Article:
White, David, Guilbert, Daniel, Varela, Victor P L et al. (2 more authors) (2022) GFMT2:A psychometric measure of face matching ability. Behavior research methods. pp. 252-260. ISSN: 1554-351X

https://doi.org/10.3758/s13428-021-01638-x

GFMT2: A psychometric measure of face matching ability

David White[1], Daniel Guilbert[2], Victor P. L. Varela[1], Rob Jenkins[3] & A. Mike Burton[3]

[1] School of Psychology, UNSW Sydney, Australia

[2] Department of Psychology, Macquarie University, Australia

[3] Department of Psychology, University of York, United Kingdom

Word Count (not including Abstract, References, Figure or Table captions): 4951

Author Note

Preparation of this chapter was supported by an Australian Research Council Linkage Project (LP160101523), and an Australian Research Council Discovery Project grant to White (DP190100957). We thank Anita Trinh for assistance with data collection. The original Glasgow Face Matching Test was developed as a collaboration between two universities in Glasgow, UK. We retain the original name despite no longer holding affiliations with those universities. The face images used in GFMT2 were created as part of the original collaboration, much of which was led by our friend and colleague, Allan McNeill, 1958-2016.

Correspondence concerning this article should be addressed to David White, School of Psychology, UNSW Sydney, 2052, Australia. Email: david.white@unsw.edu.au

ABSTRACT

We present an expanded version of a widely used measure of unfamiliar face matching ability, the Glasgow Face Matching Test (GFMT). The GFMT2 is created using the same source database as the original test but makes five key improvements. First, the test items include variation in head angle, pose, expression and subject-to-camera distance, making the new test more difficult and more representative of challenges in everyday face identification tasks. Second, short and long versions of the test each contain two forms that are calibrated to be of equal difficulty, allowing repeat tests to be performed to examine effects of training interventions. Third, the short form tests contain no repeating face identities, thereby removing any confounding effects of familiarity that may have been present in the original test. Fourth, separate short versions are created to target exceptionally high performing or exceptionally low performing individuals using established psychometric principles. Fifth, all tests are implemented in an executable program, allowing them to be administered automatically. All tests are available free for scientific use via www.gfmt2.org.

KEYWORDS

face perception; perceptual expertise; facial image comparison; super-recognizers; congenital prosopagnosia; developmental prosopagnosia; unfamiliar face matching; expertise; face recognition.

INTRODUCTION

In face matching tasks viewers compare pairs of face images and decide if they show the same person or different people. Reliable measurement of people's accuracy on this task helps researchers to understand perceptual abilities underlying face identification, provides a tool for clinical neuropsychological assessment and enables recruitment of staff to perform this task in security and forensic settings.

The Glasgow Face Matching Test (GFMT; Burton, White & McNeill, 2010) has become the most commonly used measure of unfamiliar face matching ability. The main motivation to create the original test was to provide a measure of unfamiliar face *matching* ability, as distinct from face *memory* ability, in the general population. Existing tests of face matching had been created for the purpose of neurological assessment of impaired face identification ability and so were not challenging enough to measure the broad range of ability in the general population (Benton, 1983). Prior to the GFMT, the only measures of face identification ability designed for studying the general population involved memorising faces (Cambridge Face Memory Test, Duchaine & Nakayama, 2006), rather than matching images presented together.

At the time the original test was published, researchers had only recently begun to examine individual differences in people's ability to identify faces. Consistent with the data presented in our test of face matching (Burton et al. 2010), early studies of face memory reported large individual differences in people's performance on unfamiliar face identification tasks (Duchene & Nakayama, 2006). However, most of the early work on this topic was focussed on individuals with impaired face identification abilities (see Bowles et al. 2009 for a review). Over the past decade, the study of individual differences in face identification has become a very active research area (see Wilmer, 2017 for a review) and it has become clear that these individual differences reflect a relatively stable cognitive trait (e.g. Wilmer et al. 2010; Balsdon et al. 2019) with a genetic basis (Wilmer et al. 2010; Shakeshaft & Plomin, 2015). Impairments at the low end of the ability spectrum are now known to be mirrored by extreme abilities of 'super-recognisers' at the high end (Russell et al. 2009; see Noyes, Phillips & O'Toole, 2017, Ramon, Bobak & White, 2019 for reviews).

The availability of both memory and matching tests was important because early work had shown that performance in matching tasks does not necessarily predict performance on memory tasks (Megreya & Burton, 2006). Subsequent individual difference studies have mostly confirmed that these abilities are distinct. Although there is substantial correlation between matching and memory performance, there also appears to be portions of variance that are specific to these two tasks (e.g. McCafferty et al. 2018; Wilhelm et al. 2010), and individuals who are impaired in face recognition are not necessarily impaired in matching tasks (e.g. White et al. 2017; Bowles et al. 2009). This partial dissociation suggests that these two tasks target distinct subskills, and so reliable tests of both matching and memory can help delineate component processes underlying face identification ability more precisely.

There are also practical reasons that psychometric measures of face matching are necessary. In many applied settings, people are required to compare images of unfamiliar faces to establish their identity. For example, passport officers are required to match passport images to travellers; in police investigation and criminal trials it is often necessary to compare CCTV images of culprits to known images of suspects. These tasks are not constrained by memory – viewers may examine face images presented simultaneously, without having to commit one to memory, and so it is important for practical reasons to capture this aspect of face identification in a standard test.

