Behavior Research Methods
2010, 42 (1), 286-291
doi:10.3758/BRM.42.1.286

The Glasgow Face Matching Test

A. Mike Burton and David White
University of Glasgow, Glasgow, Scotland

and

Allan McNeill
Glasgow Caledonian University, Glasgow, Scotland

We describe a new test for unfamiliar face matching, the Glasgow Face Matching Test (GFMT). Viewers are shown pairs of faces, photographed in full-face view but with different cameras, and are asked to make same/different judgments. The full version of the test comprises 168 face pairs, and we also describe a shortened version with 40 pairs. We provide normative data for these tests derived from large subject samples. We also describe associations between the GFMT and other tests of matching and memory. The new test correlates moderately with face memory but more strongly with object matching, a result that is consistent with previous research highlighting a link between object and face matching, specific to unfamiliar faces. The test is available free for scientific use.

Traditional research on face perception has tended to focus on two aspects of the problem: *recognition* of familiar faces and *memory* for unfamiliar faces. Theoretical models, such as that offered by Bruce and Young (1986), have been used for understanding familiar face recognition in typical observers and neuropsychologically impaired patients. Research on face memory, on the other hand, has tended to be led by difficult forensic problems, such as eyewitness testimony (e.g., Lane & Meissner, 2008; Malpass & Devine, 1981; Searcy, Bartlett, & Memon, 1999; Wells & Olson, 2003).

In recent years, it has become clear that *unfamiliar face matching* is a problem worthy of study in its own right. At first glance, this might appear to be a simple problem, but recent research has shown that matching unfamiliar faces is, in fact, rather difficult, even when high-quality images are used. Bruce et al. (1999) presented viewers with 1-in-10 arrays, in which a photo of a young man was accompanied by 10 possible matches. All the images were shown in a very similar pose (full face) and in good lighting and had been taken on the same day, eliminating transient differences due to hairstyle, weight, and so forth. Crucially, target and array photos were taken with different cameras (one a high-quality video camera and one a studio film camera). Under these seemingly optimal conditions, with no time constraints, and with instructions emphasizing accuracy, viewers performed surprisingly poorly. They were accurate only 70% of the time, for both target-present and target-absent arrays. This basic finding has been replicated many times and has been extended to situations in which only target-present arrays were shown, reducing the problem to a 1-in-10 forced choice, and in which viewers scored only 80% accurate (Bruce, Henderson, Newman, & Burton, 2001). These accuracy rates have also been replicated using an entirely different stimulus set, Egyptian young men as targets, with Egyptian students as viewers (Megreya & Burton, 2008).

In subsequent studies, researchers have used simple pairs of faces to measure matching ability (Clutterbuck & Johnston, 2002; Megreya & Burton, 2006, 2007). Under these circumstances, similarly poor matching rates have been observed. Typically, people have found it surprisingly difficult to match two images of an unfamiliar person, making between 10% and 25% errors, depending on the particular stimulus sets that were used. These error rates have never been experienced in matching familiar faces, where ceiling levels of performance have been observed (see Hancock, Bruce, & Burton, 2000). Indeed, a series of experiments by Clutterbuck and Johnston (2002, 2004, 2005) showed that the ability to match images of faces was a very good indicator of the viewer's level of familiarity with a face and improved predictably with increased exposure to the person depicted.

All the studies listed above employed photo-to-photo matching, rather than live-person-to-photo matching. There are a number of security-related situations in which photo-to-photo matching is important—for example, when one tries to match an image of a suspect to a surveillance camera image from a crime scene. However, it is also becoming increasingly common to ask viewers to match photos to live faces. Matching a photo to a face is required not only for passport control, but also in more commonplace settings, such as verifying one's age in order to buy alcohol. Two studies have recently demon-

A. M. Burton, mike@psy.gla.ac.uk

Figure 1. Example test items from the Glasgow Face Matching Test. (A) Mismatching pair. (B) Matching pair.

strated that matching a live person to a photo is no easier than matching two photos of the same person (Davis & Valentine, 2009; Megreya & Burton, 2008). This suggests that the psychological study of face matching addresses a problem of practical, as well as theoretical, consequence.

