Behavior Research Methods (2022) 54:2318–2333
https://doi.org/10.3758/s13428-021-01727-x

An update of the Benton Facial Recognition Test

Ebony Murray[1] · Rachel Bennetts[2] · Jeremy Tree[3] · Sarah Bate[4]

Accepted: 13 October 2021 / Published online: 16 December 2021

Abstract

The Benton Facial Recognition Test (BFRT) is a paper-and-pen task that is traditionally used to assess face perception skills in neurological, clinical and psychiatric conditions. Despite criticisms of its stimuli, the task enjoys a simple procedure and is rapid to administer. Further, it has recently been computerised (BFRT-c), allowing reliable measurement of completion times and the need for online testing. Here, in response to calls for repeat screening for the accurate detection of face processing deficits, we present the BFRT-Revised (BFRT-r): a new version of the BFRT-c that maintains the task's basic paradigm, but employs new, higher-quality stimuli that reflect recent theoretical advances in the field. An initial validation study with typical participants indicated that the BFRT-r has good internal reliability and content validity. A second investigation indicated that while younger and older participants had comparable accuracy, completion times were longer in the latter, highlighting the need for age-matched norms. Administration of the BFRT-r and BFRT-c to 32 individuals with developmental prosopagnosia resulted in improved sensitivity in diagnostic screening for the BFRT-r compared to the BFRT-c. These findings are discussed in relation to current diagnostic screening protocols for face perception deficits. The BFRT-r is stored in an open repository and is freely available to other researchers.

Keywords Face perception · Face matching · Face recognition · Prosopagnosia · Benton · Response times

The Benton Facial Recognition Test (BFRT: Benton & Van Allen, 1968; see Benton et al. (1983) for the formal reference of the test) is a face matching task that is traditionally administered face-to-face using hard copy materials. Participants are simultaneously presented with a target face above an array of six test faces. In the first six trials, one face in the array matches the identity of the target face, and in the final 16 trials, three faces in the array match the identity of the target. The task was originally developed for the assessment of individuals believed to have acquired prosopagnosia (a severe deficit in recognising familiar people from their face) following brain injury (Barton, 2008; Bate & Bennetts, 2015; Van Belle et al., 2011), but has since been widely used to assess face perception skills in a number of neurological, clinical and psychiatric conditions (Annaz et al., 2009; Rabin et al., 2005; Sachse et al., 2014).

Yet, the popularity of the BFRT has reduced in recent years, particularly for the assessment of individuals suspected to have prosopagnosia. At the turn of the century, many more people presented to researchers believing they experience a developmental form of prosopagnosia (Bate et al., 2008; De Luca et al., 2019; Geskin & Behrmann, 2017), prompting a wider individual differences perspective on human face recognition, and the belief that developmental face recognition difficulties may reside on a continuum (Barton & Corrow, 2016; Bate & Tree, 2017). These larger samples of cases have reignited long-standing questions of whether perceptual and mnemonic difficulties are dissociable (De Renzi et al., 1991), and whether subtypes of developmental prosopagnosia (DP) map onto this framework (note that the term "congenital" or "hereditary"

Sarah Bate is supported by a British Academy Mid-Career Fellowship (MD170004) and a Leverhulme Research Fellowship (RF-2020-105).

✉ Ebony Murray
emurray4@glos.ac.uk

[1] Department of Psychological Sciences, School of Natural and Social Sciences, University of Gloucestershire, Cheltenham GL50 4AZ, UK

[2] Department of Life Sciences, Brunel University London, Uxbridge, UK

[3] Department of Psychology, Swansea University, Swansea, UK

[4] Department of Psychology, Bournemouth University, Poole, UK

prosopagnosia has been used somewhat interchangeably with DP in the literature: e.g. Behrmannet al., 2005; Hasson et al., 2003; Kennerknecht et al., 2006; Palermo et al., 2011). Clearly, to address all of these questions, reliable face perception tasks are required. However, Duchaine and Weidenfeld (2003) reported that when the inner features of the faces in the BFRT were obscured, most typical participants could still achieve a typical score using the hairline and eyebrows alone. Further, Duchaine and Nakayama (2004) found that seven out of 11 DP participants achieved typical scores on the task, again suggesting that external facial cues may be used to aid performance.

Unfortunately, there is also deliberation over alternate tests of face perception, and the field still lacks a reliable task. The most widely used face perception test for the diagnosis of DP is the Cambridge Face Perception Test (CFPT: Duchaine et al., 2007), which presents participants with six morphed faces that are to be organised in order of similarity to a simultaneously presented target face. The task requires proficient use of a computer mouse within a strict time period, and the instructions are complex for online administration, particularly with clinical and older participants (Bate et al., 2018; Bate, Frowd, Bennetts, et al., 2019c; Bowles et al., 2009). Others query whether morphed faces are unnaturally similar (White et al., 2017), and whether the requirement for similarity judgements initiates higher-level cognitive processes than required for the simplistic identity matching of simultaneously presented naturalistic facial images (Rossion & Michel, 2018). Such simpler face matching tasks are typically found in the forensic face recognition literature (e.g. the Glasgow Face Matching Test: Burton et al., 2010; the Pairs Matching Test: Bate et al., 2018; Bate, Frowd, Bennetts, et al., 2019c), but are seldom used for the detection of DP due to their low sensitivity to poor performance (although see Stantic et al. (2021) for a recent face matching test that is presented as a suitable task for the entire face processing spectrum of abilities). Indeed, the chance of responding correctly on all trials is 50%, a score that is within the range achieved by typical perceivers, many of whom find these tasks particularly challenging (e.g. Robertson et al., 2016; Shah et al., 2015). In addition, White et al. (2017) reported a response bias in DP participants, where the tendency to respond "different" in a simple same/different face matching task artificially inflated their score on these trials.

Such criticisms led Rossion and Michel (2018) to return to the BFRT, citing advantages in its original paradigm. Despite the external cues to recognition that were highlighted by Duchaine and colleagues (Duchaine & Nakayama, 2006; Duchaine & Weidenfeld, 2003), the BFRT has traditionally been regarded as a difficult test with no ceiling effect (Benton & Van Allen, 1972), which is quick to administer with simple instructions. Importantly, Rossion and Michel point out that decisional response biases are avoided by the task's forced choice procedure (the number of target faces is constant across test sections), making it substantially easier to interpret test scores. Further, Rossion and Michel (2018) highlighted the importance of recording task completion time in addition to accuracy (via a computerised version of the test: the BFRT-c), as a means to detect typical scores that are achieved by compensatory mechanisms. Previous work has adopted this approach when screening for face perception deficits in acquired prosopagnosia, where apparently typical accuracy scores were found to be accompanied by the use of atypical, laboured feature-by-feature matching strategies (e.g. Bukach et al., 2006; Busigny & Rossion, 2010; Delvenne et al., 2004; Farah, 1990; Young et al., 1993).

Another way to address this issue is to administer multiple versions of a face perception task, using rather different facial stimuli. This should prohibit, or at least reduce, the transfer of compensatory strategies that are useful in one version of a task. Indeed, some DPs report the use of particular facial features when images are captured within the same photography session (e.g. no change in skin tone or appearance of the hairline, even when images are cropped), or even consider pictorial cues such as lighting conditions or the quality of the images themselves (Adams et al., 2019). Further, very recent findings also highlight the importance of repeat testing on key measures of face recognition performance when screening for DP (Bate, Bennetts, Gregory, et al., 2019a; Murray & Bate, 2020), given issues with task reliability, the occurrence of borderline scores that are difficult to reconcile, and the possibility that a particular score simply occurred by chance performance (Young et al., 1993).

Yet, no known alternate version of the BFRT exists, and an update of the task using new stimuli is certainly overdue. While the basic paradigm (with the monitoring of completion times) offers a sound means of assessing face perception, the age of the test unsurprisingly lends itself to very low image quality. Whilst face recognition can be successful even when images are of low spatial frequency (e.g. Liu et al. 2000), unfamiliar face processing, as is being assessed with the BFRT, benefits from high-quality images (Burton et al., 1999). Moreover, the findings of Duchaine and colleagues (Duchaine & Nakayama, 2004; Duchaine & Weidenfeld, 2003) indicate that extra-facial cues can be used to achieve a typical score on the existing version of the test. While more recent face-processing tasks in the neuropsychological literature have responded to Duchaine and colleagues' criticisms of the BFRT by using tightly controlled images that are captured on the same day and heavily cropped to exclude the external features (e.g. Biotti et al., 2017; Duchaine & Nakayama, 2006; Esins et al., 2016), it has been argued that this procedure actually distances the task from real-world face recognition (Burton, 2013). Rather, variability in facial

appearance is a critical feature of everyday face recognition, and should be embraced in, rather than removed from, laboratory tests (Young & Burton, 2017, 2018). In fact, even typical participants struggle to match faces of the same identity when pictured in more "ambient" images that retain the external features of the face, given image-based cues cannot be used as compensatory cues for successful performance (for further discussion, see Burton, 2013).

Here, we introduce a new version of the BFRT-c, the BFRT-revised (BFRT-r), which maintains the format of the original task but employs new, more varied, naturalistic facial images. In Experiment 1, we examine the validity of the BFRT-r in typical participants and provide norming data for comparison to clinical cases. In Experiment 2, we assess the test's diagnostic utility alongside the BFRT-c in DP.

Experiment 1

A new face matching task (the BFRT-r) was created that follows the original BFRT paradigm, but is computerised (akin to the BFRT-c) and uses new, more ambient facial images. We initially assessed the psychometric properties of the task and collected norming data from young typical adults. The reliability and validity of the BFRT-r was investigated by comparing performance on this task to the BFRT-c. In addition, content validity was assessed by comparing performance to (a) the Cambridge Face Memory Test (CFMT) and (b) to a new group of participants using an inverted version of the BFRT-r.

