Behavior Research Methods
https://doi.org/10.3758/s13428-018-1023-x

Normative accuracy and response time data for the computerized Benton Facial Recognition Test (BFRT-*c*)

Bruno Rossion[1,2,3] · Caroline Michel[1]

Abstract

We report normative data from a large ($N = 307$) sample of young adult participants tested with a computerized version of the long form of the classical Benton Facial Recognition Test (BFRT; Benton & Van Allen, 1968). The BFRT-*c* requires participants to match a target face photograph to either one or three of six face photographs presented simultaneously. We found that the percent accuracy on the BFRT-*c* (81%–83%) was below ceiling yet well above chance level, with little interindividual variance in this typical population sample, two important aspects of a sensitive clinical test. Although the split-half reliability on response accuracy was relatively low, due to the large variability in difficulty across items, the correct response times measured in this version—completed in 3 min, on average—provide a reliable and critical complementary measure of performance at individual unfamiliar-face matching. In line with previous observations from other measures, females outperformed male participants at the BFRT-*c*, especially for female faces. In general, performance was also lower following lighting changes than following head rotations, in line with previous studies that have emphasized participants' limited ability to match pictures of unfamiliar faces with important variations in illumination. Overall, this normative data set supports the validity of the BFRT-*c* as a key component of a battery of tests to identify clinical impairments in individual face recognition, such as observed in acquired prosopagnosia. However, this analysis strongly recommends that researchers consider the full test results: Beyond global indexes of performance based on accuracy rates only, they should consider the time taken to match individual faces as well as the variability in performance across items.

Keywords Face matching · Benton test · Response times · Lighting direction · Unfamiliar faces

The Benton Facial Recognition Test (BFRT, first described and used by Benton & Van Allen, 1968; see Benton, Sivan, Hamsher, Varney, & Spreen, 1983, for the official reference of the test) is one of the oldest tests measuring human face processing ability, and probably still the most widely used in clinical neuropsychological evaluations. This behavioral test was developed by the late Arthur Benton in the context of clinical and experimental studies of prosopagnosia—that is, a rare, specific and massive impairment at individual face recognition following brain damage (Bodamer, 1947; Quaglino & Borelli, 1867; see Busigny, Mayer, & Rossion, 2013, for a review), as well as of impairments at unfamiliar-face matching in groups of patients with posterior brain damage (Benton & Van Allen, 1968, 1972; De Renzi, Faglioni, & Spinnler, 1968; Tzavaras, Hécaen, & Le Bras, 1970; see also Milner, 1968). The BFRT consists of matching grayscale photographs of individual (Caucasian) faces: First, participants should match one target face appearing at the top to one of six faces presented simultaneously below (six items), and then match one target face to three of the six faces below (16 items) (Fig. 1), for a total score of 54 [i.e., (6 × 1) + (16 × 3)]. The two pictures to match are virtually identical in the first six items, which are relatively easy and can almost serve as practice trials. However, the remaining 16 items require matching across changes of either head orientation (eight items) or lighting conditions (eight items), and are much more difficult. Overall, the BFRT is a difficult test to perform, with no ceiling effect (Benton & Van Allen, 1972). The stimuli used are also

✉ Bruno Rossion
bruno.rossion@uclouvain.be

[1] Psychological Sciences Research Institute and Institute of Neuroscience, University of Louvain, Louvain-la-Neuve, Belgium

[2] Université de Lorraine, CHRU-Nancy, Service de neurologie, F-54000 Nancy, France