A TEST FOR FACE MATCHING

There are a number of tests of face recognition ability already available. However, many of these measure face memory rather than matching—for example, the Recognition Memory Test for faces (Warrington, 1984) and the Cambridge Face Memory Test (Duchaine & Nakayama, 2006). Of the available instruments for measuring matching ability, the Benton test is the most commonly used (Benton, Hamsher, Varney, & Spreen, 1983). This test requires participants to match faces across different views. However (and crucially), all images are taken with the same camera. The test we present here tackles a different problem: matching two images in the *same* view but taken with *different* cameras. No existing test of face processing incorporates this task, perhaps because it has only relatively recently become clear that it is nontrivial. Moreover, the issue of camera change is an important one in forensic settings and in everyday verification of photo ID. We have argued that it introduces important variability that discriminates familiar from unfamiliar face processing (Burton, Jenkins, Hancock, & White, 2005; Jenkins & Burton, 2008).

To summarize, the test of face matching described in the remainder of this article is intended to complement existing tests of face processing, rather than to replace any existing tests. It measures performance on a task that is not trivially easy and has been shown to correlate well with levels of familiarity. Furthermore, it mimics a situation that is commonly encountered in security settings: how to match two unfamiliar face images in similar poses but taken with different cameras.

Test Construction

To build a new database of faces, volunteers were recruited through advertising posters in student recreation areas of a university. Three hundred four individuals contributed their time in exchange for a small payment. They were 172 men and 132 women, with the mean age for men being 22.9 years (SD = 6.7), and for women 23.2 years (SD = 7.0). Over the course of a single session, each volunteer was photographed in a variety of poses, using two different digital cameras. Volunteers were also filmed moving between poses and expressions, using a digital video camera. Thus, for each volunteer, we have images from three different capture devices taken on the same day. This large database continues to expand with new volunteers and is available from the authors on request (see the Note for details).

The Glasgow Face Matching Test (GFMT) comprises 168 pairs of faces. For the construction of the test, only

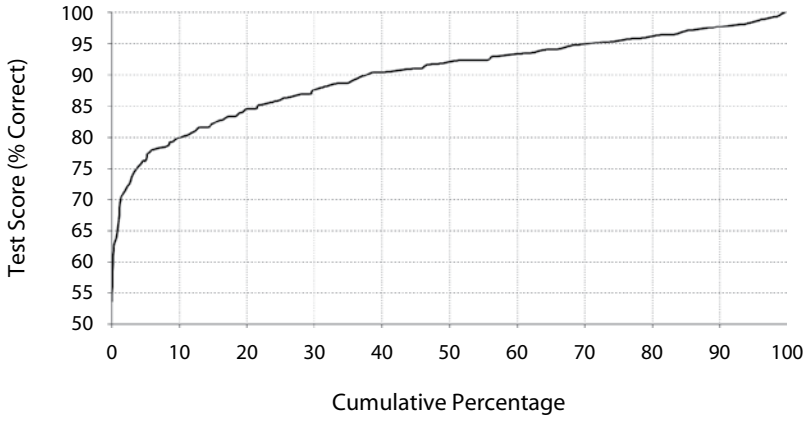

Figure 2. Cumulative frequency of accuracies for the Glasgow Face Matching Test.

