

This is a repository copy of *GFMT2:A psychometric measure of face matching ability*.

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/177641/>

Version: Accepted Version

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>

Reuse

Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

GFMT2: A psychometric measure of face matching ability

David White¹, Daniel Guilbert², Victor P. L. Varela¹, Rob Jenkins³ & A. Mike Burton³

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.

65 INTRODUCTION

66

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

72

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

83

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

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

97

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

108

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

117

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

128

1
2
3 129 This is problematic because these professionals are entrusted by the public to make
4
5 130 accurate face identification decisions. Given the clear evidence that individual differences in
6
7 131 face matching tasks are large and stable over time, selecting people that are skilled in face
8
9 132 matching provides a promising solution to this problem. Therefore, the GFMT can also be
10
11 133 used as a tool for staff selection and recruitment in roles that require people to make face
12
13 134 matching decisions (see White et al. 2014, 2015; Davis et al. 2016).

14 135

15 136 **A NEW PSYCHOMETRIC MEASURE OF FACE MATCHING ABILITY**

17 137

18
19
20 138 Despite its popularity, the GFMT has a number of limitations and would benefit from an
21
22 139 update. Over the period of its use, reported mean performance in the general population
23
24 140 has tended to increase. For example, Burton et al (2010) report mean performance of
25
26 141 around 82% for the short test, whereas more recent uses often report means of just under
27
28 142 90% (e.g. Towler et al, 2019). High mean accuracy is accompanied by a reduction in
29
30 143 variance of high scores, somewhat devaluing the test for certain uses. There are a number
31
32 144 of reasons this inflation may have occurred. The test is freely available, and many example
33
34 145 items have been published in research papers. As a result, there may be an issue of
35
36 146 familiarity for some experimental participants, especially in psychology communities.
37
38 147 Furthermore, at the time the original test was developed, the general difficulty of unfamiliar
39
40 148 face matching was poorly understood. The fact that matching is hard, even in high-quality
41
42 149 images, is now much more widely known, perhaps encouraging participants to take a more
43
44 150 studied approach to the task.

45 151

46 152 A more challenging version of the GFMT is also necessary due to the increased interest in
47
48 153 super-recognisers (e.g. Bobak et al. 2016), and professional groups displaying high levels of
49
50 154 accuracy in face recognition tasks (Phillips et al. 2018). While researchers have produced
51
52 155 challenging tests to address specific research goals (Fysh & Bindemann, 2017; White et al.
53
54 156 2015; Dunn et al. 2020), there is now a need for a standard lab-based test with known
55
56 157 psychometric properties that enables comparison across high performers, typical
57
58 158 performers and low performers. To make the GFMT2 more challenging than the original
59
60 159 test, we select image pairs that require participants to match identity across variations in
60
160 head angle, pose, expression and subject-to-camera distance. In contrast, the GFMT was

1
2
3 161 created by pairing two passport-style images of faces in neutral pose, pictured straight on
4
5 162 standing directly in front of the camera (see Figure 1). As well as making the task more
6
7 163 difficult, this change also captures a wider range of applied tasks that practitioners perform.
8

9 164

10 165 Another improvement on the original version is that, for both short and long form tests, we
11 166 provide two versions that are equated for difficulty. These paired versions enable repeat
12 167 testing of participants to examine the effectiveness of clinical interventions, professional
13 168 training and mentorship programs. For example, in a recent test of the effectiveness of
14 169 professional training, Towler and colleagues (2019) tested participants before-and-after
15 170 training using short 20 item versions of the GFMT that have been equated for difficulty –
16 171 finding no evidence of improvement. Such research points to the practical need for more
17 172 effective methods of training, and evaluation of evidence-based interventions requires
18 173 common, reliable repeated measures that are equated for difficulty (see Dowsett & Burton,
19 174 2015; White et al. 2014; Matthews & Mondloch, 2018; Towler et al. in press; c.f. Bate &
20 175 Bennetts, 2014; DeGutis et al. 2014).
21
22
23
24
25
26
27
28
29
30

31 176

32 177 Finally, the GFMT2 also provides various short forms of the tests, tailored to particular use
33 178 and selected using established psychometric principles. In the remainder of the paper, we
34 179 describe the development of these tests. We first describe the GFMT2 Long Form, which
35 180 consists of two 150-item sub-tests of equal difficulty (GFMT2-A and GFMT2-B). This long
36 181 form is not intended as the primary measure of face matching ability but was the starting
37 182 point for selecting test items that maximise desirable psychometric properties in short test
38 183 versions. The primary measure is the short form of the test (GFMT2-S), and we create two
39 184 additional short tests that are tailored to the low (GFMT2-Low) and high ends (GFMT2-High)
40 185 of the performance scale:
41
42
43
44
45
46
47
48

49 186

- 50
51 187 (i) The 80-item GFMT2 Short Form (GFMT2-S) comprises two equally difficult
52 188 40-item test forms, GFMT2-SA and GFMT2-SB. We anticipate that these will
53 189 be useful in experimental intervention studies.
54
55 190 (ii) The 40-item GFMT2-Low is designed to discriminate between low performing
56 191 participants. This version of the test will be useful in assessing acquired or
57
58
59
60 192 developmental prosopagnosia.

