 0.17$, $p = 0.96$.

In the experiment, these face sets were then rotated around blocks, across observers, in the following manner. In the feedback condition, each observer was given seven blocks of 40 trials (20 match, 20 mismatch). These corresponded to a baseline (BL) block, in which no feedback was administered. This was followed by three feedback blocks (FB1, FB2, and FB3), which provided immediate feedback for response accuracy on a trial-by-trial basis. To determine whether this feedback leads to more general improvements in task performance, observers' accuracy was then tested in three blocks in which no feedback was given. The first of these consisted of a repetition of the stimuli from the first experimental block (BLT – baseline test), to explore if the administration of feedback could improve the accuracy of responses to faces that had previously been seen without such knowledge. The second test block, called the feedback test (FBT), was a repetition of the third feedback block (FB3). This was included to determine if any benefits from feedback would continue to hold for faces that had previously been presented with this information. Finally, the third test block examined face matching with a set of new faces that had not been presented in the experiment before (NFT – the new face test), to determine if any performance gains from feedback would extend to completely new faces. By

contrasting performance in these three test blocks, we therefore hoped to determine whether any improvements in accuracy from feedback would only hold for faces that previously been presented with this information (in the FBT block), or would also generalize to faces that had already been seen without feedback (in the BLT block) and to completely new faces (in the NFT block). Finally, in the *no feedback* group, we administered exactly the same design, except that no feedback for performance was provided throughout the experiment. The experimental design is illustrated in Figure 2.

Procedure

Participants were randomly assigned to the *feedback* or *no feedback* condition. They were tested individually on a standard desktop PC, which was equipped with E-Prime software to record their responses. Each trial began with the presentation of a fixation cross, which was displayed for 1 second. A face pair was then presented onscreen until a response was registered. Participants were instructed to decide whether an identity match or mismatch was shown in each trial, by using their left and right index fingers to press two corresponding keys on the computer keyboard. Accuracy was emphasized and responses were self-spaced.

Each participant was given 280 experimental trials, comprising seven blocks of 20 match and 20 mismatch trials. Within blocks, trial order was randomised. The match and mismatch stimuli were also rotated around these blocks across observers so that, over the course of the experiment, each of the face pairs appeared in each block an equal number of times. In addition, the observers in the *feedback* group received trial-by-trial feedback in the second, third, and fourth block of the experimental procedure (see Figure 2). This feedback was provided in the form of printed words, which were presented in the centre of the screen for 1.5 seconds

immediately after a response had been made. At this stage, the faces were removed from view and were replaced with the message "Good job!" for correct responses or with "Incorrect response" when a matching error was made. In the *no feedback* group, on the other hand, exactly the same procedure was applied throughout, except that no feedback was given on task performance at any point.

Results

Accuracy

In a first step of the analysis, the percentage accuracy was calculated for match and mismatch trials for each of the experimental blocks in the *feedback* and the *no feedback* group. This data is illustrated in Figure 3. For the *no feedback* group, this data shows that observers achieved about 90% accuracy in the baseline block (BL) and performance was similar for match and mismatch trials at this point. Thereafter, accuracy gradually declines for mismatch trials throughout the remaining six blocks. This continuous decrease in accuracy appears to be accompanied by an increase in match responses. A very different response pattern is visible in the *feedback* group. Here, performance in the baseline block appeared to be comparable to the *no feedback* condition. In contrast to the *no feedback* group, however, observers maintained this initial level of accuracy throughout the experiment for both match and mismatch trials.

To analyse these observations more formally, a 2 (feedback: *feedback* versus *no feedback* group) x 2 (trial type: *match* versus *mismatch* trials) x 4 (block: *BL*, *BLT*, *FBT*, *NFT*) mixed-factor ANOVA was conducted. Note that, to simplify the analysis, the data from the three feedback blocks (FB1, FB2, FB3) is therefore omitted from this ANOVA. This makes good sense as it allows us to measure the effect of feedback by comparing performance in the baseline

block (BL) with performance in the three test blocks, for which feedback is withdrawn again (BLT, FBT, and NFT). Moreover, it allows for a direct comparison of the *feedback* and *no feedback* group, as the procedure for the BL, BLT, FBT, and NFT blocks was kept exactly identical across these two groups.

The ANOVA of the accuracy data revealed a three-way interaction between all factors, $F(3,144) = 4.32$, $p < 0.01$. To interpret this interaction, the data was analysed separately for the *no feedback* and *feedback* group with two 2 x 4 ANOVAs for the factors trial type and block. For the *no feedback* condition, this ANOVA showed no main effect of trial type, $F(1,24) = 2.96$, $p = 0.10$, or block, $F(3,72) = 1.93$, $p = 0.13$, but an interaction between both factors, $F(3,72) = 8.42$, $p = 0.001$. Analysis of simple main effects found no effect of trial type for each of the blocks, all $Fs(1,24) \leq 1.92$, all $ps \geq 0.18$. In addition, no simple main effect of block was found for match trials, $F(3,72) = 2.44$, $p = 0.07$. However, a simple main effect was found for mismatch trials, $F(3,72) = 7.74$, $p = 0.001$. Tukey HSD test revealed that this effect arises from lower mismatch accuracy in the BLT, FBT, and NFT block compared to the baseline block (BL), all $qs \geq 4.40$, all $ps \leq 0.05$. These results therefore indicate that accuracy on mismatch but not on match trials declined in the *no feedback* condition during the experiment.

An analogous analysis for the *feedback* condition found no main effect of trial type, $F(1,24) = 0.10$, $p = 0.75$, or block, $F(3,72) = 2.25$, $p = 0.09$, and no interaction between both factors, $F(3,72) = 0.41$, $p = 0.75$. In contrast to the *no feedback* condition, observers' performance on mismatch trials therefore did not decline during the face-matching task. This indicates that trial-by-trial feedback did not improve performance here, but prevented observers specifically from making increasingly more mismatch errors during the course of the experiment.

d-prime and criterion

The percentage accuracy data hints at a bias to classify faces as identity matches, which develops over the course of the experiment when no feedback on performance is given. To assess this possibility more directly, we transformed the data into signal detection measures that reflect the combined accuracy on match and mismatch trials (*d'*) and response bias (*criterion*). This data is also given in Figure 3 and shows that accuracy (*d'*) declined slightly over the course of the experiment in the *no feedback* condition compared to the *feedback group*. A 2 (feedback) x 4 (block) ANOVA showed a main effect of feedback, $F(1,48) = 4.54$, $p < 0.05$, reflecting lower accuracy in the *no feedback* condition, but no main effect of block, $F(3,144) = 1.48$, $p = 0.22$, and no interaction between both factors was found, $F(3,144) = 2.21$, $p = 0.09$. By contrast, the same analysis for the measure of response bias (*criterion*) did not show a main effect of feedback, $F(1,48) = 2.44$, $p = 0.12$, but revealed a main effect of block, $F(3,144) = 6.58$, $p < 0.001$, and an interaction between both factors, $F(3,144) = 3.99$, $p < 0.01$.

Analysis of simple main effects showed that criterion values were initially evenly matched for the *feedback* and *no feedback* condition, in the BL block, $F(1,192) = 0.61$, $p = 0.43$. By contrast, a simple main effect of feedback was observed for the FBT block, $F(1,192) = 5.16$, $p < 0.05$, and the differences for the BLT block, $F(1,192) = 3.24$, $p = 0.07$, and the NFT block, $F(1,192) = 2.97$, $p = 0.09$, were also approaching significance. These differences reflect a bias to make more match responses in the *no feedback* compared to the *feedback* condition, and suggest that this bias emerges over the course of the experiment. In addition, a simple main effect of block was also found for the *no feedback* condition, $F(3,144) = 10.26$, $p < 0.001$. Tukey HSD test revealed that criterion was lower in the BLT, FBT, and NFT block than in the BL block, all

qs ≥ 6.10, all *ps* < 0.001. None of the other comparisons were significant, all *qs* ≤ 0.48. Finally, no simple main effect of block was found for the *feedback* condition, $F(3,144) = 0.31$, $p = 0.82$.

Taken together, these results show that overall accuracy (*d'*) was lower in the *no feedback* compared to the *feedback* condition, and these groups differed in how these scores were obtained. Specifically, while the *feedback* group recorded a comparable percentage of match and mismatch responses across all blocks, and therefore *criterion* values that were close to zero, the *no feedback* group exhibited a bias to make more match responses. This bias emerged after the baseline block and was shown with this analysis in the BLT, FBT and NFT blocks. These results therefore converge with the percentage accuracy data to show that trial-by-trial feedback does not improve performance generally, but prevents observers from making increasingly more mismatch errors.

Reaction times

Although the task instructions emphasized accuracy, the mean correct response time (RTs) were also analysed to determine if any feedback effects might reflect speed-accuracy trade-offs. These RTs are also displayed in Figure 3 and show that observers initially required four to five seconds to classify the face pairs, but made increasingly faster match/mismatch decisions over the course of the experiment. In addition, observers appeared to be somewhat slower to classify mismatch than match pairs in the second half of the experiment. Generally, however, the pattern of RTs appears to be similar for the *feedback* and the *no feedback* condition.

A 2 (feedback: *feedback* versus *no feedback* group) x 2 (trial type: *match* versus *mismatch* trials) x 4 (block: *BL*, *BLT*, *FBT*, *NFT*) mixed-factor ANOVA of this data revealed a main effect of block, $F(3,144) = 40.50$, $p < 0.001$, and of trial type, $F(1,48) = 7.21$, $p < 0.01$, and

an interaction between both factors, $F(3,144) = 5.39$, $p < 0.01$. Analysis of simple main effects shows that this arises because observers were slower to classify mismatch than match pairs in the BLT block, $F(1,192) = 11.45$, $p < 0.001$, and the FBT block , $F(1,192) = 8.10$, $p < 0.01$, but not in any of the other blocks (BL and NFT), both $Fs(1,192) \leq 2.24$, $ps \geq 0.14$. Crucially, however, ANOVA found no main effect of feedback, $F(1,48) = 0.55$, $p = 0.46$, and no interaction between feedback and block, $F(3,144) = 0.39$, $p = 0.76$, feedback and trial type, $F(1,48) = 0.69$, $p = 0.41$, and no three-way interaction between all of the factors, $F(3,144) = 0.33$, $p = 0.81$. Thus, the administration of feedback did not affect the speed of observers' responses in the face-matching task.

Discussion

This experiment examined whether face-matching accuracy could be improved by providing observers with immediate trial-by-trial feedback on their performance during this task. The results failed to show an improvement in accuracy with such feedback. Remarkably, however, we found that matching accuracy declined during the experiment in the *absence* of feedback. This effect was marked specifically by reduced accuracy on mismatch trials. This indicates that, as the experiment progressed, observers increasingly perceived mismatch face pairs as identity matches when no feedback was administered. Taken together, these results suggest that feedback does not benefit observers by improving accuracy but, rather, is useful for helping to prevent a drop-off in performance.

As part of the experimental design, we also tested for the generalizability of any feedback effects by comparing face-matching accuracy for stimuli that had previously been seen with and without feedback (in the FBT and BLT blocks, respectively) and for new face pairs, that were

not encountered before (in the NFT block). The data show that the feedback advantage was found in all of these blocks, which indicates that this effect generalizes to completely new stimuli and therefore also persists after this manipulation has been removed. In addition, the response time data also indicate that any effects of feedback do not reflect a speed-accuracy trade-off whereby the observers in the *feedback* condition might have studied the face pairs for longer before making an identification decisions. Similarly, the response times show that the decrease in accuracy with mismatch pairs in the *no feedback* group cannot be attributed to such an effect. While mismatch accuracy became worse than match accuracy in this group over the course of the experiment, mismatch responses were, if anything, also made more slowly than match decisions (see Figure 3).