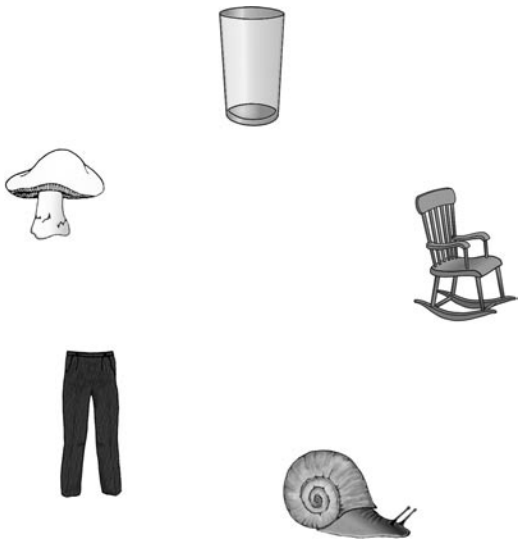

Figure 3. Example array from the visual short-term memory test.

full-face poses were used, in which volunteers displayed a neutral expression. For each person, we used the full-face image from one of the still cameras (Camera 1: Fujifilm FinePix 0800Zoom, 6 megapixel) and a frame in the same pose taken from the video camera (Camera 2: Panasonic NV-DS29B DS29). All images were captured against a background screen, from a distance of 90 cm. The fixed sequence of the photographic session ensured that these two images were taken roughly 15 min apart.

Following image capture, all the photos were edited to remove the background and any visible clothing. Images were cropped neatly around the head, using graphical software, and were resized to 350 pixels width, before being stored in grayscale at a resolution of 72 ppi. When pairs of stimuli were constructed for the test, faces were positioned in such a way that the horizontal distance between the bridge of the nose in the two images was 500 pixels.

Of the 168 test pairs, half are same-face trials, in which two images of the same person are presented side by side. These 84 people are also used in different-face trials, such that one of the person's images is presented alongside a similar face from the database. The nonmatching faces for these trials were chosen on the basis of a pilot study in which pairwise similarity measures were generated using a sorting technique (see Bruce et al., 1999). The foils for these trials were the faces most similar to each of the target identities. For *different* trials, as with *same* trials, the two photos always came from different cameras. Figure 1 shows examples of face pairs.

Performance on the Test

Subjects. Following initial pilots, the GFMT was presented to 300 subjects. This was a relatively heterogeneous sample, recruited through advertisements in the local media. There were 120 males and 180 females. Mean age was 30.8 years, with a range of 18–80 and a standard deviation of 14.

Performance. Overall accuracy ranged from 62%–100%, with a mean of 89.9% ($SD = 7.3$). Performance was slightly better on matching items (92%) than on mismatching items (88%), indicating a small response bias to respond *same*. Couched as detection measures, this gives a d' value of 2.91, with a criterion of −0.09. With this large sample size, criterion is significantly below zero [$t(299) = 4.69$, $p < .01$]. There was no correlation between accuracy and age of viewer ($r = .09$),[1] and there was no performance difference between men and women [male 89%, female 90.4%; $t(298) = 1.53$, n.s.]. In order to measure the internal reliability of the test, we examined the split-half association by correlating the subjects' performance on the first and second halves of the test items. Association was high, with $r = .81$.

Figure 2 gives the cumulative distribution of accuracies and may easily be used to establish the norm of any score against this population. As one might predict for a test of this kind, the distribution is negatively skewed (skewness = −1.33, $p < .05$). However, it is interesting to note that performance is far from perfect. Recall that the test requires the observer to match two photos of a person taken minutes apart, in the same pose, with two high-quality cameras. If we consider that the median performance is 92%, this means that half the sample make at least 8% errors—that is, 13 items wrong across the 168 items in the test. Similarly, the poorest 25% made at least 24 matching errors. In a test with no time limits, in which accuracy is emphasized, this is perhaps surprising, although it is consistent with our previous work showing rather poor levels of performance on unfamiliar face matching.

Finally, we note that the mean time to complete the self-paced test was 15 min and that there was a small, but reliable, positive correlation between overall accuracy and time taken ($r = .177$, $p < .01$).