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

196
197 The tests are free for scientific use and executable versions are available for download at
198 www.gfmt2.org [for the purpose of peer review these can be accessed at
199 <https://tinyurl.com/gfmt2review>]. Detailed item performance and metadata used to create
200 the short test versions, and normative test scores for long and short tests, are also included
201 in the test distribution [for the purpose of peer review these can be accessed at:
202 <https://tinyurl.com/gfmt2review>].

203

204 **TEST CONSTRUCTION AND RESULTS**

205

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

207

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

214

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

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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

228
229 Having selected the match and non-match identity pairs, we then created the image pairs
230 using the GUFID. The original GFMT used only passport style images, with the subject looking
231 directly at the camera, with straight-on head angle, neutral expression etc. Here, we
232 sampled other images from the GUFID that varied in both rigid movement of the head
233 relative to the camera, and non-rigid movements of the face due to talking and expression.
234 Still images showing rigid variations in head angle were captured using high quality SLR
235 cameras and non-rigid variations using a video camera to record participants facial
236 movement while speaking and making expressive gestures. We also sampled images
237 containing slight variation in pose and head angle from the digital video recording where
238 subjects had been standing a distance of two metres from the camera, to introduce
239 variation in camera-to-subject distance (Noyes & Jenkins, 2017). This variation alters
240 apparent face shape and reduces image resolution when faces are presented at the
241 standard face size for the test. Capturing these properties in our test is important
242 forensically, given the prevalence of identification from CCTV in criminal investigations and
243 trials.

244
245 Examples of the three types of image pairings that were used to create the GFMT2 are
246 shown in Figure 1 (rigid variation, non-rigid variation, distance variation). For each of the 50
247 identity pairings selected for non-match image pairs, we paired the high-quality reference
248 image of the first identity with each of three images of their foil identity to create three test
249 items. We created these same three test items for each of the 50 identities that had been
250 chosen for match image pairs, giving a total pool of 300 test items (150 match, 150 non-
251 match). As can be seen in the online version of Figure 1, the GFMT2 presented images in full
252 colour, whereas the original GFMT presented images in greyscale.

253

GLASGOW FACE MATCHING TEST 2

9

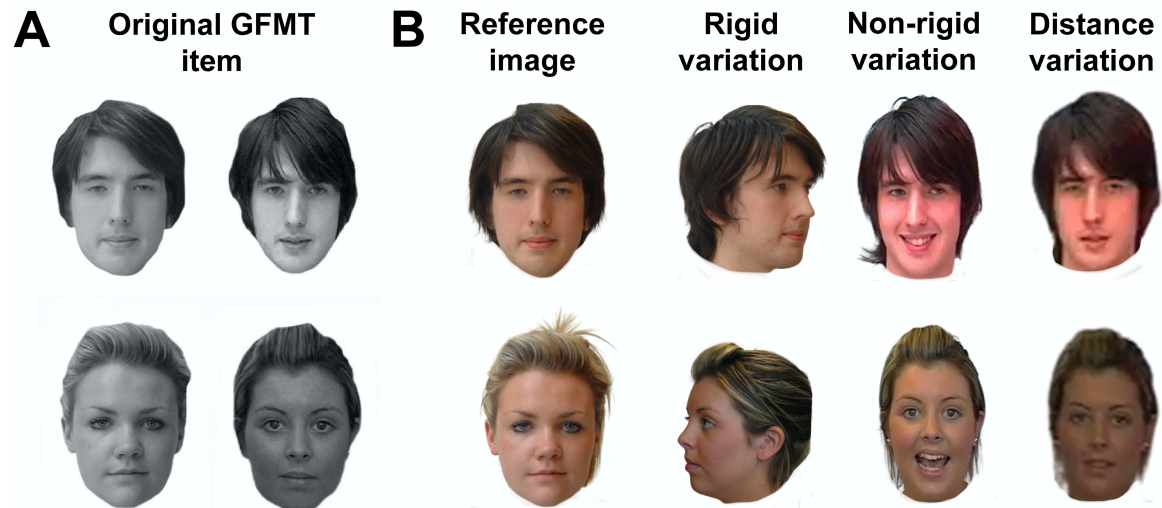


Figure 1. (A) Two items from the original GFMT. Image pairs in the GFMT were created using high quality still images from an SLR and digital video camera, with the subjects positioned directly in front of the cameras and staring straight ahead with a neutral expression. (B) Examples of the image pair 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; $M_{\text{age}} = 33.8$ 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.