[3] Université de Lorraine, CNRS, CRAN, F-54000 Nancy, France

Published online: 16 March 2018

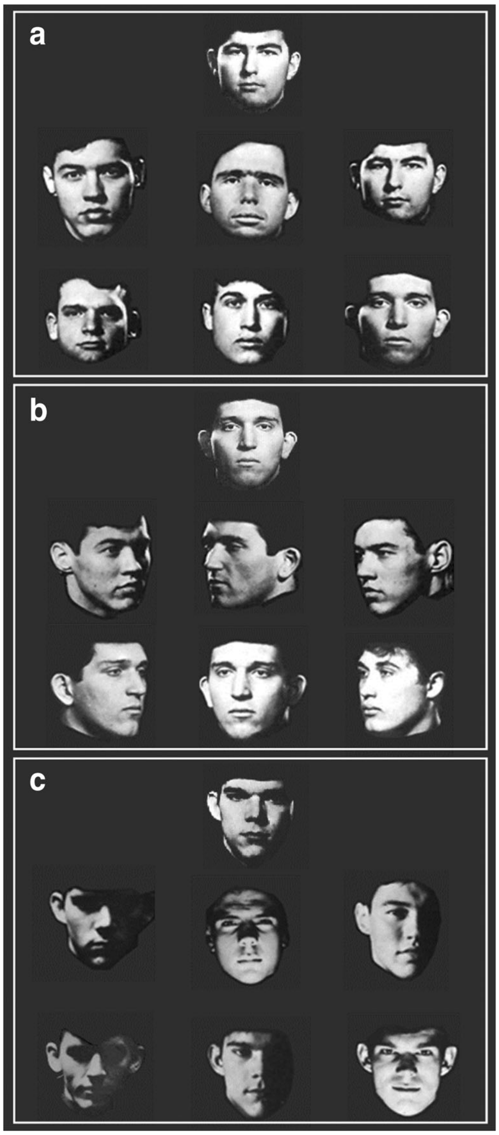

Fig. 1 Examples of items used in the BFRT-*c* (Items 4, 16, and 21 of the BFRT). In all items a target face (unfamiliar face, frontal view) presented at the top of the screen has to be found, as accurately and quickly as possible, among six face photographs presented simultaneously at the bottom of the screen. In the first part of the test (**a**; six items), only one face photograph has to be selected, which is virtually identical (i.e., a small change of size and contrast) to the target face photograph. In the second part of the test (16 items), the target face has to be found three times among the six face photographs, which vary either in terms of head orientation (**b**; eight items) or in terms of lighting conditions (**c**; eight items)

well controlled, since they do not have external features such as hair or obvious cues (facial hair, glasses, etc.), and yet the overall shape of the faces, which plays a key role in human expertise at unfamiliar-face recognition (e.g., Retter & Rossion, 2015), is well preserved.

A score above 40 out of 54 (76%) on the BFRT is considered evidence for normal individual face-matching ability; 39–40 is a borderline score; 37–38 reflects moderate impairment; and people scoring below 37 out of 54 (i.e., below 68.5%) are considered to be impaired (Benton et al., 1983).

Despite the wide usage of the test, to our knowledge, normative data for the full BFRT have not been updated since the original publication of the test.[1] Most importantly, the available normative data concern only accuracy rates. Here we provide original normative data for a computerized version of this test from a large sample—more than three hundred young Caucasian participants—for both accuracy rates and response times (RTs). We believe that reporting such data is important for at least three reasons.

A first reason is to provide complementary information about humans' behavioral ability to match pictures of unfamiliar faces, which is the main type of stimulus used in the field of face processing. Although most experimental studies have reported high performance levels at simultaneous matching of pictures of unfamiliar faces (e.g., Busigny & Rossion, 2010; Estudillo & Bindemann, 2014; Herzmann, Danthiir, Schacht, Sommer, & Wilhelm, 2008; Sergent, 1984), some studies have reported significantly lower performance (i.e., below 80%; Bruce et al., 1999; Bruce, Henderson, Newman, & Burton, 2001; Megreya & Burton, 2006, 2008), sometimes claiming that people are "poor" or "highly inaccurate" at matching images of unfamiliar faces (see the review by Young & Burton, 2018). To be fair, this claim of "poor performance" is based on a comparison with the ceiling performance obtained in the same tasks for highly familiar (i.e., famous) faces (e.g., Bruce et al., 2001), even though matching of familiar faces can be based on nonvisual information (i.e., semantic, lexical associations). Moreover, the latter studies have used either difficult tasks in which the number of targets to find in a large array of faces was undetermined and/or unknown by the participant, or same–different tasks with highly similar distractor faces, increasing the contribution of decisional responses biases (from Bruce et al., 1999). Finally, these studies have generally used pictures of individual faces that varied greatly in lighting conditions—often taken with different cameras. In this context, it seems important to provide updated normative data for such a difficult and widely used face-matching test as the BFRT, which, having been developed in a clinical context, uses a forced choice procedure, with the number of target faces being constant and known by the participant (i.e., avoiding decisional response biases). Moreover, only a subset of items on the BFRT, which can be analyzed separately, require matching faces across lighting variations.

Second, there has been relatively recent and growing interest in the interindividual variability of face-processing

[1] Since the original submission of this article, a short report providing accuracy norms for the BFRT in an Italian student population has appeared (Albonico, Malaspina, & Daini, 2017).

abilities, which is generally claimed to be much wider than was initially thought (e.g., Bowles et al., 2009; Herzmann et al., 2008; Megreya & Burton, 2006; Palermo et al., 2017; Wilmer et al., 2012; Wilmer et al., 2010). Estimates of this (high) variability have been based so far mainly on explicit memory encoding of unfamiliar faces, followed by delayed recognition (e.g., Woodhead & Baddeley, 1981), such as in the widely used Cambridge Face Memory Test (CFMT; Duchaine & Nakayama, 2006a). A "perceptual" version of this test, the Cambridge Face Perception Test (CFPT; Duchaine, Germine, & Nakayama, 2007), much less widely used, is also associated with a large interindividual variability in performance (Bowles et al., 2009; Duchaine et al., 2007; Garrido, Duchaine, & Nakayama, 2008). However, the CFPT is not based on matching different individual faces, but rather involves *similarity judgments* on morphed faces (see also Logan, Wilkinson, Wilson, Gordon, & Loffler, 2016)—that is, a task that is also likely to be highly influenced by higher-level cognitive processes. Other tests associated with high interindividual variability in unfamiliar-face-matching performance again rely on ambiguous tasks in which the number of targets to find in a large array of faces is unknown by the participant (Bruce et al., 1999; Jenkins, White, Van Montfort, & Burton, 2011; Megreya & Burton, 2006), or on same–different tasks that are particularly prone to decisional biases (e.g., the Glasgow Face Matching Test; Burton, White, & McNeill, 2010). In comparison to these tasks, the BFRT, developed in a clinical context and validated with tests in various neurological populations (Benton & Van Allen, 1968, 1972), is much simpler in its instructions, with the number of target faces being fixed and known by the participant. As for the face stimuli used in the BFRT, they are not known for being selected on the basis of physical or perceptual similarity criteria. Hence, there are reasons to believe that interindividual variability at unfamiliar-face matching—a critical component of the ability to recognize individual faces—may be estimated more accurately from the BFRT than from other tests used in laboratory settings.