This experiment therefore provides the first evidence that feedback is beneficial for *maintaining* accuracy in face matching. Without such feedback, the classification of identity mismatch pairs declines, which suggests that observers seem to find it increasingly difficult to tell faces apart. Generally, however, it is also notable that performance was rather high, at around 90% accuracy. This level of performance is consistent with the normative data for this face set (Burton et al., 2010) and with other studies that have also employed these stimuli (e.g., Bindemann et al., 2012; Megreya et al., 2011; Özbek & Bindemann, 2011). As this stimulus set was designed to provide a best-case scenario for studying face matching, this high level of accuracy is unsurprising. However, this also raises the possibility that feedback could not *improve* observers' accuracy beyond the baseline level because they were already operating close to ceiling. To explore this possibility, we attempted to increase the difficulty of this task in the following experiments.

Experiment 2

Experiment 1 demonstrates that observers' face-matching accuracy declines for mismatch pairs unless trial-by-trial feedback is administered. However, performance was also close to ceiling. This raises the possibility that more profound effects of feedback might be found, such as an improvement in performance beyond the initial baseline accuracy, when this task becomes more difficult. Experiment 2 was conducted to examine this possibility by increasing the difficulty of the matching task. For this purpose, we removed the external features (e.g., the hair and face outline) of the faces from the matching displays. Salient features, such as hairstyle, provide a context that can improve face recognition and matching (e.g., Bruce et al., 1999; Ellis, Shepherd, & Davies, 1979; Endo, Takahashi, & Maruyama, 1984), but these features can also dominate the identification of unfamiliar faces (Young, Hay, McWeeney, Flude, & Ellis, 1985; Bonner, Burton, & Bruce, 2003; Clutterbuck & Johnson, 2005; Megreya & Bindemann, 2009) and, due to their changeable nature, can provide misleading identity information (see, e.g., Frowd et al., 2012; Sinha & Poggio, 1996, 2002). The removal of these external features in this task might therefore not only increase the difficulty of face matching by removing additional identity cues, but should also serve to focus observers on the most diagnostic, internal facial features. In Experiment 2, the stimuli and procedure were therefore kept identical to those in Experiment 1, except that the external facial features were removed from all stimuli.

Method

Participants, stimuli and procedure

Fifty new students (38 female) from the University of Kent, with a mean age of 20 years (SD = 3.1), volunteered to participate in this experiment for a small fee. All reported normal or

corrected-to-normal vision. The design, stimuli and procedure were identical to Experiment 1, except that the external features were removed so that the faces now consisted only of the internal features (i.e., the eyes, nose and mouth). Example stimuli are shown in Figure 4.

Results

Accuracy

In the first step of the analysis, the percentage accuracy was again calculated for match and mismatch trials for each of the experimental blocks in the *feedback* and *no feedback* group. This data is illustrated in Figure 5 and shows that observers in the *no feedback* group again achieved a similar level of performance on match and mismatch trials in the baseline block, but accuracy declined thereafter for mismatch trials. A similar pattern can also be observed for the *feedback* group but the difference in accuracy between match and mismatch trials appears to be much smaller here than for the *no feedback* condition.

A 2 (feedback) x 2 (trial type) x 4 (block) mixed-factor ANOVA of this data showed a main effect of feedback, $F(1,48) = 5.48$, $p < 0.05$, which reflects generally lower face matching accuracy in the *no feedback* group than the *feedback* group. A main effect of trial type was also found, $F(1,48) = 14.85$, $p < 0.001$, due to generally lower accuracy on mismatch trials. In addition, an interaction of block and trial type was observed, $F(3,144) = 6.42$, $p < 0.01$. Analysis of simple main effects revealed that mismatch accuracy was lower than match accuracy in all experimental blocks, all $Fs(1,192) \geq 8.87$, all $ps \leq 0.01$, except the BL block, $F(1,192) = 0.31$, $p < 0.58$. Tukey HSD test also showed that mismatch accuracy was reduced in the BLT, FBT, and NFT blocks compared to the BL block, all $qs \geq 5.10$, all $ps \leq 0.01$. None of the remaining

comparisons between blocks, both for match and mismatch trials, reached significance, all $qs \leq 3.25$.

Overall, the accuracy data therefore shows that matching performance was reduced once again when no feedback was administered, and was also worse for mismatch than match trials. Moreover, the mismatch disadvantage once again appears to emerge during the experiment, as it is not present in the initial baseline block.

d-prime and criterion

As in Experiment 1, the accuracy data was also transformed into the signal detection measures of *d'* and *criterion* (see Figure 5). The *d'* scores show a main effect of feedback, $F(1,48) = 6.39$, $p < 0.05$, which reflects the fact that accuracy was generally higher in the *feedback* group. However, similarly to Experiment 1, no main effect of block, $F(3,144) = 1.76$, $p = 0.16$, and no interaction of block and feedback was found, $F(3,144) = 0.41$, $p = 0.74$. By contrast, *criterion* showed no main effect of feedback, $F(1,48) = 3.52$, $p = 0.07$, and no interaction of block and feedback, $F(3,144) = 1.07$, $p = 0.36$, but a main of effect of block, $F(3,144) = 6.36$, $p < 0.001$. Tukey HSD test shows that this arises from a tendency to make more match responses in the BLT and FBT blocks compared to the BL block, both $qs \geq 4.64$, $ps \leq 0.01$. None of the remaining comparisons between blocks were significant, all $qs \leq 2.98$.

Reaction times

Once again, the mean correct RTs were also analysed (see Figure 5). The RTs again show that observers initially required four to five seconds to classify the face pairs, but made increasingly faster identification decisions over the course of the experiment. In addition,

observers again appeared to be somewhat slower to classify mismatch than match pairs. A 2 (feedback) x 2 (trial type) x 4 (block) ANOVA of this data found no main effect of feedback, $F(1,48) = 0.10$, $p = 0.75$, but a main effect of trial type was found, $F(1,48) = 9.03$, $p < 0.01$, due to generally slower responses on mismatch trials. In addition, a main effect of block was found, $F(3,144) = 57.63$, $p < 0.001$. Tukey HSD test shows that responses were slower in the BL block than the BLT, FBT and NFT blocks, all $qs \geq 14.25$, $ps \leq 0.001$. None of the remaining comparisons between blocks were significant, all $qs \leq 1.33$. Finally, none of the two-way interactions or the three-way interaction were significant, all $Fs \leq 2.49$, $ps \geq 0.12$.

Discussion

This experiment sought to replicate the feedback effect that was found in Experiment 1, which showed that the administration of trial-by-trial feedback during a face matching task *prevents* a gradual decline in the classification of identity mismatches over the course of an experiment. Moreover, we sought to examine if feedback can *improve* accuracy if the matching task is more made difficult so that observers are not operating close to ceiling. For this purpose, we removed the external features of the face stimuli.

The results showed that matching accuracy was generally worse without feedback and was also reduced on mismatch trials. This effect of trial type emerged after the initial baseline block, and it was evident in the percentage accuracy scores and criterion. It also cannot be attributed to a simple speed-accuracy trade-off. These results therefore provide a conceptual replication of Experiment 1, which indicates, once again, that the administration of feedback in face matching prevents a reduction in identification accuracy on mismatch trials. However, in contrast to Experiment 1, it is notable that a direct interaction between feedback and trial type or

feedback and block was not found. Moreover, observers' accuracy was generally similar in both experiments. This is evident, for example, from the baseline block, in which mean accuracy was once again at approximately 90% in Experiment 2. While the removal of the external facial features therefore serves to ensure that these effects reflect the processing of the more important, internal facial features (see, e.g., Frowd et al., 2012; Megreya & Bindemann, 2009; Young et al., 1985), it did not have the desired effect of increasing task difficulty. As a consequence, it is possible that we once again failed to capture any improvements that feedback might incur in face matching. This was examined further in Experiment 3.

Experiment 3

It is well established that changes in view substantially reduce the identification accuracy of unfamiliar faces in recognition and matching paradigms (e.g., Bruce et al., 1987; Hill & Bruce, 1996; Hill et al., 1997). For example, faces that are learned in a frontal view are less likely to be recognized when they are subsequently seen in ¾ profile, and even less so when shown in full profile (Longmore et al., 2008). In contrast to the preceding experiments, in which observers were required to match pairs of frontal faces, we therefore sought to increase task difficulty in Experiment 3 by asking observers to match the internal features of frontal faces to profile faces. According to the results of the preceding experiments, we predicted that observers' matching accuracy would gradually decrease on mismatch trials when no feedback is given. In turn, we therefore also expected that the administration of feedback would prevent this decline in mismatch accuracy. It is less clear whether feedback can also improve observers' face matching accuracy beyond their initial baseline performance. If such improvements are possible, then one

might expect to find these under the conditions of increased task difficulty that the matching of frontal-to-profile face views should provide.

Method

Participants

Fifty students (36 female) from the University of Kent, with a mean age of 20 years (*SD* = 3.1), volunteered to participate in this experiment for a small fee. None had participated in any of the preceding experiments and all reported normal or corrected-to-normal vision. As before, participants were randomly assigned to the *feedback* or *no feedback* group.

Stimuli and procedure

The stimuli and procedure were identical to Experiment 2, except that one of the frontal faces in each stimulus pair was replaced with a profile view of the internal facial features of the same identity. As before, observers were then required to decide if these face pairs depicted an identity match or mismatch. The same experimental design was employed, so that each observer was shown seven experimental blocks of 40 trials (20 match, 20 mismatch), corresponding to a baseline block (BL), followed by three feedback blocks (FB1, FB2, FB3), and finally three blocks to measure the effect of the trial-by-trial feedback on subsequent face matching accuracy (BLT, FBT, NFT).

Results

Accuracy

This data was analysed in the same way as the preceding experiments. The mean percentage accuracy for all conditions is shown in Figure 6. A 2 (feedback) x 2 (trial type) x 4 (block) ANOVA of this data showed a main effect of trial type, $F(1,48) = 9.54$, $p < 0.01$, reflecting generally lower accuracy on mismatch trials, and a main effect of block, $F(3,144) = 3.67$, $p < 0.05$. These main effects were qualified by an interaction between these factors, $F(3,144) = 10.70$, $p < 0.001$. Analysis of simple main effects showed that mismatch accuracy was lower than match accuracy across the BLT, FBT and NFT blocks, all $Fs(1,192) \geq 5.11$, $ps \leq 0.05$, but not in the BL block, $F(1,192) = 1.35$, $p = 0.25$. In addition, mismatch accuracy was reduced in the BLT, FBT and NFT blocks compared to the BL block, all $qs \geq 6.29$, $ps \leq 0.001$. There were also some differences between blocks in the match condition. Here, accuracy was improved in the BLT and FBT blocks compared to the BL block, both $qs \geq 3.76$, $ps \leq 0.05$, and in the FBT block compared to the NFT block, $q = 4.31$, $p < 0.05$. None of the remaining comparisons were significant, all $qs \leq 3.42$.

