Journal of Experimental Psychology: Applied
2008, Vol. 14, No. 4, 364–372

1076-898X/08/$12.00 DOI: 10.1037/a0013464

Matching Faces to Photographs: Poor Performance in Eyewitness Memory (Without the Memory)

Ahmed M. Megreya
Menoufia University

A. Mike Burton
University of Glasgow

Eyewitness memory is known to be fallible. We describe 3 experiments that aim to establish baseline performance for recognition of unfamiliar faces. In Experiment 1, viewers were shown live actors or photos (targets), and then immediately presented with arrays of 10 faces (test items). Asked whether the target was present among the test items, and if so to identify the person, participants showed poor performance levels (roughly 70% accurate). Furthermore, there was no difference between immediate memory for a live person and photograph. In Experiment 2, the same targets and test items were presented simultaneously, and participants were asked to perform a matching task. Again, performance was poor (roughly 68% accurate), with no difference between matching photos and live people. In the final experiment, viewers were asked to match a live person to a single photograph. Even under these conditions, performance was poor (c. 85%), with no advantage over matching 2 photographs. We suggest that problems of eyewitness identification may involve difficulties in initial encoding of unfamiliar faces, in addition to problems of memory for an event.

Keywords: face recognition, matching unfamiliar faces, eyewitness memory

Courts in many countries place strong reliance on eyewitness identification. However, there is a very large literature demonstrating that eyewitness identification is highly error prone (for reviews, see Cutler & Penrod, 1995; Steblay, Dysart, Fulero, & Lindsay, 2001, 2003; Wells, 1993). Laboratory studies in which perpetrators are seen on video (e.g., Bradfield, Wells, & Olson, 2002; Memon & Bartlett, 2002; Semmler, Brewer, & Wells, 2004) and field experiments in which targets are experienced "live" (e.g., Pryke, Lindsay, Dysart, & Dupuis, 2004; Wells, Rydell, & Seelau, 1993; Yarmey, 2004; Yarmey, Yarmey, & Yarmey, 1996) as well as archival studies with actual criminal cases (e.g., Behrman & Davey, 2001; Fahsing, Ask, & Granhag, 2004; Wells & Seelau, 1995; Wright & McDaid, 1996) have consistently demonstrated eyewitness fallibility. Furthermore, it has been reported that a very large proportion of cases of wrongful imprisonment, in which the accused were subsequently exonerated by DNA evidence, involved eyewitness misidentifications (Huff, Rattner, & Sagarin, 1986; Scheck, Neufield, & Dwyer, 2000; Wells et al., 1998).

To address these problems, a large body of work has investigated the sources of mistaken identification, examining many variables such as the age of eye-witnesses, the race of perpetrators, the presence of weapons in crime situations, and the systems involved in line-up construction (for reviews, see Lindsay & Pozzulo, 1999; Memon, Vrij, & Bull, 2003; Narby, Cutler, & Penrod, 1996; Wells, Wright, & Bradfield, 1999; Westcott & Brace, 2002). Despite this large amount of research, one important question is rarely addressed. Specifically, what is the baseline level of performance one might expect for unfamiliar face recognition, in optimal situations with minimal memory requirement? An emphasis on memory in eyewitness research is natural because real incidents involve witnesses' memory, often over considerable time periods. However, it is also possible that part of the difficulty in identifying an unfamiliar person is tied to perception of unfamiliar faces, and not simply to general memory limitations. In this paper we examine viewers' ability to identify previously unfamiliar faces under minimal memory loads. In Experiment 1, we examine performance under immediate memory conditions, and in Experiments 2 and 3, we examine performance in a matching task, involving no requirement for participants to remember events.

Early face recognition research proposed that people are experts in recognizing unfamiliar faces. For example, recognition memory rates of more than 90% have been reported (e.g., Hochberg & Galper, 1967; Nickerson, 1965; Yin, 1969). More important however, it is now known that this high accuracy represents memory for the images of faces, rather than the faces themselves. In a recognition memory study, Bruce (1982) found that recognition rates dropped from 90% correct when the same images were used in study and test, to only 60% when different images were used. Indeed, changing images has a large detrimental effect on the recognition of identity, even when the task is based on matching, rather than memory. Bruce et al. (1999) showed participants arrays containing a target face and 10 further faces; all images were taken on the same day, but targets and test faces were photographed with different cameras. In half the trials the target was present, and in half absent, and participants were asked to pick the target if he was present. In this seemingly straightforward task, Bruce et al. found a surprisingly low level of performance: Error rates of roughly 30% occurred both for target-present and target-absent arrays.

Poor performance in matching unfamiliar faces has been replicated under different task constraints, for example when the task is reduced to a 10 alternative forced choice (10AFC) with no target-

Ahmed M. Megreya, Department of Psychology, Menoufia University; A. Mike Burton Department of Psychology, University of Glasgow.

Correspondence concerning this article should be addressed to Mike Burton, Department of Psychology, University of Glasgow, Glasgow G12 8QQ, United Kingdom. E-mail: mike@psy.gla.ac.uk

absent arrays (Bruce et al., 1999; Burton, Miller, Bruce, Hancock, & Henderson, 2001), or when the heavy demands of the 1 in 10 arrays were reduced to a 1 in 2 task (Henderson, Bruce, & Burton, 2001) or to simple match/mismatch pairs (Bruce, Henderson, Newman, & Burton, 2001; Henderson et al., 2001; Megreya & Burton, 2006b; 2007). This strongly suggests that encoding identity from unfamiliar face images is rather a difficult task.

