



OPEN Normative face recognition ability test scores vary across online participant pools

B. Popovic¹, J. D. Dunn¹, A. Towler² & D. White¹✉

Online participant recruitment is a cornerstone of modern psychology research. While this offers clear benefits for studying individual differences in cognitive abilities, test performance can vary across lab-based and web-based settings. Here we assess the stability of normative test scores across popular online recruitment platforms and in-person testing, for three standard measures of face identity processing ability: the GFMT2, CFMT+, and MFMT. Participants recruited via Amazon Mechanical Turk (MTurk) scored approximately 10 percentage points lower in all tests compared to those recruited through Prolific and university students tested in the lab. Applying stricter exclusion criteria based on attention checks resulted in notably higher exclusion rates for the MTurk group (~62%) compared to the Prolific group (~22%), yet even after exclusion, some test scores remained lower for MTurk participants. Given that the GFMT2 subtests were developed using MTurk participants, we provide updated normative scores for all subtests (GFMT2-Short, GFMT2-Low, GFMT2-High) and further recommendations for their use. We also confirm the robust psychometric properties of the GFMT2-Short and GFMT2-High, demonstrating strong test-retest reliability, convergent validity with other established tests, and high diagnostic value in identifying super-recognisers. The GFMT2 subtests are freely available for use in both online and in-person research via www.gfmt2.org.

Keywords Cambridge face memory test, Glasgow face matching test, Congenital prosopagnosia, Developmental prosopagnosia, Face processing, Face recognition, Individual differences, Prosopagnosia, Super-recognizers

Web-based testing has become an essential tool in psychological research, with Amazon Mechanical Turk (MTurk) remaining the most popular platform. Recent keyword analyses indicate that MTurk is used in 30–40% of papers in top psychology journals^{1,2}. This platform, along with others, has enabled large-scale testing of cognitive and perceptual abilities, extending beyond mere convenience. Web-based recruitment allows for diverse geographic sampling and facilitates access to specific participant groups, such as twins³, clinical populations⁴, and extreme performance groups⁵. Compared to traditional recruitment methods, such as undergraduate populations, online recruitment is also more economical and provides greater demographic diversity.

However, several studies have identified potential pitfalls in online recruitment. Data quality issues have been noted with MTurk participants, particularly in contrast to platforms like Prolific that focus on providing high-quality participants for research^{6,7}. These issues include high rates of inattentive responses⁸, meaningless data⁹, and fraudulent participation using IP masking¹⁰. Such factors can reduce the effectiveness of experimental manipulations¹⁰ and affect the mean and standard deviation of scale measurements¹¹, although some studies do not report such effects¹². While stricter exclusion criteria can mitigate these issues, their effect on experimental outcomes is often modest¹⁰.

In this study, we explore how recruitment pools influence performance on tests of face identity processing. This refers to tasks that involve comparing face images perceptually or remembering faces to later identify them. Both lab-based and online studies have demonstrated that face identity processing varies across the population¹³, from individuals with developmental prosopagnosia, who have significant deficits, to super-recognisers, who perform near-perfectly on standardised tests¹⁴. In applied contexts, assessing high performers can have real-world implications for identity verification tasks^{15,16}, and identifying deficits is clinically important for understanding disorders of social cognition¹⁷. Theoretically, reliable tests can provide insights into the cognitive systems that support face perception^{13,18}.

Initial web-based face processing tests were conducted via testmybrain.org, which has since tested over 2.5 million participants¹⁹. Performance on these tests is comparable to lab-based samples²⁰, though this platform

¹School of Psychology, The University of New South Wales, Sydney, NSW, Australia. ²School of Psychology, The University of Queensland, Brisbane, QLD, Australia. ✉email: david.white@unsw.edu.au

typically attracts motivated participants interested in learning about their cognitive abilities. Other self-selecting pools, such as those recruited via registries or targeted links (e.g., super-recogniser tests), often yield skewed performance, with participants scoring higher than average^{21,22}.

Given this, researchers should also have access to relatively unbiased online samples like those provided by MTurk and Prolific. However, some studies have reported lower accuracy on face identity tasks for MTurk participants compared to lab-based samples¹⁴. For instance, MTurk participants scored 61.4% on the Cambridge Face Memory Test (CFMT) compared to 69.3% for undergraduate students^{23–25}. Similarly, bespoke tests of face identity processing have shown lower accuracy for MTurk samples compared to student populations²⁶.

Performance on face memory tests among Prolific participants has also been lower than normative scores established in lab-based testing. For example, accuracy on the CFMT ranged from 68 to 75% in recent studies^{27–29}, which is well below the 80.4% found in lab-based samples^{30–32}. This pattern is consistent for the Australian version of the test (CFMT-Aus), with Prolific participants scoring 71.5%, compared to 80.2% in lab-based samples³³. Other Prolific studies have found mixed evidence of differences between lab and online samples, for example accuracy on the Cambridge Face Memory Test—Extended Version (CFMT+,¹⁴) was 2% lower in³⁴, with other tests in the same study showing test scores between 3 and 15% lower.

Test score variation in online settings is not limited to memory tasks but extends to perceptual matching tasks, such as the Glasgow Face Matching Test 2 (GFMT2). Normative data from MTurk participants for the GFMT2-Short subtest (GFMT2-S) showed 75% accuracy, consistent with subsequent studies on MTurk³⁵. However, Prolific participants have performed slightly better, with reported accuracies of 80–82%^{34,36,37}. Moreover, MTurk participants have shown lower performance on other face matching tasks compared to lab-based samples^{35,38}.

In this study we recruited undergraduate students and three online samples to complete the GFMT2 and other standard tests of face identity processing ability (CFMT+¹⁴, Models Face Matching Test³⁹). Details of these tests and participant groups can be found in the Methods section. To pre-empt our results, we found that M-Turk participants scored markedly lower on all tests. Because initial development of the GFMT2 relied heavily on MTurk data⁴⁰, we then decided to assess psychometric properties of the GFMT2. We find robust psychometric properties of the GFMT2 subtests that are consistent with the original test publication⁴⁰, including: high test–retest reliability, convergent validity with other face identity processing tests (GFMT⁴¹, CFMT+¹⁴, UNSW Face Test²¹) and diagnostic value in distinguishing between super-recognisers and standard participant groups. Updated normative scores for all GFMT2 subtests are provided in Table 1.

Results
Stability of GFMT2-S and GFMT2-H test scores across four recruitment methods

We administered two GFMT2 subtests (GFMT2-S, GFMT2-H) to participants from three online recruitment pools (MTurk, Prolific, UNSW Face Research Registry) and a group of UNSW undergraduates tested in-person. Participants from Prolific also completed the GFMT2-Low. Participants scoring below chance or using the same response key throughout a test were excluded prior to any analysis. Demographic details of the final participant groups are shown in Table 1, with additional information about the participant groups in the “Methods” (see: “Participant cohorts, procedure, and exclusion criteria”).