In addition, an interaction between feedback and trial type was also found, $F(1,48) = 7.61$, $p < 0.01$. Simple main effect analysis shows that this arises from an effect of trial type for the *no feedback* condition, $F(1,48) = 17.09$, $p < 0.001$, which reflects lower accuracy for mismatch than match trials. By contrast, the corresponding simple main effect for match trials was not significant, $F(1,48) = 0.06$, $p = 0.82$. Finally, the two-way interaction of feedback and block, $F(3,144) = 1.81$, $p = 0.14$, and the three-way interaction, $F(3,144) = 0.57$, $p = 0.64$, were not significant.

Overall, the percentage accuracy data therefore shows that matching accuracy was again reduced on mismatch trials compared to match trials, when no feedback for performance was

provided. In addition, mismatch accuracy again declined during the experiment, in the BLT, FBT, and NFT blocks compared to the BL block.

d-prime and criterion

The accuracy data was again also transformed into the signal detection measures of *d'* and *criterion* (see Figure 6). A 2 (feedback) x 4 (block) ANOVA of the *d'* scores show no main effect of feedback, $F(1,48) = 1.42$, $p = 0.24$, and no interaction between feedback and block, $F(3,144) = 2.08$, $p = 0.11$, but a marginally significant main effect of block was found, $F(3,144) = 2.67$, $p = 0.05$. However, Tukey HSD test revealed no differences between any of the blocks, all $qs \leq 3.52$.

An analogous analysis of criterion found a main effect of feedback, $F(1,48) = 8.35$, $p < 0.01$, which reflects a stronger bias to make match responses in the *no feedback* compared to the *feedback* group. In addition, a main effect of block was also found, $F(3,144) = 9.48$, $p < 0.001$, which reflects a stronger tendency to make match responses on the BLT, FBT and NFT blocks compared to the initial baseline block, all $qs \geq 4.47$, $ps \leq 0.05$. None of the remaining comparisons, all $qs \leq 2.15$, nor the two-way interaction, $F(3,144) = 0.23$, $p = 0.88$, were significant.

Reaction Times

The mean correct response time (RTs) were analysed again to determine if any feedback effects might reflect speed-accuracy trade-offs (see Figure 6). A 2 (feedback) x 2 (trial type) x 4 (block) ANOVA of this data found no main effect of feedback, $F(1,48) = 2.00$, $p = 0.16$, but a main effect of trial type, $F(1,48) = 14.45$, $p < 0.001$, and an interaction between both factors,

$F(1,48) = 7.88$, $p < 0.01$. This reflects the fact that observers required longer to classify identity mismatches than matches in the *no feedback* condition, $F(1,48) = 21.83$, $p < 0.001$, whereas response times for match and mismatch pairs were evenly matched in the *feedback* group, $F(1,48) = 0.49$, $p = 0.49$. Finally, a main effect of block was also found, $F(3,144) = 24.53$, $p < 0.001$. This arises because response times were slower in the BL block compared to the BLT, FBT and NFT blocks, all $qs \geq 0.55$, $ps \leq 0.001$. None of the remaining comparisons between blocks were significant, all $qs \leq 0.61$. In addition, the interaction of feedback and block, $F(3,144) = 0.31$, $p = 0.82$, and block and trial type, $F(3,144) = 1.41$, $p = 0.24$, and the three-way interaction, $F(3,144) = 1.14$, $p = 0.33$, were also not significant.

Discussion

To determine whether feedback could improve face matching when task difficulty was increased, observers were asked to match frontal to profile views of faces in Experiment 3. Accuracy was clearly reduced compared to the preceding experiments (c.f. Figures 3, 5 and 6). Despite this, the pattern of results was very similar to Experiment 2. For example, without feedback for face matching accuracy, observers again made more identification errors on mismatch than match trials, while no such differences were found in the *feedback* group. Moreover, the decrease in mismatch accuracy again was not observed in the initial baseline block but emerged during the course of the experiment. These results therefore provide further evidence that feedback can prevent a reduction in mismatch accuracy in face matching. Again, however, we also found that feedback did not *improve* performance above the initial accuracy levels of the baseline block.

Taken together, these findings indicate that trial-by-trial feedback can help observers to maintain their face matching accuracy at a standard that matches their natural performance level in this task. Without such feedback, performance gradually declines, and this appears to be the case specifically on mismatch trials. In other words, observers seem to find it increasingly difficult to dissociate two different facial identities or to "tell different faces apart" (see Jenkins et al., 2011), rather than to decide that two faces are, in fact, of the same person. The administration of feedback during face matching not only prevents this decline, but this benefit was still found several blocks after feedback was withdrawn again. This indicates that observers are acquiring information about the face-matching task during the feedback that can still be used subsequently to maintain performance.

An interesting aspect of these findings is that feedback is administered after a response has been made and the face pairs have been removed from view. The observers therefore cannot study particular faces with this knowledge to understand their identification errors. This suggests that the influence of feedback does not reflect a gain in face-specific knowledge. Rather, this indicates that the feedback effect might relate to the extent to which an observer has a more general knowledge about their level of performance in this task. One question that arises from these considerations is whether these effects are specific to *trial-by-trial* feedback, which is administered immediately after a response has been made, or whether similar benefits can be found with other forms of feedback. To explore this possibility, the next experiment examined the effect of *outcome* feedback on face matching. In this type of feedback, observers are not given information on a trial-by-trial basis about their matching performance but are provided with a summary of their general face-matching accuracy at the end of a block.

Experiment 4

In contrast to the preceding experiments, which administered feedback on a trial-by-trial basis, Experiment 4 provided observers with *outcome* feedback about their face matching performance (see, e.g., Kulhavy & Stock, 1989; Kluger & DeNisi, 1996; Mory, 2004; Shute, 2008; Narciss, 2008; Landsberg et al., 2012). This was delivered by providing each observer with a percentage score of the total number of their correct responses at the end of a block of trials. In contrast to the trial-by-trial feedback of the preceding experiments, observers therefore only received information about their general accuracy, rather than detailed information about the specific trials on which they made a correct or incorrect response. For this purpose, we retained the block design of the preceding studies but outcome feedback was administered at the end of the first four blocks. The rationale for giving feedback also for block 1 (the baseline) was that observers could only begin to utilize this feedback in block 2 (FB1), which corresponds to the stage at which feedback was first administered in Experiments 1 to 3.

Method

Participants

Twenty-five new students (18 female) from the University of Kent, with a mean age of 21 years (SD = 4.4), volunteered to participate in this experiment for a small fee. None had participated in any of the preceding experiments and all reported normal or corrected-to-normal vision.

Stimuli and procedure

The stimuli and procedure were identical to Experiment 3, except for the following changes. In contrast to the trial-by-trial feedback of the preceding studies, the feedback was now provided at the end of the experimental blocks in the form of a single summary statistic. This was presented onscreen and provided each observers with the total percentage accuracy, combined for match and mismatch trials, that they had achieved in the immediately preceding block of 40 trials. This feedback was presented to all twenty-five participants after the first four experimental blocks (BL, FB1, FB2, FB3). As in the preceding experiments, we then assessed whether any effect of feedback would continue to hold after this information had been given. Therefore, face matching accuracy was again assessed with three further blocks, corresponding to a repetition of the faces from the baseline block (BLT), the third feedback block (FBT), and a block of completely new faces (NFT). However, note that a *no feedback* group was not included in this experiment. This condition would have been identical to the *no feedback* condition of Experiment 3, so we will use that data for any between-group comparisons.

Results

Accuracy

The data for this experiment is displayed in Figure 7 and shows that observers' match and mismatch accuracy was similar in the baseline block of the outcome *feedback* condition. Thereafter, mismatch accuracy declined gradually over the course of the experiment, while a corresponding increase in match accuracy was evident. This pattern of results is comparable to the *no feedback* condition of the preceding experiments (e.g., c.f., Figures 3, 5 and 6) and suggests that *outcome* feedback is not effective in helping to maintain observers' accuracy in this face-matching task.

To analyse these observations formally, we compared this data to the *no feedback* condition of Experiment 3. A 2 (feedback) x 2 (trial type) x 4 (block) ANOVA of this data showed a main effect of block, $F(3,144) = 3.64$, $p < 0.05$, and an interaction of block and trial type, $F(3,144) = 23.74$, $p < 0.001$. Simple main effect analysis of this interaction shows a main effect of trial type for the BLT, FBT and NFT blocks, all $Fs(1,192) \geq 26.17$, $ps \leq 0.001$, due to higher accuracy on match than mismatch trials, except for the BL block , $F(1,192) = 1.33$, $p = 0.25$, in which performance for these trial types was comparable. In addition, Tukey HSD test showed that accuracy for match trials was substantially better in the BLT and FBT blocks compared to the BL block, both $qs \geq 6.08$, $ps \leq 0.001$. By contrast, accuracy declined for mismatch face pairs in the BLT, FBT and NFT blocks compared to the baseline, all $qs \geq 9.96$, $ps \leq 0.001$. None of the remaining comparisons were significant, all $qs \leq 3.59$.

These differences show that accuracy declined on mismatch trials over the course of the experiment and also improved on match trials. Crucially, however, no main effect of feedback was found, $F(1,48) = 1.44$, $p = 0.51$, and no interaction between feedback and block, $F(3,144) = 0.28$, $p = 0.84$, feedback and trial type, $F(1,48) = 0.07$, $p = 0.78$, and also no three-way interaction, $F(3,144) = 0.93$, $p = 0.43$. This indicates that the performance in the outcome *feedback* condition did not differ from the *no feedback* group of Experiment 3.

d-prime and criterion

These observations were confirmed by the analysis of signal detection measures. Firstly, a 2 (feedback) x 4 (block) ANOVA was conducted of the *d'* scores, which showed no main effect of feedback, $F(1,48) = 0.43$, $p = 0.52$, or block, $F(3,144) = 1.16$, $p = 0.33$, and no interaction between these factors, $F(3,144) = 0.76$, $p = 0.52$. By contrast, the analogous analysis

of *criterion* showed no main effect of feedback, $F(1,48) = 0.15$, $p = 0.70$, and no interaction, $F(3,144) = 1.34$, $p = 0.26$, but a main effect of block, $F(3,144) = 22.31$, $p < 0.001$. Tukey HSD test showed that criterion was lower in the BLT, FBT and NFT blocks compared to the BL block, all $qs \geq 7.68$, $ps \leq 0.001$. None of the remaining comparisons were significant, all $qs \leq 2.58$.

Taken together, these results indicate that overall accuracy (*d'*) did not improve during the course of this experiment, but observers developed a gradual response bias to make more match responses (*criterion*). This leads to a decrease in mismatch performance and the increase in match performance that can be observed in the percentage accuracy scores in Figure 7.

Reaction times

For completeness, the mean correct RTs were analysed again (see Figure 7). A 2 (feedback) x 2 (trial type) x 4 (block) ANOVA of this data found a main effect of trial type, $F(1,48) = 23.72$, $p < 0.001$, reflecting generally longer RTs on mismatch trials, and a main effect of block, $F(3,144) = 23.70$, $p < 0.001$. Tukey HSD test showed that response times were faster in the BLT, FBT and NFT blocks compared to the BL block, all $qs \geq 9.15$, $ps \leq 0.001$. None of the remaining comparisons were significant, all $qs \leq 1.22$. Despite all of these differences, no main effect of feedback was found, $F(1,48) = 0.05$, $p = 0.83$, and no interaction between feedback and block, $F(3,144) = 0.27$, $p = 0.85$, feedback and trial type, $F(1,48) = 1.70$, $p = 0.20$, and block and trial type, $F(3,144) = 2.52$, $p = 0.06$. The three-way interaction was also not significant, $F(3,144) = 1.48$, $p = 0.22$.