The difficulty of unfamiliar face matching has been explained by suggesting that this task relies on pictorial, or image-based, processes; whereas recognition of familiar faces engages a more specialized, and robust, type of processing (for a review, see Hancock, Bruce, & Burton, 2000). For example, Megreya and Burton (2006b) found no correlation between viewers' ability to match familiar faces and their ability to match unfamiliar faces. Furthermore, matching inverted faces was the best predictor of individual differences in unfamiliar face matching, regardless of whether the inverted faces were familiar or unfamiliar. These results suggest that the processes involved in upright unfamiliar face processing appear to be qualitatively similar to those underlying the recognition of inverted familiar and unfamiliar faces, but very different to those responsible for upright familiar face processing. This dissociation between familiar and unfamiliar face processing and this surprising association between matching upright unfamiliar faces and inverted familiars was further replicated by a more recent study (Megreya & Burton, 2007), in which we found that a short familiarization procedure was successful in producing the mirror effect (a negative correlation between hits and false positives) in matching upright unfamiliar faces. However, this effect disappeared when targets were presented upside down. Accordingly, we suggest that faces in the unfamiliar face matching task are treated as "images" or "simple visual patterns," and matched on this basis without domain-specific expertise (Megreya & Burton, 2006b).

The work described so far uses photo-matches only. How would these accuracy rates translate to matching live faces to photos, in situations such as passport control? Although most previous work using "live" targets has been in eyewitness memory settings, the few matching studies as do exist suggest similarly poor levels of performance. For example, Kemp, Towell, and Pike (1997) conducted a field experiment to explore matching accuracy of credit-card photographs to their "live" bearers. The experiment was run in a real supermarket, and participants were six highly experienced cashiers, who had to verify the identities of shoppers (a mix of White and non-White, men and women) by matching them to photo-ID mounted on credit cards. The photographs were small in size (2×2 cm), showing a full-face view, and were taken by a color Polaroid passport camera a few days prior to the experiment. Kemp et al. reported very high false positive error rates: between 34% and 64% of fraudulent cards were accepted, depending on the similarity of the photo to the bearer (i.e., 34% errors for matched age, sex, and race).

In this paper, we aim to test an important hypothesis in the area of face recognition. How accurate are viewers' when asked to match high quality photographs to live individuals, under minimal constraints? We report three experiments examining recognition of previously unfamiliar faces in a "live" situation. In the first experiment, using an immediate memory task, participants were shown individual targets either "live" or in a static video image for roughly 30 s. After a 5-s gap, they were shown photo line-ups of 10 faces, 1 of which might be the target. In a second experiment, live versus static image targets were presented simultaneously with the 10-face photo line-ups, and participants had to reject or to find the correct match. In a final experiment, we used a task similar to that used by Kemp et al. (1997), but with a much larger number of participants, and a more homogeneous group of faces. In this task, we asked participants to match a target seen "live" or in static image to a high-quality digital photograph showing the target or a different foil.

Experiment 1

In this experiment, we examined how accurately people could remember unfamiliar faces seen "live," using an immediate memory paradigm. There are many studies demonstrating poor performance following live eyewitness exposure, but these tend to use relatively long intervals between the event and test (e.g., Pryke et al., 2004; Wells et al., 1993; Yarmey, 2004; Yarmey et al., 1996). Here we reduce this to a minimum to establish baseline performance. In previous studies with immediate memory for unfamiliar faces, we have demonstrated poor performance, though these have been based on memory for photographic stimuli (Megreya & Burton, 2006a; 2006b). Here we compare immediate memory for photographs with memory for people seen live.

Targets and Stimuli

For this and all subsequent experiments, photographic stimuli were taken from a large database of Egyptian faces. The database comprises images of 230 volunteers, all young men with no facial hair or distinguishing marks. These young men were volunteers from a graduating class, and were no more or less homogenous than any student cohort. Images show full-face, neutral expression poses, and were taken with a high quality digital camera (Cyber-short Sony, 7.2 Megapixel resolutions). For 130 of these volunteers, the database also contains a 30-s high quality video clip, recorded with a different capture device (Digital 8 Sony Handicam). Volunteers stood in front of a large window providing a high level of natural light and below two nondirectional fluorescent strip lights. Still and video images were taken on the same day, under the same lighting conditions. For the purpose of the present experiments, a full-face, neutral-expression still-image was captured from the video-sequences, providing a set of faces for which two images are available, in similar pose and lighting, but having been derived from two cameras.

This database has been used to construct matching arrays, similar to those used by Bruce et al. (1999). Each of the 130 still images derived from video serves as a target, and for each of these, two 10-item arrays were derived from the still images. The size of all faces was approximately 5×7 cm, captured in 8-bit gray scale at a resolution of 216×298, and jpeg-compressed at 27 pixels per cm. For consistency with previous matching studies (Bruce et al., 1999, 2001), all digital line-up images were cropped carefully using a graphic software to remove any background and clothes, hence eliminating cues other than faces. Target-present and target-absent arrays were constructed such that foils were subjectively similar in appearance, given the rage of natural variability in the student population from which the faces were taken. Examples are shown in Figures 1 and 2. Across target-present arrays, the target