ASSOCIATION BETWEEN THE GFMT AND OTHER TESTS OF FACE AND OBJECT PROCESSING

The matching test described above reveals substantial individual differences in a task that, at first glance, might appear relatively easy. In order to establish whether this variation reflects more general variation in visual-processing abilities, we also examined our subjects' performance on three more commonly used tests of visual matching and memory. Each of the 300 subjects who took part in the study above also contributed measures on three further tests: (1) recognition memory for faces, (2) the Matching Familiar Figures Test (MFFT), and (3) a visual short-term memory test.

Table 1
Performance on Four Tests of Matching and Memory

	GFMT	Recognition Memory for Faces	Matching Familiar Figures	Visual Short-Term Memory
Mean (% correct)	89.9	62.4	66.3	62.9
SD	7.3	10.0	21.9	9.4

Table 2
Correlations Between Tests: Pearson's *r*

	Recognition Memory for Faces	Matching Familiar Figures	Visual STM	Age
GFMT	.285**	.420**	.050	.090
Recognition memory for faces	–	.158*	.186*	−.209**
Matching Familiar Figures Test	–	–	.176*	−.023
Visual STM	–	–	–	−.177*

Note—STM, short-term memory; GFMT, Glasgow Face Matching Test. *$p < .01$. **$p < .001$.

1. *Recognition memory for faces*. For this test, a further 40 people's faces from the same database were used (20 men and 20 women). Images were prepared in exactly the same way as described above, were presented to the subjects in grayscale, at the same size and resolution as those in the GFMT, and were cropped of background in the same way.

To test recognition memory, the subjects were shown images of 20 of the faces, all taken with Camera 1. The subjects sat in front of a computer screen and were instructed to pay close attention to the faces, since they would be asked to identify them later. The images appeared in sequence for 2 sec each, preceded by a fixation cross for 750 msec. Once all 20 images had been presented, a message appeared instructing the subjects to wait for further instructions. After a 20-sec interval, test phase instructions appeared. During test, the viewers were presented with 40 faces, all taken with Camera 2 (i.e., not the same camera as that used for images in the first phase). They were told that they should decide, independently for each face, whether it had appeared in the earlier phase. Testing was self-paced.

2. *Matching Familiar Figures Test*. The MFFT is a common technique for measuring cognitive style, impulsivity versus reflexivity (Kagan, 1965). The test consists of 20 standard line drawings of common objects (targets) and six variants of each object, one of which is identical to the target image. Performance on this test has been shown to correlate with performance on unfamiliar-face-matching tests in previous research using a lineup task (Megreya & Burton, 2006).

3. *Visual short-term memory for objects test*. For this test, circular visual arrays of objects were constructed. Forty-five common objects were taken from the database of Rossion and Pourtois (2004). These were used to create six circular arrays of 5, 6, 7, 8, 9, and 10 objects. An example is given in Figure 3. Testing followed the procedure described by Miller (1956), in his highly influential account of memory span. The subjects were presented with each array in turn, starting with the array with the fewest objects (5 items) and ending with the array with the most objects (10 items). Each array was presented on the screen for 5 sec, after which the subjects were asked to write as many of the items as they could remember on a sheet of paper provided to them.

Results and Discussion

Table 1 shows the overall performance levels for the GFMT and the three tests described here. Table 2 shows the association between the tests (Pearson's *r*), as well as the correlation between performance on the test and the subjects' ages.

There are a number of points to note from these data. First, the highest correlation with the GFMT is the MFFT. This is consistent with the notion that *unfamiliar* faces tend to be processed as general visual objects, without recruiting the perceptual processes that lead to very robust performance with *familiar* faces (e.g., Hancock et al.,