Over the past decade, the GFMT has been used alongside other face matching tasks to test the accuracy of people who perform these types of task in their daily work. These assessments span a range of professions. In published studies, the GFMT had been used to test 450 practitioners including passport officers (White et al. 2014; Towler et al. 2019), police officers (Davis et al. 2016), facial forensic examiners (White et al. 2015) and police 'super-recognisers' (Robertson et al. 2015). White, Towler and Kemp (2020) recently presented a meta-analysis of 29 studies that have compared face matching accuracy of face identification practitioners to participants sampled from the general public. Surprisingly, half of those tests show no accuracy difference between professionals and novices, with both groups showing large error rates.

This is problematic because these professionals are entrusted by the public to make accurate face identification decisions. Given the clear evidence that individual differences in face matching tasks are large and stable over time, selecting people that are skilled in face matching provides a promising solution to this problem. Therefore, the GFMT can also be used as a tool for staff selection and recruitment in roles that require people to make face matching decisions (see White et al. 2014, 2015; Davis et al. 2016).

A NEW PSYCHOMETRIC MEASURE OF FACE MATCHING ABILITY

Despite its popularity, the GFMT has a number of limitations and would benefit from an update. Over the period of its use, reported mean performance in the general population has tended to increase. For example, Burton et al (2010) report mean performance of around 82% for the short test, whereas more recent uses often report means of just under 90% (e.g. Towler et al, 2019). High mean accuracy is accompanied by a reduction in variance of high scores, somewhat devaluing the test for certain uses. There are a number of reasons this inflation may have occurred. The test is freely available, and many example items have been published in research papers. As a result, there may be an issue of familiarity for some experimental participants, especially in psychology communities. Furthermore, at the time the original test was developed, the general difficulty of unfamiliar face matching was poorly understood. The fact that matching is hard, even in high-quality images, is now much more widely known, perhaps encouraging participants to take a more studied approach to the task.

A more challenging version of the GFMT is also necessary due to the increased interest in super-recognisers (e.g. Bobak et al. 2016), and professional groups displaying high levels of accuracy in face recognition tasks (Phillips et al. 2018). While researchers have produced challenging tests to address specific research goals (Fysh & Bindemann, 2017; White et al. 2015; Dunn et al. 2020), there is now a need for a standard lab-based test with known psychometric properties that enables comparison across high performers, typical performers and low performers. To make the GFMT2 more challenging than the original test, we select image pairs that require participants to match identity across variations in head angle, pose, expression and subject-to-camera distance. In contrast, the GFMT was

created by pairing two passport-style images of faces in neutral pose, pictured straight on standing directly in front of the camera (see Figure 1). As well as making the task more difficult, this change also captures a wider range of applied tasks that practitioners perform.

Another improvement on the original version is that, for both short and long form tests, we provide two versions that are equated for difficulty. These paired versions enable repeat testing of participants to examine the effectiveness of clinical interventions, professional training and mentorship programs. For example, in a recent test of the effectiveness of professional training, Towler and colleagues (2019) tested participants before-and-after training using short 20 item versions of the GFMT that have been equated for difficulty – finding no evidence of improvement. Such research points to the practical need for more effective methods of training, and evaluation of evidence-based interventions requires common, reliable repeated measures that are equated for difficulty (see Dowsett & Burton, 2015; White et al. 2014; Matthews & Mondloch, 2018; Towler et al. in press; c.f. Bate & Bennetts, 2014; DeGutis et al. 2014).

Finally, the GFMT2 also provides various short forms of the tests, tailored to particular use and selected using established psychometric principles. In the remainder of the paper, we describe the development of these tests. We first describe the GFMT2 Long Form, which consists of two 150-item sub-tests of equal difficulty (GFMT2-A and GFMT2-B). This long form is not intended as the primary measure of face matching ability but was the starting point for selecting test items that maximise desirable psychometric properties in short test versions. The primary measure is the short form of the test (GFMT2-S), and we create two additional short tests that are tailored to the low (GFMT2-Low) and high ends (GFMT2-High) of the performance scale:

(i) The 80-item GFMT2 Short Form (GFMT2-S) comprises two equally difficult 40-item test forms, GFMT2-SA and GFMT2-SB. We anticipate that these will be useful in experimental intervention studies.

(ii) The 40-item GFMT2-Low is designed to discriminate between low performing participants. This version of the test will be useful in assessing acquired or developmental prosopagnosia.

(iii) The 40-item GFMT2-High is designed to discriminate between high performing participants. This version of the test will be useful in assessing super-recognisers and certain professional groups.

The tests are free for scientific use and executable versions are available for download at www.gfmt2.org [*for the purpose of peer review these can be accessed at https://tinyurl.com/gfmt2review*]. Detailed item performance and metadata used to create the short test versions, and normative test scores for long and short tests, are also included in the test distribution [*for the purpose of peer review these can be accessed at: https://tinyurl.com/gfmt2review*].

TEST CONSTRUCTION AND RESULTS

Test item creation, item accuracy screening and long-form test construction

The GFMT2 was constructed using the same source as the original GFMT – the Glasgow Unfamiliar Face Database (GUFD). This database consists of multiple images of 304 people taken on two digital SLR cameras, and a digital video camera. We removed two identities that were determined to be either duplicate entries or twins of existing identities in the database, and another identity who had withdrawn consent for their image to be used, leaving a total of 301 identities.