Method

Participants

A total of 165 participants took part in Experiment 1. One hundred and nine participants aged between 18 and 35 years (mean age = 24.7 years, SD = 3.5; 55 female) completed the full string of tests in their upright format. To avoid re-exposure effects, 56 different participants aged between 18 and 35 years (mean age = 24.9 years, SD = 3.5; 27 male) completed only the inverted version of the BFRT-r. All participants were recruited via the online participant recruitment website *Prolific*, in exchange for a small financial incentive. All were Caucasian and lived within the UK, reported no history of socio-emotional, neurological or psychiatric disorder, and had normal or corrected-to-normal vision. This project was approved by the institutional Research Ethics Committee.

Materials

*BFRT-c (*Rossion & Michel*,* 2018*):* The BFRT-c is the original version of the BFRT, in a computerised format. The test contains a total of 22 trials in which an unfamiliar Caucasian target face (shown from a frontal viewpoint with a neutral expression) has to be found among a simultaneously presented array of six Caucasian probe faces, also showing neutral expressions. For the first six trials (half male), the target face has to be found only once within each array, where all faces are shown from a frontal viewpoint, such that the corresponding probe image is very similar to the target image. For the remaining 16 trials (half male), the target face is again presented from a frontal viewpoint. The participant is required to find three images within the six-image array that match the identity of the target. The six faces in each array vary either in terms of head orientation (the second section of the test: eight trials, half female) or lighting (the third section of the test: eight items, half female). Some target faces are repeated: four of the seven female targets appear in two separate sections, one of the seven male targets appears in all three sections, and three male targets are used in two sections. All target identities are also used as distractors in at least one trial of the task.

In each trial, target faces are presented at a slightly different size than those in the array (target faces were 156 x 232 pixels; faces in the array were 201 x 234 pixels, in order to minimise successful matching based on low-level, image-based visual cues: Rossion & Michel, 2018). All images are grayscale and display the overall shape of the face, but are cropped below the chin and beyond the hairline. As in the original version of the task, the order of the trials is not randomised and participants have an unlimited length of time to complete each trial. There is an inter-stimulus interval of 800 ms. Information screens at the beginning of each section instruct the participant how many responses to make for each trial, and inform them that response time is recorded.

Participants are required to select their responses by clicking on the appropriate face(s) in each array. For trials that require three responses, participants are able to select faces in any order, but cannot change a response once a face has been selected. The maximum score on the task is 54. Participants can receive one point in each of the six trials that compose Section 1 (where one response is required per trial), and between 0 and 3 points for each of the trials in Section 2 (where three responses are required per trial). Trial completion times are measured to aid data processing (see below), and overall task completion times are monitored for analysis.

BFRT-r The basic paradigm of the BFRT-c is retained, with the same number of trials. However, the facial stimuli are replaced throughout. As gender biases have been shown for the recognition of female but not male faces (e.g. Herlitz & Lovén, 2013; Lovén et al., 2011), we followed the precedent of more recent tests by only using male faces (e.g. the CFMT: Duchaine & Nakayama, 2006; the CFPT: Duchaine

et al., 2007). We initially acquired facial images from a total of 130 Caucasian males (aged 18–34 years: M = 21.9 years, SD = 3.2) in exchange for course credit or a small financial incentive. Images were captured within the laboratory, and/or were existing photographs provided by the participant that had been taken within the space of a single year. Thus, different images of the same person had been captured on different days, often months apart, and in many cases, using different cameras. However, images of the same person had all been captured within the same year, preventing any major ageing effects. Blemishes, skin tone and hairstyle varied from image to image, as well as lighting conditions. No image had been manipulated, and all were of sufficiently high quality (no less than 96 DPI). They displayed the target without spectacles or very heavy facial hair to the extent that the faces were obscured. There were variations in viewpoint due to their capture in naturalistic settings.

A unique target was used in each of the 22 trials, and no target was re-used as a distractor. Ten distractor identities were repeated over the 22 trials, but different images of each individual were used where possible; only two images were repeated twice through the test. No distractor identity was repeated in the same array. Distractors were allocated to each trial based on their perceived similarity to the target, as judged by a member of the research team. Pilot testing supported these judgements: the trials included in the final BFRT-r did not elicit ceiling nor floor effects. In total, the test used images from 76 different individuals.

All images were presented in greyscale. This decision was made based on initial pilot testing/materials analyses, which indicated that ceiling effects in the typical population could be achieved when images were in colour. To prevent low-level image matching, target faces were not cropped to exclude any part of the head, hair or ears, but array images were cropped around the hairline (see Fig. 1). Target images were larger (166 x 232 pixels) than those in the array (approximately 153 x 200 pixels). As in the BFRT-c, only one of the array faces matched the identity of the target in the first six trials, and three in the remaining 16 trials. In the first 12 trials of the task, all faces are displayed from frontal viewpoints. In the final ten trials, faces are displayed from frontal, but more naturalistic, viewpoints. The rotation of most faces is small; approximately 10–30 degrees to the left or right. A small number of images ($N = 7$) are displayed at a larger rotation (less than 45 degrees), but the whole face can be viewed in every photograph (i.e. both eyes are clearly visible; see Fig. 1).

As for the BFRT-c, trials were presented in the same order for each participant, with an inter-stimulus interval of 800 ms, and responses were made and scored in the same manner. Instructions were identical to the BFRT-c, but additionally informed participants that some images were taken some time apart, and some aspects of the target's appearance (e.g. hairstyle) may have changed during this time. The BFRT-r test materials are available in an open repository: https://osf.io/vza3m/?view_only=404f6d1971924759b126d46cba1d25b7. A fully programmed version can also be shared with researchers on request, via Testable. The test and its materials are protected by a Creative Commons Attribution-Non-Commercial license.

BFRT-r inverted The BFRT-r was also prepared in an inverted format to assess the content-validity of the test. All stimuli and parameters were identical, with the exception that all images were rotated 180 degrees.

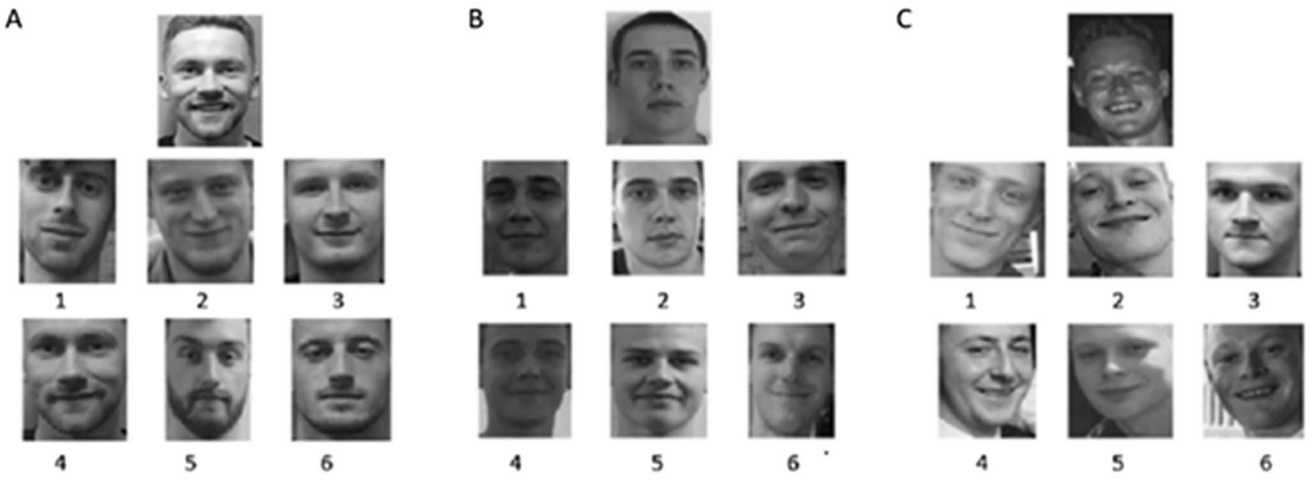

Fig. 1 Example trials from the BFRT-r. *Note.* Panel **A** shows the trial format for trials 1–6 where all faces are presented from a frontal viewpoint. There is only one correct response (4). Panel **B** shows the trial format for trials 7–12 where all faces are again shown from a frontal viewpoint, but there are three correct responses (1, 2, 4). Panel **C** shows the format for trials 13–23, where the target face is shown from a frontal viewpoint and the probe images show some rotation. There are three correct responses (2, 5, 6)

CFMT (Duchaine & Nakayama, 2006) The CFMT is an unfamiliar face memory test. The overall objective of the CFMT is to introduce unfamiliar, young male faces to the participant and then test their recognition of those faces. It contains three test stages which increase in difficulty as the test progresses: (a) Learn: Participants view a target face from three viewpoints for 3 s per image. They then choose which of three faces, presented simultaneously, is the target. This is repeated for six faces, resulting in a maximum score of 18. (b) Test stage: Thirty triads of faces are presented, where one face is a novel image of a target identity intermixed with two distractors. (c) Noise stage: Twenty-four new triads are displayed with added visual noise. Again, each trial contains any one of the targets and two distractors. The entire test is scored out of 72, and chance is 24. More information about this test can be found in the associated publication. We include it here as a means to check the content validity of the BFRT-r (i.e. that it correlates well with a dominant face-processing task that is already known to have high content validity).

Procedure

All tasks were completed online, using the *Testable* platform (www.testable.org; see Rezlescu et al., 2020). Participants were required to initially calibrate the tests for screen size, ensuring uniform presentation. The 109 participants that completed the main string of tests completed the CFMT, BFRT-r, and BFRT-c, in that order. This enabled us to collect accurate norming data for the new task without introducing practice effects from the repeated use of the same paradigm. The 56 participants who only took part in the inverted version of the BFRT-r did not complete any other tests.