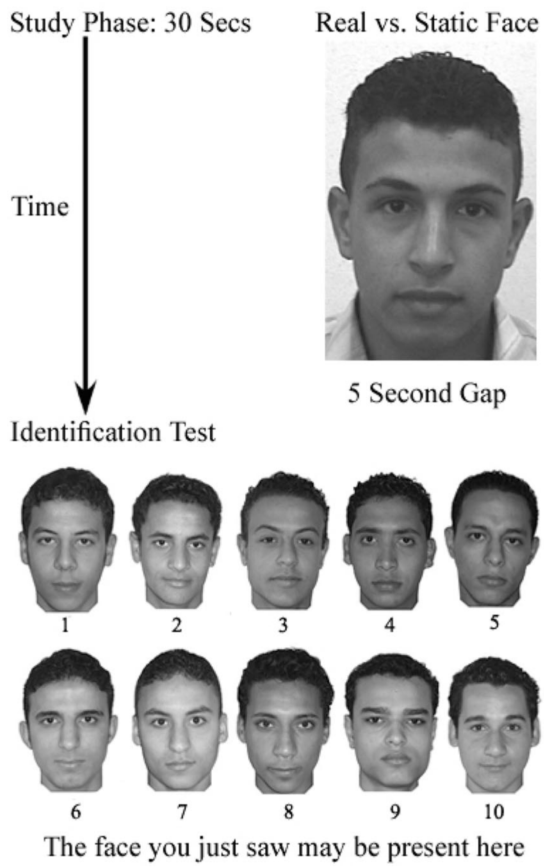

Figure 1. A schematic representation of the procedure used in Experiment 1. In the figure the target is not present. For illustration, all stimulus examples presented were chosen on the basis of the by-item data across participants, such that performance on these stimuli represent overall mean performance.

appears in random locations, and target absent arrays are identical with the exception that the target face has been replaced with another.

Actors participating in the present experiments were 36 volunteers from the face matching database, chosen simply by being available to take part. Their photos were no more than 8 months old at the time of testing. The corresponding 72 arrays (target present and target absent) were used as stimuli in the experiments described here.

Method

Participants. A total of 92 volunteers from the student population of Menoufia University, Egypt participated in the experiment (56 women, 36 men). All participants were Egyptian, and had normal or corrected to normal vision.

Design and procedure. The experiment was conducted in a teaching room in Menoufia University, equipped with data projection facilities. Using a between-participants design, testing was conducted with four separate groups of 23 participants: two groups being shown targets live, and two groups being shown static targets. Seating position of participants was designed so that each had a good full-face view of the live faces.

In the live condition, a target person entered the teaching room, and stood in front of the data projection screen for 30 s. Targets were instructed not to stare at or speak to any of the participants, to keep a neutral expression, and to keep their heads pointing toward a specific location at the back of the room. The target person then left the room, and a 5-s blank slide was presented. Following this, a 10-face image-array was projected onto the screen. Participants had been informed that the target person may or may not be one these 10 (he would be present on 50% of trials). Their task was to decide whether he was present, and if present to identify him by recording the appropriate image number in their response sheets. There was no time constraint for making the identification. Once all participants in the group indicated that they had completed the trial, the next target person was brought into the room. This process was repeated for all 36 target people.

Testing for groups in the image condition was very similar. The sole exception was that targets were presented as static images, rather than live. Participants were presented with a video still of an unfamiliar face for 30 s. After a 5-s gap, they were shown a target-present or target-absent line-up, as above. Figure 1 shows a schematic representation of the procedure.

Two sets of test array slides were prepared to counterbalance the presence/absence of targets. So, across the experiment, each face was followed equally often by a target-present and a target-absent array. Furthermore, note that the same faces were used as live and as image targets, so across the experiment each face was seen equally often live and static. The location of targets among the 10 faces was varied systematically across trials.

Results

Table 1 shows overall recognition performance. Four submeasures are calculated from the array methodology. For target-present items we report: (a) hits (identifying the target); (b) misses (claiming falsely that the target is absent); and (c) misidentifications (Misid; identifying a foil, despite the presence of a target). For target-absent items, we report false positives (FPs; claiming falsely that the target is present). In addition, the overall accuracy is calculated by combining the scores of hits and correct rejections (the complement of FPs). Independent-means t tests showed no differences between identifying faces seen live or through static images on any of these measures. Power for this comparison was high, at 0.66 for effect size $d_z = 0.5$ (calculated using G*Power; Faul, Erdfelder, Lang, & Buchner, 2007). Subsequent analyses showed no effects of participant gender for this or any of the following experiments.

Discussion

In this experiment, participants were asked to view faces of unfamiliar people seen in either static video images or live. After a very short gap, they were given a photo line-up identification test, consisting of 10 high-quality, full-face images, and were informed that targets might or might not be present in the line-ups. The results showed very poor overall performance. Accuracy rates for target-present items were roughly 60%, and for target-absent items roughly 80%. These poor levels of performance replicate our