276 Participants achieved an average accuracy score of 75.9% (SD = 11.3). Item accuracy is
 277 shown in Table 1, separately for image pair type (match, non-match) and image variation
 278 type (rigid, non-rigid, distance). Overall item accuracy was roughly equivalent for non-rigid
 279 and rigid variation but was notably poorer in the distance variation. In addition, accuracy
 280 was slightly better for match than non-match pairs, but this pattern varied markedly in the
 281 three image conditions. In both the rigid variation and non-rigid variation conditions,
 282 accuracy was greater on match than non-match pairs. Conversely, in the distance condition,
 283 accuracy was greater on non-match relative to match pairs. This finding is consistent with
 284 earlier work showing that a change in subject-to-camera distance is associated with an
 285 increased tendency to view images of the same identity as being different people (Hahn,
 286 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

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

291 We used this initial item accuracy screening data to split the items into two equally difficult
 292 forms of 75 match and 75 non-match pairs containing an equal number of each image type.
 293 Based on these previously collected item accuracy data, difficulty of the two tests was
 294 precisely matched (Long Form A: M = 75.9%, SD = 13.5; Long Form B: M = 75.9%, SD = 13.2).
 295 Nevertheless, it was necessary to test this in a study where participants completed these
 296 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.

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

300

301 **Normative scores and overall test reliability of long form tests**

302

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

307

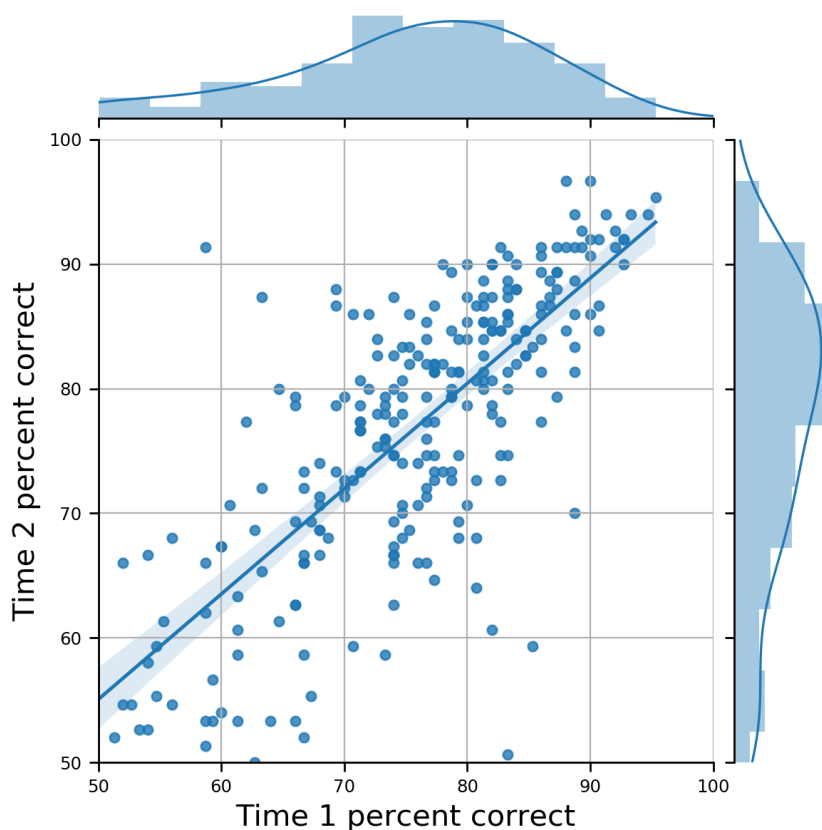
308 A total of 371 participants were recruited via M-Turk for the study and completed the test in
309 the first test session (131 female; $M_{\text{age}} = 36.6$ years, $SD = 10.7$; 81% self-identified as
310 Caucasian, 9% African American, 5% Asian, 4% Hispanic, 1% other ethnicity). Participant
311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

320

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

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

331



332

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

334

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

341

342 **Creating psychometrically calibrated short form tests**

343

344 We next created three short versions of the test that provide more efficient test options.

345 We found some evidence of improvements in accuracy with repeated testing in the long

1
2
3 346 form test-retest data (see Supplementary Material), which could potentially be caused by
4
5 347 repeating identities across the two test forms. In response, identities did not repeat within
6
7 348 any of the short form tests described below. The main short version consists of two 40-item
8
9 349 test forms that are selected to be of equal difficulty to enable repeated testing (GFMT2-S
10
11 350 A/B). Two additional versions are calibrated for discriminating among low performers
12
13 351 (GFMT2-Low) and high performers respectively (GFMT2-High).