Normative test scores for the GFMT2 subtests are displayed in Table 1. We found substantial variation in group averages for both the GFMT2-S (74–89%) and GFMT2-H (67–82%). MTurk participants showed scores equivalent to those reported in⁴⁰ and³⁵. However, MTurk scores were lower than those of other groups, suggesting MTurk may not provide generalizable normative measures. We found no demographic differences to explain these results (see Supplementary Material 2; Figures S1-2 and Tables S2-4).

Accuracy for Prolific and in-person samples was equivalent, but were both approximately 5 percentage points lower than the UNSW Face Research Registry group. This elevated performance is consistent with the self-selection in recruitment of this group, as evidenced by their scores on other standard tests of face recognition [CFMT+: $M = 78.8\%$, $SD = 12.6\%$; $t(712) = 20.0$, $p < 0.001$, Cohen’s $d = 0.75$; GFMT: $M = 92.8\%$, $SD = 7.74\%$; $t(761) = 41.0$, $p < 0.001$, Cohen’s $d = 1.48$; UNSW Face Test: $M = 64.7\%$, $SD = 6.51\%$; $t(995) = 18.90$, $p < 0.001$, Cohen’s $d = 0.92$].

	n	Age	GFMT2-S	GFMT2-H	GFMT2-L
MTurk (web)	91	31.8 (8.4)	74.1 (10.9)	67.1 (10.3)	–
Prolific (web)	100	41.0 (15.2)	82.9 (7.46)	73.9 (9.56)	90.4 (6.67)
UNSW Registry (web)	1388	46.8 (13.9)	88.5 (7.61)	81.6 (9.70)	–
UNSW students (in-person)	92	19.8 (4.3)	84.9 (8.20)	75.9 (9.98)	–
White et al. 2022 (MTurk)	108	38.0 (11.3)	75.0 (10.0)	–	–

Table 1. Mean GFMT2 subtest percent correct scores across participant samples (standard deviations in parenthesis). Original normative data from White et al.⁴⁰ is included for comparison, but we suggest that researchers adopt normative scores from the Prolific sample, given the relative demographic heterogeneity and high task engagement we observed in this cohort (see main text in this and next sections). See Table S1 in Supplementary Material for test scores separated by match and non-match trials.

	n	Age	CFMT+ percent correct	MFMT percent correct
Amazon MTurk (all devices)	95	35.4(11.3)	58.9 (14.7)	62.3 (8.8)
Amazon MTurk (computer only)	76	30.8 (3.59)	59.8 (14.6)	61.9 (8.5)
Prolific (computer only)	100	44.8 (14.0)	67.9 (13.2)	71.3 (8.2)
Amazon MTurk (all devices + attention check)	68	35.5 (12.0)	62.2 (14.4)	64.6 (8.69)
Amazon MTurk (computer only + attention check)	35	29.8 (4.5)	67.7 (13.4)	62.2 (8.6)
Prolific (computer only + attention check)	78	45.6(14.3)	68.2(13.1)	72.1(7.4)

Table 2. CFMT+ and MFMT scores for different web-based participant samples and exclusion criteria. See main text for details.

	GFMT2-S	GFMT2-H	GFMT	CFMT+
GFMT2-S	–	***	***	***
GFMT2-H	0.721	–	***	***
GFMT	0.604	0.649	–	***
CFMT+	0.341	0.358	0.362	–

Table 3. Spearman’s correlations between performance on face recognition tests for UNSW Students. (n = 92; ***p < 0.001).

Stability of CFMT+ and MFMT normative test scores across recruitment methods

To assess whether the lower performance of MTurk participants was specific to the GFMT2, we recruited three additional groups. We recruited two groups from MTurk—one that could use any device to complete the tests and another that was limited to participants using a desktop or laptop computer⁴². An additional group of Prolific participants were recruited and limited to using desktop or laptop computers. They completed the Cambridge Face Memory Test—Extended Version (CFMT+¹⁴), which measures ability to remember unfamiliar faces, and the Models Face Matching Test (MFMT²⁸), which measures ability to perceptually match the identity of images showing faces of unfamiliar male models.

Table 2 shows average test scores. The top three rows show accuracy when the same general exclusion criteria as for the GFMT2 participant groups were applied. Consistent with the GFMT2 results, MTurk participants scored approximately 10 percentage points lower than Prolific participants, and there was no significant difference between participants using any device and those that only used computers (see Supplementary Material 3 for statistical comparisons).

For tests in Table 2, we also included attention checks that were designed to assess participants’ engagement with the task. First, participants completed three practice trials using easily recognisable cartoon faces (*The Simpsons* characters). Additionally, free text responses to demographic questions regarding country of birth and the number of countries lived in were manually reviewed. We then excluded participants who failed our attention check criteria, by either failing to correctly answer all three practice questions or providing inappropriate or incorrect responses to demographic questions (e.g., entering a year instead of a country).

The effectiveness of including these attention checks is shown in the bottom 3 rows of Table 3. The attention checks led to a 7% improvement in CFMT+ scores, but little change in MFMT scores, despite removing over 60% of MTurk participants from the sample. Applying the same criteria to Prolific led to fewer exclusions (22%) and no effect on average scores, suggesting low engagement contributed to lower scores for MTurk participants. For more details on these participant groups and exclusion criteria see “Methods” (“Participant cohorts, procedure and exclusion criteria”).

Psychometric properties of GFMT2 subtests

Given we developed the GFMT2 using MTurk participants, the observed discrepancies in MTurk results prompted us to reevaluate the psychometric properties of the GFMT2-S and GFMT2-H, including: (i) test–retest reliability, (ii) internal reliability, (iii) convergent validity, (iv) sensitivity to participant age, and (v) diagnostic value of subtests for identifying super-recognisers.

Test–retest reliability

High test–retest reliability for the GFMT2-S ($r = 0.774$) was reported in⁴⁰ but was not measured for the GFMT2-H. In our study, 80 UNSW students and 713 UNSW Face Research Registry (UNSW-FRR) participants completed both tests twice, with UNSW students having a similar retest interval to⁴⁰ ($M = 7.3$ days; $SD = 2.16$), and UNSW Face Research Registry participants having a longer interval ($M = 50.0$ days; $SD = 6.08$).