GFMT2 test items are image pairs that either show the same person (match) or two different people (non-match). We first created a pool of 150 matching and 150 non-matching image pairs for the long form test. To create non-matching pairs, we found similar looking faces in the database. First, we identified closely matching faces using a leading open-source face recognition algorithm (Cao et al. 2018). Second, we collected human similarity ratings via Amazon Mechanical Turk between each face in the database (target) and the four different faces that the algorithm rated as being most similar to the target (foils). Using these similarity ratings, we selected 50 non-match pairs containing 100 unique identities. Fifty identities were then selected from the remaining set of 201 for use in match pairs, so that identities used in non-match pairs were not used in the match pairs. In

addition, we ensured that all match and non-match pairs were different from the pairings used in the original GFMT. A full description of this process is provided in Supplementary Materials.

Having selected the match and non-match identity pairs, we then created the image pairs using the GUFD. The original GFMT used only passport style images, with the subject looking directly at the camera, with straight-on head angle, neutral expression etc. Here, we sampled other images from the GUFD that varied in both rigid movement of the head relative to the camera, and non-rigid movements of the face due to talking and expression. Still images showing rigid variations in head angle were captured using high quality SLR cameras and non-rigid variations using a video camera to record participants facial movement while speaking and making expressive gestures. We also sampled images containing slight variation in pose and head angle from the digital video recording where subjects had been standing a distance of two metres from the camera, to introduce variation in camera-to-subject distance (Noyes & Jenkins, 2017). This variation alters apparent face shape and reduces image resolution when faces are presented at the standard face size for the test. Capturing these properties in our test is important forensically, given the prevalence of identification from CCTV in criminal investigations and trials.

Examples of the three types of image pairings that were used to create the GFMT2 are shown in Figure 1 (rigid variation, non-rigid variation, distance variation). For each of the 50 identity pairings selected for non-match image pairs, we paired the high-quality reference image of the first identity with each of three images of their foil identity to create three test items. We created these same three test items for each of the 50 identities that had been chosen for match image pairs, giving a total pool of 300 test items (150 match, 150 non-match). As can be seen in the online version of Figure 1, the GFMT2 presented images in full colour, whereas the original GFMT presented images in greyscale.

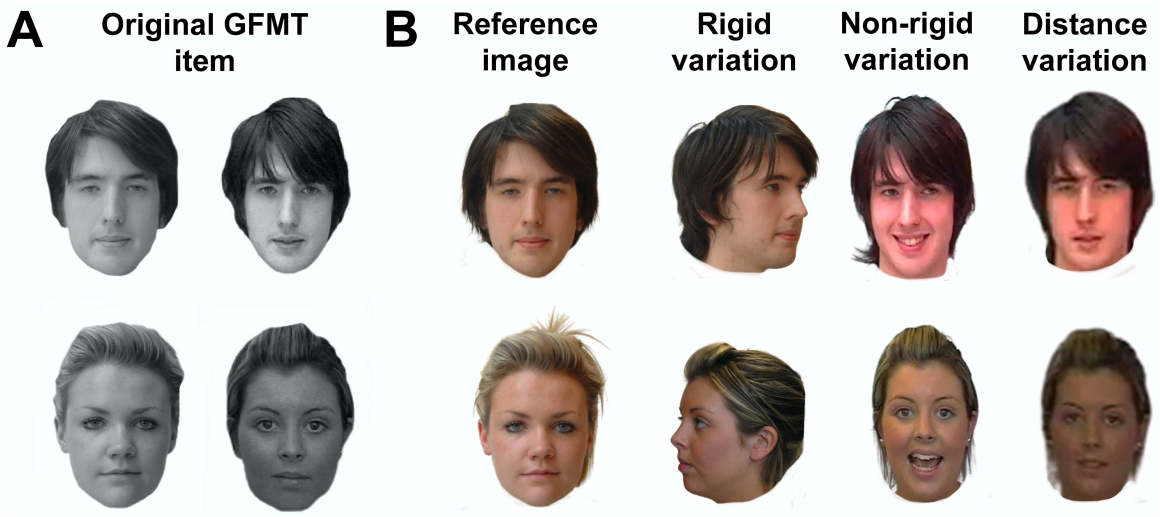

Figure 1. *(A) Two items from the original GFMT. Image pairs in the GFMT were created using high quality still images from an SLR and digital video camera, with the subjects positioned directly in front of the cameras and staring straight ahead with a neutral expression. (B) Examples of the image pair types used for the GFMT2. Test pairs were created by pairing colour 'reference' images of the type used in the original GFMT (left) with colour images containing variation in either (i) head angle (rigid variation), (ii) pose and expression (non-rigid variation) or (iii) camera-to-subject distance (distance variation). Test pairs used in the GFMT were not included in the GFMT2, and so the people shown here are used for illustration only.*

To enable repeated testing on different long forms of the GFMT2 (GFMT2-A/ B), we then separated the 300 test items described above into two subsets equated for difficulty. To do this, we first conducted an item difficulty screening of the 300 items by recruiting 320 participants via M-Turk (117 females; M_{age} = 33.8 years, SD = 9.19; 65% self-identified as Caucasian, 18% African American, 9% Asian, 6% Hispanic, 2% other ethnicity). Participants were randomly assigned to complete one of six versions of the test, each containing 50 trials presented in a random order (25 match, 25 non-match). For each image pair, participants were instructed to decide whether the two faces were of the same person or of different people. The tasks were self-paced.