14 352

15
16 353 To select items for the short versions, we computed the item-to-test correlation for each
17
18 354 item in the long version using data from Time 1 described in the previous section. This
19
20 355 measure provides an estimate of the item's contribution to the overall test reliability and is
21
22 356 a standard approach to subsampling test items that are most predictive of overall test
23
24 357 performance (Guilford, 1954; see Wilmer et al. 2012). Item-test correlations were Pearson's
25
26 358 correlations between participants' response to that particular item (correct, incorrect) and
27
28 359 participants' d-prime computed for all other items in the test. Given the pattern of response
29
30 360 bias observed in our data (see Table 1), we used d-prime to avoid patterns of decision
31
32 361 criterion in our data influencing item selection.

33 362

34
35 363 To select the final items for the GFMT-S, we first computed item-test correlations using
36
37 364 responses from all participants that completed Long Form A (n = 262) and Long Form B (n =
38
39 365 241). Average item-test correlations were substantially above zero [match: M= .266; SD =
40
41 366 0.76; non-match: M= .226; SD = .112], but there was a large range of values showing that
42
43 367 some image pairs predicted overall test performance more than others. We then excluded
44
45 368 image pairs for which accuracy differed by more than 15% between Time 1 and Time 2. This
46
47 369 ensured that all image pairs produced relatively stable accuracy on repeated testing. We
48
49 370 also excluded pairs that were answered correctly by more than 85% of participants. Because
50
51 371 test-item correlation and item accuracy were correlated (nonmatch: $r(150) = 0.408$; match:
52
53 372 $r(150) = 0.746$) this step avoided creating a test that was too easy. We then ranked the
54
55 373 remaining image pairs by item-test correlation and selected the 40 match and 40 nonmatch
56
57 374 pairs with the highest correlation, excluding any repeating identities and aiming – so far as
58
59 375 possible – to equate accuracy for match and non-match pairs. These image pairs were then
60
376 divided to create two equally difficult forms of the GFMT2-S.

377

378 We followed the same procedure to select the final image pairs for the two 40 item tests
 379 that were specifically designed to test low and high performing participants – GFMT2-Low
 380 and GFMT2-High. However, item-correlations for these tests were computed separately
 381 from low performing (scoring below median accuracy, GFMT2-Low) and high performing
 382 participants (scoring above median accuracy, GFMT2-high) respectively. Additionally, for the
 383 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

385
 386 **Table 2.** Mean item accuracy and number of each image pairing type (rigid, non-rigid,
 387 distance variation) in the short forms of the GFMT2. Counts of pair type are out of 80 for the
 388 GFMT2-S and 40 for the other tests. Normative test scores for the GFMT-S are provided in
 389 Table 3.

390
 391 Summary item accuracy for all the short tests are shown in Table 2. Overall item accuracy
 392 for the GFMT2-S (76%) is near the midpoint of the measurement scale which ranges from
 393 chance (50%) to perfect accuracy (100%). GFMT2-SA and GFMT2-SB versions both match
 394 the difficulty of the GFMT2-S precisely in terms of overall and match/ non-match item
 395 accuracy. Item accuracy of the GFMT2-Low and GFMT-High are calibrated to the target
 396 populations of these tests. Notably, item accuracy is higher for match items than non-match
 397 in all versions of the test. This was due to lower overall item-test correlations for match
 398 compared to non-match pairs, allied with the correlation between item difficulty and item-
 399 test correlation. These two constraints meant that it was not possible to choose as many
 400 difficult match pairs as one would like to, without compromising the reliability of the test.

1
2
3 401 We provide more detailed description of item-test correlation results in the Supplementary
4
5 402 Material.