Third, and most importantly, there are currently no normative data for the *time* taken to perform the BFRT, which is generally administered as a paper-and-pencil test. The lack of RT data is a critical issue: Without time constraints, a simultaneous face-matching task, even across changes of views and lighting, can probably always be performed at a high level of accuracy by human participants (see Özbek & Bindemann, 2011). This issue was noted early on by Benton himself, who mentioned that even some patients with prosopagnosia could reach normal performance on his test (Benton, 1980, 1990; Benton & Van Allen, 1972; see, e.g., De Renzi, Faglioni, Grossi, & Nichelli, 1991; McNeil & Warrington, 1991; and the review by Gainotti, 2013). For this reason, Benton always stated that low performance on the BFRT was not diagnostic of (acquired) prosopagnosia (Benton, 1980). Duchaine and Nakayama (2004, 2006b) reached the same conclusion for people diagnosed with life-long difficulties in individual face recognition—that is, so-called developmental prosopagnosia (or *prosopdysgnosia*; Rossion, in press)—showing that 73% of their group of participants could perform at normal accuracy levels on the BFRT (see also Albonico et al., 2017).

However, it has long been noticed that the RTs of patients with (acquired) prosopagnosia are unusually prolonged for the BFRT or other face-matching tasks, and thus that an analysis of RTs can reveal their disorder at individual face recognition (e.g., Bukach, Bub, Gauthier, & Tarr, 2006; Busigny & Rossion, 2010; Davidoff & Landis, 1990; Delvenne, Seron, Coyette, & Rossion, 2004; Farah, 1990; Jansari et al., 2015). For instance, patient P.S., a case of severe prosopagnosia following brain damage documented in about 30 publications (e.g., Rossion et al., 2003; for a review, see Rossion, 2014; for a recent study, see Ramon, Busigny, Gosselin, & Rossion, 2016), achieved a score of 39/54 (72.2%, "borderline score") at the BFRT as tested in 2006 (reported by Busigny & Rossion, 2010). However, P.S. took more than 30 min to perform the test, whereas age-matched participants in that study performed the test with an average accuracy rate of 84.7%, in 6–8 min. Another case of acquired prosopagnosia, N.S., also had a borderline score (40/54) on the BFRT, but took about 20 min to perform the test (Delvenne et al., 2004; see also Bukach et al., 2006; Jansari et al., 2015, for other cases with prolonged RTs; and Young, Newcombe, de Haan, Small, & Hay, 1993, for slow RTs on the BFRT in right posterior brain-damaged patients).

In a similar vein, Duchaine and Weidenfeld (2003) found that normal participants could achieve reasonably high scores on the BFRT even when a large portion of the face was masked, leaving only the eyebrows and hairlines as matching cues. However, even though these authors did not provide RTs in their study, they mentioned that participants provided with the eyebrows and hairline only were particularly slow at performing the test. More generally, it is becoming increasingly clear that the speed of processing is critical in evaluating individual face recognition: Faces are thought to be individually recognized or discriminated within one or two gaze fixations (Hsiao & Cottrell, 2008; Prieto, Van Belle, Liu, Norcia, & Rossion, 2013), successful behavioral matching of individual faces is typically performed in a few hundreds of milliseconds from stimulus onset to motor response (e.g., Jacques, d'Arripe, & Rossion, 2007; Sergent, 1984), and electrophysiological studies have indicated that the repetition of a picture of the same face is detected less than 200 ms following stimulus onset, even across size and viewpoint changes (i.e., a reduction of the N170 face-sensitive component; see, e.g., Caharel, d'Arripe, Ramon, Jacques, & Rossion, 2009; Jacques et al., 2007; see Rossion & Jacques, 2011, for a review). Hence, it is important to provide updated norms of performance on the BFRT that include the time taken to perform this unfamiliar-face-matching test.

For all these reasons, we report here the detailed results of a large sample of participants (307; 202 females, 105 males) tested with an electronic version of this test, the BFRT-*c*.

Method

Participants

A total of 307 Caucasian participants (mean age 22.62 years, range 18.25–39.75 years, but only 5% of participants above 26 years of age), with 202 females (mean age 22.66 years, range 18.25–39.75 years) and 105 males (mean age: 22.53 years, range 18.75–30.75 years) were tested on a computerized version of the BFRT between March 2013 and April 2016. They were tested before a longer, unrelated experiment recording their electroencephalogram (EEG). They were part of a larger sample of participants, from which we excluded non-Caucasian participants or participants older than the mean age +3 *SD*s. Participants reported normal or corrected-to-normal vision and did not report difficulties at face identity recognition, as was required for the selection of participants in the experiment. Participation was voluntary. Signed informed consent was given, and monetary compensation was provided for the whole study, including the BFRT evaluation.