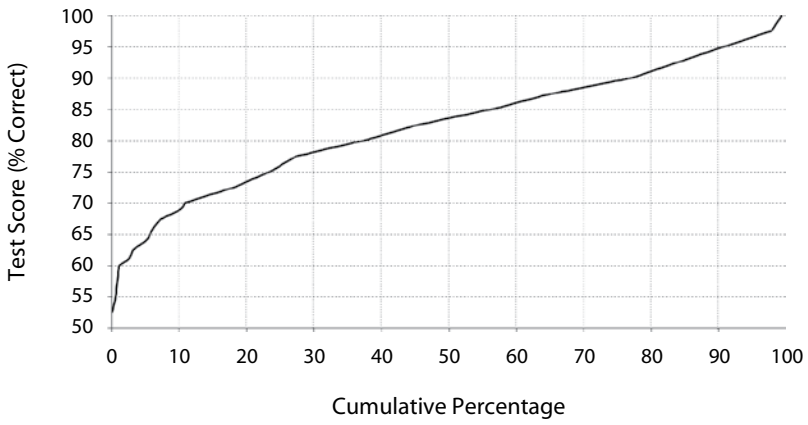

Figure 4. Cumulative frequency of accuracies for the short version of the Glasgow Face Matching Test.

2000; Megreya & Burton, 2006). Note that the high association between the GFMT and MFFT occurs despite some large differences in the format of the tests. Notably, the GFMT involves a yes/no response to pairs of faces, whereas the MFFT involves a lineup of six options. Furthermore, the MFFT contains only target-present items; a match always exists. Nevertheless, there is a striking association here.

There is a smaller association between face matching and face memory, using these tests. Nevertheless, there is a substantial effect here, suggesting some shared processing. Note that the recognition memory test for unfamiliar faces is very difficult ($M = 62\%$, with chance being 50%), in contrast to many similar tests in the literature that use the identical image at learning and at test. This inevitably skews the memory data positively and, therefore, may lead to an underestimation of the correlations with other measures. Nevertheless, it is noticeable that this is the only measure that correlates with all the other tests. Perhaps more interesting is the pattern of associations between the tests and the subjects' ages. It is clear that both tests of memory show a decline in performance with age. This is the case despite large differences in style between the two tests of memory (faces or objects, delayed vs. immediate memory). However, the association with age is completely absent in the two rather different tests of matching. This observation appears interesting and will be followed up in future research.

A SHORT VERSION OF THE GFMT

The full GFMT comprises 168 pairs of faces and is self-paced. We anticipated that some users would prefer a briefer test, and so we developed a shortened version comprising only 40 face pairs. Items for this test were selected as being the most difficult items from the full version. Using data from the test of 300 subjects above, the 20 matching and 20 nonmatching items were chosen that had resulted in the most errors. Scores on this subset of items correlated very highly with overall scores on the full test ($r = .91$), making this a potentially useful version of the test.

The short version of the GFMT was tested on 194 new volunteers, none of whom had taken part in the studies described above. These were young adult subjects with a mean age of 26 years (range, 18–46). There were 121 men and 73 women. The test was run self-paced and typically took between 3 and 4 min to complete, making it appreciably shorter than the full version.

Mean performance on the short test was 81.3%, with $SD = 9.7$ and range = 51%–100%. This is substantially lower than performance on the full test, confirming the choice of difficult items. Mean performance on match and mismatch trials was 79.8% and 82.5%, respectively. Figure 4 shows the cumulative distribution of accuracies and may easily be used to establish the norm of any score against this population. The test is significantly negatively skewed (skewness $= -0.45$, $p < .05$), although rather less so than the full version.

GENERAL DISCUSSION

We have presented a new test for face matching. Unlike other available tests, the GFMT presents two images taken in the same pose, minutes apart, with high-quality cameras. Despite these apparently optimal conditions, this task is not trivially easy, and we have demonstrated that there is large interindividual variation in performance.

We note that modern security measures mean that people are commonly asked to prove their identity with a photograph. Correspondingly, there are very many people whose daily activity requires them to confirm somebody's identity in this way. Previous research has established that unfamiliar face matching is a surprisingly difficult task, and we have recently demonstrated that matching a live person to their photo is no easier than matching two photos (Megreya & Burton, 2008). With this in mind, we have constructed a test that does not make the task artificially difficult—for example, by covering people's hair or requiring a match across different poses. Instead, we have examined a commonplace match, two full-face views in good lighting, in an attempt to mimic situations in which one is trying to *optimize* the accuracy of a photo ID, not to make it difficult.