As shown in Fig. 1, test–retest correlations were high for both tests in both groups, with slightly lower correlations in the UNSW-FRR sample (GFMT2-S *Spearman’s rho* = 0.734; GFMT2-H *Spearman’s rho* = 0.675)

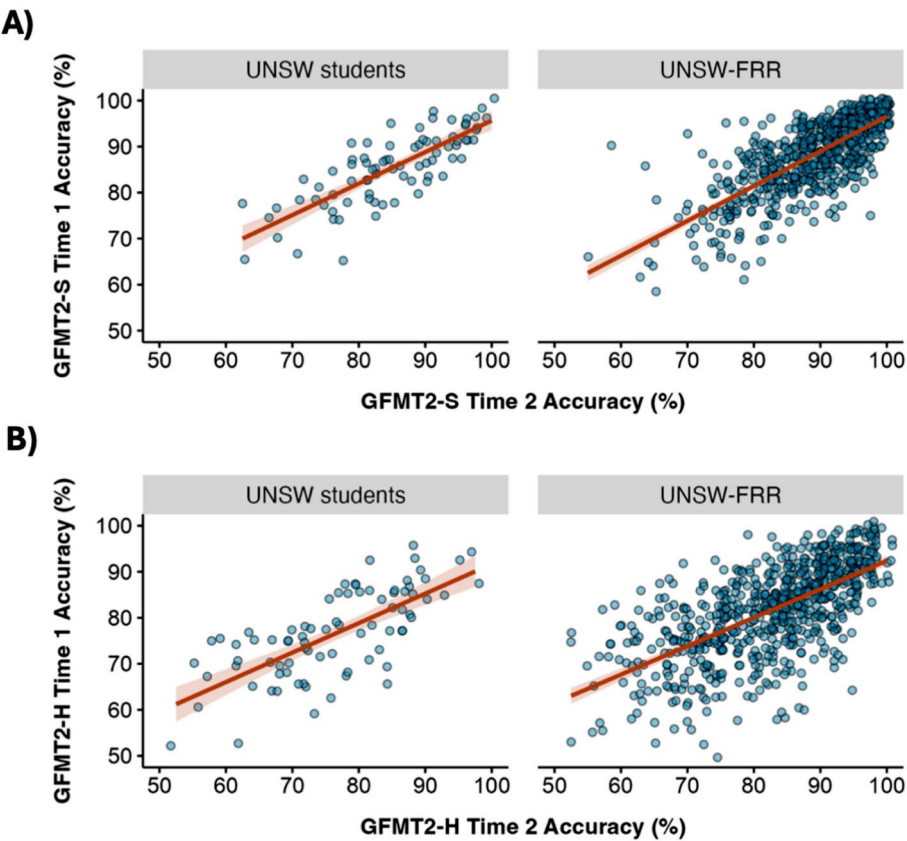


Fig. 1. Correlations between accuracy at time 1 and time 2 for the GFMT2-S (A) and GFMT2-H (B). The orange shaded region around each regression line represents the 95% confidence interval of the regression estimate.

	GFMT2-S	GFMT2-H	GFMT	CFMT+	UNSWFT
GFMT2-S	–	***	***	***	***
GFMT2-H	0.707	–	***	***	***
GFMT	0.570	0.581	–	***	***
CFMT+	0.532	0.544	0.566	–	***
UNSWFT	0.444	0.457	0.499	0.537	–

Table 4. Spearman’s correlations between performance on face recognition tests for UNSW Face Research Registry participants (n = 687; *** p < 0.001).

than the UNSW student sample (GFMT2-S *Spearman’s rho* = 0.801; GFMT2-H *Spearman’s rho* = 0.714), presumably due to the longer retest interval. Nevertheless, high reliability over 7 weeks provides compelling evidence that individual differences persist over extended periods and can be reliably measured by GFMT2 subtests.

Internal reliability

We calculated internal reliability of the GFMT2 subtests, in addition to the CFMT+ and MFMT, using Cronbach’s alpha. This showed good internal reliability for all subtests on Prolific samples (GFMT-S: 0.851; GFMT-L: 0.738; GFMT-H: 0.841). These scores were higher than for the CFMT+ in our Prolific sample (0.520) and comparable to the MFMT (0.888). We note however that this estimate of CFMT+ internal reliability is markedly lower than that reported in previous publications (e.g. 0.9 in²²). Interestingly, despite the data quality issues associated with our MTurk sample noted above, the measures of internal reliability computed from this sample were consistently higher for all tests (GFMT-S: 0.929; GFMT-H: 0.910; CFMT+: 0.838; MFMT: 0.929).

Convergent validity

We computed Spearman’s correlations between GFMT2 subtest scores and three standardized face identification tests (UNSWFT, CFMT+, GFMT). Tables 3 and 4 show strong convergence between the GFMT2 and GFMT, with

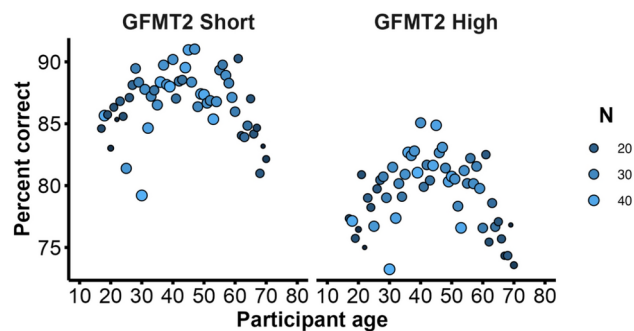


Fig. 2. Average accuracy for each participant age on the GFMT2-S (left) and GFMT2-H (right) where the size and shade of each data point shows the number of participants in each age group. A formula for age-correcting scores is provided in Supplementary Material 4.

slightly lower correlations for memory-based tasks (CFMT+, UNSWFT). Interestingly, for the UNSW students tested in-person this pattern is especially pronounced, showing a statistically higher correlation between the GFMT2 subtests and GFMT compared to the correlation between the GFMT2 and CFMT+ [GFMT2-S: Fisher's z -test = 2.66, $p = 0.004$; GFMT2-H: Fisher's z -test = 3.20, $p = 0.001$]. These same contrasts were non-significant for the UNSW-FRR participants.

We speculate this difference relates to the relative homogeneity of the UNSW student sample in terms of age and testing conditions. Controlling for these sources of variation in accuracy may have provided greater precision in isolating variance attributable to the processing differences between the CFMT+ and GFMT tests.

Sensitivity of test scores to participant age

We examined age-related accuracy patterns by aggregating test scores from all cohorts. Figure 2 shows a pattern consistent with prior research^{21,43}, with peak accuracy estimated at age 36 for both GFMT2-S and GFMT2-H. A formula for age-correcting scores is provided in Supplementary Material 4.

Diagnostic value of the GFMT2 for identifying super-recognisers

The original GFMT has been used extensively to test staff in applied settings, with a view to selecting high performers for specialist face identity roles^(44,45 for a review see 46). It is also used as part of test batteries to identify 'super-recognisers' in research settings (e.g.,^{15,47}). However, it suffers from high average accuracy making it poorly suited for this purpose⁴⁰. Because of this, we specifically designed a subtest of the GFMT2—the GFMT2-High—to be more challenging than the primary GFMT2-S subtest, and therefore better calibrated for identifying high-performers.