Data processing

As data were collected online, responses were initially screened for task engagement. Each individual's mean response time (and SD) was calculated for each group of trials in both the BFRT-c and BFRT-r (i.e. the trials which required one response, and the trials which required three responses). Any responses which were greater that than 3 SDs above the mean were removed, as were any responses that were quicker than 150 ms. In addition, trials that required three unique responses were screened to ensure correct completion (i.e. to remove trials that had received duplicate responses). If participants made more than two duplications on only one of the tests, their data were removed from the analysis. Participants who made a similar number of duplications (no more than four; equivalent of 10% of responses) on both tests were retained. As a duplication was scored as 0, we did this to ensure that scores on both tasks were similarly affected by duplications, without artificially affected the mean scores. Overall accuracy scores were also screened for outliers across the dataset, using a three SDs from the mean criterion.

For the participants that completed both the BFRT-r and BFRT-c, 15 were removed for failing to complete enough trials (i.e. too many duplications), and one for achieving accuracy scores on both tasks that surpassed three SDs from the mean score. No participant responded quicker than 150 ms on any trial, and no participant was excluded for giving too many abnormally slow responses. Of the remaining participants, none were identified as outliers on the CFMT. The final sample consisted of 93 participants aged 18–30 years (49 female; mean age = 24.84 years, SD = 3.44). The same screening procedures were applied to the participants who only completed the inverted version of the BFRT-r, resulting in the exclusion of one individual. A final sample of 55 participants aged 18–30 years (27 male; mean age = 24.9 years, SD = 3.6) therefore proceeded to the analysis phase.

Results

Mean accuracy performance on the upright version of the BFRT-r was 78.83% (SD = 8.72). Given different numbers of responses were required in different sections of each test, we followed the precedent of Rossion and Michel (2018) in analysing overall task completion times, rather than the average response time per trial. The mean overall task completion time for the BFRT-r was 253.75 s (SD = 130.97).

Mean accuracy performance on the BFRT-c was 80.82% (SD = 8.17), and mean task completion time was 168.34 s (SD = 69.82 s). The performance of this sample on the BFRT-c is therefore comparable to the norms presented by Rossion and Michel (2018), who reported a mean accuracy of 82.98% (SD = 6.37), and a mean task completion time of 180.85 s (SD = 59.86).

To fully compare the two tasks, a mixed 2 (test: BFRT-c, BFRT-r) x 2 (gender: male, female) ANOVA was carried out. There was a main effect of test, $F(1,91) = 6.760$, $p = .011$, $\eta\rho^2 = .069$, in that individuals performed significantly better on the BFRT-c (M = 80.82%, SD = 8.17) than the BFRT-r (M = 78.83%, SD = 8.72; see Table 1). There was also a significant main effect of gender over the two tests, $F(1,91) = 4.469$, $p = .037$, $\eta\rho^2 = .047$ (see Fig. 2), where females (M = 81.39%, SD = 8.14) outperformed males (M = 78.09%, SD = 8.50). However, there was no significant interaction between test and gender, $F(1,91) = 1.173$, $p = .282$.

A 2 (test: BFRT-c, BFRT-r) x 2 (gender) ANOVA on task completion times revealed a main effect of test, $F(1,91) = 87.616$, $p < .001$, $\eta\rho^2 = .491$, with participants taking longer to complete the BFRT-r (M = 253.75s, SD = 130.98) than the BFRT-c (M = 168.34s, SD = 69.82; see Table 1). There was no main effect of gender, $F(1,91)$

Table 1 Descriptive data (means and standard deviations) for the upright versions of the BFRT-c and BFRT-r for younger controls (Experiment 1), and older controls and DPs (Experiment 2); accuracy is presented as a percentage, and completion times in seconds

	BFRT-c accuracy	BFRT-c completion times	BFRT-r accuracy	BFRT-r completion times
Younger controls (Mean age = 24.8; $N = 93$)	80.82 (8.17)	168.34 (69.82)	78.83 (8.72)	253.75 (130.97)
Older controls (Mean age = 48.4; $N = 218$)	82.71 (8.76)	247.55 (108.30)	78.98 (10.46)	348.22 (165.20)
DPs (Mean age = 52.0; $N = 32$)	74.83 (7.83)	341.16 (139.87)	67.48 (8.23)	489.13 (202.29)

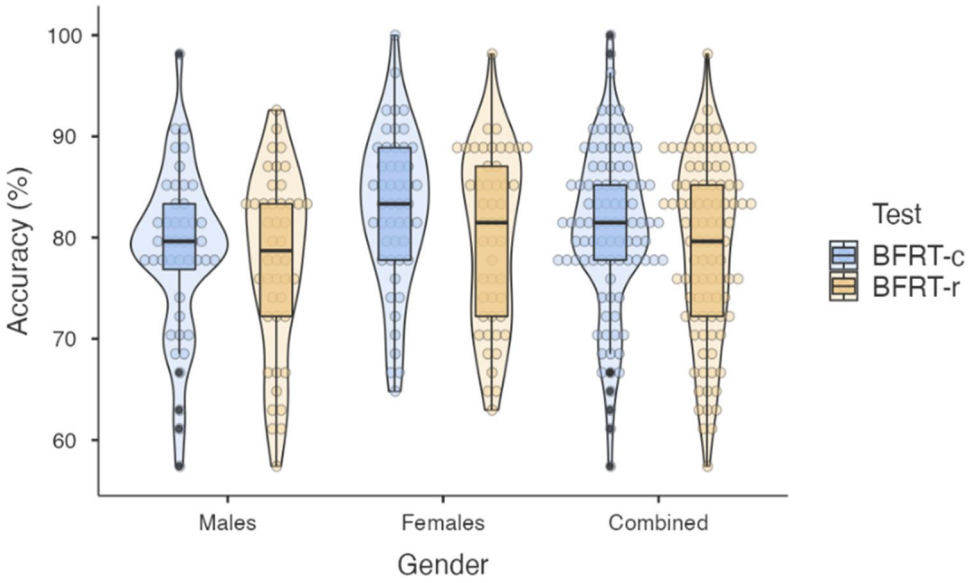

Fig. 2 Violin plots indicating scores on the BFRT-r and BFRT-c for males and females separately, and the overall sample

= 0.081, $p = .776$ (see Fig. 3), nor an interaction between test and gender, $F(1,91) = 0.003$, $p = .954$. Completion times on the BFRT-r strongly correlated with the BFRT-c ($r = .787$, $p < .001$).

Following the precedent of Rossion and Michel (2018), the task's internal reliability was assessed by correlating performance on even versus odd items, considering only the second part of the test in which three responses are made per trial. The inter-item correlation was significant for accuracy rates (mean score for the eight even items = 20.41/24, SD = 2.34; mean score odd items = 18.78/24, SD = 2.18; r_{SB} [Spearman–Brown] = .735, $p < .001$). The interitem correlation was even higher for trial completion times (mean trial completion times for the eight even items = 89.50s, SD = 46.05s; mean trial completion times for the eight odd items = 98.91s, SD = 55.48s; $r_{SB} = .963$, $p < .001$).

To further explore the reliability and validity of the BFRT-r, scores were correlated against performance on the CFMT (see Table 2 for the full correlation matrix). Accuracy performance on the BFRT-r strongly correlated with the BFRT-c ($r = .636$, $p < .001$). Both the BFRT-c and BFRT-r also correlated significantly and moderately (BFRT-c) or strongly (BFRT-r) with the CFMT.

Finally, comparison between overall accuracy scores on the upright (M = 78.83%, SD = 8.72) and inverted (M = 56.66%, SD = 10.64) versions of the BFRT-r revealed a substantial inversion effect, $t(146) = 13.746$, $p < .001$, $d = 2.28$. However, there was no significant difference between upright BFRT-r (M =253.75s, SD = 130.97) task completion times and inverted task completion times (M = 249.90 s, SD = 146.88), $t(146) = 0.165$, $p = .869$.

Summary

Here, we present the BFRT-r: a new test of face perception that adopts the same paradigm as the original BFRT (as per the BFRT-c) but uses more naturalistic images to accommodate within-person variation in facial images. As the BFRT-r follows the procedure of the BFRT-c, the test continues to be simple and quick to administer, with an approximate completion time of four minutes in typical young adults. Initial analyses reveal that the BFRT-r has good internal

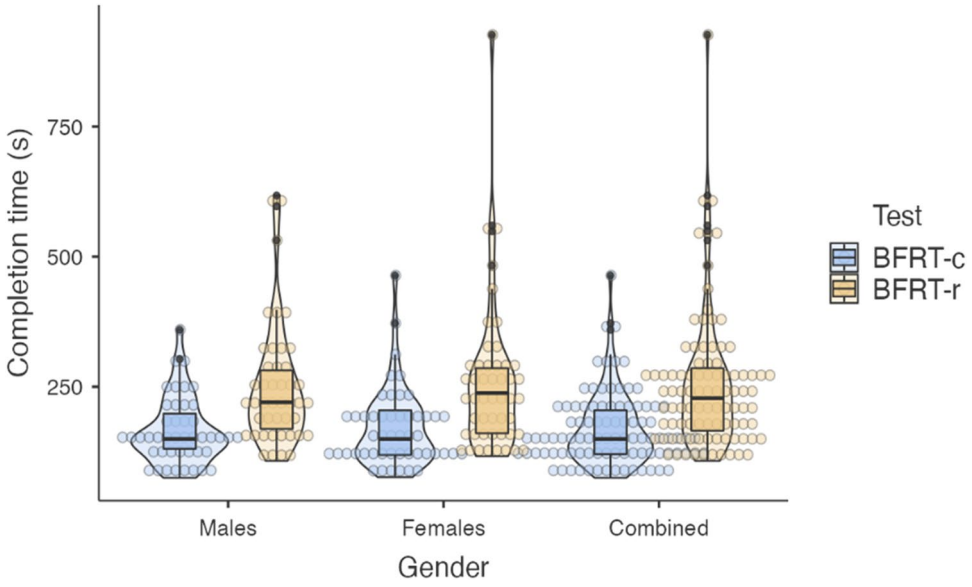

Fig. 3 Violin plots indicating completion times on the BFRT-r and BFRT-c for males and females separately, and overall sample

Table 2 Correlation matrix for overall scores on the three tests (*r* (*p*)). Cut-offs for significance are Bonferroni corrected. $N = 93$

	BFRT-r	CFMT
BFRT-c	.636 (< .001)	.432 (.001)
BFRT-r		.510 (< .001)

reliability with strong inter-item correlations. It also has a strong inversion effect according to accuracy (although not completion times), suggesting that it taps face- rather than image-processing mechanisms. A strong correlation with the CFMT further supports this.