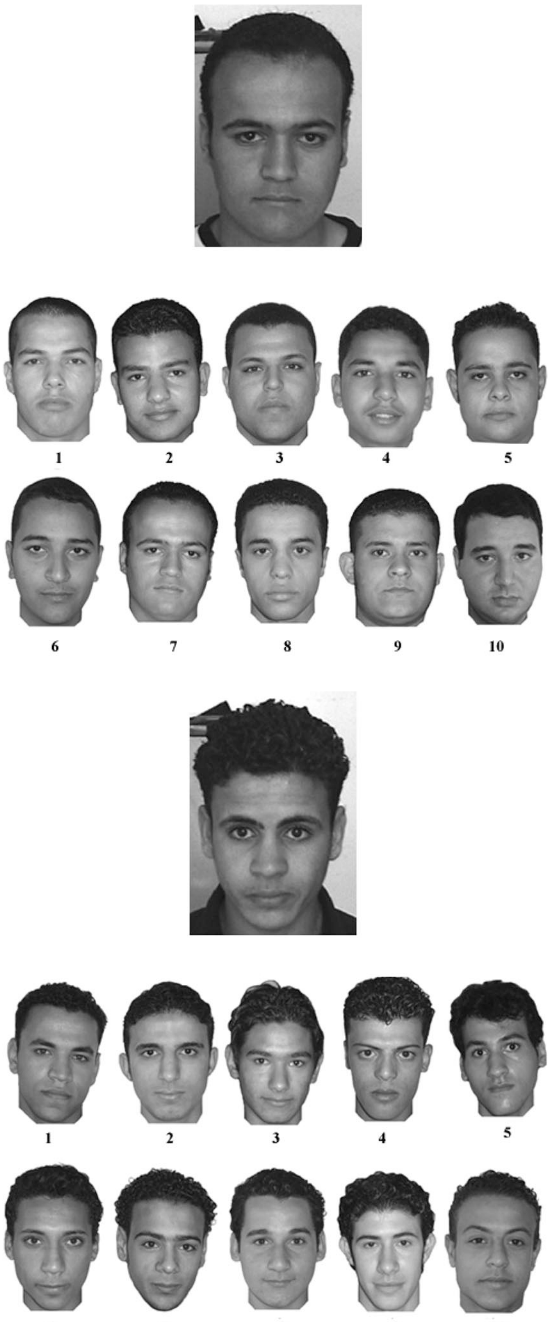

Figure 2. Examples of arrays used in Experiment 2. The person shown at the top may or may not be one of the 10 below. Participants' task is to decide if he is present, and if so, which he is. In this Experiment, half the participants saw the target live and half as an image. In Figure 2a, the target is No. 7, and the target is absent in 2b. For illustration, all stimulus examples presented were chosen on the basis of the by-item data across participants, such that performance on these stimuli represent overall mean performance.

previous research with a different database of photographic targets (Megreya & Burton, 2006a, 2006b), but what is striking about the present results is that there is no improvement at all when presented with live targets.

We should emphasize the simplicity of the experimental context here. In many ways, the present setting represents an ideal case for estimating baseline performance in an identification task. The memory requirement was minimal, with only 5 s between the disappearance of the target and the presentation of the test array. The experiment did not involve many of the factors thought to be significant sources of eyewitness misidentifications, such as long retention intervals (e.g., Behrman & Davey, 2001), biased instructions (e.g., Malpass & Devine, 1981), verbal description of targets (e.g., Schooler & Engstler-Schooler, 1990), emotional stress (e.g., Deffenbacher, Bornstein, Penrod, & McGorty, 2004), an unexpected identification test (e.g., Kerstholt, Raaijmakers & Valenton, 1992), alcohol intoxication (e.g., Dysart, Lindsay, MacDonald & Wicke, 2002), seeing targets with others (e.g., Megreya & Burton, 2006a), aging (e.g., Memon & Bartlett, 2002), disguise (e.g., Patterson & Baddeley, 1977), or degraded environment (e.g., for a review, see Narby et al., 1996). Even in the present optimal conditions, which could never be met in a forensic setting, participants performed very poorly. This suggests that processing of unfamiliar faces, rather than effects on memory or interventions between encoding and test might form a significant component in understanding the difficulty of eyewitness memory. To test this hypothesis, the following experiment eliminates the 5-s memory load altogether, and uses a face matching task.

Experiment 2

Previous laboratory research shows that matching images of unfamiliar faces is a difficult task (Bruce et al., 1999; 2001; Henderson et al., 2001; Megreya & Burton, 2006a, 2006b, 2007). Here, we examined the extent to which this standard finding might extend to real world situations. Experiment 1 showed very similar rates of immediate face memory between image and live conditions. However, perhaps an advantage for live presentation can be shown when participants can continuously examine the real person in the presence of the test images. In this experiment, we repeat Experiment 1 very closely, with the sole difference being that targets and test arrays are presented simultaneously.

Method

Participants. Participants were 100 Egyptian undergraduate students at Menoufia University (63 women, and 37 men). All had

Table 1
Recognition Performance for Static and Live Faces in Experiment 1

	Static		Live		
	M	*SD*	*M*	*SD*	t ($df = 90$)
Overall					
Accuracy (/36)	25.4	4.5	25.5	4.5	.02 ($p = .98$)
Target present					
Hits (/18)	11.1	2.5	11.0	2.4	.21 ($p = .83$)
Miss (/18)	5.1	2.2	5.0	2.6	.22 ($p = .83$)
Misid (/18)	1.8	1.9	2.0	1.8	.56 ($p = .58$)
Target absent					
FPs (/18)	3.7	3.3	3.5	3.5	.18 ($p = .85$)

Note. Misid = misidentification; FPs = false positives.

normal or corrected to normal vision. None of these participants participated in Experiment 1.

Design and procedure. Testing was performed with four groups of participants: 25 participants per group. Targets, stimuli, and procedure were the same as in Experiment 1, except that targets were presented simultaneously with the line-ups.