BFRT, computerized version

The original panels of the test (long form) were scanned and saved in high resolution in .BMP format. All photographs are grayscale and depict (Caucasian) faces only, that is with the overall shape preserved but without the hair and without any clothing or other obvious cues (e.g., earring, moustache, beard, makeup), with a neutral expression (Fig. 1). The 22 panels/items (6 + 16) of the original test are presented in a computerized format, in the same fixed order as in the original version, using Matlab 7.8.0 (R2009a).[2] The task is the same as in the original test: For each item, an unfamiliar target face (full-front view) has to be found among six probe faces presented below this target face. In the first part of the test (six items, half of them males), the target face has to be found only once among six probe faces that are all full-front views. The face photograph to be selected is virtually identical to the target face photograph (Fig. 1a). These items are scored 1 or 0, for correct or incorrect matches, respectively. In the second part of the test (16 items, half males), the target face has to be found three times (i.e., three different exemplars of the same identity) among the six probe faces, which vary either in terms of head orientation (eight items, half with female faces; Fig. 1b) or in terms of lighting condition (eight items, half with female faces; Fig. 1c). These 16 items are scored 3, 2, 1, or 0, depending on the number of faces correctly selected.

There are three differences between the original test and the computerized test. First, whereas the target and the probe faces are presented in two different panels in the original test, in the computerized test the target and the six probe faces are presented on a single, dark gray background panel subtending approximately 26.5° × 20° of visual angle when viewed from 60 cm. Second, in the computerized version of the test, the target faces are displayed at a slightly larger size (133 × 200 pixels; 3.53° × 5.25° of visual angle) than the probe faces (129 × 150 pixels; 3.34° × 4° of visual angle), to further reduce matching based on low-level, image-based visual cues. Third, in the computerized test, participants are instructed to select the probe face(s) not only as accurately as possible, but also as fast as possible.

Face photographs have to be selected with the computer mouse, without any constraints regarding the order of selection for the items requiring three choices, and without any chance of correcting one's responses (i.e., a probe face that has been selected cannot be deselected). The six face photographs remain on the computer screen until the participant's (completed) response (three choices), or for a maximum of 30 s. After each item, a black screen is presented until the participant presses a key to continue.

Procedure

Participants were tested individually at a distance of approximately 60 cm from the computer screen. After having completed an electronic demographic questionnaire, they were presented with the BFRT-computerized version (BFRT-*c*).

Results

Global performance

Accuracy (total maximum score: 54) The average score on the BFRT-*c* was 44.81/54 (≈ 83%), with a standard deviation of 3.44. The scores ranged from 35/54 to 52/54 (65%–96%), with none of the participants performing at ceiling and all participants scoring well above chance level (i.e., >25/54). The frequency distribution of total scores (Fig. 2a) deviated from normality (Kolmogorov–Smirnov = .106, $df = 307$, $p < .001$; Shapiro–Wilk = .975, $df = 307$, $p < .001$). Therefore, we calculated the 5th percentile, which falls between 40/54 (74.1%) and 39/54 (72.22%). Note that only 12 participants (≈ 4%) presented with a score equal to or below two *SD*s from the mean (37.93 ≈ 38/54, or 70.2%), and none of the

[2] The test in its electronic version is available on request (via e-mail to the first author).