Comparison of performance on the two tasks indicates that the BFRT-r is slightly more difficult than the original version. More importantly, typical participants are able to score well above chance on the BFRT-r (the lowest score was 55.56%; chance performance is 46.30%). Further, the norming data reported here (M = 78.24%, SD = 9.20) would enable clinical participants to score two SDs below the mean without performing at chance level. Indeed, those with DP often show impaired face perception skills, but these skills are not completely abolished to the point that they are scoring at chance level (e.g. Bate et al., 2019a, b, c; Biotti et al., 2019; Righart & de Gelder, 2007). Thus, the task is suitably calibrated to detect variations in performance between chance and the control mean.

It is of note that a gender difference was found for accuracy (but not completion time) across both versions of the BFRT. A similar effect has previously been reported for completion time but not accuracy on the BFRT-c (Rossion & Michel, 2018). Previous work suggests a small gender difference in face recognition, with a meta-analysis confirming that females outperform males (Herlitz & Lovén, 2013), even when only male faces are presented. The authors of the meta-analysis observed that the female advantage may be accentuated in tasks which involve generalisation across different viewpoints or images (as is the case in both BFRT tests). Thus, the gender effect observed here was not surprising; we continued to explore whether it persevered in our second experiment using older adult control participants.

Experiment 2

Having explored the validity of the BFRT-r in younger participants, we next sought to examine the diagnostic utility of the updated version in individuals with DP. In particular, we examine (a) the additional benefit of evaluating response times as well as accuracy in atypical participants, and (b) whether there is a case for administration of multiple versions of the same task when screening for face perception deficits.

Method

Participants

Thirty-two participants with a prior diagnosis of DP took part in this study. They had previously taken part in an objective screening session and scored atypically on at least two of three diagnostic tests: the CFMT (Duchaine & Nakayama, 2006), the CFPT (Duchaine et al., 2007), and a famous faces test (e.g. Bate, Bennetts, Gregory, et al., 2019a;

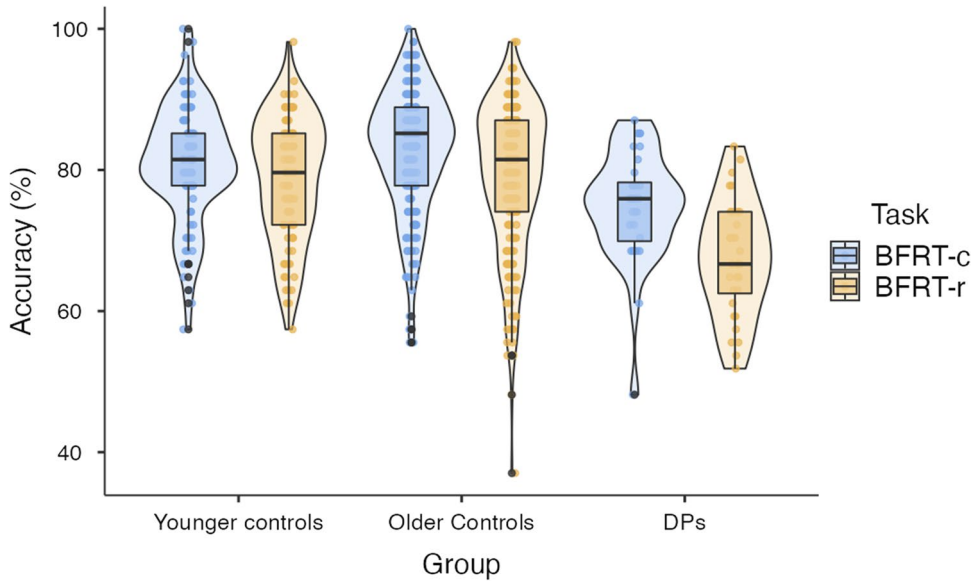

Fig. 4 Violin plots indicating scores on the BFRT-c and BFRT-r for younger control participants, older control participants, and the DP group

Bennetts et al., 2015; Murray & Bate, 2019), following existing diagnostic protocols (Dalrymple & Palermo, 2015; Bate & Tree, 2017: see supplementary material for their diagnostic results). Eight were male, and they were aged between 40 and 59 years (M = 52.0 years, SD = 5.6).

Because our DP sample were older than the younger adults reported in Experiment 1, a new set of 243 older control participants (M age = 48.4 years, SD = 5.9; 119 females) were recruited for age-matched comparison. These individuals were again recruited via the *Prolific* online recruitment platform, in exchange for a small financial incentive. All DP and control participants were Caucasian, and reported no history of socio-emotional, neurological or psychiatric disorder (including mild cognitive impairment) and had normal or corrected-to-normal vision.

Following the same data-processing strategies as in Experiment 1, data from 25 control participants were removed: 23 provided too many duplications in their responses, and an additional two participants took an abnormally long time to complete both the BFRT-r and BFRT-c. This resulted in a final sample of 218 (116 male) control participants, aged between 40 and 60 years (M = 48.4 years, SD = 5.9). The same exclusion criteria were applied to the DP data as for the control participants; no DP data were removed from the analysis.

Materials and procedure

All participants completed the upright version of the BFRT-r and BFRT-c in that order, online, via the testing platform *Testable*.

Results

Age and gender

For the older controls, high correlations were observed between performance on the two versions of the Benton on both the accuracy ($r = .743$, $p < .001$) and task completion time ($r = .844$, $p < .001$) measures (as seen in Experiment 1 in the younger control data). Further, BFRT-r accuracy performance did not differ between the new set of older control participants (M = 78.98%, SD = 10.46) and the younger sample reported in Experiment 1 (M = 78.83%, SD = 8.7), $t(309) = -.115$, $p = .908$. However, overall task completion times were slower in older (M = 348.22 s, SD = 165.20) compared to younger (M = 253.75 s, SD = 130.97) controls, $t(309) = -4.896$, $p < .001$, $d = .63$ (see Fig. 4). The same pattern emerged for the BFRT-c: younger (M = 80.82%, SD = 8.17) and older (M = 82.71%, SD = 8.76) controls performed similarly in terms of accuracy, $t(309) = -1.776$, $p = .077$, but younger controls (M = 168.34 s, SD = 69.81) completed the test significantly faster than older controls (M = 247.55 s, SD = 108.30), $t(309) = 6.497$, $p < .001$, $d = .87$ (see Fig. 5). Thus, subsequent analyses only compared the performance of DPs to the older control group. No gender effects were found on either the BFRT-r or BFRT-c in this age group ($ps > .05$).

DP performance: Group analyses

A mixed 2 (test: BFRT-c, BFRT-r) x 2 (group: DP, older controls) ANOVA was conducted to explore overall group differences in accuracy scores (see Table 1). There was a

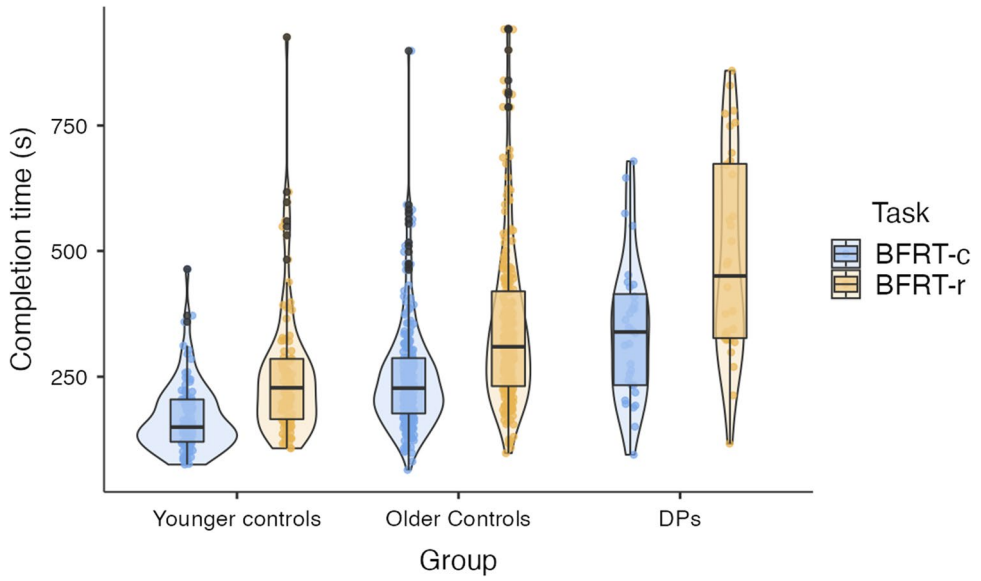

Fig. 5 Violin plots indicating completion times on the BFRT-c and BFRT-r for younger control participants, older control participants, and the DP group

significant main effect of group, whereby DP participants scored significantly poorer (M = 71.15%, SD = 8.03) than control participants (M = 80.84%, SD = 9.61), $F(1,248) = 34.207$, $p < .001$, $\eta\rho^2 = .121$ (see Fig. 4). There was also a main effect of test, $F(1,248) = 66.702$, $p < .001$, $\eta\rho^2 = .212$: scores on the BFRT-c (M = 87.70%, SD = 9.02) were higher than those on the BFRT-r (M = 77.50%, SD = 10.89). There was also a significant interaction between test and group, $F(1,248) = 7.079$, $p = .008$, $\eta\rho^2 = .028$. Pairwise comparisons revealed that DPs scored significantly lower than controls on both the BFRT-c (M = 74.83%, SD = 8.64 and M = 82.71%, SD = 8.65 respectively: $p < .001$) and BFRT-r (M = 67.48%, SD = 10.21 and M = 78.98%, SD = 10.20 respectively: $p < .001$).