Participants achieved an average accuracy score of 75.9% (SD = 11.3). Item accuracy is shown in Table 1, separately for image pair type (match, non-match) and image variation type (rigid, non-rigid, distance). Overall item accuracy was roughly equivalent for non-rigid and rigid variation but was notably poorer in the distance variation. In addition, accuracy was slightly better for match than non-match pairs, but this pattern varied markedly in the three image conditions. In both the rigid variation and non-rigid variation conditions, accuracy was greater on match than non-match pairs. Conversely, in the distance condition, accuracy was greater on non-match relative to match pairs. This finding is consistent with earlier work showing that a change in subject-to-camera distance is associated with an increased tendency to view images of the same identity as being different people (Hahn, O'Toole & Phillips, 2016; Noyes & Jenkins, 2017)[1].

	Rigid variation		**Non-rigid variation**		**Distance variation**		**Total**	
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
Match	83.9	8.0	80.3	11.1	66.6	14.7	77.0	13.7
Non-match	72.4	13.0	78.2	12.9	74.1	12.3	74.9	12.9
Overall	78.1	12.2	79.2	12.0	70.4	14.0	75.9	13.3

Table 1. Item accuracy screening data used in initial item selection for the long form test, separated by image pair type and image variation type.

We used this initial item accuracy screening data to split the items into two equally difficult forms of 75 match and 75 non-match pairs containing an equal number of each image type. Based on these previously collected item accuracy data, difficulty of the two tests was precisely matched (Long Form A: M = 75.9%, SD = 13.5; Long Form B: M = 75.9%, SD = 13.2). Nevertheless, it was necessary to test this in a study where participants completed these tests in full and so we conducted an additional study in which participants completed the

[1] While this is potentially of theoretical interest, it also produces some challenges for test construction, because test item difficulty was correlated with test item response bias. This introduces some complexity when selecting test items for shorter versions that we describe in subsequent sections.

full version of the long-form tests. This also enabled us to compute reliable item accuracy and item-to-test correlation statistics for the purpose of selecting items for shorter forms of the test.

Normative scores and overall test reliability of long form tests

Next we recruited a group of participants to establish normative data for GFMT2 long form tests (GFMT2-A, GFMT2-B). This also provided the opportunity to examine the reliability of our test by asking participants to perform two test sessions one week apart and measuring the correlation between performance on their test scores.

A total of 371 participants were recruited via M-Turk for the study and completed the test in the first test session (131 female; M_{age} = 36.6 years, SD = 10.7; 81% self-identified as Caucasian, 9% African American, 5% Asian, 4% Hispanic, 1% other ethnicity). Participant attrition meant that 270 of these completed the test in both the first and second test sessions, approximately one week later. We varied the test forms that participants completed at Time 1 and Time 2 such that some participants completed the same form of the test both times, whereas other participants completed two different forms of the test. This approach enabled us to determine whether completing the same form of the test twice is likely to cause improvement in accuracy. In one of the four groups, we found evidence of a slight improvement in accuracy between Time 1 and Time 2 (see Supplementary Material). Image pairs were presented in a different random order for each participant and the tasks were self-paced.

We first established normative performance on the test using Time 1 accuracy data. Average accuracy on Form A was 74.4% (SD = 10.9; Min = 46.0, Max = 92.7; n = 182) and on Form B it was 75.4% (SD = 10.1; Min = 46.7, Max = 95.3; n = 189). The small difference in accuracy between the test forms was non-significant, t (369) = 0.89, p = 0.38. Both tests show higher accuracy on match image pairs (Form A = 76.9%, Form B = 79.1%) than on non-match image pairs (Form A = 72.0%, Form B = 71.7%), consistent with a tendency to respond 'match' in the original GFMT. Kurtosis scores were close to 3 suggesting the scores were normally distributed (Form A: 3.10; Form B = 2.95). Both forms show moderate negative

skew (Form A: -0.793, $p < 0.05$; Form B = -0.571, $p < 0.05$), although this is substantially less skewed than the original GFMT (-1.33).