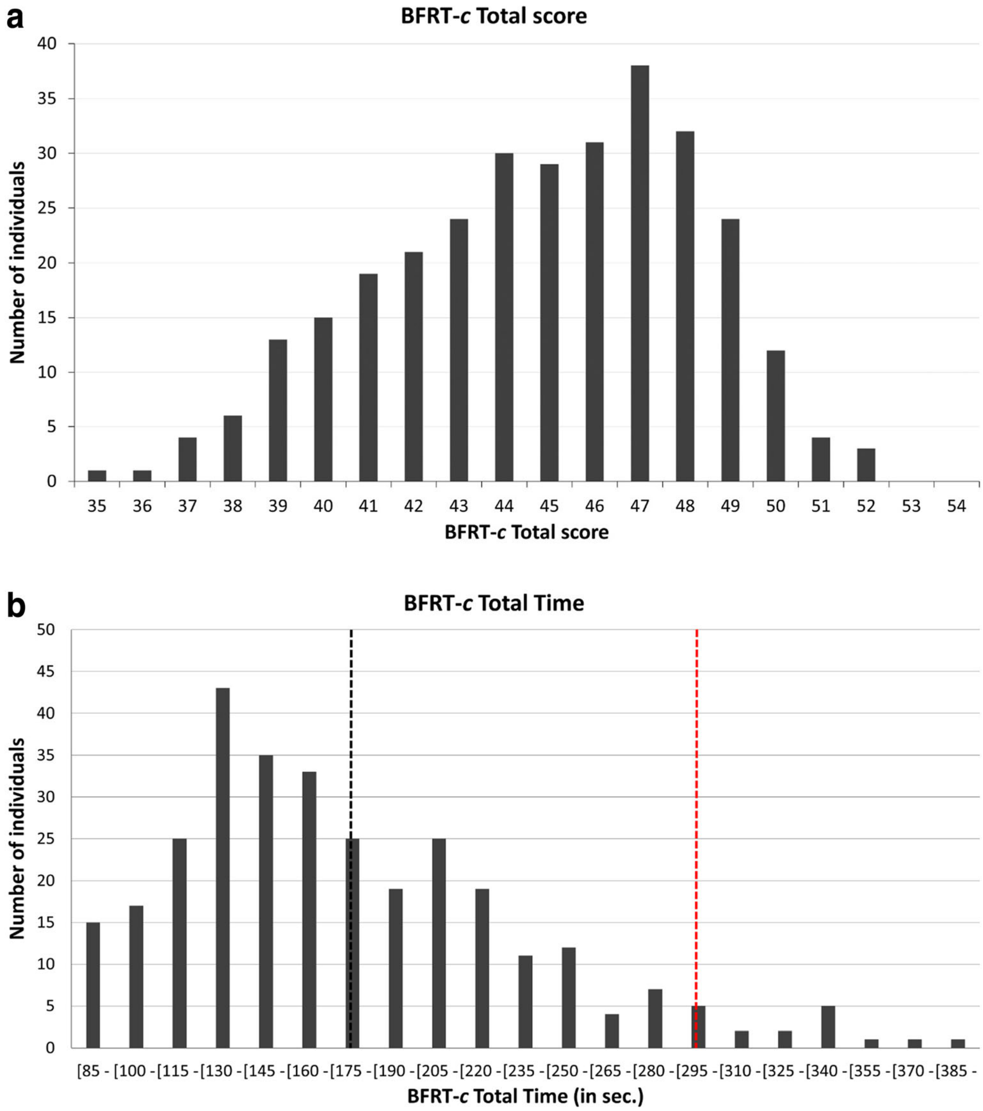

Fig. 2 Frequency distributions of BFRT-*c* total scores (**a**) and time to complete the test (**b**). In panel B, the vertical gray dashed line indicates approximately the average total time to complete the BFRT-*c*, and the vertical red dashed line indicates approximately the 95th percentile (P95)

participants presented with a score below three *SD*s (34.49/54, or 63.9%) from the mean. Note also that the relatively poor performance in terms of accuracy rates of the 12 participants could not be attributed to generally poor cognitive abilities, since these participants had all obtained at least their secondary school certificate, and most of them had already obtained a certificate for higher education.

Response times Since the first six items of the test required only one response (i.e., one face photograph to select) and viewing conditions between the target face and the response face remained the same, the RTs for these items were shorter (about 3 s, on average; see below) than for the remaining 16 items requiring three responses (i.e., three face photographs to select) and with a change of viewing conditions (head orientation or lighting) between the target face and the response faces (about 10 s, on average; see below). Hence, we did not consider the mean RT/item for the whole test, but rather considered the *total time* taken by participants to complete the test. Across individuals, this measure ranged between 1.42 min (85.09 s) and 6.54 min (392.22 s), with an average of 3.01 min (180.85 s; $SD = 59.87$ s, or 0.99 min). As is shown in Fig. 2b, the time to perform the test was not normally distributed (Kolmogorov–Smirnov = .092, $df = 307$, $p < .001$; Shapiro–Wilk = .941, $df = 307$, $p < .001$). The 95th percentile was 300.57 s.

The time to complete the test did not correlate with accuracy rates ($r = -.086$, $p = .067$). In fact, as is shown in Fig. 3, among the 15 participants who needed more than 300.57 s (95th percentile) to complete the test, only two scored below the normal range (≤38/54). Yet the participants who were slower than the 95th percentile (300.57 s) at the test were

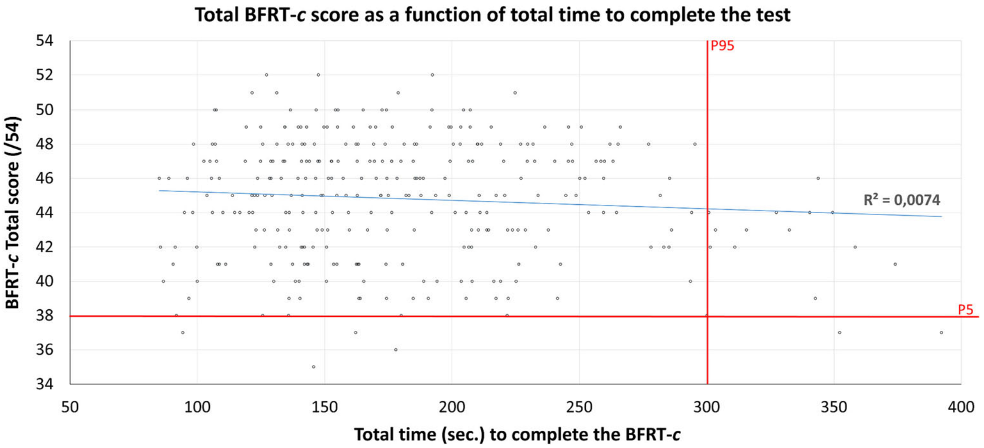

Fig. 3 Relationship between accuracy (total score) and RT (total time) for the BFRT-*c*. Among the 15 slowest participants (total time to complete the BFRT-*c* > P95), only two scored below the 5th percentile (P5)