6
7 403
8
9 404 The numbers of items for each image pairing type are also shown in Table 2. As expected,
10 405 our item selection method produced a higher incidence of the most difficult distance
11 406 variation items for the GFMT2-High and a higher incidence of the easier rigid variation items
12 407 in the GFMT2-Low. There was some overlap between test items used in GFMT-S and these
13 408 tests (GFMT Low $n = 11$; GFMT High $n = 18$), but only 2 items were included in both the
14 409 GFMT High and GFMT Low, highlighting the importance of calibrating tests separately when
15 410 targeting upper and lower quartiles of the performance distribution (see also Wilmer et al.
16 411 2012).

17 412

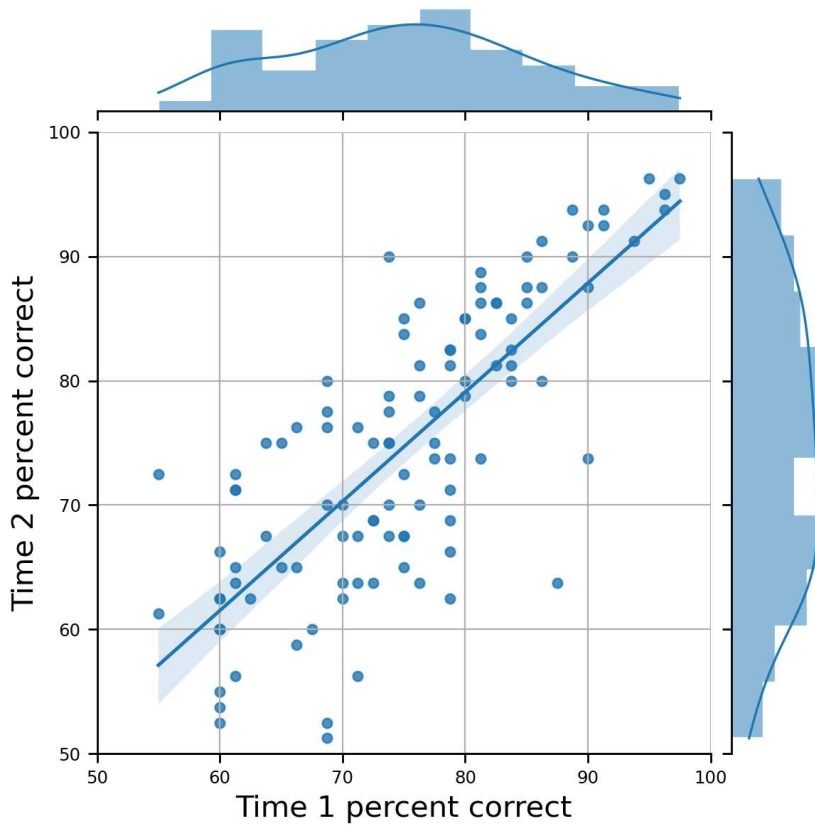
18 413 **Normative test scores and test-retest reliability for the GFMT2 short version (GFMT2-S)**

19 414
20 415 The GFMT2-S is intended to be the primary measure of face matching ability. To provide
21 416 normative scores and test reliability of the GFMT2-S, we recruited a further 153 participants
22 417 to perform the GFMT2-S twice, with an interval of one week between tests. The final
23 418 dataset contained 108 participants (42 female; $M_{\text{age}} = 38.0$ years, $SD = 11.3$; 76% self-
24 419 identified as Caucasian, 15% African American, 6% Asian, 3% Hispanic), after removing
25 420 participants who did not complete both test sessions (27), performed below chance (10) or
26 421 entered non-matching demographic details in the two tests (8).

27 422
28 423 Participants completed the same version of the GFMT2-S in both test sessions. Unlike
29 424 previous studies, we fixed the order of image pairs so that it was the same for every
30 425 participant and was consistent across test sessions. This fixed order is desirable in
31 426 psychometric tests of ability (e.g. Mollon, Bosten, Peterzell, & Webster, 2017) and is
32 427 adopted in other popular tests of face identity processing (Duchaine & Nakayama, 2006).
33 428 We recommend using a fixed trial order for all versions of the GFMT2 and this is the default
34 429 setting in executable versions of the test.

35 430
36 431 Normative accuracy on the GFMT2-S was 75.0% ($SD = 10.0$, $Min = 55.0$, $Max = 97.5$). As with
37 432 the long form test, we found higher accuracy on match image pairs (78.9%) than non-match

433 image pairs (71.1%). Kurtosis was roughly normal, but slightly less than the expected value
 434 of 3 for a normal distribution (2.38), suggesting a slightly higher proportion of extreme
 435 values in this dataset. The skewness of the distribution was non-significant (-0.108, $p >$
 436 0.05).