The difference in BFRT-r and BFRT-c scores was larger for DPs (mean difference = 7.35%, SD = 7.85) than for control participants (mean difference = 3.74%, SD = 7.07), $t(248) = 2.661$, $p = .008$, $d = .48$.

To investigate any differences in task completion times, a 2 (test: BFRT-c, BFRT-r) x 2 (group: DP, controls) ANOVA was conducted (see Table 1). A significant main effect of group indicated that DPs took longer to complete the tests (M = 415.14 s, SD = 171.08) than controls (M = 297.88 s, SD = 136.75), $F(1,248) = 20.812$, $p < .001$, $\eta\rho^2 = .077$ (see Fig. 5). A significant main effect of test indicated that participants took longer to complete the BFRT-r (M = 366.26 s, SD = 176.36) than the BFRT-c (M = 259.53 s, SD = 116.79), $F(1,248) = 178.426$, $p < .001$, $\eta\rho^2 = .418$. There was also a significant interaction between test and group, $F(1,248) = 6.455$, $p = .012$, $\eta\rho^2 = .025$. Pairwise comparisons revealed that DPs took significantly longer than controls to complete both the BFRT-c (M = 341.16, SD = 112.74 and M = 247.55, SD = 112.73 respectively: $p < .001$) and BFRT-r (M = 489.13, SD = 170.28 and M = 348.22, SD = 170.28 respectively: $p < .001$).

The difference in BFRT-r and BFRT-c scores were larger for DPs (mean difference = 147.67 s, SD = 124.55) than for control participants (mean difference = 100.67 s, SD = 93.98), $t(248) = 2.54$, $p = .012$, $d = .43$.

Because we also held CFPT upright accuracy scores for all our DPs as part of their background diagnostic profiles (see supplementary material), we were able to investigate whether accuracy or completion time on both versions of the BFRT were associated with this indicator. However, no significant correlations were observed between CFPT performance and any of the four BFRT measures (all $ps > .490$), a finding which has recently been reported elsewhere (Mishra et al., 2020). As such, this may not be a surprising finding, especially given the BFRT and CFPT have considerable differences in their paradigms. On the other hand, accuracy ($r = 0.538$, $p = .002$) and completion time ($r = 0.788$, $p = .001$; sequential Bonferroni correction for multiple correlations applied) on the two versions of the BFRT were highly correlated in DP participants, as was found for control participants in both experiments. Because of the lack of association between CFPT and BFRT scores, we did not proceed to use CFPT scores to further interpret individual patterns of performance on either version of the BFRT (see below). Indeed, it is not possible to infer from the current methodology whether either the CFPT or BFRT offers a "true" indicator of face perception, and we instead focus on

consistency of individual performance across the two versions of the BFRT.

Single-case analyses

To examine the importance of assessing both accuracy and response times on face matching tests, each DP's performance on the BFRT-r was evaluated on both parameters on a case-by-case basis (see Table 3). As all participants were over the age of 40, their scores and completion times were compared to that of the older control group (see Table 1). The *z*-score used as a cut-off for typicality varies within the DP literature, with some authors using two SDs from the mean (Bate, Bennetts, Tree, et al., 2019b; Biotti et al., 2019; Bowles et al., 2009) and others 1.7 SDs (DeGutis et al., 2012, 2014; Palermo et al., 2017; White et al., 2017). Here, to allow for recording error, and to err on the conservative side when determining face perception is intact (given it is currently assumed that the process is impaired in most DPs: Bate, Bennetts, Gregory, et al., 2019a; Biotti et al., 2019), we present the findings in terms of a 1.7 SD cut-off.

Fifteen of the 32 DPs (46.88%) performed within the typical range on both the BFRT-r and BFRT-c according to both accuracy and completion time measures (see Table 3).

Table 3 Normalised accuracy scores and task completion times for the 32 DP participants on the BFRT-r and BFRT-c

Participant ID	BFRT-r Accuracy	BFRT-r Completion Time	BFRT-c Accuracy	BFRT-c Completion Time
DPM01	– 0.11	– 0.82	0.30	– 0.96
DPM02	– 0.47	– 0.15	– 0.45	1.07
DPM03	– 2.24 *	– 0.09	– 0.64	– 0.24
DPM04	– 0.47	– 0.48	– 1.01	0.79
DPM05	– 1.89 *	2.01 *	– 0.64	1.54
DPM06	– 1.00	– 0.05	– 1.39	– 0.52
DPM07	– 1.89 *	2.43 *	– 0.64	4.31 *
DPM08	– 1.53	– 0.30	– 0.77	– 0.04
DPF01	– 1.18	0.48	– 1.62	0.83
DPF02	– 0.82	0.75	0.07	1.19
DPF03	– 0.65	0.79	– 0.56	1.37
DPF04	– 0.82	2.47 *	0.28	1.67
DPF05	– 0.82	2.47 *	– 0.35	3.98 *
DPF06	– 1.18	1.34	– 0.56	3.02 *
DPF07	– 2.24 *	– 1.40	– 3.95 *	– 1.41
DPF08	– 1.35	2.61 *	– 1.62	1.71 *
DPF09	– 0.47	1.84 *	– 0.14	1.89 *
DPF10	– 2.42 *	1.96 *	– 1.41	1.27
DPF11	– 0.47	1.04	– 0.99	1.49
DPF12	– 0.82	– 0.18	– 1.62	– 0.13
DPF13	– 0.42	2.10 *	0.28	1.07
DPF14	– 1.53	1.22	– 0.56	0.93
DPF15	– 1.35	3.09 *	– 1.62	0.25
DPF16	0.06	– 0.13	0.07	– 0.55
DPF17	– 2.59 *	0.20	– 2.47 *	– 0.48
DPF18	– 1.53	1.22	– 0.77	1.76 *
DPF19	– 1.35	3.09 *	– 0.99	2.79 *
DPF20	0.06	– 0.13	– 0.77	0.07
DPF21	– 1.18	– 0.08	0.49	– 0.14
DPF22	– 1.71 *	– 0.15	– 1.20	0.14
DPF23	– 2.06 *	– 0.02	– 1.62	0.60
DPF24	0.11	– 0.14	– 0.99	– 0.41

Negative *z*-scores represent poorer performance for accuracy, and positive scores indicate slower completion times.* denotes an atypical z-score (+/– 1.7)

Notably few borderline scores were detected: the closest score to a cut-off was a *z*-score of -1.62 on the BFRT-c accuracy (DPF01; DPF12) and – 1.53 on the BFRT-r accuracy (DPM08; DPF14), with the vast majority of other scores occurring within 1.25 SDs of the control mean. Of the 17 DPs who showed at least some impairment, seven exceeded cut-off on both tasks, according to at least one measure. An additional eight participants were only impaired on the BFRT-r, and two participants were only impaired on the BFRT-c. Interestingly, only five DPs displayed impairments on accuracy alone, whereas nine DPs only showed impairments on the completion time measure. Thus, task completion time was the primary indicator of impairment on both versions of the test.

False Alarms & Sensitivity

"Typical" performance on each task was classified as scoring above – 1.7 SD from the age-matched control mean for accuracy, and less than 1.7 SD above the control mean for RT. For younger control participants, 86.02% (80/93) were correctly classified as typical performers on the BFRT-c and on the BFRT-r. For older control participants, 88.07% (192/218) were correctly classified as typical performers on the BFRT-c and 86.70% (189/218) were correctly classified as typical performers on the BFRT-r. For the DP participants, there was a false alarm rate of 71.88% (23/32 classified as typical performers) for the BFRT-c, and 53.12% (17/32 classified as typical performers) for the BFRT-r (see Figs. 6 and 7).

To examine how well the BFRT-c and BFRT-r discriminate between control participants and individuals with DP, we calculated *d'* for each test. *d'* is a bias-free measure of sensitivity that combines information about hit rates (in this case, the number of older control participants who were correctly classified as "typical" performers on each test) and false alarms (participants with DP who were classified as "typical" performers on each test) (Macmillan & Creelman, 2005). A *d'* of 0 would indicate chance discrimination; and a *d'* of 5.00 would represent perfect discrimination between control and DP participants in this sample. The *d'* for the BRFT-c is 0.60, whereas the *d'* for the BFRT-r is 1.03.

Summary

This study examined performance on the BFRT-r and BFRT-c in DPs and matched older adult control participants. Comparison of the control data to the younger participants tested in Experiment 1 indicates that accuracy performance is consistent across the two age groups, but older participants took longer to complete both tasks. Thus, performance of atypical participants needs to be compared to an age-matched control group. While a small gender effect was found for younger controls in Experiment 1, this did not emerge for the older controls in this Experiment, and we therefore recommend