In the static image condition, a video still face was presented above a line-up of 10 faces until all participants had reached a decision. There was a 2-s interstimulus interval (ISI) before presentation of the next array. Figure 2 shows examples of face matching arrays. In the live condition, each target stood beside the screen, on to which the 10-face test array was projected. As in Experiment 1, each condition consisted of 36 trials: 18 target-present and 18 target-absent and the presence/absence of targets was counterbalanced across the experiment. Participants were informed that targets would be present only on half trials, and were encouraged to perform as accurately as possible.

Results

Table 2 shows mean performance across conditions, with accuracy being divided into separate components, as above. Overall performance was a little above 66% accurate, but there were no differences between photo-to-photo and live-to-photo conditions, on any of the components of performance. This pattern holds in analyses both by-participants and by-items. Once again, the power of these comparisons was high (0.70 by-participant s; 0.83 by-items; for effect size $d_z = 0.5$).

To examine whether poor overall performance was due to a small number of very difficult faces, Figure 3 shows the distribution of response accuracies across the face set. It is clear from these data that some faces are consistently well-recognized, and some are not, but there is no suggestion that the effect arises because of a small number of outlying faces. To establish whether the same faces are consistently recognized in live and photo trials, the by-item data was subject to a Pearson correlation. This showed that the number of participants correctly identifying a face was highly associated across the two types of match (overall accuracy: $r = .63$, $p < .01$; hits: $r = .51$, $p < .01$; false positives: $r = .46$, $p < .01$; all $n = 36$).

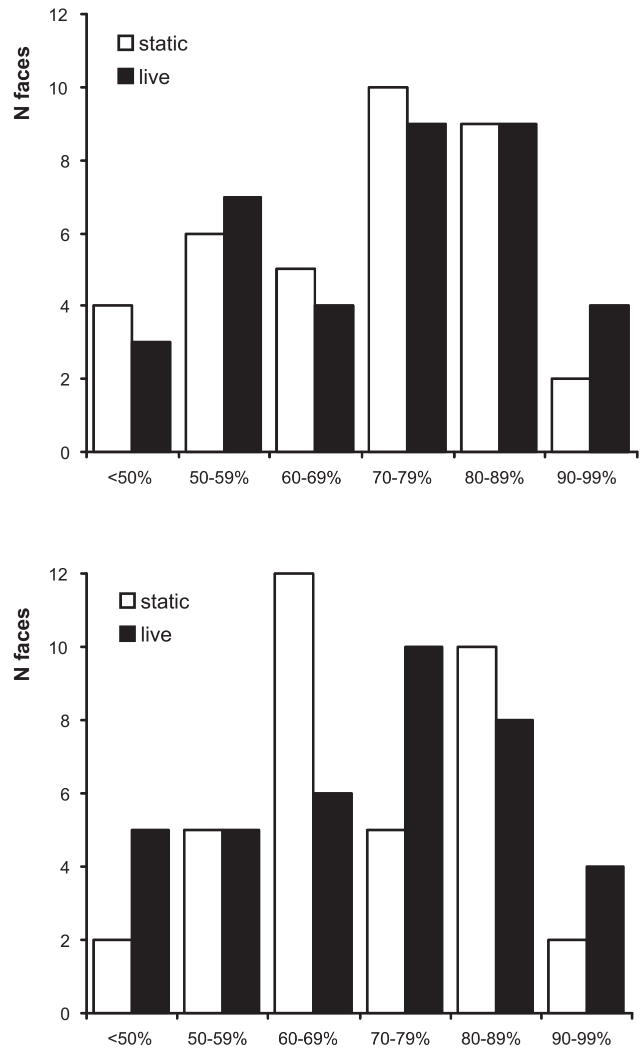

Figure 3. Number of faces recognized with varying levels of accuracy across Experiment 2 (target-present arrays top; target-absent arrays bottom).

Discussion

This experiment examined how well participants could match faces seen live or in static video images using a 1 in 10 matching procedure. Participants' performance in both conditions was very low. When a target was present, participants picked the correct face on only roughly 70% of occasions. When it was absent, they identified a wrong match on about 35% of occasions. These rates are very similar to those previously found using image matches with a different face database (Bruce et al., 1999; Megreya & Burton, 2006a, 2006b, 2007), confirming the fact that matching unfamiliar face images is a very difficult task. What is perhaps more surprising is that there is no advantage for matching live faces as compared to matching photos. Furthermore, it appears that there is quite high correspondence between the faces when presented as photos or live. The high correlation between the formats suggests a degree of consistency across faces across format presentation.

Although matching was generally poor, there were large individual differences in both conditions, as shown by the variances in Table 2. To give examples of the range of performance of partic-

Table 2
Matching Performance for Static and Live Faces in Experiment 2

	Static		Live		*t* (By-participants)	*t* (By-items)
	M	*SD*	*M*	*SD*	(df = 98)	(df = 35)
Accuracy (/36)	24.1	5.5	24.2	4.5	0.18 (p = .86)	0.30 (p = .76)
Hits (/18)	12.8	2.5	13.1	2.8	0.41 (p = .68)	0.50 (p = .62)
Miss (/18)	2.4	1.9	2.0	1.9	1.21 (p = .23)	1.46 (p = .15)
Misid (/18)	2.7	2.4	3.0	2.1	0.53 (p = .60)	0.78 (p = .44)
FPs (/18)	6.8	4.1	6.8	3.5	0.05 (p = .96)	0.08 (p = .93)

Note. Misid = misidentification; FPs = false positives.

ipants: for live targets, hit rates in target-present trials ranged from 44% to 94%. In target-absent trials, the rates of FPs ranged from 5% to 72% across participants. Our previous research with photo matching suggests that these individual differences are likely to be stable over time, reflecting genuine differences in ability with the task (Megreya & Burton, 2006b, 2007). It is interesting to note that these large differences also exist in live matching.