We tested the effectiveness of the GFMT2-S and GFMT2-H in discriminating super-recognisers from a standard participant group. Super-recognisers were the sample of UNSW Face Research Registry super-recognisers featured in the convergent validity analyses above ($n = 97$; 58 female, 37 male, 2 prefer a different term; $M_{age} = 39.3$, $SD = 9.69$, 1 missing age), who scored at least 1.7 standard deviations above established norms on each of three standard face recognition tests: UNSW Face Test (UNSWFT²¹), Glasgow Face Matching Test (GFMT⁴¹), and the Cambridge Face Memory Test – Extended Version (CFMT+¹⁴). The standard participant group combined the Prolific participants and UNSW students who completed the GFMT2-S and GFMT2-H ($n = 192$; 107 female, 85 male; $M_{age} = 30.8$, $SD = 15.6$). We did not include the UNSW-FRR cohort here because they self-selected for participation in the tests and show above average accuracy. For completeness, we report the same analysis using this group in Supplementary Material 5.

Super-recognisers scored substantially higher on the GFMT2-S ($M = 95.7\%$, $SD = 3.15\%$) and the GFMT2-H ($M = 91.5\%$, $SD = 4.81\%$) compared to the standard participant group (GFMT2-S: $M = 83.8\%$, $SD = 7.86\%$; GFMT2-H: $M = 74.9\%$, $SD = 9.79\%$). To quantify how well each GFMT2 subtest discriminated between super-recognisers and the standard participant group, we use Area Under the Receiver Operator Characteristic curve (AUC) which measures the extent to which test scores can differentiate between two classes (i.e., super-recognisers vs standard participant group). The larger the AUC score, the better the face matching test is at differentiating between the groups. The GFMT2-H had slightly higher discriminative power ($AUC = 0.9368$) compared to the GFMT2-S ($AUC = 0.9297$). Combining the GFMT2-S and the GFMT2-H scores produced even higher diagnostic value ($AUC = 0.949$). Overall, this analysis shows high diagnostic value of the GFMT2 subtests for categorising super-recognisers, with the GFMT2-H providing an efficient alternative where short test sessions are desirable.

Discussion

We found systematic differences in face identity processing test scores collected on Amazon Mechanical Turk (MTurk) compared with other web-based and in-person samples. On average, MTurk participants performed approximately 10 percentage points below mean accuracy established from other participant groups. This difference was observed for standard tests of both face memory and perceptual matching. Using stricter data quality cleaning processes, that screened for inattentive participants, brought some average test scores closer to those observed in other cohorts—but at the cost of removing around 60% of the sample.

The reasons for this are uncertain, but it is consistent with a more general decline in the quality of MTurk participant data since 2018¹¹. It is possible that some of this is attributable to non-human ‘bot’ participants⁴⁸. However, our screening procedures for all testing included reCAPTCHA, and subsequent tests used manual checking of qualitative responses and catch trials that have been shown to be effective in identifying bots. This suggests that differences in accuracy are due to differences in human respondents. These differences might include motivation, task engagement, testing conditions or potentially underlying differences in the ability level of MTurk participants.

Overall test completion times showed that MTurk participants spent substantially longer than Prolific participants, which may suggest they were completing the task in parallel to other tasks or in multiple sessions (see Supplementary Material 6; Tables S5–6). The GFMT2 is self-paced, and response time is collected automatically in the desktop versions of the test but trial-level response time data was not collected in our online study due to limitations of the Qualtrics survey software. Users may wish to monitor response times in future online testing and so we have since created new online versions of the GFMT2 that collect this data and can be shared with researchers on request.

Other studies have found lower accuracy of MTurk participants on face identity processing tasks. In a recent study, GFMT2-S accuracy was 75%³⁵ which is in line with normative measures reported in the MTurk sample used in the initial GFMT2 publication⁴⁰. This study also reported lower accuracy in MTurk participants for the Expertise in Facial Comparison Test than had been reported in a group of university students tested in-person (White et al. 2015). Interestingly,³⁵ also reported higher scores on the self-report measure of face recognition ability using MTurk participants (PI-20⁴⁹). This suggests MTurk participants also believe they had poorer-than-average face recognition ability. MTurk workers score markedly higher on scales that are associated with poor face recognition ability compared to the overall population (see⁵⁰ for a review), such as those measuring traits of social anxiety⁵¹ and autism^{52,53}. This might suggest that poorer accuracy in MTurk participants is due to differences in this cohort that extend beyond simple motivation.

Our results also provide a more complete set of psychometric tools for researchers that use the GFMT2 to measure unfamiliar face matching ability. Normative data is now available for all subtests (GFMT2-S, GFMT2-H, GFMT2-L), however our results clearly show that MTurk is not suitable for estimating normative accuracy on face identity processing tests. Normative test scores acquired from MTurk could lead to underdiagnosis of impaired ability and overestimates of the prevalence of people with extremely high ability. Given the relative heterogeneity of the Prolific sample, in addition to evidence of high task engagement, we suggest researchers adopt these scores as normative measures of test performance on the GFMT2-S ($M=82.9\%$; $SD=7.46\%$), GFMT2-H ($M=73.9\%$; $SD=9.56\%$) and GFMT2-L ($M=90.4\%$; $SD=6.67\%$). Where appropriate for diagnostic purposes, individual scores on the GFMT2-S and GFMT2-H can be adjusted for age using the formulas found in Supplementary Material 4.

We found the GFMT2-S and GFMT2-H were highly reliable and valid measures of unfamiliar face matching ability. Test–retest reliability remained high even when we conducted follow-up tests from 6 weeks after an initial test. Convergent validity analysis shows the test correlates highly with other tests of face identity processing ability. There was also evidence of some discriminant validity with face memory tasks, which is consistent with studies showing face memory and matching tasks rely on somewhat different abilities (e.g.^{41,54–57}). This suggests that GFMT2 subtests are a complementary tool for use alongside face memory tests to provide an overall picture of face identity processing ability.

We assessed the value of the GFMT2-S and GFMT2-H in discriminating between super-recognisers—that we have identified in our prior work (e.g., see^{21,47})—from standard performers. We found both tests were able to classify super-recognisers with high diagnostic accuracy, and so both tests will be useful in applied settings where super-recognisers are selected for specialist face identity processing roles^{45,58–60}, or for theoretically motivated studies of the underlying perceptual mechanisms^{47,61–66}. Although the best results were found when combining both the GFMT2-S and the GFMT2-H, we also found the 40-item GFMT2-H was slightly better than the 80-item GFMT2-S in identifying super-recognisers, suggesting it is a more suitable tool for short test sessions aimed at finding super-recognisers.

Future studies could aim to determine optimal uses of the GFMT2-S and GFMT2-L for identifying individuals at the opposite end of the ability spectrum that have impairments in face identity processing. Given that impaired perceptual encoding of face identity appears to be a key deficit in developmental prosopagnosia³², it is important to include perceptual matching tasks in diagnostic testing. Our analysis of psychometric properties shows the GFMT2 is able to reliably measure—and specifically target—perceptual identity processing ability, supporting its inclusion in comprehensive test batteries of face identity processing ability.

More generally, our results show that normative scores on popular cognitive measures of face processing ability vary across different participant cohorts. Some of this variation is attributable to differences in participant demographics, for example age, and some can be removed by screening participant responses with quality control measures. But normative scores are also likely to covary with factors that we have not measured or controlled here. Further, our study has focussed on differences between lab-based testing of undergraduate students and online testing, but it is also likely that accuracy of undergraduate samples differs across different test sites. Together this suggests that researchers should view published normative scores on tests of face identity processing ability not as properties of the test alone, but in the context of the specific cohort that was tested.