437
 438 Figure 3. Test scores for 108 participants on the GFMT2 short form (GFMT2-S) taken one
 439 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)

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

444

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

452

453 Normative scores for the GFMT2-S in our online test are provided in Table 3. The mean
454 score of the overall test is centred – with surprising precision – on the midpoint of the scale
455 ranging from chance (50%) to perfect accuracy (100%). Differences in accuracy scores for
456 GFMT2-SA and GFMT2-SB were non-significant (Time 1: $t(107) = 1.32$, $p = 0.189$; Time 2:
457 $t(107) = 0.87$, $p = 0.386$), suggesting that they can be used in experimental intervention
458 studies. Along with the reliability analysis, normative scores show the GFMT2-S to have
459 attractive psychometric properties and we hope it will be used widely. As researchers begin
460 to use the GFMT2-S, GFMT2-High and GFMT2-Low in different settings, and with different
461 cohorts, we encourage them to share their test score data with the scientific community, to
462 assess the generality of these normative data.

463

464 DISCUSSION

465

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

471

472 Until relatively recently, the theoretical importance of variation in people's face matching
473 ability was not appreciated. Theoretical work tended to emphasise *intra*-individual
474 differences, for example between familiar and unfamiliar faces (Johnston & Edmonds,
475 2009), but little attention was paid to *inter*-individual differences in ability within the typical

1
2
3 476 population. However, over the past decade there has been a growing acknowledgement
4
5 477 that individual differences can be highly informative in contributing to our theoretical
6
7 478 understanding of face recognition (e.g. Yovel, Wilmer & Duchaine, 2014; Bruce, Bindemann
8
9 479 & Lander, 2018). Patterns of recognition performance in patients with acquired
10
11 480 prosopagnosia have, for many years, informed models of face processing. However, it is
12
13 481 only more recently that these have been linked theoretically to developmental difficulties
14
15 482 with face perception, as well as to exceptionally good face recognition performance. The
16
17 483 past decade has, finally, seen the inclusion of variation among typical populations.

18 484
19
20 485 At the same time as individual differences have become important in theoretical studies of
21
22 486 face perception, it has become increasingly clear that they play a critical role in a number of
23
24 487 applied settings. For example, natural variations in ability vastly exceed any effects of
25
26 488 professional training for face matching tasks (Towler et al, 2019; White et al, 2014). In
27
28 489 applied settings such as passport control or surveillance, the importance of personnel
29
30 490 selection is becoming widely recognised. There are even suggestions that members of the
31
32 491 public making eyewitness statements may have their testimony qualified by tests of their
33
34 492 face recognition ability (Bindemann, Brown, Koyas & Russ, 2012).

35 493
36 494 For all the reasons listed here, we hope that a standard test of face matching will be
37
38 495 valuable to the community. The GFMT2 has many useful properties, including simplicity of
39
40 496 administration. As a self-paced test of matching, with no requirement to remember faces, it
41
42 497 mimics many real-world tasks. The test is available for download via www.gfmt2.org.

43 498

45 499 **OPEN PRACTICES STATEMENT**

46 500 Detailed item performance and metadata used to create the test are available as part of the
47
48 501 test distribution folder available via www.gfmt2.org [for the purpose of peer review these
49
50 502 can be accessed at: <https://tinyurl.com/GFMT2peerReview>].

51 503

54 504 **REFERENCES**

55 505

56 506 Balsdon, T., Summersby, S., Kemp, R. I., & White, D. (2018). Improving face identification
57
58 507 with specialist teams. *Cognitive Research: Principles and Implications*, 3(1), 25.

GLASGOW FACE MATCHING TEST 2

19

1
2
3 508
4

5 509 Bate, S., & Bennetts, R. J. (2014). The rehabilitation of face recognition impairments: a
6
7 510 critical review and future directions. *Frontiers in Human Neuroscience*, 8, 491.

8
9 511

10 512 Benton, A. L., Hamsher, K. S., Varney, N. R., & Spreen, O. (1983). *Contributions to*
11
12 513 *neuropsychological assessment*. New York: Oxford University Press.

13
14 514

15 515 Bindemann, M., Brown, C., Koyas, T., & Russ, A. (2012). Individual differences in face
16
17 516 identification predict eyewitness accuracy. *Journal of Applied Research in Memory and*
18
19 517 *Cognition*, 1, 96-103.

20
21 518

22
23 519 Bobak, A. K., Bennetts, R. J., Parris, B. A., Jansari, A., & Bate, S. (2016). An in-depth cognitive
24
25 520 examination of individuals with superior face recognition skills. *Cortex*, 82, 48-62.