These results are important for the real-life situations in which the procedure of checking photo IDs (as in airports or security sittings) is necessary, and also important for the eyewitness identification field. They seem to suggest that some of the problems associated with eyewitness memory involve initial unfamiliar face encoding, in addition to those well-documented problems associated with memory itself. To emphasize this point, consider the mean misidentification rate in target-present items for the live condition. On average, participants misidentified a target on 3 out of 18 occasions. In other words, in the presence of the live target person, with no time constraints, and good quality images taken within the previous 8 months, participants nevertheless failed to choose the correct photo of the person in front of them, and chose the photo of someone else. This is perhaps a surprising result, though it is consistent with recent work on photographic matching. It does emphasize that viewers are markedly poor at processing unfamiliar faces. This is in stark contrast to their very high ability to recognize familiar faces, even in very poor quality images (e.g., Burton, Wilson, Cowan & Bruce, 1999).

To explore this performance further, the following experiment reduces the heavy demands of the 1 in 10 matching paradigm, instead using a simple match/mismatch task. Once again, we compare live and photographic targets.

Experiment 3

One possible explanation for the generally poor performance in Experiments 1 and 2 is the presence of multiple distractors. Although a choice of 10 possible matches is common in some forensic settings (notably line-ups), it is possible that this in itself puts particular strain on the unfamiliar face processing system. The faces chosen as foils here were not contrived to be similar to the targets; all volunteers for this database came from a graduating university class, and so there will naturally be some similar and some distinctive faces among the set, but no greater homogeneity than in any other set of young men of similar ages. Nevertheless, the large number of distractors may in some way overload the face processing system.

In this experiment we compare faces using a simple two-item match/mismatch task. In the live condition, participants are simply shown a person at the same time as an image, and asked whether the image matches the person. In the control condition participants are shown two different images, and asked whether they are the same person. The live condition therefore represents a very common task, routinely carried out by passport officers and security officers.

Method

Participants. Participants were 100 Egyptian undergraduate students at Menoufia University (68 women, and 32 men). None had taken part in the previous experiments, and all had normal or corrected to normal vision. None of these participants participated in the previous experiments.

Design and procedure. Testing was performed in groups of 25 participants. Targets, stimulus presentation, and procedure were the same as in Experiment 2, with the sole exception that participants had to match a face (real or static) to only 1 photographic image instead of the array of 10 used in Experiment 2. As in previous experiments, there were 36 trials: 18 match and 18 mismatch. Figure 4 shows examples. The real/live comparison was manipulated between participants, and items were rotated around conditions such that each face was seen equally often in match and mismatch trials across the experiment. Participants were self-paced, and were encouraged to perform as accurately as possible.

Results

Table 3 shows mean matching performance in Experiment 3. Overall accuracy showed no significant difference between live and static conditions, either by-participants or by-items. However, once the data are divided into match and mismatch trials, there are reliable differences in each. Participants in the live condition make more hits when the target is present, and more FPs when he is absent than participants in the static condition. Once again, this pattern holds for by-participants and by-items analyses (power = 0.70 by-participants, 0.83 by-items; for effect size $d_z = 0.5$).

Figure 4. Two examples of pairs used in Experiment 3. Participants' task was to decide whether the two faces belong to the same person. Targets were presented either as static images or live. The top pair do not match, whereas the bottom pair match. For illustration, all stimulus examples presented were chosen on the basis of the by-item data across participants, such that performance on these stimuli represent overall mean performance.

Table 3
Matching Performance for Static and Live Faces in Experiment 3

	Static		Live		*t* (By-participants)	*t* (By-items)
	M	*SD*	*M*	*SD*	(df = 98)	(df = 35)
Accuracy (/36)	30.4	2.5	29.9	2.4	1.06 (p = .29)	0.68 (p = .50)
Hits (/18)	15.3	2.1	16.1	2.1	2.11 (p <.05)	2.43 (p <.05)
FPs (/18)	2.8	1.8	4.2	2.3	3.39 (p <.01)	2.07 (p <.05)

Note. FPs = false positives.

Discussion

The present experiment provides good evidence that encoding unfamiliar faces is rather a difficult task. Given no time constraints, nothing to remember, and only a match/mismatch judgment to make, participants nevertheless make over 15% errors. Furthermore, this figure seems to be roughly similar for match and mismatch trials. This is a very high rate. For example, those checking photo-ID in security settings would probably not find this an acceptable error rate.

The results for matching two images replicate previous work (Megreya & Burton, 2006b, 2007). However, the fact that overall accuracy is the same for matching a real person to a photo, as it is for matching two images, is perhaps surprising given the common use of photo-ID. The analysis of hits and FPs shows a response bias in the live condition: the first time we have observed any differences at all between static and live conditions in this set of experiments. The response bias is for participants to claim that two items match. This bias was also reported by Kemp et al. (1997) in their supermarket photo-ID study. If this observation, elicited in experimental settings, reflects a general bias in real-world settings, it may go some way to explaining the well-documented problems of false convictions based on identification errors (Huff et al., 1986; Scheck et al., 2000; Wells et al., 1998).