Methods

The studies reported were all approved by UNSW Human Research Ethics Advisory Panel. As a condition of this ethical approval, informed consent was obtained from all subjects and all methods were performed in accordance with the relevant guidelines and regulations.

Tests of face identity processing ability

Glasgow Face Matching Test 2 (GFMT2⁴⁰)

The GFMT2 is an expanded version of the original GFMT⁴¹ designed to assess unfamiliar face matching ability. It comprises three subtests created from a pool of 300 pairs of face images: the 80-item GFMT-Short (GFMT2-S), the GFMT2-High (GFMT2-H), and the GFMT2-Low (GFMT2-L). Each pair consists of a high-quality frontal image and either a same-identity or different-identity face. The difficulty levels of the pairs vary by type of variation: rigid (e.g., head angle), non-rigid (e.g., expression), or distance variation. For this study, we focus on the psychometric properties of the GFMT2-S and GFMT2-H, although we also provide normative test scores for the GFMT-Low (GFMT2-L; see Table 1). Further details of the GFMT2 subtest development and example face pairs are provided in the original publication⁴⁰.

Glasgow Face Matching Test (GFMT⁴¹)

The format and design of the original GFMT is the same as the GFMT2, and we included the GFMT in this study to assess convergent validity of the GFMT2. Test images are frontal with a neutral expression and consistent subject-to-camera distance making the task somewhat easier than the GFMT2. The most widely used version of this test contains 40 test items of unfamiliar face matching ability (20 match, 20 non-match).

Cambridge Face Memory Test—Long Form (CFMT¹⁴)

The CFMT+ evaluates memory for unfamiliar faces. Participants learn a series of faces and later identify the learned faces from arrays containing distractors. As the test progresses, visual noise is added to obscure the faces, thus increasing task difficulty. In total there are 102 test items.

UNSW Face Test (UNSWFT²¹)

The UNSWFT is a challenging test of general face identity processing ability that was originally developed as a screening test for high performing ‘super-recognisers’. Again, we included this in the battery of tests used to assess convergent validity of the GFMT2. It consists of two tasks which are completed in a fixed order. The first task is a standard recognition memory paradigm where participants memorise studio-quality face images and are later asked to recognise these faces in social-media-style photos. The second task is a match-to-sample sorting task where participants memorise a face image presented for 3 s. Immediately after the face image disappears from screen, they must sort a ‘pile’ of new, unseen face images as either belonging to the identity they just saw or not belonging to the identity.

Models Face Matching Test (MFMT³⁹)

The MFMT uses 90 pairs of face images of White male models, half of which are same-identity pairs. Participants judge whether each pair represents the same person or different people. The MFMT is designed to be a challenging test of face matching and so the face images are unconstrained and contain a lot of natural variability (e.g., head angle, lighting).

Participant cohorts, procedure and exclusion criteria

Amazon Mechanical Turk participant groups

Mechanical Turk was used in the initial GFMT2 test development and normative test data in⁴⁰. As in this prior work, we selected participants based on a screening procedure aimed at ensuring high-quality data. Eligible workers had to achieve a HIT approval rate of over 99%, indicating a strong track record of submitting high-quality work. Additionally, they were required to have completed more than 100 HITs to confirm sufficient experience with MTurk norms and expectations, and a *reCAPTCHA* security check was administered on Qualtrics to confirm human participation. This screening procedure was applied to all MTurk samples described below.

The first MTurk sample were 99 participants that completed the GFMT2-S and GFMT2-H. After excluding participants for pressing the same response key repeatedly ($n=2$) or performing below chance ($n=6$), 91 participants remained ($M_{age}=31.8$, $SD=8.40$; 28 female, 63 male). Performance of this group is presented in Table 1.

A second sample of 106 participants completed the CFMT+ and MFMT. This sample was not restricted by device type and could complete these two tests using any device, including mobile devices. Participants were excluded from analysis if they repeatedly pressed the same or different response buttons for at least one of the tests ($n=1$) or performed below chance on the CFMT+ or the MFMT ($n=10$), resulting in a final sample of 95 participants ($M_{age}=35.4$, $SD=11.3$; 31 female, 64 male). This sample is referred to as ‘Amazon MTurk (all devices)’ in Table 2.

The third sample of 98 MTurk participants completed the CFMT+ and MFMT using computers only (i.e. laptop and desktop PCs), with exclusions due to repeated pressing of the same button ($n=4$) or below chance accuracy ($n=18$) resulting in a final sample of 76 ($M_{age}=30.8$, $SD=3.59$; 21 female, 55 male). This sample is referred to as ‘Amazon MTurk (computer only)’ in Table 2.

In addition to the general exclusion criteria detailed above, we also identified groups of individuals who failed strict attention checks from both the Amazon MTurk (all devices) and Amazon MTurk (computer only) samples to create subsamples including high quality data only. These attention checks involved passing catch trials in both the CFMT+ and MFMT, as well as a strict manual review of free-text responses (e.g., country of birth, number of countries lived in). In the Amazon MTurk (all devices) sample, 27 participants failed the attention checks resulting in a strict sample of 68 participants ($M_{age}=35.5$, $SD=12.0$; 22 female, 46 male). In the Amazon MTurk (computer only) sample, 41 participants failed the attention checks resulting in a strict sample

of 35 participants ($M_{age} = 29.8$, $SD = 4.5$; 12 female, 23 male). These samples are denoted by the label ‘+ attention check’ in Table 2.

Prolific participant groups

Prolific was used to establish normative data as it is a commonly used online research platform and reported to have superior data quality than MTurk for a range of measures (e.g., attention, reliability, comprehension, honesty⁶). All Prolific participant groups had to meet screening criteria, including using a computer device only and residing in the UK and the same basic exclusion criteria used for the Amazon MTurk samples were applied (i.e., removing repeated pressing of same key or scoring below-chance accuracy).

One sample of 100 participants completed the GFMT2-S and GFMT2-H, with no participants failing basic exclusion criteria ($M_{age} = 41.0$, $SD = 15.2$; 47 female, 53 male). We recruited a separate sample of 100 Prolific participants to complete the GFMT2-L with no participants failing basic exclusion criteria ($M_{age} = 43.6$, $SD = 13.5$; 55 female, 44 male, 1 prefers a different term). Finally, a third group of 100 participants completed the CFMT+ and MFMT ($M_{age} = 44.8$, $SD = 14.0$; 46 female, 54 male). This sample is referred to as ‘Prolific (computer only)’ in Table 2. Additionally, we applied the same strict attention checks as used for the Amazon MTurk samples, resulting in the exclusion of 22 participants. This sample is referred to as ‘Prolific (computer only) + attention check’ in Table 2 ($n = 78$; $M_{age} = 45.6$, $SD = 14.3$; 38 female, 40 male).