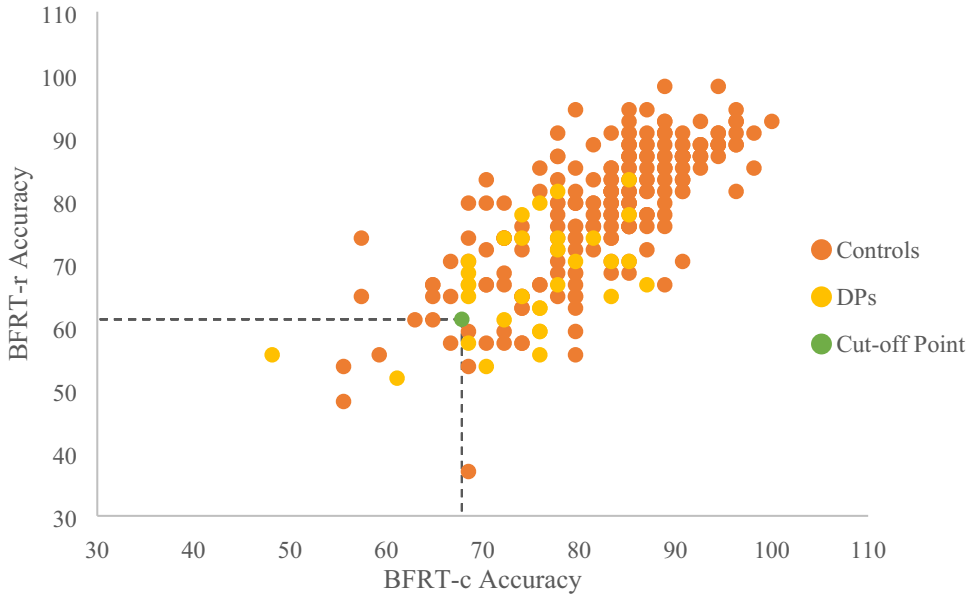

Fig. 6 Scatterplot indicating the individual scores, in percentage, for DPs and older control participants. *Note*. The *dashed line* indicates the cut-offs for DP criteria (1.7 SDs from the mean). Thus, those below the dashed line on the *y*-axis were atypical on the BFRT-r and those to the left of the dashed line on the *x*-axis were atypical on the BFRT-c. Note that two of the DP data plots represent two DPs each; they scored the same on both tests (two scored 68.52 and 66.67%, and two scored 75.93 and 55.56%, on the BFRT-c and BFRT-r, respectively)

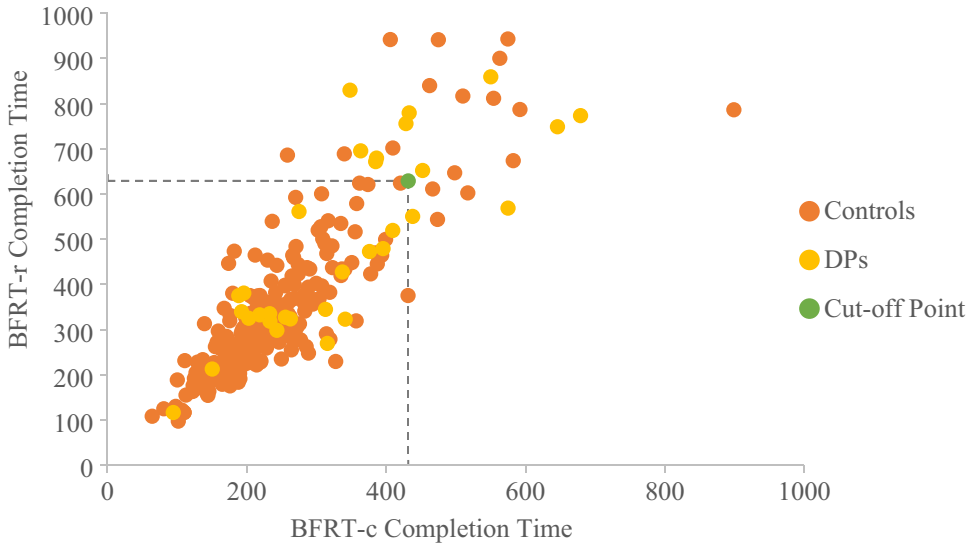

Fig. 7 Scatterplot indicating the individual completion times for DPs and older control participants. *Note.* The *dashed line* indicates the cut-offs for DP criteria (1.7 SDs from the mean). Thus, those above the dashed line on the *y*-axis were atypical on the BFRT-r and those to the right of the dashed line on the *x*-axis were atypical on the BFRT-c

that further investigation is required to confirm whether gender differences persist on the task.

As a group, the DPs performed more poorly than controls on both the BFRT-r and BFRT-c according to both accuracy and completion time measures (see Table 1). However, akin to previous work (Bate, Bennetts, Gregory, et al., 2019a; Burns et al., 2017; Le Grand et al., 2006; Minnebusch et al., 2007; Stantic et al., 2021), case-by-case analyses indicated considerable heterogeneity in DPs' face perception performance, with just under half of the sample displaying intact face perception skills on both tests. Consistent deficits in face perception were noted across both versions of the test in seven DPs. However, only the BFRT-r detected impairment for eight DPs, and only the BFRT-c for the remaining two, suggesting some cases may be missed by administration of a single face perception task (see also Murray & Bate, 2020).

General discussion

In this paper, we introduce a new version of the BFRT (the BFRT-r), with updated stimuli that address recent theoretical progress in the face recognition literature. We sought to examine the validity of the BFRT-r in typical participants, and provide norming data for younger (aged 18–35 years) and older (aged 40–60 years) adult populations. Although our control samples did not include individuals aged 36–39 years, patterns observed in previous work (e.g. Bate, Bennetts, Gregory, et al., 2019a; Bowles et al., 2009) suggest clinical cases within this age range should be compared to the younger control group. That is, individuals aged 36–39 years perform similarly to those ages 35 and younger on face processing tasks. Finally, we investigated the test's utility in identifying face perception difficulties in DP.

First, we replicated several known advantages of the BFRT. As the BFRT-r procedure is identical to the computerised version of the original BFRT (the BFRT-c: Rossion & Michel, 2018), the test is known to follow a simple procedure and is quick to administer. Here, we found that the BFRT-r takes approximately 4–6 min for typical participants to complete (with longer completion times in older adults), and approximately 8 min for older adults with DP. Further, we noted the particular importance of monitoring task completion times for accurate diagnosis of face perception impairments. This is clearly facilitated by the use of a computerised rather than pencil-and-paper format (see Rossion & Michel, 2018), and also lends the task to online administration – a particularly important concern in very recent times. Notably, online administration was used here in both experiments, and the resulting strong internal reliability and inter-item correlations directly support this mode of implementation. Moreover, the BFRT-r elicits a strong inversion effect and strongly correlates with the BFRT-c and CFMT, evidencing content-validity akin to other tests of face processing (Busigny & Rossion, 2010; Duchaine et al., 2007; Duchaine & Nakayama, 2006; McKone et al., 2011). Most strikingly, the sensitivity of the new BFRT-r in identifying face perception impairments improved substantially than that for the BFRT-c.

It should be noted that no measure of BFRT performance was found to be associated with CFPT accuracy scores in the DP sample. This finding replicates a similar recent

report for the BFRT-c (Mishra et al., 2020), and also fits with other findings of low correlations between four different face matching tests (Fysh et al., 2020). Further investigation is required to decipher whether these poor associations between tasks that are thought to operate under the umbrella of "face perception" represent differences in process or content, or even participant inconsistency (see Bate et al., 2018; Bate, Frowd, Bennetts, et al., 2019c). This could potentially have important implications for DP screening protocols as it appears from the present data that these tasks should not be used interchangeably, and, pending further investigation, it may be prudent for researchers to administer all three tests to their DP participants. Indeed, investigations into (acquired) prosopagnosia traditionally investigated face processing abilities using several tasks which assessed the same subprocess (e.g. De Renzi & di Pellegrino, 1998; Rossion et al., 2003; Takahashi et al., 1995; Wada & Yamamoto, 2001) and more recently, researchers are urging the use of multiple assessments to examine the consistency of performance in DP (see Murray and Bate (2019) for a discussion). At the same time, it is often encouraged that time-effective tests are used to do this and, whilst the CFPT can take up to 16 minutes to complete, the BFRT-r and BFRT-c are typically much quicker to administer and, consequently, offer a more practical way to assess the consistency of one's face perception skills. That is, face perception could be thoroughly assessed (in the vast majority of cases) in less than 30 min.

The main adaptation of the new BFRT-r concerns the new images, both in terms of visual quality, and in addressing important theoretical concerns within the field. While the original BFRT images were highly constrained and were presumably captured in the same setting on the same day and using the same camera, our images embraced the natural variability which typically occurs when viewing the same person on different occasions in everyday life. The photographs were captured over different days (sometimes months apart), using a variety of cameras, showing the person from varying viewpoints and distances from the camera, and in different lighting conditions. By moving away from the tightly controlled conditions that prevail in existing tests of face perception, it is likely that we move closer towards the circumstances of everyday face perception, providing a more ecologically valid diagnostic test (Burton, 2013). In addition, the use of more ambient facial images also overcomes previous concerns that extra-facial or distinguishing features could be used by clinical participants to achieve typical scores on the BFRT (Duchaine & Weidenfeld, 2003). Taken together, the new stimuli have likely gone some way towards addressing this issue.

Importantly, our data also indicated that face perception skills are heterogeneous in DP – a factor that has been highlighted in previous work (Bate, Bennetts, Tree, et al., 2019b; Burns et al., 2017; Le Grand et al., 2006; Minnebusch et al., 2007; Stantic et al., 2021). Here, we found that just less than half of our DP sample presented with no impairments on either version of the BFRT. While seven of the remaining 17 DPs consistently displayed deficits on both versions of the BFRT, eight were only detected on the BFRT-r and two by the BFRT- c (note that we did not attempt to further clarify these patterns using CFPT scores given the lack of association between the two paradigms). Together, these patterns of performance highlight the importance of administering more than one task when screening for face perception deficits. This is particularly true for a condition such as DP, where face recognition difficulties appear to mostly be lifelong and do not accompany any other form of dysfunction. This allows many people with DP to develop elaborate compensatory strategies that may help them with particular facial stimuli or task paradigms, allowing them to obscure their difficulties (Adams et al., 2019). The case for repeat-testing aligns with our recent demonstration of the importance of repeat-screening for face memory deficits in DP, given the possibility that "typical" scores can be achieved by chance or due to low task reliability (Murray & Bate, 2020).

One further way to address this issue, particularly in tasks of face perception, is to place more emphasis on completion times, given accurate scores may be obtained by spending a long time on a task. Consistent with existing work (e.g. Bukach et al., 2006; Busigny & Rossion, 2010; Delvenne et al., 2004; Jansari et al., 2015; Rossion & Michel, 2018), the finding reported here that nine DPs were only impaired on completion time (but not accuracy) on either, or both tests, highlights the importance of assessing both measures. Indeed, longer completion times may reflect the use of laboured face processing strategies and methods which ultimately lead to a correct response. However, it is important to note that our older adult controls took longer to complete the task than younger adults, although the same effect did not emerge for accuracy. Thus, we strongly suggest that age-matched norms are used for identifying impaired performance on this task. Additionally, with this finding in mind, the BFRT-r likely offers itself to be a suitable task for examining age-related changes in face processing within the typical population.