General Discussion

Recent studies have reported very low accuracy rates for tasks involving identification of previously unfamiliar faces, even when these relied on simultaneous matching, rather than memory (Bruce et al., 1999; 2001; Henderson et al., 2001; Megreya & Burton, 2006a, 2006b, 2007).The main aim of the present paper was to examine the extent to which this standard laboratory finding might extend to a realistic setting. Experiment 1 examined how well people could identify a face shown in a static photo or live using a simple 1 in 10 immediate memory task. Identification rates in these two conditions were very similar, and were very low. More interesting, this same finding was observed when targets were presented simultaneously with the 10-face photo line-ups (Experiment 2). Even when the task was reduced to a simple paired matching task, the overall accuracy of matching a live target was very similar to that for matching two images (Experiment 3). Despite this, participants adopted a lax criterion in the live condition compared to the static one. Seeing targets live significantly increased hits in match pairs and FPs in mismatch trials.

The results have both theoretical and practical implications. Theoretically, it is important to establish the relatively poor levels of unfamiliar face performance, in the context of general approaches to face recognition. In recent years there has been much debate about the specialization of neural mechanisms for face processing, and the degree to which these capture general visual expertise or face-specific processes (e.g., Gauthier & Bukach, 2007; McKone, Kanwisher, & Duchaine, 2007). However, such theoretical approaches often ignore the very large differences between familiar and unfamiliar face processing. Appeals to the notion that human perceivers are face experts rely on our abilities with familiar faces. In fact, we are demonstrably poor at matching tasks with unfamiliar faces, and the experiments presented here demonstrate this using real faces: A demonstration that is difficult to perform because of the practical constraints of conducting such experiments but is nevertheless very important.

In fact, this failure properly to discriminate between familiar and unfamiliar face processing also pervades the computer-based face recognition literature. Engineering approaches to the problem often try to tackle exactly the problem that human perceivers find hardest: How to match two images of an individual in the absence of any further information. We have recently attempted to build familiarity in to automatic face recognition procedures, with quite profound corresponding increases in performance (Burton, Jenkins, Hancock, & White, 2005; Jenkins & Burton, 2008).

These data also have implications for forensic face recognition. It is well established that eyewitness identification is error-prone (e.g., for reviews, see Cutler & Penrod, 1995; Steblay et al., 2001, 2003; Wells, 1993). However, a great deal of research in this field concentrates on factors involving imperfect memory for a witnessed event (e.g., for reviews, see Lindsay & Pozzulo, 1999; Memon et al., 2003; Narby et al., 1996; Wells et al., 1999; Westcott & Brace, 2002). The results described here, in seeking to establish a baseline level of performance, suggest that a significant part of the problem of eyewitness memory may involve problems of unfamiliar face encoding in the first place. The very large literature on how to improve witnesses' memory for an event may therefore benefit from a complementary experimental program on how to improve initial unfamiliar face processing. In previous work (Megreya & Burton 2006a), we demonstrated very large individual differences in facial image matching ability, and the present study suggests that there may be similar variability in live face matching. One way forward to improving accuracy on these tasks would be to establish what psychological factors underlie good versus poor performance.

The implications of the results go beyond eyewitness memory. High levels of security world-wide mean that there is an increasing use of photo-ID cards, and these are viewed by a great range of people, from passport officers to those selling age-restricted goods. The results presented here suggest that even in ideal conditions (no time constraints, good viewing conditions, within-race identifications, and recently taken photographs) matching people to their photos is a difficult task. There remain, of course, many differences between a laboratory setting and a practical identification setting. For example, when checking identity, the checker normally has the opportunity to interact with the person offering the photo ID, an opportunity not available to participants here. In other ways, the laboratory setting might be thought to be easier than a real setting because there is no pressure to make a decision in either direction. If the response bias reported in Experiment 3 turns out to be general in real-world situations (as was the case in the

Kemp et al., 1997 study), then the suggestion is that there will be many future failures to detect fraudulent ID.

References

Behrman, B. W., & Davey, S. L. (2001). Eyewitness identification in actual criminal cases: An archival analysis. *Law & Human Behavior, 25,* 475–491.

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.

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., Henderson, Z., Newman, C., & Burton, A. M. (2001). Matching identities of familiar and unfamiliar faces caught on CCTV images. *Journal of Experimental Psychology: Applied, 7,* 207–218.

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

Burton, A. M., Miller, P., Bruce, V., Hancock, P. J. B., & Henderson, Z. (2001). Human and automatic face recognition: A comparison across image format. *Vision Research, 41,* 3185–3195.

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.

Cutler, B. L., & Penrod, S. D. (1995). *Mistaken identification: The eyewitness, psychology, and the law.* New York: Cambridge University Press.

Deffenbacher, K. A., Bornstein, B. H., Penrod, S. D., & McGorty, E. K. (2004). A meta-analytic review of the effects of high stress on eyewitness memory. *Law & Human Behavior, 28,* 687–706.

Dysart, J. E., Lindsay, R. C. L., MacDonald, T. K., & Wicke, C. (2002). The intoxicated witness: Effects of alcohol on identification accuracy. *Journal of Applied Psychology, 87,* 170–175.

Fahsing, I. A., Ask, K., & Granhag, P. A. (2004). The man behind the mask: Accuracy and predictors of eyewitness offender descriptions. *Journal of Applied Psychology, 89,* 722–729.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39,* 175–191.

Gauthier, I., & Bukach, C. (2007). Should we reject the expertise hypothesis? *Cognition, 103,* 322–330.