26
27 521

28
29 522 Bowles, D. C., McKone, E., Dawel, A., Duchaine, B., Palermo, R., Schmalzl, L., ... & Yovel, G.
30
31 523 (2009). Diagnosing prosopagnosia: Effects of ageing, sex, and participant–stimulus ethnic
32
33 524 match on the Cambridge Face Memory Test and Cambridge Face Perception Test. *Cognitive*
34
35 525 *Neuropsychology*, 26, 423-455.

36
37 526

38 527 Bruce, V., & Young, A. W. (1986). Understanding face recognition. *British Journal of*
39
40 528 *Psychology*, 77, 305-327.

41
42 529

43
44 530 Burton, A. M., White, D., & McNeill, A. (2010). The Glasgow face matching test. *Behavior*
45
46 531 *Research Methods*, 42, 286-291.

47
48 532

49 533 Cao, Q., Shen, L., Xie, W., Parkhi, O. M., & Zisserman, A. (2018, May). Vggface2: A dataset for
50
51 534 recognising faces across pose and age. In *2018 13th IEEE International Conference on*
52
53 535 *Automatic Face & Gesture Recognition (FG 2018)* (pp. 67-74).

54
55 536

56 537 Davis, J. P., Lander, K., Evans, R., & Jansari, A. (2016). Investigating predictors of superior
57
58 538 face recognition ability in police super-recognisers. *Applied Cognitive Psychology*, 30, 827-
59
60 539 840.

- 1
2
3 540
4
5 541 DeGutis, J. M., Chiu, C., Grosso, M. E., Cohan, S. (2014). Face processing improvements in
6
7 542 prosopagnosia: Successes and failures over the last 50 years. *Frontiers in Human*
8
9 543 *Neuroscience*, 8, 561.
10
11 544
12 545 Dowsett, A. J., & Burton, A. M. (2015). Unfamiliar face matching: Pairs out-perform
13
14 546 individuals and provide a route to training. *British Journal of Psychology*, 106, 433-445.
15
16 547
17
18 548 Duchaine, B. C., & Nakayama, K. (2006). The Cambridge face memory test: Results for
19
20 549 neurologically intact individuals and an investigation of its validity using inverted face stimuli
21
22 550 and prosopagnosic participants. *Neuropsychologia*, 44, 576– 585.
23
24 551
25 552 Dunn, J. D., Summersby, S., Towler, A., Davis, J. P., & White, D. (2020). UNSW Face Test: A
26
27 553 screening tool for super-recognizers. *PloS one*, 15(11), e0241747.
28
29 554
30
31 555 Fysh, M. C., & Bindemann, M. (2018). The Kent face matching test. *British Journal of*
32
33 556 *Psychology*, 109(2), 219-231.
34
35 557
36 558 Guilford, J. P. (1954). *Psychometric methods*. New York: McGraw-Hill.
37
38 559
39
40 560 Hahn, C. A., O'Toole, A. J., & Phillips, P. J. (2016). Dissecting the time course of person
41
42 561 recognition in natural viewing environments. *British Journal of Psychology*, 107(1), 117-134.
43
44 562
45 563 Matthews, C. M., & Mondloch, C. J. (2018). Improving identity matching of newly
46
47 564 encountered faces: effects of multi-image training. *Journal of Applied Research in Memory*
48
49 565 *and Cognition*, 7, 280-290.
50
51 566
52
53 567 McCaffery, J. M., Robertson, D. J., Young, A. W., & Burton, A. M. (2018). Individual
54
55 568 differences in face identity processing. *Cognitive Research: Principles and Implications*, 3, 21.
56
57 569
58 570 Megreya, A. M., & Burton, A. M. (2006). Unfamiliar faces are not faces: Evidence from a
59
60 571 matching task. *Memory & Cognition*, 34, 865-876.

GLASGOW FACE MATCHING TEST 2

21

1
2
3 572
4

5 573 Mollon, J. D., Bosten, J. M., Peterzell, D. H., & Webster, M. A. (2017). Individual differences
6
7 574 in visual science: What can be learned and what is good experimental practice?. *Vision*
8
9 575 *Research*, 141, 4-15.

10 576

11
12 577 Murray, E., & Bate, S. (2020). Diagnosing developmental prosopagnosia: repeat assessment
13
14 578 using the Cambridge Face Memory Test. *Royal Society Open Science*, 7, 200884.

15 579

16
17
18 580 Noyes, E., & Jenkins, R. (2017). Camera-to-subject distance affects face configuration and
19
20 581 perceived identity. *Cognition*, 165, 97-104.