generally *abnormally* slow—that is, up to 392 s, when considering all participants who fell in the extreme 5% for RTs (and up to 374 s, when removing the two participants who scored below the 5th percentile).

Performance on the first part of the test (six first items)

Accuracy (maximum score: 6) An overwhelming majority of the participants (274 out of 307, 89.25%) performed at ceiling (i.e., scored 6/6) for the first part of the BFRT-*c* (i.e., the first six items). Only three participants scored 4/6, the lowest score. The mean score was thus as high as 5.88/6 (98%), with a small standard deviation (0.35). Interestingly, all three participants who missed two items on the first part (4/6) scored below the normal range in total (i.e., scoring in the range 35/54–37/54), suggesting that failure on the first six items (≤4/6) of the BFRT-*c* already reveals difficulties with individual face matching. However, good performance on these items (which only required matching virtually identical face photographs) does not guarantee good performance on the remaining part of the test.

Response times The mean RT for an item on the first part of the test (irrespective of accuracy) was 2.99 s on average, with a standard deviation of 1.54 s (range: 1.32–17.40 s; distribution nonnormal in shape: Kolmogorov–Smirnov = .188, df = 307, p < .001; Shapiro–Wilk = .656, df = 307, p < .001; percentile 95 = 5.43 s). This mean was significantly correlated with the mean RT for an item on the second part of the test (r = .492, p < .001) and the total time to complete the test (r = .605, p < .001). However, as with accuracy, normal performance on the first part of the BFRT-*c* did not predict normal performance on the whole test (and thus, preserved individual face-matching ability). Indeed, as is shown in Fig. 4, among the 15 participants who took more time than average to

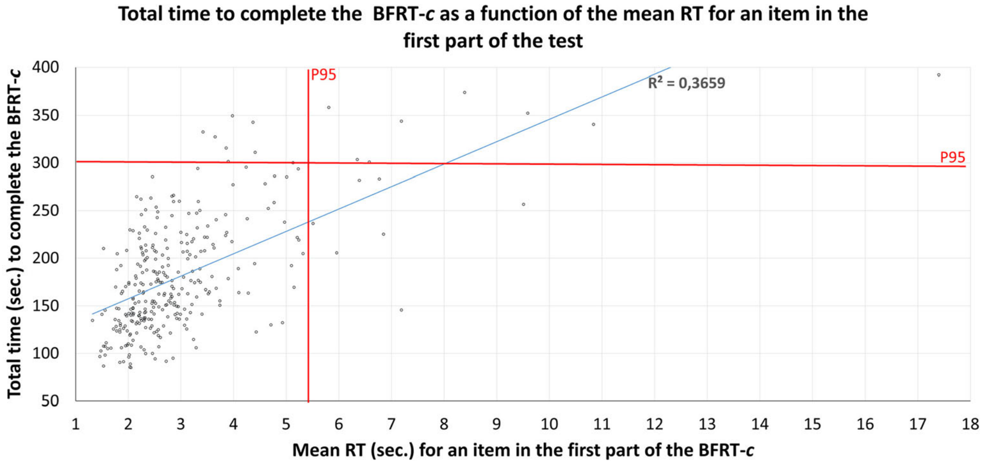

Fig. 4 Relationship between the total time to complete the BFRT-*c* and the mean RT for an item in the first part of the test. Although the positive correlation between the two measures was significant, normal performance (i.e., <P95) on the first part of the BFRT-*c* does not necessarily predict normal performance (i.e., <P95) on the whole test