In conclusion, this paper presents an updated version of the BFRT with new theoretically motivated stimuli. As researchers continue to recommend repeat assessment of face processing deficit is to explore consistency of impairment, the field will continue to benefit from more tasks which to assess sub-processes of face processing. As such, the BFRT-r offers an opportunity for repeat screening for consistency of performance that improves substantially on the sensitivity offered by the BFRT-c. The task can be shared with other researchers on request.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.3758/s13428-021-01727-x.

References

Adams, A., Hills, P., Bennetts, R., & Bate, S. (2019). Coping strategies for developmental prosopagnosia. *Neuropsychological Rehabilitation*, 1–20. https://doi.org/10.1080/09602011.2019.1623824

Annaz, D., Karmiloff-Smith, A., Johnson, M., & Thomas, M. (2009). A cross-syndrome study of the development of holistic face recognition in children with autism, Down syndrome, and Williams syndrome. *Journal of Experimental Child Psychology*, *102*(4), 456–486. https://doi.org/10.1016/j.jecp.2008.11.005

Barton, J. (2008). Structure and function in acquired prosopagnosia: Lessons from a series of 10 patients with brain damage. *Journal of Neuropsychology*, *2*(1), 197–225. https://doi.org/10.1348/174866407x214172

Barton, J., & Corrow, S. (2016). The problem of being bad at faces. *Neuropsychologia*, *89*, 119–124. https://doi.org/10.1016/j.neuropsychologia.2016.06.008

Bate, S., & Bennetts, R. (2015). The independence of expression and identity in face-processing: evidence from neuropsychological case studies. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.00770

Bate, S., & Tree, J. (2017). The Definition and Diagnosis of Developmental Prosopagnosia. *Quarterly Journal of Experimental Psychology*, *70*(2), 193–200. https://doi.org/10.1080/17470218.2016.1195414

Bate, S., Haslam, C., Tree, J., & Hodgson, T. (2008). Evidence of an eye movement-based memory effect in congenital prosopagnosia. *Cortex*, *44*(7), 806–819. https://doi.org/10.1016/j.cortex.2007.02.004

Bate, S., Frowd, C., Bennetts, R., Hasshim, N., Murray, E., Bobak, A.K., Wills, H., & Richards, S. (2018). Applied screening tests for the detection of superior face recognition. *Cognitive Research: Principles and Implications*, *3*(1). https://doi.org/10.1186/s41235-018-0116-5

Bate, S., Bennetts, R., Gregory, N., Tree, J., Murray, E., & Adams, A., Bobak, A.K., Penton, T., Yang, T., & Bannisy, M.J. (2019a). Objective Patterns of Face Recognition Deficits in 165 Adults with Self-Reported Developmental Prosopagnosia. *Brain Sciences*, *9*(6), 133. https://doi.org/10.3390/brainsci9060133

Bate, S., Bennetts, R., Tree, J., Adams, A., & Murray, E. (2019b). The domain-specificity of face matching impairments in 40 cases of developmental prosopagnosia. *Cognition*, *192*, 104031. https://doi.org/10.1016/j.cognition.2019.104031

Bate, S., Frowd, C., Bennetts, R., Hasshim, N., Portch, E., Murray, E., & Dudfield, G. (2019c). The consistency of superior face recognition skills in police officers. *Applied Cognitive Psychology*, *33*(5), 828–842. https://doi.org/10.1002/acp.3525

Behrmann, M., Avidan, G., Marotta, J., & Kimchi, R. (2005). Detailed Exploration of Face-related Processing in Congenital Prosopagnosia: 1. Behavioral Findings. *Journal of Cognitive Neuroscience*, *17*(7), 1130–1149. https://doi.org/10.1162/0898929054475154

Bennetts, R., Butcher, N., Lander, K., Udale, R., & Bate, S. (2015). Movement cues aid face recognition in developmental prosopagnosia. *Neuropsychology*, *29*(6), 855–860. https://doi.org/10.1037/neu0000187

Benton, A. L., & Van Allen, M. W. (1968). Impairment in facial recognition in patients with cerebral disease. *Transactions of the American Neurological Association, 93*, 38–42.

Benton, A., & Van Allen, M. (1972). Prosopagnosia and facial discrimination. *Journal Of The Neurological Sciences, 15*(2), 167–172. https://doi.org/10.1016/0022-510x(72)90004-4

Benton, A. L., Sivan, A. B., Hamsher, K. D. S., Varney, N. R., & Spreen, O. (1983). Facial recognition: Stimulus and multiple-choice pictures. In A. L. Benton, A. B. Sivan, K. D. S. Hamsher, N. R. Varney, & O. Spreen (Eds.), Contribution to neuropsychological assessment (pp. 30–40). Oxford University Press.

Biotti, F., Wu, E., Yang, H., Jiahui, G., Duchaine, B., & Cook, R. (2017). Normal composite face effects in developmental prosopagnosia. *Cortex*, *95*, 63–76. https://doi.org/10.1016/j.cortex.2017.07.018

Biotti, F., Gray, K., & Cook, R. (2019). 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

Bowles, D., McKone, E., Dawel, A., Duchaine, B., Palermo, R., Schmalzl, L., Rivolta, D., Wilson, E., & Yovel, G. (2009). Diagnosing prosopagnosia: Effects of ageing, sex, and participant–stimulus ethnic match on the Cambridge Face Memory Test and Cambridge Face Perception Test. *Cognitive Neuropsychology*, *26*(5), 423–455. https://doi.org/10.1080/02643290903343149

Bukach, C. M., Bub, D. N., Gauthier, I., & Tarr, M. J. (2006). Perceptual expertise effects are not all or none: Spatially limited perceptual expertise for faces in a case of prosopagnosia. *Journal of Cognitive Neuroscience, 18*, 48–63. https://doi.org/10.1162/089892906775250094

Burns, E., Martin, J., Chan, A., & Xu, H. (2017). Impaired processing of facial happiness, with or without awareness, in developmental prosopagnosia. *Neuropsychologia*, *102*, 217–228. https://doi.org/10.1016/j.neuropsychologia.2017.06.020

Burton, M.A. (2013). Why has research in face recognition progressed so slowly? The importance of variability. *Quarterly Journal of Experimental Psychology*, *66*(8), 1467–1485. https://doi.org/10.1080/17470218.2013.800125

Burton, A., Bruce, V., & Hancock, P. (1999). From Pixels to People: A Model of Familiar Face Recognition. *Cognitive Science, 23*(1), 1–31. https://doi.org/10.1207/s15516709cog2301_1

Burton, M.A., White, D., & McNeill, A. (2010). The Glasgow Face Matching Test. *Behavior Research Methods, 42*, 286–291. https://doi.org/10.3758/BRM.42.1.286

Busigny, T., & Rossion, B. (2010). Acquired prosopagnosia abolishes the face inversion effect. *Cortex, 46*, 965–981. https://doi.org/10.1016/j.cortex.2009.07.004

Dalrymple, K., & Palermo, R. (2015). Guidelines for studying developmental prosopagnosia in adults and children. *Wiley Interdisciplinary Reviews: Cognitive Science*, *7*(1), 73–87. https://doi.org/10.1002/wcs.1374

De Luca, M., Pizzamiglio, M., Di Vita, A., Palermo, L., Tanzilli, A., Dacquino, C., & Piccardi, L. (2019). First the nose, last the eyes in congenital prosopagnosia: Look like your father looks. *Neuropsychology*, *33*(6), 855–861. https://doi.org/10.1037/neu0000556

De Renzi, E., & di Pellegrino, G. (1998). Prosopagnosia and Alexia Without Object Agnosia. *Cortex*, *34*(3), 403–415. https://doi.org/10.1016/s0010-9452(08)70763-9

De Renzi, E., Faglioni, P., Grossi, D., & Nichelli, P. (1991). Apperceptive and Associative Forms of Prosopagnosia. *Cortex*, *27*(2), 213–221. https://doi.org/10.1016/s0010-9452(13)80125-6

DeGutis, J., Chatterjee, G., Mercado, R., & Nakayama, K. (2012). Face gender recognition in developmental prosopagnosia: Evidence for holistic processing and use of configural information. *Visual Cognition*, *20*(10), 1242–1253. https://doi.org/10.1080/13506285.2012.744788

DeGutis, J., Cohan, S., & Nakayama, K. (2014). Holistic face training enhances face processing in developmental prosopagnosia. *Brain*, *137*(6), 1781–1798. https://doi.org/10.1093/brain/awu062

Delvenne, J.F., Seron, X., Coyette, F., & Rossion, B. (2004). Evidence for perceptual deficits in associative visual (prosop)agnosia: A

single-case study. *Neuropsychologia, 42,* 597–612. https://doi.org/10.1016/j.neuropsychologia.2003.10.008

Duchaine, B., & Nakayama, K. (2004). Developmental prosopagnosia and the Benton Facial Recognition Test. *Neurology*, *62*(7), 1219–1220. https://doi.org/10.1212/01.wnl.0000118297.03161.b3

Duchaine, B., & Nakayama, K. (2006). The Cambridge Face Memory Test: Results for neurologically intact individuals and an investigation of its validity using inverted face stimuli and prosopagnosic participants. *Neuropsychologia*, *44*(4), 576–585. https://doi.org/10.1016/j.neuropsychologia.2005.07.001

Duchaine, B., & Weidenfeld, A. (2003). An evaluation of two commonly used tests of unfamiliar face recognition. *Neuropsychologia*, *41*(6), 713–720. https://doi.org/10.1016/s0028-3932(02)00222-1

Duchaine, B., Germine, L., & Nakayama, K. (2007). Family resemblance: Ten family members with prosopagnosia and within-class object agnosia. *Cognitive Neuropsychology*, *24*(4), 419–430. https://doi.org/10.1080/02643290701380491

Esins, J., Schultz, J., Stemper, C., Kennerknecht, I., & Bülthoff, I. (2016). Face Perception and Test Reliabilities in Congenital Prosopagnosia in Seven Tests. *I-Perception*, *7*(1), 204166951562579. https://doi.org/10.1177/2041669515625797

Farah, M. J. (1990). *Visual Agnosia: Disorders of object recognition and what they tell us about normal vision*. MIT Press.