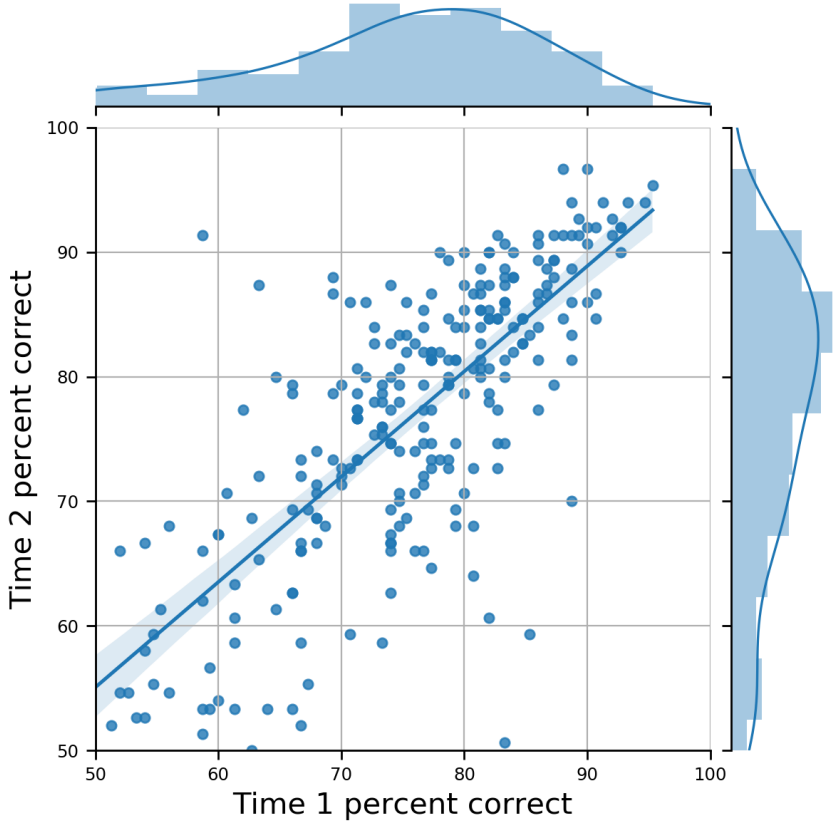

Figure 2. Test scores for 270 participants on the GFMT2 long form taken one week apart.

The correlation between test scores of individual participants at Time 1 and Time 2 is shown in Figure 2. Test-retest reliability of the overall test was high, r (270) = 0.778 and exceeded reliability measures of other leading tests (e.g. CFMT: Test-retest correlation = 0.68 in Murray & Bate 2020 and 0.7 in Wilmer et al., 2010). Internal test reliability computed using responses from all participants was also high for both long test forms (Form A: n = 262, Cronbach's alpha = .899; Form B: n = 241, Cronbach's alpha = .903).

Creating psychometrically calibrated short form tests

We next created three short versions of the test that provide more efficient test options. We found some evidence of improvements in accuracy with repeated testing in the long

form test-retest data (see Supplementary Material), which could potentially be caused by repeating identities across the two test forms. In response, identities did not repeat within any of the short form tests described below. The main short version consists of two 40-item test forms that are selected to be of equal difficulty to enable repeated testing (GFMT2-S A/B). Two additional versions are calibrated for discriminating among low performers (GFMT2-Low) and high performers respectively (GFMT2-High).

To select items for the short versions, we computed the item-to-test correlation for each item in the long version using data from Time 1 described in the previous section. This measure provides an estimate of the item's contribution to the overall test reliability and is a standard approach to subsampling test items that are most predictive of overall test performance (Guilford, 1954; see Wilmer et al. 2012). Item-test correlations were Pearson's correlations between participants' response to that particular item (correct, incorrect) and participants' d-prime computed for all other items in the test. Given the pattern of response bias observed in our data (see Table 1), we used d-prime to avoid patterns of decision criterion in our data influencing item selection.

To select the final items for the GFMT-S, we first computed item-test correlations using responses from all participants that completed Long Form A (n = 262) and Long Form B (n = 241). Average item-test correlations were substantially above zero [match: M= .266; SD = 0.76; non-match: M= .226; SD = .112], but there was a large range of values showing that some image pairs predicted overall test performance more than others. We then excluded image pairs for which accuracy differed by more than 15% between Time 1 and Time 2. This ensured that all image pairs produced relatively stable accuracy on repeated testing. We also excluded pairs that were answered correctly by more than 85% of participants. Because test-item correlation and item accuracy were correlated (nonmatch: r (150) = 0.408; match: r (150) = 0.746) this step avoided creating a test that was too easy. We then ranked the remaining image pairs by item-test correlation and selected the 40 match and 40 nonmatch pairs with the highest correlation, excluding any repeating identities and aiming – so far as possible – to equate accuracy for match and non-match pairs. These image pairs were then divided to create two equally difficult forms of the GFMT2-S.

We followed the same procedure to select the final image pairs for the two 40 item tests that were specifically designed to test low and high performing participants – GFMT2-Low and GFMT2-High. However, item-correlations for these tests were computed separately from low performing (scoring below median accuracy, GFMT2-Low) and high performing participants (scoring above median accuracy, GFMT2-high) respectively. Additionally, for the GFMT-Low only, we did not exclude easier pairs.

	Overall accuracy	Match accuracy	Non-match accuracy	Rigid items (n)	Non-rigid items (n)	Distance items (n)
GFMT2-S	76.4	79.0	73.8	25	29	26
GFMT2-SA	76.4	79.0	73.8	14	12	14
GFMT2-SB	76.4	79.0	73.8	11	17	12
GFMT2-Low	82.7	85.6	79.8	19	14	7
GFMT2-High	67.5	69.1	65.9	6	15	19

Table 2. Mean item accuracy and number of each image pairing type (rigid, non-rigid, distance variation) in the short forms of the GFMT2. Counts of pair type are out of 80 for the GFMT2-S and 40 for the other tests. Normative test scores for the GFMT-S are provided in Table 3.

Summary item accuracy for all the short tests are shown in Table 2. Overall item accuracy for the GFMT2-S (76%) is near the midpoint of the measurement scale which ranges from chance (50%) to perfect accuracy (100%). GFMT2-SA and GFMT2-SB versions both match the difficulty of the GFMT2-S precisely in terms of overall and match/ non-match item accuracy. Item accuracy of the GFMT2-Low and GFMT-High are calibrated to the target populations of these tests. Notably, item accuracy is higher for match items than non-match in all versions of the test. This was due to lower overall item-test correlations for match compared to non-match pairs, allied with the correlation between item difficulty and item-test correlation. These two constraints meant that it was not possible to choose as many difficult match pairs as one would like to, without compromising the reliability of the test.