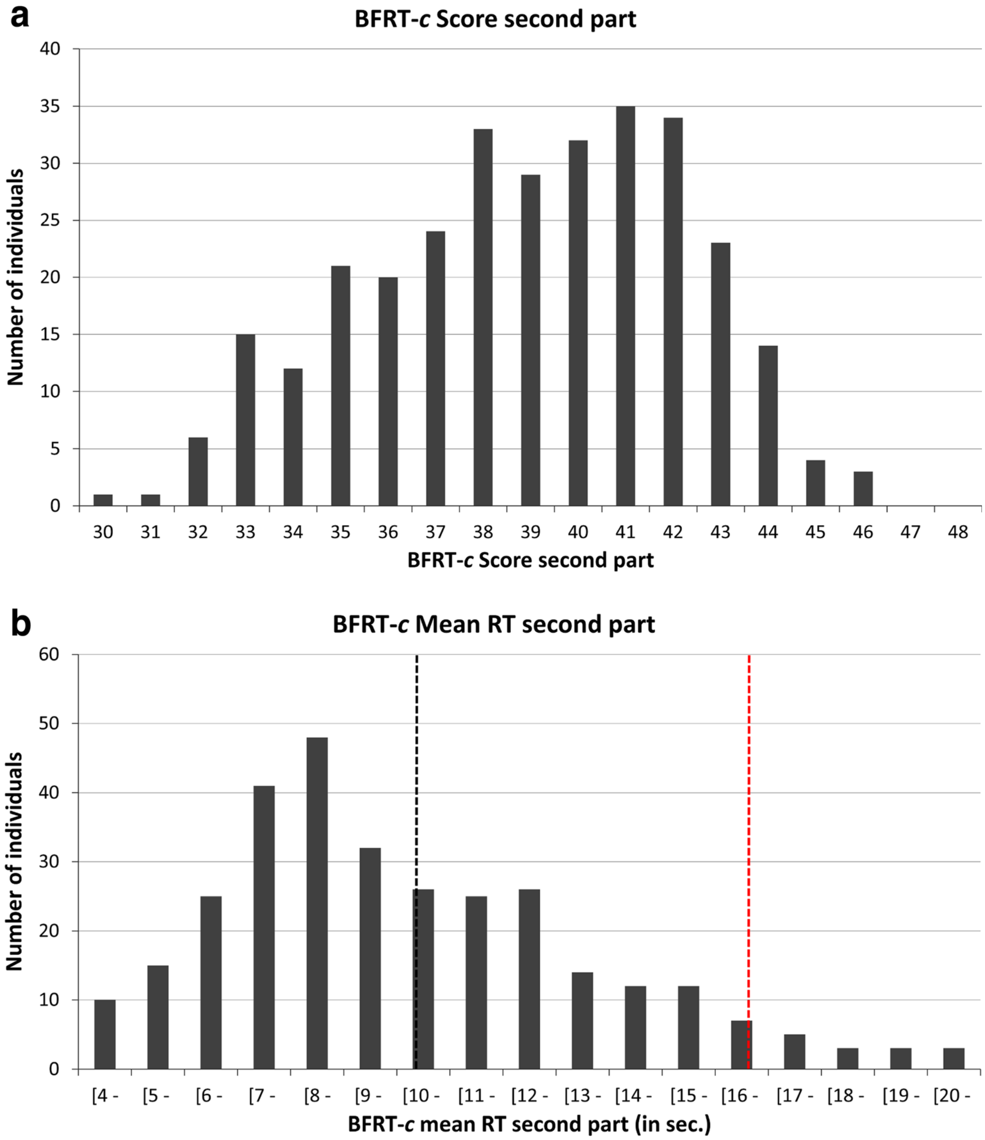

Fig. 5 Frequency distributions of scores (**a**) and average RTs (**b**) on the second part of the BFRT-*c*. In panel **b**, the vertical gray dashed line indicates approximately the average time for an item of the BFRT-*c* second part, and the vertical red dashed line indicates approximately the P95

perform the first six items, only eight were also too slow to complete the test (>300.57 s).

Performance on the second part of the test (16 remaining items)

Accuracy (maximum score: 48) On the second part of the BFRT-*c*, scores ranged from 30/48 to 46/48, with a mean of 38.92/48 (≈ 81%) and a standard deviation of 3.34. The frequency distribution of scores significantly deviated from normality (Kolmogorov–Smirnov = .101, *df* = 307, *p* < .001; Shapiro–Wilk = .975, *df* = 307, *p* < .001; see Fig. 5a). The 5th percentile for this second part of the test fell between 33 and 34/48 (eight out of the 307 participants, or 2.6%, scored <33; 23 out of 307, or 7.5%, scored <34). Note that all the participants who scored below 33/48 still performed well above chance level.

Response times The mean RT for an item on the second part of the test (irrespective of accuracy) was 10.17 s on average, with a standard deviation of 3.44 s (range: 4.55–20.35 s). The frequency distribution of mean RTs also deviated from normality (Fig. 5b; Kolmogorov–Smirnov = .096, *df* = 307, *p* < .001; Shapiro–Wilk = .952, *df* = 307, *p* < .001). The 95th percentile was 16.90 s. Mean RT did not correlate with accuracy (*r* = –.056, n.s.), and among the 15 participants who had a mean RT >16.90 s (cutoff), only one also presented a score below the normal range (<33/48) on this part of the test. Again, participants who were slower than the 95th percentile (16.90 s.) were generally *abnormally* slow—that is, up to 20.35 s.

Internal reliability

To evaluate the BFRT-*c*'s internal reliability, we measured how participants' performance on even items correlated with their performance on odd items, considering only the second (i.e., the main) part of the test. Importantly, there were as many trials involving head rotation changes as changes in lighting direction (see below) in the odd and even samples of trials. The interitem correlation was significant for accuracy rates (mean score even items = 19.66/24, *SD* = 1.73; mean score odd items = 19.27/24, *SD* = 2.19; $p < .001$; r_{SB} [Spearman–Brown] = .606). The interitem correlation was even higher for RTs (mean RT for the eight even items = 9.69 s, *SD* = 3.29 s; mean RT for the eight odd items = 10.67 s, *SD* = 3.78 s; r_{SB} = .883, $p < .001$).