Discussion

This experiment examined whether the effect of feedback on face matching depends on the type of feedback that is administered. In contrast to the preceding experiments, which administered trial-by-trial feedback, this experiment therefore provided observers with *outcome* feedback. This was administered in the form of a single summary statistic at the end of the first four experimental blocks, which corresponded to an observer's mean face-matching accuracy.

The results show that observers' accuracy was comparable for match and mismatch trials in the baseline block, but then declined on mismatch trials. This decline was accompanied by an increase in match accuracy. The signal detection measures indicate that this pattern of results reflects a response bias to make more match responses over the course of the experiment. We compared this pattern of results with the *no feedback* condition from Experiment 3 and found no differences between these groups. These results therefore suggest that outcome feedback cannot prevent the decline in face matching accuracy that also occurs in the absence of any feedback. These findings therefore suggest that the *type* of feedback is crucial, and specifically trial-by-trial feedback, is important for maintaining observers' accuracy in face matching.

However, an alternative explanation for these results also remains possible. In Experiment 4, feedback was administered as an overall score that provided a combined index of match and mismatch accuracy. In contrast to the trial-by-trial feedback, which provides separable information for match and mismatch trials, the combined outcome score cannot inform observers of the specific decline in mismatch accuracy that we have observed in all of the experiments here. Moreover, in Experiment 4 this decline in mismatch accuracy was seemingly offset by a bias to make more match responses, and the combined feedback score would have failed to capture these differences. As a result, face-matching accuracy may have *appeared* to be relatively stable across blocks to observers from the feedback, which could have undermined the

efficacy of this manipulation. To address this issue, we conducted a further experiment, in which observers were given separate outcome feedback for match and mismatch trials. This experiment is described below.

Experiment 5

In Experiment 4, participants were given outcome feedback for their face-matching performance at the end of the first four experimental blocks. This type of feedback was ineffective in helping participant maintain their initial mismatch accuracy over the course of the experiment. However, we also found that observers' match responses increased as mismatch accuracy decreased. A possible explanation for the absence of a feedback effect in Experiment 4 is therefore that the decline in mismatch accuracy was offset by the increase in match responses. Crucially, the combined outcome feedback score would have failed to capture this changing response pattern, and this could have undermined the efficacy of this manipulation. To address this issue, observers were provided with two outcome feedback scores in the next experiment, which corresponded to the percentage of correct match and mismatch responses, respectively. As in Experiment 4, these scores were provided onscreen after the completion of the first four blocks.

Method

Participants, stimuli and procedure

Twenty-five new students (18 female) from the University of Kent, with a mean age of 20 years (SD = 2.9), volunteered to participate in this experiment for a small fee. None had participated in any of the preceding experiments and all reported normal or corrected-to-normal

vision. The stimuli and procedure were identical to Experiment 4, except that the feedback summary statistics, which were displayed at the end of the first four blocks, now provided separate accuracy measures for match and mismatch trials.

Results

Accuracy

Figure 8 shows the percentage accuracy data for this experiment. Once again, this data shows that match and mismatch accuracy was similar in the baseline block of the *feedback* condition, but mismatch accuracy declined thereafter. To analyse these observations formally, we again compared this data to the *no feedback* condition of Experiment 3. A 2 (feedback) x 2 (trial type) x 4 (block) ANOVA of this data did not show a main effect of block, $F(3,144) = 2.59$, $p = 0.06$, but a main effect of trial type, $F(1,48) = 17.29$, $p < 0.0001$, and an interaction between both factors, $F(3,144) = 9.29$, $p < 0.001$. Simple main effect analysis of this interaction revealed a main effect of trial type for the BLT, FBT and NFT block, all $Fs(1,192) \geq 16.82$, $ps \leq 0.001$, but not the BL block, $F(1,192) = 0.02$, $p = 0.88$. In addition, Tukey HSD test showed that accuracy for match trials was substantially better in the BLT, FBT and NFT block, compared to the baseline block, all $qs \geq 3.78$, $ps \leq 0.05$. By contrast, performance declined for mismatch face pairs. Here, accuracy was reduced in the BLT, FBT and NFT block compared to the baseline, all $qs \geq 5.85$, $ps \leq 0.001$. None of the remaining comparisons were significant, all $qs \leq 2.44$. Notably, however, no main effect of feedback was found, $F(1,48) = 0.44$, $p = 0.51$, and no interaction between feedback and block, $F(3,144) = 1.73$, $p = 0.16$, feedback and trial type, $F(1,48) = 0.44$, $p = 0.51$, and also no three-way interaction, $F(3,144) = 0.82$, $p = 0.48$.

d-prime and criterion

This data was analysed further with signal detection measures. A 2 (feedback) x 4 (block) ANOVA was conducted of *d'* scores, which showed no main effect of feedback, $F(1,48) = 0.32$, $p = 0.57$, or block, $F(3,144) = 0.83$, $p = 0.48$, and no interaction between these factors, $F(3,144) = 1.77$, $p = 0.16$. The analogous analysis of *criterion* also showed no main effect of feedback, $F(1,48) = 0.61$, $p = 0.44$, and no interaction, $F(3,144) = 0.99$, $p = 0.40$. However, a main effect of block was found, $F(3,144) = 7.26$, $p < 0.001$. Tukey HSD test showed that criterion was lower in the BLT, FBT and NFT blocks compared to the baseline, all $qs \geq 5.21$, $ps \leq 0.01$. None of the remaining comparisons were significant, all $qs \leq 0.39$. Similar to Experiment 4, these results therefore indicate that overall accuracy (*d'*) did not decline or improve during the course of this experiment. However, observers developed a gradual response bias to make more match responses (*criterion*), which lead to a decrease in mismatch performance.

Reaction times

A 2 (feedback) x 2 (trial type) x 4 (block) ANOVA of the response times found a main effect of trial type, $F(1,48) = 21.07$, $p < 0.001$, reflecting generally longer response times on mismatch trials, and a main effect of block, $F(3,144) = 26.37$, $p < 0.001$. Tukey HSD test showed that response times were faster in the BLT, FBT and NFT block compared to the baseline block, all $qs \geq 8.83$, $ps \leq 0.001$. None of the remaining comparisons were significant, all $qs \leq 1.94$. Neither the main effect of feedback, $F(1,48) = 0.24$, $p = 0.63$, and none of the interactions were significant, all $Fs \leq 2.31$, $ps \geq 0.08$.

Discussion

Despite providing separate outcome feedback for match and mismatch trials, this experiment once again failed to show a benefit from this type of feedback compared to the *no feedback* condition. Thus it appears that, even when observers are given separate information to summarize their performance for match and mismatch face pairs, outcome feedback does not influence identification accuracy. Moreover, a match bias once again emerged during the experiment, whereby mismatch accuracy declined because observers made more match responses. Therefore, these findings not only show that the type of feedback that is provided is crucial for maintaining face-matching accuracy, but they also strengthen the data from the no feedback condition of Experiments 1 to 3, in which mismatch accuracy declined after the baseline block, by replicating a similar effect when outcome feedback is given.

In all of the experiments, this decline in accuracy emerged in the *no feedback* condition (Experiments 1 to 3) or with outcome feedback (Experiments 4 and 5) after the initial baseline block. Moreover, all of these experiments suggest that mismatch accuracy continues to decline throughout the task. For example, in Experiment 5 mismatch accuracy was highest in block 1, at 77%, and was lowest in the final block, at 66%. This raises the question of whether mismatch accuracy would continue to decline if the matching task is extended. To investigate this possibility, we conducted a final experiment, in which observers were given 25 successive blocks of the matching task, equating to a total of 1000 trials. Our aim here was to determine whether mismatch performance would continue to decrease over the duration of this prolonged experiment, or whether observers' accuracy would eventually stabilise at a particular level during this task.

Experiment 6

The final experiment investigates whether mismatch accuracy continues to decline below the performance levels that were observed in the preceding experiments if the face-matching task is extended. For this purpose, we asked observers to complete 25 successive blocks of this task. No feedback of any kind was administered throughout.

Method

Participants

Twenty-five new students (21 female) from the University of Kent, with a mean age of 20 years (*SD* = 3.9), volunteered to participate in this experiment for a small fee. None had participated in any of the preceding experiments and all reported normal or corrected-to-normal vision.

Stimuli and procedure

The stimuli consisted of the same 100 match and 100 mismatch pairs that were used in Experiment 5. These stimuli were administered in 5 blocks of 40 trials (20 match and 20 mismatch). This sequence of blocks was then repeated four more times, to give a total of 25 experimental blocks. The presentation of the stimuli was randomized within all blocks for each observer. However, the order of the blocks was counterbalanced across participants over the course of the experiment, so that each face pair was equally likely to appear in all of the blocks.

The trial procedure was kept identical to all of the preceding experiments. Thus, each trial began with a fixation cross, which was displayed for 1 second. This was followed by a face pair, which was presented until a response was registered. Participants were instructed to decide whether an identity match or mismatch was shown in each trial, by pressing two corresponding

buttons on a computer keyboard. Accuracy was emphasized and responses were self-spaced. No feedback was administered throughout the experiment.

Results

Accuracy

In a first step of the analysis, the mean percentage accuracy for match and mismatch trials was calculated. This data is shown in Figure 9 for each block of the experiment. An inspection of this data shows that performance was initially highly comparable for match and mismatch trials, in the first block of the experiment. Thereafter, mismatch accuracy begins to decline, while accuracy on match trials gradually increases. This pattern persists throughout the experiment. For example, the highest mean accuracy for mismatch trials in any of the blocks was actually achieved in block 1 (65%), while the lowest score (of 41%) is recorded in the final block. By contrast, the lowest match score, of 67%, was obtained in block 1, while the score for the final block, of 72%, is close to the highest match score that was obtained in any of the 25 blocks here.

To analyse this data, we correlated the match and mismatch scores with the block number to determine if accuracy was likely to increase or decrease over time. This analysis revealed a positive correlation of block and accuracy on match trials, $r(24) = 0.708$, $p < 0.01$, which shows that performance on match trials gradually improved over the course of the experiment. The opposite pattern was found for mismatch trials, which showed a continuous decline in accuracy, $r(24) = -0.953$, $p < 0.01$. This decline appears to be more marked than the increase seen for match trials. In order to assess this, we also correlated overall accuracy (the average of match and mismatch accuracy) with block. This analysis revealed a negative correlation, $r(24) = -0.932$, $p < 0.01$, which shows that overall accuracy generally decreased during the experiment.

As for all previous experiments, we also calculated *d'* prime and *criterion* (see Figure 9). The *d'* scores declined throughout the task, which corresponds to the gradual decrease in overall accuracy during the experiment that is already noted above. *Criterion* was initially close to zero, which indicates that observers were initially equally likely to make a correct match or mismatch response. However, *criterion* then begins to decline in block 2, and continues to decrease throughout the experiment, which indicates a consistently growing bias to make more match responses. Taken together, the percentage accuracy scores and the signal detection measures therefore show that match and mismatch accuracy is similar at the start of the experiment. Thereafter, match accuracy increases while mismatch accuracy begins to decline, and these behavioural patterns persist over the course of the entire experiment.