We provide more detailed description of item-test correlation results in the Supplementary Material.

The numbers of items for each image pairing type are also shown in Table 2. As expected, our item selection method produced a higher incidence of the most difficult distance variation items for the GFMT2-High and a higher incidence of the easier rigid variation items in the GFMT2-Low. There was some overlap between test items used in GFMT-S and these tests (GFMT Low n = 11; GFMT High n = 18), but only 2 items were included in both the GFMT High and GFMT Low, highlighting the importance of calibrating tests separately when targeting upper and lower quartiles of the performance distribution (see also Wilmer et al. 2012).

Normative test scores and test-retest reliability for the GFMT2 short version (GFMT2-S)

The GFMT2-S is intended to be the primary measure of face matching ability. To provide normative scores and test reliability of the GFMT2-S, we recruited a further 153 participants to perform the GFMT2-S twice, with an interval of one week between tests. The final dataset contained 108 participants (42 female; M_{age} = 38.0 years, SD = 11.3; 76% self-identified as Caucasian, 15% African American, 6% Asian, 3% Hispanic), after removing participants who did not complete both test sessions (27), performed below chance (10) or entered non-matching demographic details in the two tests (8).

Participants completed the same version of the GFMT2-S in both test sessions. Unlike previous studies, we fixed the order of image pairs so that it was the same for every participant and was consistent across test sessions. This fixed order is desirable in psychometric tests of ability (e.g. Mollon, Bosten, Peterzell, & Webster, 2017) and is adopted in other popular tests of face identity processing (Duchaine & Nakayama, 2006). We recommend using a fixed trial order for all versions of the GFMT2 and this is the default setting in executable versions of the test.

Normative accuracy on the GFMT2-S was 75.0% (SD = 10.0, Min = 55.0, Max = 97.5). As with the long form test, we found higher accuracy on match image pairs (78.9%) than non-match

image pairs (71.1%). Kurtosis was roughly normal, but slightly less than the expected value of 3 for a normal distribution (2.38), suggesting a slightly higher proportion of extreme values in this dataset. The skewness of the distribution was non-significant (-0.108, $p > 0.05$).

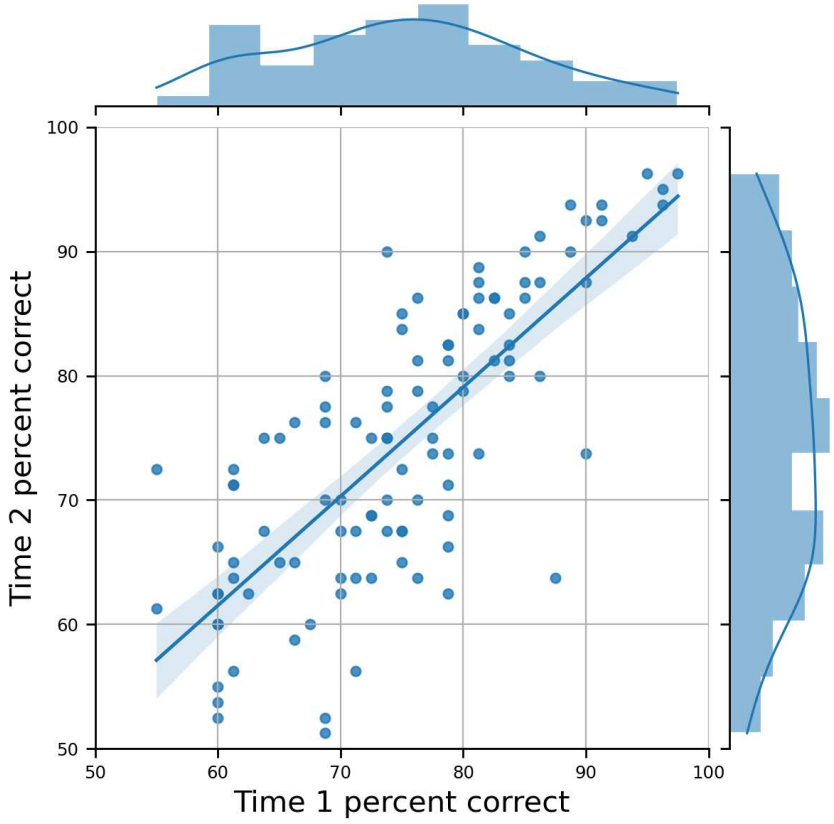

Figure 3. Test scores for 108 participants on the GFMT2 short form (GFMT2-S) taken one week apart.

	Overall accuracy	Match item accuracy	Non-match item accuracy	Sensitivity (d-prime)	Response criterion (C)
GFMT2-S	75.0 (10.0)	78.9 (17.1)	71.1 (19.7)	1.68 (0.77)	-0.167 (0.599)
GFMT2-SA	74.5 (10.1)	77.8 (17.4)	71.2 (21.0)	1.64 (0.78)	-0.121 (0.617)
GFMT2-SB	75.5 (11.4)	80.0 (18.5)	71.0 (20.2)	1.75 (0.87)	-0.205 (0.594)

Table 3. Normative scores on the GFMT2-S calculated from a group of 108 participants that completed the test online via Amazon Mechanical Turk. Standard deviations are in parenthesis.