Fysh, M., Stacchi, L., & Ramon, M. (2020). Differences between and within individuals, and subprocesses of face cognition: implications for theory, research and personnel selection. *Royal Society Open Science, 7*(9), 200233. https://doi.org/10.1098/rsos.200233

Geskin, J., & Behrmann, M. (2017). Congenital prosopagnosia without object agnosia? A literature review. *Cognitive Neuropsychology*, *35*(1–2), 4–54. https://doi.org/10.1080/02643294.2017.1392295

Hasson, U., Avidan, G., Deouell, L., Bentin, S., & Malach, R. (2003). Face-selective activation in a Congenital Prosopagnosic subject. *Journal Of Cognitive Neuroscience*, *15*(3), 419–431. https://doi.org/10.1162/089892903321593135

Herlitz, A., & Lovén, J. (2013). Sex differences and the own-gender bias in face recognition: A meta-analytic review. *Visual Cognition*, *21*(9–10), 1306–1336. https://doi.org/10.1080/13506285.2013.823140

Jansari, A., Miller, S., Pearce, L., Cobb, S., Sagiv, N., Williams, A. L., Tree, J., & Hanley, J. R. (2015). The man who mistook his neuropsychologist for a popstar: When configural processing fails in acquired prosopagnosia. *Frontiers in Human Neuroscience, 9*, 390. https://doi.org/10.3389/fnhum.2015.00390

Kennerknecht, I., Plümpe, N., Edwards, S., & Raman, R. (2006). Hereditary prosopagnosia (HPA): the first report outside the Caucasian population. *Journal of Human Genetics*, *52*(3), 230–236. https://doi.org/10.1007/s10038-006-0101-6

Le Grand, R., Cooper, P., Mondloch, C., Lewis, T., Sagiv, N., de Gelder, B., & Maurer, D. (2006). What aspects of face processing are impaired in developmental prosopagnosia? *Brain and Cognition*, *61*(2), 139–158. https://doi.org/10.1016/j.bandc.2005.11.005

Liu, C., Collin, C., Rainville, S., & Chaudhuri, A. (2000). The effects of spatial frequency overlap on face recognition. *Journal Of Experimental Psychology: Human Perception And Performance, 26*(3), 956–979. https://doi.org/10.1037/0096-1523.26.3.956

Lovén, J., Herlitz, A., & Rehnman, J. (2011). Women's own-gender bias in face recognition memory. The role of attention at encoding. *Experimental Psychology, 58*, 333–340. https://doi.org/10.1027/1618-3169/a000100

McKone, E., Hall, A., Pidcock, M., Palermo, R., Wilkinson, R., Rivolta, D., Yovel, G., David, J.M., & O'Connor, K.B. (2011). Face ethnicity and measurement reliability affect face recognition performance in developmental prosopagnosia: Evidence from the Cambridge Face Memory Test–Australian. *Cognitive Neuropsychology*, *28*(2), 109–146. https://doi.org/10.1080/02643294.2011.616880

Minnebusch, D., Suchan, B., Ramon, M., & Daum, I. (2007). Event-related potentials reflect heterogeneity of developmental prosopagnosia. *European Journal of Neuroscience*, *25*(7), 2234–2247. https://doi.org/10.1111/j.1460-9568.2007.05451.x

Mishra, M., Fry, R., Saad, E., Arizpe, J., Ohashi, Y., & DeGutis, J. (2020). *Comparing the sensitivity of face matching assessments to detect face perception deficits.* PsyArXiv. https://doi.org/10.31234/osf.io/68gbm

Murray, E., & Bate, S. (2019). Self-ratings of face recognition ability are influenced by gender but not prosopagnosia severity. *Psychological Assessment*, *31*(6), 828–832. https://doi.org/10.1037/pas0000707

Murray, E., & Bate, S. (2020). Diagnosing developmental prosopagnosia: repeat assessment using the Cambridge Face Memory Test. *Royal Society Open Science*, *7*(9), 200884. https://doi.org/10.1098/rsos.200884

Palermo, R., Willis, M., Rivolta, D., McKone, E., Wilson, C., & Calder, A. (2011). Impaired holistic coding of facial expression and facial identity in congenital prosopagnosia. *Neuropsychologia*, *49*(5), 1226–1235. https://doi.org/10.1016/j.neuropsychologia.2011.02.021

Palermo, R., Rossion, B., Rhodes, G., Laguesse, R., Tez, T., Hall, B., Albonico, A., Malaspina, M., Daini, R., Irons, J., Al-Janabi, S., Taylor, L.C., Rivolta, D., & McKone, E. (2017). Do People Have Insight into their Face Recognition Abilities? *Quarterly Journal of Experimental Psychology*, *70*(2), 218–233. https://doi.org/10.1080/17470218.2016.1161058

Rabin, L., Barr, W., & Burton, L. (2005). Assessment practices of clinical neuropsychologists in the United States and Canada: A survey of INS, NAN, and APA Division 40 members. *Archives of Clinical Neuropsychology*, *20*(1), 33–65. https://doi.org/10.1016/j.acn.2004.02.005

Rezlescu, C., Danaila, I., Miron, A., & Amariei, C. (2020). More time for science: Using Testable to create and share behavioral experiments faster, recruit better participants, and engage students in hands-on research. *Progress in Brain Research, 253*, 243–262. https://doi.org/10.1016/bs.pbr.2020.06.005

Righart, R., & de Gelder, B. (2007). Impaired face and body perception in developmental prosopagnosia. *Proceedings of The National Academy of Sciences*, *104*(43), 17234–17238. https://doi.org/10.1073/pnas.0707753104

Robertson, D., Noyes, E., Dowsett, A., Jenkins, R., & Burton, A. (2016). Face Recognition by Metropolitan Police Super-Recognisers. *PLOS ONE*, *11*(2), e0150036. https://doi.org/10.1371/journal.pone.0150036

Rossion, B., & Michel, C. (2018). Normative accuracy and response time data for the computerized Benton Facial Recognition Test (BFRT-c). *Behavior Research Methods*, *50*(6), 2442–2460. https://doi.org/10.3758/s13428-018-1023-x

Rossion, B., Caldara, R., Seghier, M., Schuller, A., Lazeyras, F., & Mayer, E. (2003). A network of occipito-temporal face-sensitive areas besides the right middle fusiform gyrus is necessary for normal face processing. *Brain, 126*(11), 2381–2395. https://doi.org/10.1093/brain/awg241

Sachse, M., Schlitt, S., Hainz, D., Ciaramidaro, A., Walter, H., Poustka, F., Bolte, S., & Freitag, C.M. (2014). Facial emotion recognition in paranoid schizophrenia and autism spectrum disorder. *Schizophrenia Research*, *159*(2–3), 509–514. https://doi.org/10.1016/j.schres.2014.08.030

Shah, P., Sowden, S., Gaule, A., Catmur, C., & Bird, G. (2015). The 20-item prosopagnosia index (PI20): relationship with the Glasgow face-matching test. *Royal Society Open Science*, *2*(11), 150305. https://doi.org/10.1098/rsos.150305

Stantic, M., Brewer, R., Duchaine, B., Banissy, M.J., Bate, S., Susilo, T., Catmur, C. & Bird, G. (2021). The Oxford Face Matching Test: A non-biased test of the full range of individual differences in face perception. *Behavior Research Methods*.https://doi.org/10.3758/s13428-021-01609-2

Takahashi, N., Kawamura, M., Hirayama, K., Shiota, J., & Isono, O. (1995). Prosopagnosia: A Clinical and Anatomical Study of Four Patients. *Cortex, 31*(2), 317–329. https://doi.org/10.1016/s0010-9452(13)80365-6

Van Belle, G., Busigny, T., Lefèvre, P., Joubert, S., Felician, O., Gentile, F., & Rossion, B. (2011). Impairment of holistic face perception following right occipito-temporal damage in prosopagnosia: Converging evidence from gaze-contingency. *Neuropsychologia*, *49*(11), 3145–3150. https://doi.org/10.1016/j.neuropsychologia.2011.07.010

Wada, Y., & Yamamoto, T. (2001). Selective impairment of facial recognition due to a haematoma restricted to the right fusiform and lateral occipital region. *Journal Of Neurology, Neurosurgery & Psychiatry, 71*(2), 254–257. https://doi.org/10.1136/jnnp.71.2.254

White, D., Rivolta, D.A., Burton, M., Al-Janabi, S., & Palermo, R. (2017). Face Matching Impairment in Developmental Prosopagnosia. *Quarterly Journal of Experimental Psychology, 70*(2), 287–297. https://doi.org/10.1080/17470218.2016.1173076

Young, A., & Burton, A. (2017). Recognizing Faces. *Current Directions in Psychological Science*, *26*(3), 212–217. https://doi.org/10.1177/0963721416688114

Young, A. W., Newcombe, F., de Haan, E. H., Small, M., & Hay, D. C. (1993). Face perception after brain injury. Selective impairments affecting identity and expression. *Brain, 116*, 941–959. https://doi.org/10.1093/brain/116.4.941

Open Practices Statement None of the experiments were preregistered. The data are available as supplementary material. The BFRT-r stimuli and dataset are available via the Open Science Framework, and can be accessed here: https://osf.io/vza3m/?view_only=404f6d1971924759b126d46cba1d25b7