Response times

For completeness, Figure 9 also shows the mean correct RTs. For match and mismatch trials, these data correlated negatively with block, $r(24) = -0.880$, $p < 0.01$, and, $r(24) = -0.948$, $p < 0.01$, respectively. This shows that response times for both trial types declined over the course of the experiment.

Discussion

Experiment 6 employed a long version of the face-matching task to determine if accuracy continues to decline if this task is extended beyond the duration of the preceding experiments. To assess this possibility, we extended the experimental design from seven blocks and 280 trials to 25 blocks and a total of 1000 trials. The results show that performance does indeed continue to decline throughout this task. As in all of the preceding experiments, match and mismatch

accuracy was similar in the first block of the experiment. Thereafter, mismatch accuracy began to decline immediately, while accuracy on match trials increased. This pattern persisted throughout the entire experiment. Indeed, the lowest mismatch accuracy for any of the blocks was obtained in the final block of the experiment, which implies that performance would have decreased still if the task had been extended even further.

Essentially, this response pattern represents a response bias whereby observers are making increasingly more match responses during the course of the experiment. This indicates that observers find it increasingly difficult to tell the faces of different people *apart* or, in other words, that different facial identities are increasingly looking the *same* with continuous exposure to this task. These findings are discussed in detail in the General Discussion.

General Discussion

Matching photographs of faces is a difficult task that gives rise to many identification errors (e.g., Bindemann & Sandford, 2011; Burton et al., 2010; Henderson et al., 2001; Megreya & Burton, 2008). A possible explanation for why these errors continue to occur is that observers typically do not receive feedback for the identification of unfamiliar people, and may therefore be unaware mistaken identifications are made (see Burton & Jenkins, 2011; Jenkins et al., 2011). In this study, we therefore investigated whether face-matching performance could be improved when feedback is administered.

Across several different experiments, we consistently found that face-matching accuracy decreased when *no* feedback is given. This decline in accuracy specifically affected mismatch trials, so observers were more likely to respond by mistake that the faces of two different people belonged to the same person, than to mistake two photographs of the same person as different

people. In addition, in contrast to the decline in mismatch accuracy, we also found that accuracy seemed to increase on match trials. While this increase was only reliable in Experiments 4 to 6, it indicates that observers develop a response bias during face matching when no feedback on performance is given, whereby they become generally more likely to classify both match and mismatch face pairs as identity matches over the course of an experiment. These observations receive further support from the signal detection analysis, which indicated in all experiments that the absence of feedback does not affect overall accuracy (i.e., *d'*) *per se*, but results in a response bias (*criterion*) to classify face pairs as identity matches. It is noteworthy that our results also rule out a speed-accuracy trade-off for this pattern of results, whereby observers might make more errors on mismatch trials as response times speed up over the course of the experiment; if anything, response times were consistently slower on mismatch than match trials. This indicates that observers generally found it more difficult to classify identity mismatch pairs.

A markedly different pattern emerged when observers were given trial-by-trial feedback for their face-matching performance (Experiment 1-3). Similarly to the *no feedback* condition, accuracy was initially very similar for match and mismatch trials. However, with the administration of this type of feedback, observers were able to maintain this accuracy level in both match and mismatch trials throughout the task. Thus, trial-by-trial feedback helps to prevent the bias to make more match responses. It is remarkable that this effect was found after the feedback had been withdrawn again and with faces that were previously seen with and without feedback as well as with new faces. This indicates that the effects of trial-by-trial feedback are relatively long lasting, in that they persist for several blocks after feedback is removed, and generalize across different face stimuli.

The results also show that the type of feedback that is given is crucial for maintaining accuracy. When outcome feedback was administered in Experiment 4, which informed observers of their mean face matching accuracy at the end of a block of trials, performance was indistinguishable from the *no feedback* condition. These differences between the effectiveness of trial-by-trial and outcome feedback raise the question of what observers were able to learn from feedback in this task. Even when outcome feedback was administered as separate scores for match and mismatch trials in Experiment 5, a decline in mismatch accuracy was still found. This indicates that effective feedback needs to be administered online, immediately after an identification decision has been made.

A possible explanation for these differences between trial-by-trial and outcome feedback might reflect the different reinforcing properties of these manipulations (see, e.g., Annett, 1969; Chapanis, 1964; Gibbs & Brown, 1955). This notion is appealing considering that trial-by-trial feedback served only to *maintain* observers' initial face-matching accuracy throughout the task. Thus, this type of feedback seems to *reinforce* the original, intrinsic criteria that observers use to make match and mismatch decisions, but does not improve or change these criteria. For feedback to serve as an effective reinforcer, it has been suggested that its frequency has to correlate positively with that of correct responses on a performance task (Anderson, Kulhavy, & Andre, 1971; Hundal, 1969; Cook, 1968, Ivancevich, Donnelly, & Lyon, 1970). So the more frequently feedback is provided, the more accurate participants generally seem to be in performing a given task (Anderson et al., 1971; Cook, 1968; Ivancevich et al., 1970; Ilgen, Fisher, & Taylor, 1979). In this framework, the trial-by-trial feedback may help to maintain face-matching accuracy because it presents a continuous, and therefore a much more frequent, reinforcement than the outcome feedback manipulation.

The notion of reinforcement may provide a useful framework for interpreting some of the differences between trial-by-trial and outcome feedback here. On its own, however, this framework still seems to provide limited explanatory power. For example, an explanation solely in terms of reinforcement may be difficult to reconcile with the observation that trial-by-trial feedback also helped to maintain performance for several blocks *after* it was withdrawn and with completely new faces. At this stage, both the trial-by-trial and outcome feedback conditions provide observers with a measure of their face matching accuracy prior to the three final test blocks (BLT, FBT, NFT). Indeed, one could argue that the outcome feedback of Experiment 5 might have provided the observers with a more concrete performance index at this stage than the trial-by-trial feedback, by providing exact percentage accuracy scores. The effectiveness of feedback therefore does not seem to depend on providing observers with an accurate index of their accuracy *per se*, but on their mental approach to the test blocks after feedback has been given.

One way to explain performance at this stage in the trial-by-trial feedback condition could be based on the effect that this feedback might have on observers' motivation or task engagement. To maintain engagement in a task, observers need to obtain a sense that the task demands can be met (e.g., Fisher & Ford, 1998; Ford et al., 1998). Trial-by-trial feedback provides an immediate and accessible confirmatory measure for observers that may encourage such engagement. By contrast, the summative format of the outcome feedback may prevent observers from obtaining a similar appreciation that they really can meet the task demands, because this type of feedback does not provide information about the specific instances, or trials, on which a correct or incorrect identification decision was actually made. Thus, outcome

feedback might fail to maintain observers' task engagement because it fails to convey *how* a particular mean level of accuracy is achieved during the task.

Similarly, it is also possible that trial-by-trial feedback serves as an intrinsic motivator by providing observers with a *feeling of competence*. According to Deci (1972, 1975) and White (1959), observers seek a sense of competence when performing a task, which can in itself be seen as a powerful reward for the individual. However, to feel a sense of competence, an individual has to be able to judge their own performance. Feedback provides such an index to observers, and the more feedback is provided, the higher may be the motivating potential of a task (Hackman & Oldham, 1976). Thus, trial-by-trial feedback may be effective in maintaining observers' accuracy in the test blocks by providing such a sense of competence. Overall, we therefore suggest that the effect of trial-by-trial feedback may be two-fold here. One side of this effect may reflect the reinforcement of participants' intrinsic, initial response criteria in face matching tasks. In addition, such feedback may help to maintain motivation or task engagement by providing a direct and accessible measure of a person's accuracy that generates an increased feeling of competence in the observer.

These explanations are clearly still speculative. For example, considering an explanation in terms of motivation, one might predict that the less motivated participants of the *no feedback* group might be prone to making less conscientious and therefore faster matching decisions here. However, the response time data suggest that these observers were often less accurate *and* slower in this task, which indicates that mismatch accuracy may have been poor despite observers' increased efforts in the *no feedback* conditions. These details, in addition to a need to understand the effects of feedback on face matching more generally, clearly require further investigation. We are critically aware that our findings only provide a starting point here.

To this point, it is also notable that trial-by-trial feedback did not improve accuracy above the initial baseline levels in any of the experiments here. This shows that feedback did not help observers to produce any actual improvements in their face matching accuracy. In Experiment 1 and 2, overall accuracy was generally high, so such improvements may have been prevented by possible ceiling effects. However, trial-by-trial feedback also failed to improve overall performance when the task demands were increased in Experiment 3, by asking observers to match frontal to profile views of faces. Despite these changes to the stimulus material, it is possible that the experimental conditions were still insufficient for enabling observers to raise their face-matching ability. It is notable, for example, that feedback was always administered *after* the faces had been removed from view. This greatly limits the extent to which observers could study their own face matching mistakes and learn from these errors during this task.

Moreover, we also did not attempt to direct observers to facial information that might help them to improve, for example, such as key facial features. One reason for this is that we simply do know what these key facial features might be for face matching. However, it has already been shown that eye movements are functional during face learning (Henderson, Williams, & Falk, 2005) and specific eye movement patterns have also been linked to improved memory for new faces (Sekiguchi, 2011). An interesting extension of our work might therefore combine task-relevant information, which is extracted from eye movements to faces, with feedback for face-matching accuracy, to see if this can help observers improve in this task beyond their initial performance levels. This could be achieved, for example, by imposing eye scanpaths from a correct matching decision by one observer on the corresponding face pair after a matching error has been made by a different observer, to direct the latter person to appropriate

task-relevant information (for a similar approach from outside of the face domain, see, e.g., Litchfield et al., 2010).

While this is an interesting avenue for future research, we have previously also found that observers are consistently drawn to only a few central regions in faces in matching tasks (Özbek & Bindemann, 2011). This could also indicate that accuracy in face matching might not depend on the extent to which observers are drawn to specific facial features *per se*, but in *how* these features are processed. The current data may speak to this issue too by converging with a recent report, which showed that the same observers often respond differently to the same face pairs on different encounters. Bindemann et al. (2012) showed observers the same set of face pairs on several consecutive days and found that only very few identification errors were made consistently. This suggests that observers might vary in how they perceive the same faces on different encounters. While we did not analyse individual items in the current experiments, our results are, nonetheless, entirely consistent with these observations. For example, the *no feedback* conditions of Experiments 1 to 3 indicate clearly that observers varied in how they responded to the face pairs from the baseline block (BL) in a subsequent encounter (in the BLT block). Moreover, in Experiment 6, the same sequence of blocks was shown five times over the course of the experiment. Despite this repetition, mismatch accuracy continued to decline, which indicates that observers varied in how they responded to the same face pairs in different instances. This is an interesting finding as research on unfamiliar face identification has predominantly focused on external factors, outside of an observer's control, such as changes in lighting, viewpoint, and facial expression (e.g., Bruce, 1982; Johnston et al., 1992; Longmore et al., 2008), to explain limitations in performance. The current findings add to an increasing body of research that highlights the role of *internal* factors, within observers, for face matching (see,

e.g., Bindemann et al., 2012; Burton et al., 2010; Megreya & Burton, 2006; Megreya & Bindemann, 2013).