Given the substantial individual differences in face matching demonstrated here, we anticipate that one potential use of the test may be in personnel selection for particular tasks requiring face matching. There is clearly also a potential for use in training: Since almost no one we tested showed perfect performance, it would be interesting to use difficult items in training regimes. There is also a clear potential for neuropsychological use of the test.

AUTHOR NOTE

This work was supported by Grant 000-23-1348 from the ESRC to A.M.B. and A.M. The full GFMT and the short version are available for download from the authors' Web site at www.psy.gla.ac.uk/gfmt. The test is free for research use, and the download package includes instructions, scoring sheets, and the norm data presented here. All those who volunteered use of their faces for this test have provided written permission for the images to be used for any research purposes, including scientific publication. The full database of images (Glasgow Unfamiliar Face Database) from which the test was derived is available at the same site.
Correspondence concerning this article should be addressed to A. M. Burton, Department of Psychology, University of Glasgow, Glasgow G12 8QQ, Scotland (e-mail: mike@psy.gla.ac.uk).

REFERENCES

BENTON, A. L., HAMSHER, K. S., VARNEY, N. R., & SPREEN, O. (1983). *Contributions to neuropsychological assessment*. New York: Oxford University Press.

BRUCE, V., HENDERSON, Z., GREENWOOD, K., HANCOCK, P., 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.

BRUCE, V., & YOUNG, A. W. (1986). Understanding face recognition. *British Journal of Psychology*, **77**, 305-327.

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.

CLUTTERBUCK, R., & JOHNSTON, R. A. (2002). Exploring levels of face familiarity by using an indirect face-matching measure. *Perception*, **31**, 985-994.

CLUTTERBUCK, R., & JOHNSTON, R. A. (2004). Matching as an index of face familiarity. *Visual Cognition*, **11**, 857-869.

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.

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

DUCHAINE, B., & NAKAYAMA, K. (2006). The Cambridge Face Memory Test: Results for neurologically intact individuals and an investigation of its validity using inverted face stimuli and prosopagnosic participants. *Neuropsychologia*, **44**, 576-585.

HANCOCK, P. J. B., BRUCE, V., & BURTON, A. M. (2000). Recognition of unfamiliar faces. *Trends in Cognitive Sciences*, **4**, 330-337.

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

KAGAN, J. (1965). Reflection-impulsivity and reading ability in primary grade children. *Child Development*, **36**, 609-628.

LANE, S. M., & MEISSNER, C. A. (2008). A "middle road" approach to bridging the basic–applied divide in eyewitness identification research. *Applied Cognitive Psychology*, **22**, 779-787.

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

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

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.

MILLER, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, **63**, 81-97.

ROSSION, B., & POURTOIS, G. (2004). Revisiting Snodgrass and Vanderwart's object set: The role of surface detail in basic-level object recognition. *Perception*, **33**, 217-236.

SEARCY, J. H., BARTLETT, J. C., & MEMON, A. (1999). Age differences in accuracy and choosing in eyewitness identification and face recognition. *Memory & Cognition*, **27**, 538-552.

WARRINGTON, E. K. (1984). *Recognition Memory Test*. Windsor, U.K.: NFER-Nelson.

WELLS, G. L., & OLSON, E. (2003). Eyewitness identification. *Annual Review of Psychology*, **54**, 277-295.

NOTE

1. Previous research (Searcy et al., 1999) suggests that adult age may be more strongly associated with false positives than with hits. However, that association was not present here: Correlations with age were $r = .197$ and $-.023$ for hits and false positives, respectively.

(Manuscript received April 7, 2009;
revision accepted for publication May 24, 2009.)