UNSW Face Research Registry (UNSW-FRR; online testing group 1)

Research volunteers have the option to join this participant registry after completing the UNSW Face Test (www.unswfacetest.com). People typically complete the UNSW Face Test via weblinks posted on media coverage of our research on super-recognisers (e.g., see⁶⁷). As a result, this cohort displays a self-selecting bias towards higher than average scores on standardised tests of face identity processing ability^{21,22,68}. While they are instructed to complete the study using a computer, there are no technical restrictions preventing the use of other devices.

The UNSW-FRR cohort comprised 1393 individuals who had responded to our recruitment email and completed both the GFMT2-S and GFMT2-H. Participants were excluded from analysis if they repeatedly pressed the same or different response buttons for at least one of the subtests ($n = 3$), or performed below chance on the GFMT2-S or GFMT2-H ($n = 2$), resulting in a final sample of 1388 participants (890 female, 480 male, 10 prefer a different term, 4 prefer not to answer, 4 individuals with missing data; $M_{age} = 46.8$, $SD = 13.9$, 8 individuals with missing age data).

To assess the test–retest reliability of the GFMT2-S and GFMT2-H, we contacted participants who completed these tests in the first session and invited them to repeat the test 6 weeks later, with 725 participants completing the tests a second time. Participants were excluded from test–retest analysis if they repeatedly pressed the same button ($n = 1$), or if they did not have a valid time 1 score ($n = 11$). This resulted in a final sample of 713 participants (454 female, 251 male, 4 prefer a different term; $M_{age} = 48.4$, $SD = 14.0$).

To assess convergent validity of GFMT2-S and GFMT2-H, we used existing performance data on 3 other standardised tests of face identity processing ability (UNSWFT, CFMT+, GFMT). Approximately half of the UNSW participants that completed the first test session ($n = 703$) had completed all three of these other tests. After excluding participants who performed below chance on at least one of these standardised tests ($n = 16$), this left a final sample of 687 participants ($M_{age} = 46.7$, $SD = 12.6$, 5 missing values; 440 female, 241 male, 4 prefer a different term, 1 preferred not to answer, 1 missing value).

UNSW students (in-person testing)

Ninety-four first year psychology students (undergraduate) from UNSW Sydney participated in return for course credit. Participants completed the study under experimenter supervision in our research lab on a desktop computer. Two participants were excluded from analysis due to technical issues while completing the study, so this resulted in a final lab sample of 92 participants ($M_{age} = 19.8$, $SD = 4.31$; 60 female, 32 male). We also applied general exclusion criteria to this sample (i.e., repeatedly pressing the same button and/or below chance accuracy), however, all participants passed these checks.

To assess test–retest reliability and convergent validity, participants in this cohort were invited to complete a second testing session one week after the first. Eighty-four participants returned to complete the second session. Four participants were excluded from analysis because of either technical issues ($n = 2$), repeatedly pressing the same response key for at least one of the tests ($n = 1$), or scoring below chance on at least one of the tests ($n = 1$). This resulted in a final sample of 80 participants ($M_{age} = 20.0$, $SD = 4.55$; 51 female, 29 male).

In each session the participants completed four tests: GFMT2-S, GFMT2-H, GFMT, and CFMT+. The order in which these four tests were completed was counterbalanced across participants, but each participant completed the tests in the same order in test session 1 and 2.

Data availability

Participant-level normative accuracy for the GFMT subtests and test materials are available on request via www.gfmt2.org. Other raw data is available via request directly to the corresponding author.

Received: 25 October 2024; Accepted: 28 February 2025

Published online: 14 March 2025

References

1. Anderson, C. A. et al. The MTurkification of social and personality psychology. *Pers. Soc. Psychol. Bull.* **45**, 842–850. <https://doi.org/10.1177/0146167218798821> (2019).