While face matching is increasingly at the focus of research, a final theoretical question that arises here is why the decrease in accuracy that was found in the current experiments has not been documented previously. In this field, face-matching accuracy is typically illustrated by combining performance across blocks. Moreover, face-matching is frequently studied in best-case scenarios, which are designed to produce the highest possible level of performance (see, e.g., Burton et al., 2010; Bindemann et al., 2010, 2012; Megreya & Burton, 2006, 2008), or, conversely, under conditions that make this task particularly difficult (e.g., Bindemann & Sandford, 2011; Henderson et al., 2001; Kemp et al., 1997). Both of these approaches may be insensitive to the gradual decline in mismatch accuracy that was observed here, especially when performance data is also pooled across blocks.

In addition to the theoretical issues raised by this data, this research also has applied implications. Face matching is frequently used as a laboratory analogue to study important identification tasks in security settings, such as passport control (see, e.g., Bindemann & Sandford, 2011; Kemp et al., 1997; Jenkins & Burton, 2008, 2011). While it has been known for many years that this task is highly susceptible to identification errors, rather little is still known about how performance might be improved (for a possibility, see Bindemann et al., 2012). The present experiments also fall short of revealing a method to improve face-matching performance. However, our experiments are important nonetheless, for two reasons.

Firstly, our findings show that accuracy declines in this task over time. Moreover, this decline is characterized by a striking error, whereby observers' accuracy decreases specifically on mismatch trials. This indicates that people find it increasingly difficult to tell different faces

apart in face matching. This effect is particularly striking in Experiment 6, in which face matching was assessed over the course of 1000 trials. Surprisingly, we found that matching accuracy continuously declined throughout this prolonged task. In fact, mismatch accuracy declined below an accuracy level of 50%, which reflects *chance* in this binary decision task, and showed no signs of reaching a floor level of performance. Thus, by the end of Experiment 6, mismatch face pairs were more likely to be classified as identity matches than mismatches. This is an astonishing finding considering its potential applied relevance. Passport control, for example, requires the routine matching of facial identities on an enormous scale. In the UK alone, hundreds of thousands of people travel through airports every day and are subject to passport identification. If the decline in mismatch accuracy that was observed in the experiments here is also found in such real-life security settings, then the practical problem of face matching may be much graver than was previously thought.

If this proves to be the case, then our data might also provide a potential means to reduce such errors. While we were unable to improve face-matching performance with feedback here, the application of feedback consistently reduced the decline in mismatch accuracy. Moreover, these effects persisted after feedback had been removed again. This raises the possibility of administering such feedback intermittently in practical settings, perhaps for known 'decoy' face identities or particularly clear identity matches, to help security personnel maintain their face matching accuracy. Comparable approaches already exist to improve human performance during security baggage screening at airports. In this field, the importance of providing feedback for maintaining performance has been acknowledged for some time (see, e.g., Harris, 2002) and has led to developments such as Threat Image Projection, in which pictures of threat objects are occasionally inserted electronically into ongoing airport baggage scans to maintain operator

vigilance (see, e.g., von Bastian, Schwaninger, & Michel, 2010; Hofer & Schwaninger, 2005). If field studies confirm that face-matching performance also declines over time in corresponding security tasks outside the laboratory, then the application of a similar approach to person identification, administered with immediate feedback, might provide a potential method for maintaining accuracy in these settings.

References

Anderson, R. C., Kulhavy, R. W., & Andre, T. (1971). Feedback procedures in programmed instruction. *Journal of Educational Psychology, 62*, 148–156.

Annett, J. (1969). *Feedback and human behavior*. Middlessex, England: Penguin Books.

Bindemann, M., & Sandford, A. (2011). Me, myself, and I: Different recognition rates for three photo-IDs of the same person. *Perception, 40,* 625–627.

Bindemann, M., Avetisyan, M., & Blackwell, K. (2010). Finding needles in haystacks: Identity mismatch frequency and facial identity verification. *Journal of Experimental Psychology: Applied, 16,* 378−386.

Bindemann, M., Avetisyan, M., & Rakow, T. (2012). Who can recognize unfamiliar faces? Individual differences and observer consistency in person identification. *Journal of Experimental Psychology: Applied, 18,* 277-291.

Birney, R. C., Burdick, H., & Teevan, R. C. (1969). *Fear of failure*. New York, NY: Van Nostrand-Reinhold.

Bonner, L., Burton, A. M., & Bruce, V. (2003). Getting to know you: How we learn new faces. *Visual Cognition, 10,* 527-536.

Bradfield, A. L., Wells, G. L., & Olson, E. A. (2002). The damaging effect of confirming feedback on the relation between eyewitness certainty and identification accuracy. *Journal of Applied Psychology, 87,* 112–120.

Brown, E., Deffenbacher, K., & Sturgill, W. (1977). Memory for faces and the circumstances of encounter. *Journal of Applied Psychology, 62*,311-318.

Bruce, V. (1982). Changing faces: Visual and non-visual coding processes in face recognition. *British Journal of Psychology, 73,* 105–116.

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, 5,* 339-360.

Bruce, V., Valentine, T., & Baddeley, A. D. (1987). The basis of the ¾ advantage in face recognition. *Applied Cognitive Psychology, 1*, 109–120.

Burton, A. M., & Jenkins, R. (2011). Unfamiliar face perception. In A. J. Calder, G. Rhodes, M. H. Johnson & J. V. Haxby (Eds.), *The Oxford handbook of face perception*, Oxford: Oxford University Press.

Burton, A. M., Jenkins, R., & Schweinberger, S.R. (2011). Mental representations of familiar faces. *British Journal of Psychology, 102*, 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,* 243–248.

Chapanis, A. (1964). Knowledge of performance as an incentive in repetitive monotonous tasks. *Journal of Applied Psychology, 48*, 263–267.

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

Cook, D. M. (1968). The impact on managers of frequency of feedback. *Academy of Management Journal*, *11,* 263–277.

Costigan, R. (2007). Identification from CCTV: the risk of injustice. *Criminal Law Review,* 591-608.

Davis, J., & Valentine, T. (2009). CCTV on trial: Matching video images with the defendant in the dock. *Applied Cognitive Psychology, 23,* 482–505.

Deci, E. L. (1972). Intrinsic motivation, extrinsic reinforcement, and inequity. *Journal of Personality and Social Psychology, 22,* 113–120.

Deci, E. L. (1975). *Intrinsic motivation.* New York, NY: Plenum Press.

Douglass, A. B., & Steblay, N. M. (2006). Memory distortion in eyewitnesses: A meta-analysis of the post-identification feedback effect. *Applied Cognitive Psychology, 20,* 859–869.

Duchaine, B., & Nakayama, K. (2004). Developmental prosopagnosia and the Benton Facial Recognition Test. *Neurology, 62,* 1219–1220.

Duchaine, B., & Weidenfeld, A. (2003). An evaluation of two commonly used tests of unfamiliar face recognition. *Neuropsychologia, 41,* 713–720.

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,* 431-439.

Endo, M., Takahashi, K., & Maruyama, K. (1984). Effects of observer's attitude on the familiarity of faces: Using the difference in cue value between central and peripheral facial elements as an index of familiarity. *Tohoku Psychologica Folia, 43,* 23–34.

Fisher, S. L., & Ford, J. K. (1998). Differential effects of learner effort and goal orientation on two learning outcomes. *Personnel Psychology, 51,* 397–420.

Ford, J. K., Smith, E. M., Weissbein, D. A., Gully, S. M., & Salas, E. (1998). Relationships of goal orientation, metacognitive activity, and practice strategies with learning outcomes and transfer. *Journal of Applied Psychology, 83*, 218–233.

Frowd, C. D., Skelton, F., Atherton, C., Pitchford, M., Hepton, G., Holden, L., McIntyre, A., & Hancock, P. J. B. (2012). Recovering faces from memory: The distracting influence of external facial features. *Journal of Experimental Psychology. Applied, 18,* 224-238.

Gibbs, C. B., & Brown, I. D. (1955). *Increased production from the information incentive in a repetitive task*. Cambridge: MRC.

Hackman, J. R., & Oldham, G. R. (1976). Motivation through the design of work: Test of a theory. *Organizational Behavior and Human Performance, 16*, 250–279.

Hancock, P. J., Bruce, V. V., & Burton, A. M. (2000). Recognition of unfamiliar faces. *Trends in Cognition Sciences, 4*, 330–337.

Harmon, L. D. (1973). The recognition of faces. *Scientific American, 229,* 70–82.

Harris, D. H. (2002). How to really improve airport security. *Economics in Design: The Quarterly of Human Factors Applications, 10*, 17-22.

Henderson, J. M., Williams, C. C., & Falk, R. J. (2005). Eye movements are functional during face learning. *Memory & Cognition, 33,* 98–106.

Henderson, Z., Bruce, V., & Burton, A. M. (2001). Matching the faces of robbers captured on video. *Applied Cognitive Psychology, 15,* 445–464.

Hill, H., & Bruce, V. (1996). Effect of lighting on perception of facial surfaces. *Journal of Experimental Psychology: Human Perception and Performance, 22,* 986–1004.

Hill, H., Schyns, P. G., & Akamatsu, S. (1997). Information and viewpoint dependence in face recognition. *Cognition, 62*, 201–222.

Hochberg, J., & Galper, R. E. (1967). Recognition of faces: I. An exploratory study. *Psychonomic Science, 9*, 619-620.

Hofer, F., & Schwaninger, A. (2005). Using threat image projection data for assessing individual screener performance. *Safety and Security Engineering, 82*, 417-426.

Hundal, P. S. (1969). Knowledge of performance as an incentive in repetitive industrial work. *Journal of Applied Psychology, 53,* 224–226.

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

Ilgen, D. R., Fisher, C. D., & Taylor, M. S. (1979). Consequences of individual feedback on behavior in organizations. *Journal of Applied Psychology, 64*, 349–371.

Ivancevich, J. M., Donnelly, J. H., & Lyon, H. L. (1970). A study of the impact of management by objectives on perceived need satisfaction. *Personnel Psychology, 23*, 139–151.

Jenkins, R. & Burton, A. M. (2008). Limitations in facial identification: The evidence. *Justice of the Peace, 172*, 4-6.

Jenkins, R., & Burton, A. M. (2011). Stable face representations. *Philosophical Transactions 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.

Johnston, R. A., Hill, H., & Carman, N. (1992). Recognizing faces: Effects of lighting direction, inversion, and brightness reversal. *Perception, 21*, 365–375.

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

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*, 254–284.

Kulhavy, R. W., & Stock, W. A. (1989). Feedback in written instruction: The place of response certitude. *Educational Psychology Review, 1*, 279–308.

Landsberg, C. R., Astwood, R. S., Van Buskirk, W. L., Townsend, L. N., Steinhauser, N. B., & Mercado, A. D. (2012). Review of adaptive training system techniques. *Military Psychology, 24*, 96–113.

Litchfield, D., Ball, L. J., Donovan, T., Manning, D. J., & Crawford, T. (2010). Viewing another person's eye movements improves identification of pulmonary nodules in chest x-ray inspection. *Journal of Experimental Psychology: Applied, 16,* 251–262.

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

Megreya, A. M., & Bindemann, M. (2009). Revisiting the processing of internal and external features of unfamiliar faces: The headscarf effect. *Perception, 38,*1831−1848.

Megreya, A. M., & Bindemann, M. (2013). Individual differences in personality and face identification. *Journal of Cognitive Psychology, 25*, 30-33.

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

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

Megreya, A. M., Bindemann, M., & Harvard, C. (2011). Sex differences in unfamiliar face identification: Evidence from matching tasks. *Acta Psychologica, 137,* 83–89.