21 582

22
23 583 Noyes, E., Phillips, P. J., & O'Toole, A. J. (2017). What is a super-recogniser? In *Face*
24
25 584 *processing: Systems, disorders and cultural differences*. M. Bindemann & A. Megreya (Eds.).
26
27 585 Nova Science.

28 586

29
30
31 587 Phillips, P. J., Yates, A. N., Hu, Y., Hahn, C. A., Noyes, E., Jackson, K., ... & O'Toole, A. J.
32
33 588 (2018). Face recognition accuracy of forensic examiners, superrecognizers, and face
34
35 589 recognition algorithms. *Proceedings of the National Academy of Sciences*, 115(24), 6171-
36
37 590 6176.

38 591

39
40 592 Ramon, M., Bobak, A. K., & White, D. (2019). Super-recognizers: From the lab to the world
41
42 593 and back again. *British Journal of Psychology*, 110, 461-479.

43 594

44
45 595 Robertson, D. J., Noyes, E., Dowsett, A. J., Jenkins, R., & Burton, A. M. (2016). Face
46
47 596 recognition by metropolitan police super-recognisers. *PloS one*, 11, e0150036.

48 597

49
50
51 598 Russell, R., Duchaine, B., & Nakayama, K. (2009). Super-recognizers: People with
52
53 599 extraordinary face recognition ability. *Psychonomic Bulletin & Review*, 16, 252-257.

54 600

55
56 601 Shakeshaft, N. G., & Plomin, R. (2015). Genetic specificity of face recognition. *Proceedings of*
57
58 602 *the National Academy of Sciences*, 112, 12887-12892.

59 603
60

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

- 604 Towler, A., Kemp, R. I., Burton, A. M., Dunn, J. D., Wayne, T., Moreton, R., & White, D.
605 (2019). Do professional facial image comparison training courses work?. *PloS one*, 14,
606 e0211037.
- 607
- 608 Towler, A., Keshwa, M., Ton, B., Kemp, R. I., & White, D. (in press). Diagnostic feature
609 training improves face matching accuracy. *Journal of Experimental Psychology: Learning,*
610 *Memory & Cognition.*
- 611
- 612 White, D., Kemp, R. I., Jenkins, R., & Burton, A. M. (2014). Feedback training for facial image
613 comparison. *Psychonomic Bulletin & Review*, 21, 100-106.
- 614
- 615 White, D., Kemp, R. I., Jenkins, R., Matheson, M., & Burton, A. M. (2014). Passport officers'
616 errors in face matching. *PloS one*, 9(8), e103510.
- 617
- 618 White, D., Phillips, P. J., Hahn, C. A., Hill, M., & O'Toole, A. J. (2015). Perceptual expertise in
619 forensic facial image comparison. *Proceedings of the Royal Society B: Biological*
620 *Sciences*, 282, 20151292.
- 621
- 622 White, D., Rivolta, D., Burton, A. M., Al-Janabi, S., & Palermo, R. (2017). Face matching
623 impairment in developmental prosopagnosia. *Quarterly Journal of Experimental*
624 *Psychology*, 70(2), 287-297.
- 625
- 626 White, D., Towler, A., & Kemp, R. I. (2020). Understanding professional expertise in
627 unfamiliar face matching. In M. Bindemann (Ed.), *Forensic face matching: Research and*
628 *practice*. Oxford University Press.
- 629
- 630 Wilhelm, O., Herzmann, G., Kunina, O., Danthiir, V., Schacht, A., & Sommer, W. (2010).
631 Individual differences in perceiving and recognizing faces—One element of social
632 cognition. *Journal of Personality and Social Psychology*, 99, 530.

633

GLASGOW FACE MATCHING TEST 2

23

- 1
2
3 634 Wilmer, J. B., Germine, L., Chabris, C. F., Chatterjee, G., Williams, M., Loken, E., ... &
4
5 635 Duchaine, B. (2010). Human face recognition ability is specific and highly heritable.
6
7 636 *Proceedings of the National Academy of Sciences*, 107, 5238-5241.
8
9 637
10 638 Wilmer, J. B., Germine, L., Chabris, C. F., Chatterjee, G., Gerbasi, M., & Nakayama, K. (2012).
11
12 639 Capturing specific abilities as a window into human individuality: The example of face
13
14 640 recognition. *Cognitive Neuropsychology*, 29(5-6), 360-392.
15
16 641
17
18 642 Wilmer, J. B. (2017). Individual differences in face recognition: A decade of
19
20 643 discovery. *Current Directions in Psychological Science*, 26(3), 225-230.
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

For Review Only