2. Uittenhove, K., Jeanneret, S. & Vergauwe, E. From lab-testing to web-testing in cognitive research: who you test is more important than how you test. *J. Cogn.* **6**, 259. <https://doi.org/10.5334/joc.259> (2023).
3. Wilmer, J. B. et al. Human face recognition ability is specific and highly heritable. *Proc. Natl. Acad. Sci. USA* **107**, 5238–5241. <https://doi.org/10.1073/pnas.0913053107> (2010).
4. Gehdu, B. K., Press, C., Gray, K. L. H. & Cook, R. Autistic adults have insight into their relative face recognition ability. *Sci. Rep.* **14**, 17802. <https://doi.org/10.1038/s41598-024-67649-8> (2024).
5. Bell, L., Duchaine, B. & Susilo, T. Dissociations between face identity and face expression processing in developmental prosopagnosia. *Cognition* **238**, 105469. <https://doi.org/10.1016/j.cognition.2023.105469> (2023).
6. Peer, E., Rothschild, D., Gordon, A., Evernden, Z. & Damer, E. Data quality of platforms and panels for online behavioral research. *Behav. Res. Methods* **54**, 1643–1662. <https://doi.org/10.3758/s13428-021-01694-3> (2022).
7. Douglas, B. D., Ewell, P. J. & Brauer, M. Data quality in online human-subjects research: Comparisons between MTurk, Prolific, CloudResearch, Qualtrics, and SONA. *PLoS ONE* **18**, e0279720. <https://doi.org/10.1371/journal.pone.0279720> (2023).
8. Albert, D. A. & Smilek, D. Comparing attentional disengagement between Prolific and MTurk samples. *Sci. Rep.* **13**, 20574. <https://doi.org/10.1038/s41598-023-46048-5> (2023).
9. Barends, A. J. & de Vries, R. E. Noncompliant responding: comparing exclusion criteria in MTurk personality research to improve data quality. *Pers. Individ. Differ.* **143**, 84–89. <https://doi.org/10.1016/j.paid.2019.02.015> (2019).
10. Kennedy, R. et al. The shape of and solutions to the MTurk quality crisis. *Polit. Sci. Res. Methods* **8**, 614–629. <https://doi.org/10.1017/psrm.2020.6> (2020).
11. Chmielewski, M. & Kucker, S. C. An MTurk crisis? Shifts in data quality and the impact on study results. *Soc. Psychol. Personal. Sci.* **11**, 464–473. <https://doi.org/10.1177/1948550619875149> (2020).
12. Keith, M. G., Stevenor, B. A. & McAbee, S. T. Scale mean and variance differences in MTurk and non-MTurk samples: a meta-analysis. *J. Pers. Psychol.* **22**, 1–12. <https://doi.org/10.1027/1866-5888/a000309> (2023).
13. White, D. & Burton, A. M. Individual differences and the multidimensional nature of face perception. *Nat. Rev. Psychol.* **1**, 287–300. <https://doi.org/10.1038/s44159-022-00041-3> (2022).
14. Russell, R., Duchaine, B. & Nakayama, K. Super-recognizers: people with extraordinary face recognition ability. *Psychon. Bull. Rev.* **16**, 252–257. <https://doi.org/10.3758/PBR.16.2.252> (2009).
15. Phillips, P. J. et al. Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Proc. Natl. Acad. Sci. USA* **115**, 6171–6176. <https://doi.org/10.1073/pnas.1721355115> (2018).
16. Towler, A. et al. Diverse types of expertise in facial recognition. *Sci. Rep.* **13**, 11396. <https://doi.org/10.1038/s41598-023-28632-x> (2023).
17. Webb, S. J., Neuhaus, E. & Faja, S. Face perception and learning in autism spectrum disorders. *Q. J. Exp. Psychol.* **70**, 970–986. <https://doi.org/10.1080/17470218.2016.1151059> (2017).
18. Bruce, V., Bindemann, M. & Lander, K. Individual differences in face perception and person recognition. *Cogn. Res. Princ. Implic.* **3**, 1–3. <https://doi.org/10.1186/s41235-018-0109-4> (2018).
19. Singh, S. et al. The TestMyBrain digital neuropsychology toolkit: development and psychometric characteristics. *J. Clin. Exp. Neuropsychol.* **43**, 786–795. <https://doi.org/10.1080/13803395.2021.2002269> (2021).
20. Germine, L. et al. Is the Web as good as the lab? Comparable performance from Web and lab in cognitive/perceptual experiments. *Psychon. Bull. Rev.* **19**, 847–857. <https://doi.org/10.3758/s13423-012-0296-9> (2012).
21. Dunn, J. D., Summersby, S., Towler, A., Davis, J. P. & White, D. UNSW face test: a screening tool for super-recognizers. *PLoS ONE* **15**, e0241747. <https://doi.org/10.1371/journal.pone.0241747> (2020).
22. Lara, A. & Leue, A. Extraordinary face recognition performance in laboratory and online testing. *Appl. Cogn. Psychol.* **35**, 579–589. <https://doi.org/10.1002/acp.3805> (2021).
23. Bobak, A. K., Bennetts, R. J., Parris, B. A., Jansari, A. & Bate, S. An in-depth cognitive examination of individuals with superior face recognition skills. *Cortex* **82**, 48–62. <https://doi.org/10.1016/j.cortex.2016.05.003> (2016).
24. Sunday, M. A., Richler, J. J. & Gauthier, I. Limited evidence of individual differences in holistic processing in different versions of the part-whole paradigm. *Atten. Percept. Psychophys.* **79**, 1453–1465. <https://doi.org/10.1167/18.1.2> (2017).
25. Sunday, M. A., Lee, W. Y. & Gauthier, I. Age-related differential item functioning in tests of face and car recognition ability. *J. Vis.* **18**, 2. <https://doi.org/10.1167/18.1.2> (2018).
26. Davis, S. D. & Peterson, D. J. Simulated viewing distance impairs the confidence–accuracy relationship for long, but not moderate distances: support for a model incorporating the role of feature ambiguity. *Cogn. Res. Princ. Implic.* **7**, 55. <https://doi.org/10.1186/s41235-022-00406-5> (2022).
27. Stantić, M. et al. Individuals with developmental prosopagnosia show independent impairments in face perception, face memory and face matching. *Cortex* **157**, 266–273. <https://doi.org/10.1016/j.cortex.2022.09.012> (2022).
28. Stantić, M. et al. Independent measurement of face perception, face matching, and face memory reveals impairments in face perception and memory, but not matching, in autism. *Psychon. Bull. Rev.* **30**, 2240–2249. <https://doi.org/10.3758/s13423-023-02304-3> (2023).
29. Tsantani, M., Vestner, T. & Cook, R. The Twenty Item Prosopagnosia Index (PI20) provides meaningful evidence of face recognition impairment. *R. Soc. Open Sci.* **8**, 202062. <https://doi.org/10.1098/rsos.202062> (2021).
30. Duchaine, B. & Nakayama, K. 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. <https://doi.org/10.1016/j.neuropsychologia.2005.07.001> (2006).
31. Bowles, D. C. et al. Diagnosing prosopagnosia: effects of ageing, sex, and participant-stimulus ethnic match on the Cambridge Face Memory Test and Cambridge Face Perception Test. *Cogn. Neuropsychol.* **26**, 423–455. <https://doi.org/10.1080/02643290903343149> (2009).
32. Biotti, F., Gray, K. L. H. & Cook, R. Is developmental prosopagnosia best characterised as an apperceptive or mnemonic condition? *Neuropsychologia* **124**, 285–298. <https://doi.org/10.1016/j.neuropsychologia.2018.11.014> (2019).
33. McKone, E. et al. Face ethnicity and measurement reliability affect face recognition performance in developmental prosopagnosia: evidence from the Cambridge Face Memory Test-Australian. *Cogn. Neuropsychol.* **28**, 109–146. <https://doi.org/10.1080/02643294.2011.616880> (2011).
34. Bobak, A. K. et al. Data-driven studies in face identity processing rely on the quality of the tests and data sets. *Cortex* **166**, 348–364 (2023).
35. Kramer, R. S. Face matching and metacognition: Investigating individual differences and a training intervention. *PeerJ* **11**, e14821. <https://doi.org/10.7717/peerj.14821> (2023).
36. Grown, B., Dunn, J. D., Mattijssen, E. J., Quigley-McBride, A. & Towler, A. Match me if you can: evidence for a domain-general visual comparison ability. *Psychon. Bull. Rev.* **29**, 866–881. <https://doi.org/10.3758/s13423-021-02044-2> (2022).
37. Grown, B. et al. Statistical feature training improves fingerprint-matching accuracy in novices and professional fingerprint examiners. *Cogn. Res. Princ. Implic.* **7**, 1–21. <https://doi.org/10.1186/s41235-022-00413-6> (2022).
38. White, D., Phillips, P. J., Hahn, C. A., Hill, M. & O’Toole, A. J. Perceptual expertise in forensic facial image comparison. *Proc. R. Soc. B* **282**, 20151292. <https://doi.org/10.1098/rspb.2015.1292> (2015).
39. Dowsett, A. J. & Burton, A. M. Unfamiliar face matching: Pairs out-perform individuals and provide a route to training. *Br. J. Psychol.* **106**, 433–445. <https://doi.org/10.1111/bjop.12103> (2015).