Test-retest reliability of the GFMT2-S is shown in Figure 3. We found a high correlation of test scores across repeated tests, r (107) = 0.774, which was very similar to the test-retest correlation of the long form test, suggesting that we were successful in reducing the length of the test without compromising test reliability. Again, this compares favourably to test-retest reliability of other leading tests (e.g. CFMT: Test-retest correlation = 0.68 in Murray & Bate 2020 and 0.7 in Wilmer et al., 2010). Internal test reliability computed using responses from all participants was also very high (Cronbach's alpha = .938).

Normative scores for the GFMT2-S in our online test are provided in Table 3. The mean score of the overall test is centred – with surprising precision – on the midpoint of the scale ranging from chance (50%) to perfect accuracy (100%). Differences in accuracy scores for GFMT2-SA and GFMT2-SB were non-significant (Time 1: t(107) = 1.32, p =0.189; Time 2: t(107) = 0.87, p =0.386), suggesting that they can be used in experimental intervention studies. Along with the reliability analysis, normative scores show the GFMT2-S to have attractive psychometric properties and we hope it will be used widely. As researchers begin to use the GFMT2-S, GFMT2-High and GFMT2-Low in different settings, and with different cohorts, we encourage them to share their test score data with the scientific community, to assess the generality of these normative data.

DISCUSSION

We have presented a new face matching test, representing a significant update of the Glasgow Face Matching Test. The test includes various sub-tests, designed to facilitate experimental research on interventions and studies of those with exceptionally good or poor unfamiliar face matching ability. The tests have stable psychometric properties and deliver patterns of performance that can support research in individual differences.

Until relatively recently, the theoretical importance of variation in people's face matching ability was not appreciated. Theoretical work tended to emphasise *intra*-individual differences, for example between familiar and unfamiliar faces (Johnston & Edmonds, 2009), but little attention was paid to *inter*-individual differences in ability within the typical

population. However, over the past decade there has been a growing acknowledgement that individual differences can be highly informative in contributing to our theoretical understanding of face recognition (e.g. Yovel, Wilmer & Duchaine, 2014; Bruce, Bindemann & Lander, 2018). Patterns of recognition performance in patients with acquired prosopagnosia have, for many years, informed models of face processing. However, it is only more recently that these have been linked theoretically to developmental difficulties with face perception, as well as to exceptionally good face recognition performance. The past decade has, finally, seen the inclusion of variation among typical populations.

At the same time as individual differences have become important in theoretical studies of face perception, it has become increasingly clear that they play a critical role in a number of applied settings. For example, natural variations in ability vastly exceed any effects of professional training for face matching tasks (Towler et al, 2019; White et al, 2014). In applied settings such as passport control or surveillance, the importance of personnel selection is becoming widely recognised. There are even suggestions that members of the public making eyewitness statements may have their testimony qualified by tests of their face recognition ability (Bindemann, Brown, Koyas & Russ, 2012).

For all the reasons listed here, we hope that a standard test of face matching will be valuable to the community. The GFMT2 has many useful properties, including simplicity of administration. As a self-paced test of matching, with no requirement to remember faces, it mimics many real-world tasks. The test is available for download via www.gfmt2.org.

OPEN PRACTICES STATEMENT

Detailed item performance and metadata used to create the test are available as part of the test distribution folder available via www.gfmt2.org [for the purpose of peer review these can be accessed at: https://tinyurl.com/GFMT2peerReview].

REFERENCES

Balsdon, T., Summersby, S., Kemp, R. I., & White, D. (2018). Improving face identification with specialist teams. *Cognitive Research: Principles and Implications*, 3(1), 25.

Bate, S., & Bennetts, R. J. (2014). The rehabilitation of face recognition impairments: a critical review and future directions. *Frontiers in Human Neuroscience*, 8, 491.

Benton, A. L., Hamsher, K. S., Varney, N. R., & Spreen, O. (1983). *Contributions to neuropsychological assessment*. New York: Oxford University Press.

Bindemann, M., Brown, C., Koyas, T., & Russ, A. (2012). Individual differences in face identification postdict eyewitness accuracy. *Journal of Applied Research in Memory and Cognition*, *1*, 96-103.

Bobak, A. K., Bennetts, R. J., Parris, B. A., Jansari, A., & Bate, S. (2016). An in-depth cognitive examination of individuals with superior face recognition skills. *Cortex,* 82, 48-62.

Bowles, D. C., McKone, E., Dawel, A., Duchaine, B., Palermo, R., Schmalzl, L., ... & Yovel, G. (2009). Diagnosing prosopagnosia: Effects of ageing, sex, and participant–stimulus ethnic match on the Cambridge Face Memory Test and Cambridge Face Perception Test. *Cognitive Neuropsychology*, *26*, 423-455.

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

Burton, A. M., White, D., & McNeill, A. (2010). The Glasgow face matching test. *Behavior Research Methods*, *42*, 286-291.

Cao, Q., Shen, L., Xie, W., Parkhi, O. M., & Zisserman, A. (2018, May). Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)* (pp. 67-74).

Davis, J. P., Lander, K., Evans, R., & Jansari, A. (2016). Investigating predictors of superior face recognition ability in police super-recognisers. *Applied Cognitive Psychology,* 30, 827-840.

DeGutis, J. M., Chiu, C., Grosso, M. E., Cohan, S. (2014). Face processing improvements in prosopagnosia: Successes and failures over the last 50 years. *Frontiers in Human Neuroscience*, 8, 561.