Meinhardt-Injac, B., Persike, M., & Meinhardt, G. (2010). The time course of face matching by internal and external features: Effects of context and inversion. *Vision Research, 50,* 1598-1611.

Meinhardt-Injac, B., Persike, M., & Meinhardt, G. (2011). The context effect in face matching: Effects of feedback. *Vision Research, 51,* 2121-2131.

Mory, E. H. (2004). Feedback research revisited. In D. H. Jonassen (Ed.), *Handbook of research on educational communications and technology.* (2nd ed.) (pp. 745–783). Mahwah, NJ: Erlbaum.

Narciss, S. (2008). Feedback strategies for interactive learning tasks. In J. M. Spector, M. D. Merrill, J. J. G. Van Merriënboer, & M. P. Driscoll (Eds.), *Handbook of research on educational communications and technology* (3rd ed.) (pp. 125-143) Mahwah, NJ: Erlbaum.

Nickerson, R. S. (1965). Short-term memory for complex meaningful visual configurations: A demonstration of capacity. *Canadian Journal of Psychology, 19,* 155–160.

Norman, D. A., & Bobrow, D. G. (1975). On data-limited and resource-limited processes. *Cognitive Psychology, 7,* 44-64.

Özbek, M., & Bindemann, M. (2011). Exploring the time course of face matching: Temporal constraints impair unfamiliar face identification under temporally unconstrained viewing. *Vision Research, 51,* 2145-2155.

Palmer, M. A., Brewer, N., & Weber, N. (2010). Postidentification feedback affects subsequent eyewitness identification performance. *Journal of Experimental Psychology. Applied, 16,* 387–398.

Sekiguchi, T. (2011). Individual differences in face memory and eye fixation patterns during face learning. *Acta Psychologica, 137*, 1-9.

Shepard, R. N. (1967). Recognition memory for words, sentences, and pictures. *Journal of Verbal Learning and Verbal Behavior, 6,* 156‑163.

Shute, V. (2008). Focus on formative feedback. *Review of Educational Research, 78,* 153–189.

Sinha, P., & Poggio, T. A. (1996). Role of learning in three-dimensional form perception. *Nature, 384*, 460‑463.

Sinha, P., & Poggio, T. A. (2002). High-level learning of early visual tasks. In M. Fahle & T. Poggio (Eds.), *Perceptual learning.* Cambridge, MA: MIT Press.

Standing, L. (1973). Learning 10,000 pictures. *Quarterly Journal of Experimental Psychology, 25,* 207-222.

Standing, L., Conezio, J., & Haber, R. N. (1970). Perception and memory for pictures: Single-trial learning of 2500 visual stimuli. *Psychonomic Science, 19*, 73-74.

von Bastian, C. C., Schwaninger, A., & Michel, S. (2010). Colour impact on security screening. *IEEE Aerospace and Electronic Systems Magazine, 25*, 33-38.

Wells, G. L., & Bradfield, A. L. (1998). "Good, you identified the suspect": Feedback to eyewitnesses distorts their reports of the witnessing experience. *Journal of Applied Psychology, 83,* 360–376.

Wells, G. L., & Bradfield, A. L. (1999). Distortions in eyewitnesses' recollections: Can the postidentification-feedback effect be moderated? *Psychological Science, 10,* 138–144.

White, R. W. (1959). Motivation reconsidered: The concept of competence. *Psychological Review, 66*, 297–333.

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

Young, A. W., Hay, D. C., McWeeny, K. H., Flude, B. M., & Ellis, A. W. (1985). Matching familiar and unfamiliar faces on internal and external features. *Perception, 14,* 737-746.

FIGURE 1. Examples of the face pairs used in Experiment 1, depicting an identity mismatch (A) and match (B). All stimuli were sourced from the Glasgow University Face Database (Burton, White, & McNeill, 2010).

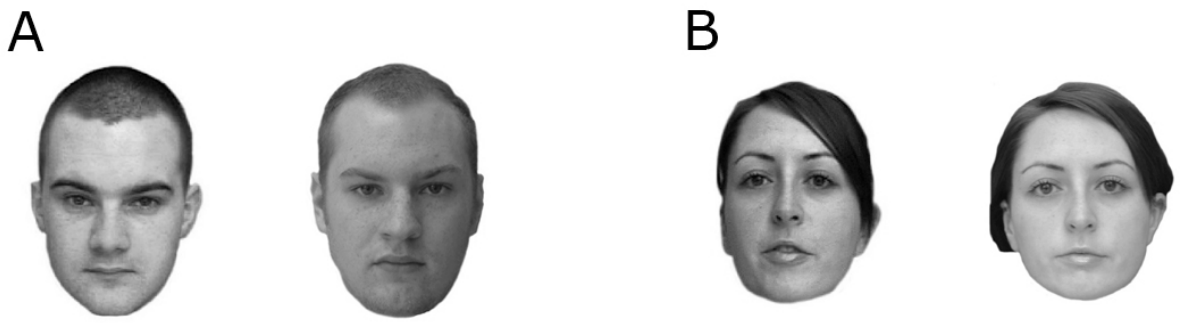

FIGURE 2. An illustration of the block procedure. In the *feedback* condition, initial face-matching performance was measured in the baseline (BL) block, before observers were given trial-by-trial feedback for three consecutive blocks (FB1, FB2, and FB3). The effect of feedback was then measured with three more blocks; for face pairs that were previously presented without feedback, in the baseline test (BLT); for face pairs that were previously presented with feedback, in the feedback test (FBT); and with a block of new faces, in the new faces test (NFT). In the *no feedback* condition, the same seven blocks were administered, except that no feedback was provided during the presentation of the face pairs in blocks FB1, FB2 and FB3.

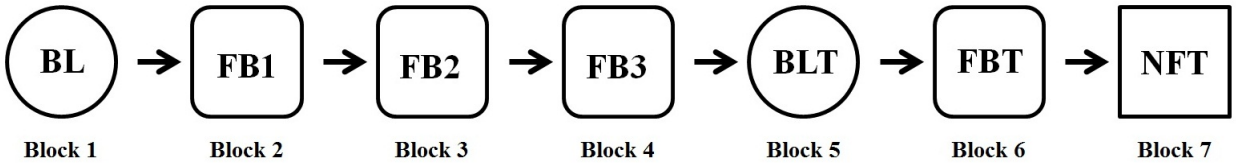

FIGURE 3. Face-matching performance for the *no feedback* condition, shown in row A, and the trial-by-trial *feedback* condition, shown in row B, in Experiment 1. Individual graphs show percentage accuracy (left column) and response times (right column), while open symbols denote match trials and grey-filled symbols denote mismatch trials. Row C provides *d'* and *criterion* measures, with circles denoting the *no feedback* condition and squares the *feedback* condition.

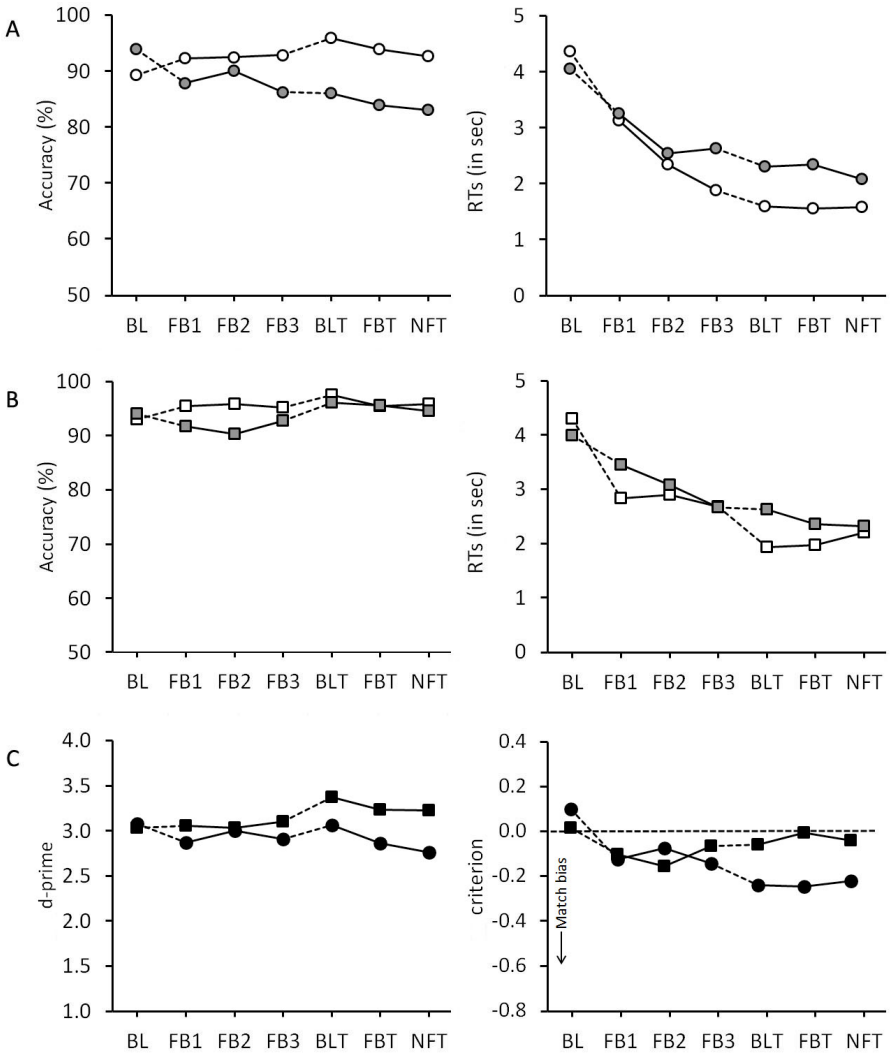

FIGURE 4. Examples of the face pairs used in Experiment 2, depicting an identity mismatch (A) and match (B).

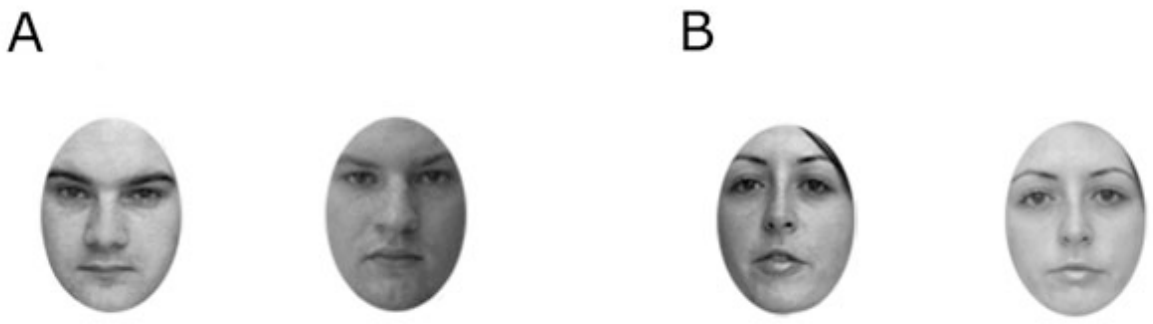

FIGURE 5. Face-matching performance for the *no feedback* condition, shown in row A, and the trial-by-trial *feedback* condition, shown in row B, in Experiment 2. Individual graphs show percentage accuracy (left column) and response times (right column), while open symbols denote match trials and grey-filled symbols denote mismatch trials. Row C provides *d'* and *criterion* measures, with circles denoting the *no feedback* condition and squares the *feedback* condition.