40. White, D., Guilbert, D., Varela, V. P. L., Jenkins, R. & Burton, A. M. GFMT2: A psychometric measure of face matching ability. *Behav. Res. Methods* **54**, 1318–1332. <https://doi.org/10.3758/s13428-021-01638-x> (2022).
41. Burton, A. M., White, D. & McNeill, A. The glasgow face matching test. *Behav. Res. Methods* **42**, 286–291. <https://doi.org/10.3758/BRM.42.1.286> (2010).
42. Passell, E. et al. Cognitive test scores vary with choice of personal digital device. *Behav. Res. Methods* **53**, 2544–2557. <https://doi.org/10.3758/s13428-021-01597-3> (2021).
43. Germine, L. T., Duchaine, B. & Nakayama, K. Where cognitive development and aging meet: face learning ability peaks after age 30. *Cognition* **118**, 201–210. <https://doi.org/10.1016/j.cognition.2010.11.002> (2011).
44. Davis, J. P., Lander, K., Evans, R. & Jansari, A. Investigating predictors of superior face recognition ability in police super-recognisers. *Appl. Cogn. Psychol.* **30**, 827–840. <https://doi.org/10.1002/acp.3260> (2016).
45. Dunn, J. D., Towler, A., Kemp, R. I. & White, D. Selecting police super-recognisers. *PLoS ONE* **18**, e0283682. <https://doi.org/10.1371/journal.pone.0283682> (2023).
46. White, D., Towler, A. & Kemp, R. I. Understanding professional expertise in unfamiliar face matching. In *Forensic Face Matching: Research and Practice* (ed. Bindemann, M.) 62–144 (Oxford Univ. Press, 2021). <https://doi.org/10.1093/oso/9780198837749.003.0004>.
47. White, D., Wayne, T. & Varela, V. P. Partitioning natural face image variability emphasizes within-identity over between-identity representation for understanding accurate recognition. *Cognition* **219**, 104966. <https://doi.org/10.1016/j.cognition.2021.104966> (2022).
48. Webb, M. A. & Tangney, J. P. Too good to be true: bots and bad data from Mechanical Turk. *Perspect. Psychol. Sci.* <https://doi.org/10.1177/17456916221120027> (2022).
49. Shah, P., Gaulle, A., Sowden, S., Bird, G. & Cook, R. The 20-item prosopagnosia index (PI20): a self-report instrument for identifying developmental prosopagnosia. *R. Soc. Open Sci.* **2**, 140343. <https://doi.org/10.1098/rsos.140343> (2015).
50. Chandler, J. & Shapiro, D. Conducting clinical research using crowdsourced convenience samples. *Annu. Rev. Clin. Psychol.* **12**, 53–81. <https://doi.org/10.1146/annurev-clinpsy-021815-093623> (2016).
51. Arditte, K. A., Çek, D., Shaw, A. M. & Timpano, K. R. The importance of assessing clinical phenomena in Mechanical Turk research. *Psychol. Assess.* **28**, 684–691. <https://doi.org/10.1037/pas0000217> (2016).
52. Eriksson, K. Autism-spectrum traits predict humor styles in the general population. *Humor* **26**, 461–475. <https://doi.org/10.1515/humor-2013-0030> (2013).
53. Lewis, G. J., Shakeshaft, N. G. & Plomin, R. Face identity recognition and the social difficulties component of the Autism-Like phenotype: evidence for phenotypic and genetic links. *J. Autism Dev. Disord.* **48**, 2758–2765. <https://doi.org/10.1007/s10803-018-3539-4> (2018).
54. Fysh, M. C. & Bindemann, M. The Kent face matching test. *Br. J. Psychol.* **109**, 219–231. <https://doi.org/10.1111/bjop.12260> (2018).
55. Fysh, M. C., Stacchi, L. & Ramon, M. Differences between and within individuals, and subprocesses of face cognition: implications for theory, research and personnel selection. *R. Soc. Open Sci.* **7**, 200233. <https://doi.org/10.1098/rsos.200233> (2020).
56. McCaffery, J. M., Robertson, D. J., Young, A. W. & Burton, A. M. Individual differences in face identity processing. *Cogn. Res. Princ. Implic.* **3**, 1–15. <https://doi.org/10.1186/s41235-018-0112-9> (2018).
57. Bobak, A. K., Mileva, V. R. & Hancock, P. J. Facing the facts: Naïve participants have only moderate insight into their face recognition and face perception abilities. *Q. J. Exp. Psychol.* **72**(4), 872–881 (2019).
58. Baldson, T., Summersby, S., Kemp, R. I. & White, D. Improving face identification with specialist teams. *Cogn. Res. Princ. Implic.* **3**, 1–13. <https://doi.org/10.1186/s41235-018-0114-7> (2018).
59. Mayer, M. & Ramon, M. Improving forensic perpetrator identification with Super-Recognizers. *Proc. Natl. Acad. Sci. USA* **120**, e2220580120. <https://doi.org/10.1073/pnas.2220580120> (2023).
60. Ramon, M., Bobak, A. K. & White, D. Super-recognizers: from the lab to the world and back again. *Br. J. Psychol.* **110**, 461–479. <https://doi.org/10.1111/bjop.12368> (2019).
61. Abudarham, N., Bate, S., Duchaine, B. & Yovel, G. Developmental prosopagnosics and super recognisers rely on the same facial features used by individuals with normal face recognition abilities for face identification. *Neuropsychologia* **160**, 107963. <https://doi.org/10.1016/j.neuropsychologia.2021.107963> (2021).
62. Bobak, A. K., Parris, B. A., Gregory, N. J., Bennetts, R. J. & Bate, S. Eye-movement strategies in developmental prosopagnosia and “super” face recognition. *Q. J. Exp. Psychol.* **70**, 201–217. <https://doi.org/10.1080/17470218.2016.1161059> (2017).
63. Dunn, J. D., Miellet, S. & White, D. Information sampling differences supporting superior face identity processing ability. *Psychon. Bull. Rev.* <https://doi.org/10.3758/s13423-024-02579-0> (2024).
64. Tardif, J. et al. Use of face information varies systematically from developmental prosopagnosics to super-recognizers. *Psychol. Sci.* **30**, 300–308. <https://doi.org/10.1177/0956797618811338> (2019).
65. Nador, J. D., Zoia, M., Pachai, M. V. & Ramon, M. Psychophysical profiles in super-recognizers. *Sci. Rep.* **11**, 13184. <https://doi.org/10.1038/s41598-021-92549-6> (2021).
66. Jenkins, R. E. et al. Are super-face-recognisers also super-voice-recognisers? Evidence from cross-modal identification tasks. *Appl. Cogn. Psychol.* **35**, 590–605. <https://doi.org/10.1002/acp.3813> (2021).
67. White, D., Dunn, J. D. & Summersby, S. Are you a super-recogniser? *Medium*. <https://medium.com/@whitelab/are-you-a-super-recogniser-ffa78390fc44> (2017).
68. Davis, J. P., Bretfelean, D., Belanova, E. & Thompson, T. Assessing the long-term face memory of highly superior and typical-ability short-term face recognisers. *PsyArXiv* (2019). <https://doi.org/10.31234/osf.io/var4m>.

Acknowledgements

This project was supported by funding from the Australian Research Council (DP190100957, FT200100353). The authors thank Dr. Anita Trinh for help collecting data in Table 1.

Author contributions

BP: project design, data collection, data analysis, manuscript preparation and revision. JD: project design, manuscript preparation and revision; AT: project design, manuscript preparation and revision; DW: acquiring funding, project design, data analysis, manuscript preparation and revision.

Competing interests

The author(s) declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-025-92907-8>.

Correspondence and requests for materials should be addressed to D.W.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025