Dowsett, A. J., & Burton, A. M. (2015). Unfamiliar face matching: Pairs out-perform individuals and provide a route to training. *British Journal of Psychology,* 106, 433-445.

Duchaine, B. C., & 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.

Dunn, J. D., Summersby, S., Towler, A., Davis, J. P., & White, D. (2020). UNSW Face Test: A screening tool for super-recognizers. *PloS one*, *15*(11), e0241747.

Fysh, M. C., & Bindemann, M. (2018). The Kent face matching test. *British Journal of Psychology,* 109(2), 219-231.

Guilford, J. P. (1954). *Psychometric methods.* New York: McGraw-Hill.

Hahn, C. A., O'Toole, A. J., & Phillips, P. J. (2016). Dissecting the time course of person recognition in natural viewing environments. *British Journal of Psychology*, *107*(1), 117-134.

Matthews, C. M., & Mondloch, C. J. (2018). Improving identity matching of newly encountered faces: effects of multi-image training. *Journal of Applied Research in Memory and Cognition*, *7*, 280-290.

McCaffery, J. M., Robertson, D. J., Young, A. W., & Burton, A. M. (2018). Individual differences in face identity processing. *Cognitive Research: Principles and Implications*, 3, 21.

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

Mollon, J. D., Bosten, J. M., Peterzell, D. H., & Webster, M. A. (2017). Individual differences in visual science: What can be learned and what is good experimental practice?. *Vision Research*, *141*, 4-15.

Murray, E., & Bate, S. (2020). Diagnosing developmental prosopagnosia: repeat assessment using the Cambridge Face Memory Test. *Royal Society Open Science*, *7*, 200884.

Noyes, E., & Jenkins, R. (2017). Camera-to-subject distance affects face configuration and perceived identity. *Cognition*, 165, 97-104.

Noyes, E., Phillips, P. J., & O'Toole, A. J. (2017). What is a super-recogniser? In *Face processing: Systems, disorders and cultural differences.* M. Bindemann & A. Megreya (Eds.). Nova Science.

Phillips, P. J., Yates, A. N., Hu, Y., Hahn, C. A., Noyes, E., Jackson, K., ... & O'Toole, A. J. (2018). Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Proceedings of the National Academy of Sciences*, 115(24), 6171-6176.

Ramon, M., Bobak, A. K., & White, D. (2019). Super-recognizers: From the lab to the world and back again. *British Journal of Psychology,* 110, 461-479.

Robertson, D. J., Noyes, E., Dowsett, A. J., Jenkins, R., & Burton, A. M. (2016). Face recognition by metropolitan police super-recognisers. *PloS one*, 11, e0150036.

Russell, R., Duchaine, B., & Nakayama, K. (2009). Super-recognizers: People with extraordinary face recognition ability. *Psychonomic Bulletin & Review*, *16*, 252-257.

Shakeshaft, N. G., & Plomin, R. (2015). Genetic specificity of face recognition. *Proceedings of the National Academy of Sciences*, 112, 12887-12892.

Towler, A., Kemp, R. I., Burton, A. M., Dunn, J. D., Wayne, T., Moreton, R., & White, D. (2019). Do professional facial image comparison training courses work?. *PloS one,* 14, e0211037.

Towler, A., Keshwa, M., Ton, B., Kemp, R. I., & White, D. (in press). Diagnostic feature training improves face matching accuracy. J*ournal of Experimental Psychology: Learning, Memory & Cognition.*

White, D., Kemp, R. I., Jenkins, R., & Burton, A. M. (2014). Feedback training for facial image comparison. *Psychonomic Bulletin & Review*, 21, 100-106.

White, D., Kemp, R. I., Jenkins, R., Matheson, M., & Burton, A. M. (2014). Passport officers' errors in face matching. *PloS one,* 9(8), e103510.

White, D., Phillips, P. J., Hahn, C. A., Hill, M., & O'Toole, A. J. (2015). Perceptual expertise in forensic facial image comparison. *Proceedings of the Royal Society B: Biological Sciences*, *282*, 20151292.

White, D., Rivolta, D., Burton, A. M., Al-Janabi, S., & Palermo, R. (2017). Face matching impairment in developmental prosopagnosia. *Quarterly Journal of Experimental Psychology*, *70*(2), 287-297.

White, D., Towler, A., & Kemp, R. I. (2020). Understanding professional expertise in unfamiliar face matching. In M. Bindemann (Ed.), *Forensic face matching: Research and practice.* Oxford University Press.

Wilhelm, O., Herzmann, G., Kunina, O., Danthiir, V., Schacht, A., & Sommer, W. (2010). Individual differences in perceiving and recognizing faces—One element of social cognition. *Journal of Personality and Social Psychology*, *99*, 530.

Wilmer, J. B., Germine, L., Chabris, C. F., Chatterjee, G., Williams, M., Loken, E., ... & Duchaine, B. (2010). Human face recognition ability is specific and highly heritable. *Proceedings of the National Academy of Sciences,* 107, 5238-5241.

Wilmer, J. B., Germine, L., Chabris, C. F., Chatterjee, G., Gerbasi, M., & Nakayama, K. (2012). Capturing specific abilities as a window into human individuality: The example of face recognition. *Cognitive Neuropsychology*, *29*(5-6), 360-392.

Wilmer, J. B. (2017). Individual differences in face recognition: A decade of discovery. *Current Directions in Psychological Science*, *26*(3), 225-230.