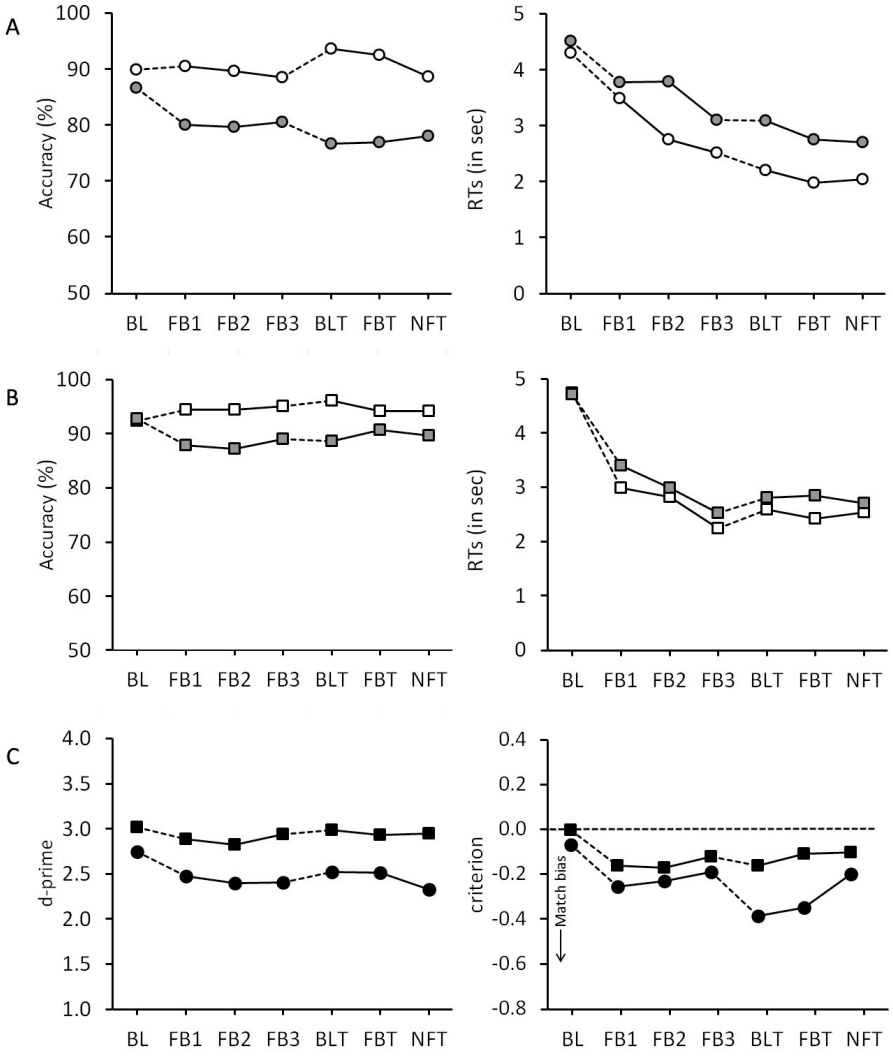

FIGURE 6. Face-matching performance for the *no feedback* condition, shown in row A, and the trial-by-trial *feedback* condition, shown in row B, in Experiment 3. Individual graphs show percentage accuracy (left column) and response times (right column), while open symbols denote match trials and grey-filled symbols denote mismatch trials. Row C provides *d'* and *criterion* measures, with circles denoting the *no feedback* condition and squares the *feedback* condition.

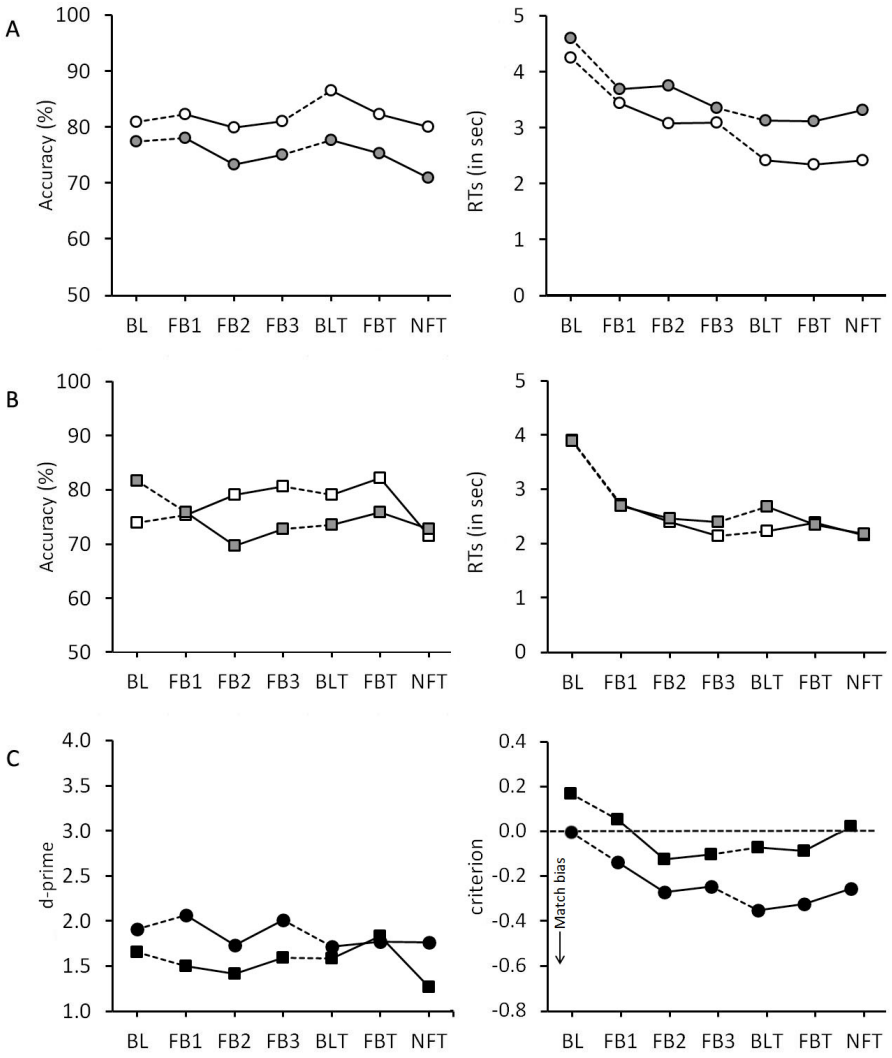

FIGURE 7. Face-matching performance for the outcome *feedback* condition, shown in row A, in Experiment 4. Individual graphs show percentage accuracy (left column) and response times (right column). In addition, row B provides *d'* and *criterion* measures for this *feedback* condition (denoted by squares) and, for comparison, the corresponding data from the *no feedback* condition of Experiment 3 (circles).

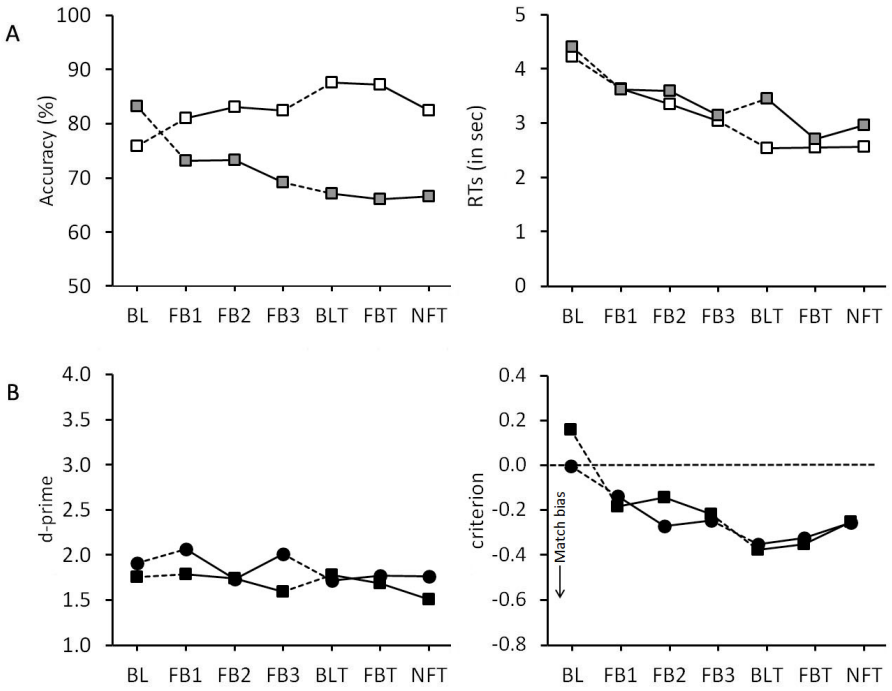

FIGURE 8. Face-matching performance for the outcome *feedback* condition, shown in row A, in Experiment 5. Individual graphs show percentage accuracy (left column) and response times (right column). In addition, row B provides *d'* and *criterion* measures for this *feedback* condition (denoted by squares) and, for comparison, the corresponding data from the *no feedback* condition of Experiment 3 (circles).

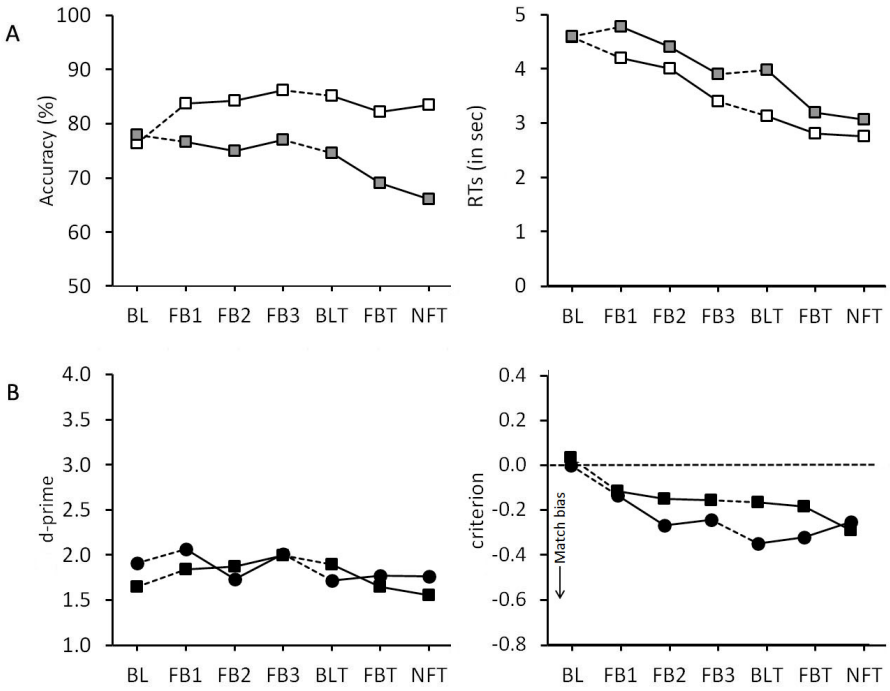

FIGURE 9. Face-matching performance for Experiment 6. Individual graphs show percentage accuracy, response times, *d'* and *criterion*. Open symbols denote match trials and grey-filled symbols denote mismatch trials.