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

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

Hochberg, J., & Galper, R. E. (1967). Recognition of faces: 1. An exploratory study. *Psychonomic Science, 9,* 619–620.

Huff, C. R., Rattner, A., & Sagarin, E. (1986). Guilty until proved innocent: Wrongful conviction and public policy. *Crime & Delinquency, 32,* 518–544.

Jenkins, R., & Burton, A. M. (2008). 100% accuracy in automatic face recognition. *Science, 319,* 435.

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

Kerstholt, J. H., Raaijmakers, J. G. W., & Valenton, M. J. (1992). The effect of expectation on the identification of known and unknown persons. *Applied Cognitive Psychology, 6,* 173–180.

Lindsay, R. C. L., & Pozzulo, J. D. (1999). Sources of eyewitness identification errors. *International Journal of Law and Psychiatry, 22,* 347–360.

Malpass, R. S., & Devine, P. G. (1981). Eyewitness identification: Lineup instructions and the absence of the offender. *Journal of Applied Psychology, 66,* 482–489.

McKone, E., Kanwisher, N., Duchaine, B. C. (2007). Can generic expertise explain special processing for faces? *Trends in Cognitive Science, 11,* 8–15.

Megreya, A. M., & Burton, A. M. (2006a). Recognising faces seen alone or with others: When two heads are worse than one. *Applied Cognitive Psychology, 20,* 957–972.

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

Megreya, A. M., & Burton, A. M. (2007). Hits and false positives in face matching: A familiarity based dissociation. *Perception & Psychophysics, 69,* 1175–1184.

Memon, A., & Bartlett, J. (2002). The effects of verbalization on face recognition in young and older adults. *Applied Cognitive Psychology, 16,* 635–650.

Memon, A., Vrij, A., & Bull, R. (2003). *Psychology and law: Truthfulness, accuracy and credibility* (2nd ed.). Chichester, England: Wiley.

Narby, D. J., Cutler, B. L., & Penrod, S. D. (1996). The effects of witness, target, and situational factors on eyewitness identifications. In S. L. Sporer, R. S. Malpass & G. Koehnken (Eds.), *Psychological issues in eyewitness identification* (pp. 23–52). 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.

Patterson, K. E., & Baddeley, A. D. (1977). When face recognition fails. *Journal of Experimental Psychology: Human Learning & Memory, 3,* 406–417.

Pryke, S., Lindsay, R. C. L., Dysart, J. E., Dupuis, P. (2004). Multiple independent identification decisions: A method of calibrating eyewitness identification. *Journal of Applied Psychology, 89,* 73–84.

Scheck, B., Neufield, P., & Dwyer, J. (2000). *Actual innocence: Five days to execution and other dispatches from the wrongly convicted.* New York: Doubleday.

Schooler, J. W., & Engstler-Schooler, T. Y. (1990). Verbal overshadowing of visual memories: Some things are better left unsaid. *Cognitive Psychology, 22,* 36–71.

Semmler, C., Brewer, N., & Wells, G. (2004). Effects of postidentification feedback on eyewitness identification confidence. *Journal of Applied Psychology, 89,* 334–346.

Steblay, N., Dysart, J., Fulero, S., & Lindsay, R. C. L. (2001). Eyewitness accuracy rates in sequential and simultaneous lineup presentations: A meta-analytic comparison. *Law & Human Behavior, 25,* 459–473.

Steblay, N., Dysart, J., Fulero, S., & Lindsay, R. C. L. (2003). Eyewitness accuracy rates in police showup and lineup presentation: A meta-analytic comparison. *Law & Human Behavior, 27,* 523–540.

Wells, G. L. (1993). What do we know about eyewitness identification? *American Psychologist, 48,* 553–571.

Wells, G. L., Rydell, S. M., & Seelau, E. P. (1993). The selection of distractors for eyewitness lineups. *Journal of Applied Psychology, 78,* 835–844.

Wells, G. L., & Seelau, E. P. (1995). Eyewitness identification: Psychological research and legal policy on lineups. *Psychology, Public Policy & Law, 1,* 765–791.

Wells, G. L., Small, M., Penrod, S., Malpass, R. S., Fulero, S. M., & Brimacombe, C. A. E. (1998). Eyewitness identification procedures: Recommendations for lineups and photospreads. *Law & Human Behavior, 22,* 1–39.

Wells, G. L., Wright, E. F., & Bradfield, A. L. (1999). Witnesses to crime: Social and cognitive factors governing the validity of people's reports. In R.

Roesch, S. D. Hart, & J. R. P. Ogloff (Eds.), *Psychology and law: The state of the discipline* (pp. 53–87). New York: Kluwer Academic/Plenum.

Westcott, H., & Brace, N. (2002). Psychological factors in witness evidence and identification. In N. Brace & H. Westcott (Eds.), *Applying psychology* (pp. 117–178). Milton Keynes, England: Open University.

Wright, D. B., & McDaid, A. T. (1996). Comparing system and estimator variables using data from real line-ups. *Applied Cognitive Psychology, 10,* 75–84.

Yarmey, A. D. (2004). Eyewitness recall and photo identification: A field experiment. *Psychology, Crime, & Law, 10,* 53–68.

Yarmey, A. D., Yarmey, M. J., & Yarmey, A. L. (1996). Accuracy of eyewitness identifications in showups and lineups. *Law & Human Behavior, 20,* 459–477.

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

Received June 28, 2007
Revision received June 27, 2008
Accepted July 1, 2008 ■