Sex differences in performance

About two-thirds of the 307 participants tested on the BFRT-*c* were females (202 females and 105 males, no age difference [$t(305) = 0.404$, $p > .5$]). Considering the whole test, there was no main effect of participant's sex on accuracy rates [$F(1, 305) = 2.74$, $p > .05$], but a significant interaction with face sex was observed [$F(1, 305) = 8.95$, $p < .01$]: As is shown in Fig. 6a, female faces were recognized better by female participants (mean = 22.63/27, *SD* = 2.14) than by male participants (mean = 21.89/27; *SD* = 2.37) [$t(305) = 2.79$, $p < .01$], whereas male faces were recognized equally well by females (mean = 22.41/27, *SD* = 1.84) and males (mean = 22.48/27, *SD* = 1.92) [$t(305) = 0.29$, n.s.]. The same pattern of results was observed when the scores on the second part of the test only were considered [no significant main effect of participant's sex: $F(1, 305) = 2.59$, n.s.; no significant main effect of face's sex: $F(1, 305) = 2.65$, n.s.; significant interaction between the two factors: $F(1, 305) = 10.50$, $p < .01$].

In terms of RTs, female participants outperformed male participants irrespective of the sex of the face. This was valid when considering the total times to respond to female items ($N = 3 + 8 = 11$) versus male items ($N = 3 + 8 = 11$), as revealed by a significant main effect of participants' sex [$F(1, 305) = 14.97$, $p < .001$] that did not interact with the faces' sex [$F(1, 305) = 0.49$, n.s.; see Fig. 6b]. This was also valid when considering the mean RTs for female items ($N = 8$) versus male items ($N = 8$) in the second part of the test [main effect of participants' sex: $F(1, 305) = 15.16$, $p < .001$; interaction with faces' sex: $F(1, 305) = 0.67$, n.s.]. Note that the analyses of variance (ANOVAs) conducted on RTs also showed a main effect of face sex, with male faces being responded to faster than female faces, for both male participants [total time: 95.71 s (33.19) vs. 103.11 s (36.38), $t(104) = 4.695$, $p < .001$; mean time second part: 10.84 s (3.69) vs. 11.59 s (4.05), $t(104) = 4.00$, $p < .001$] and female participants [total time: 82.72 s (26.08) vs. 88.80 s (29.35), $t(201) = 5.58$, $p < .001$; mean time second part: 9.36 s (3.08) vs. 9.93 s (3.42), $t(201) = 4.47$, $p < .001$].

As is shown in Table 1, the very same pattern of results (with a female advantage restricted to female faces in terms of accuracy, and observed for both male and female faces in terms of RTs) was observed when the samples were equal in number (105 males vs. 105 females). Here, contrary to the norms collected on the CFMT (Bowles et al., 2009; Duchaine & Nakayama, 2006a), the advantage in time-related performance was not associated with a larger standard deviation for female participants. In fact, the standard deviations were smaller for females than for males, even for equal sample sizes, with the difference reaching or approaching significance for most measures [total time, female faces: significant difference in one out of the two comparisons[3] (Levene's test: $F = 7.74$, $p < .01$ and $F = 2.16$, n.s., respectively); total time male faces: significant difference in one out of the two comparisons, with the difference approaching significance in the other one (Levene's test: $F = 9.81$, $p < .01$, and $F = 2.97$, $p = .09$, respectively); mean RT on second part, female faces: significant difference in one of the two comparisons (Levene's test: $F = 5.17$, $p < .05$, and $F = 1.88$, n.s., respectively); mean RT on second part, male faces: significant difference in one of the two comparisons, with the difference approaching significance in the other (Levene's test: $F = 8.39$, $p < .01$, and $F = 2.75$, $p = .09$, respectively)]. Hence, regarding the time-related measures, the cutoff scores (in terms of the 95th percentile [P95]) to be used on the BFRT-*c* in a diagnostic approach should ideally be determined separately for males and females. The cutoff scores and RTs determined for each sex are presented in Table 2.

Performance on "lighting direction" versus "head rotation" items

We also compared participants' performance on items requiring matching a target face with three face photographs taken from different angles ("head rotation" [HR] items, $N = 8$) versus those requiring matching a target face with three face photographs taken under different lighting conditions ("lighting direction" [LD], $N = 8$; see the examples in Fig. 1). The results revealed substantial advantages for HR items, in terms of both accuracy [HR items: mean = 21.34/24, *SD* = 1.84; LD items: mean = 17.59/24, *SD* = 2.088; $t(306) = 31.395$, $p < .001$] and mean RTs [HR items: mean = 8.57 s, *SD* = 3.01 s; LD items: mean = 11.80 s, *SD* = 4.21 s; $t(306) = 22.178$, $p < .001$; see the distributions of scores and mean RTs in Fig. 7]. In fact, although participants' performance scores on HR and LD items were significantly correlated, in terms of both accuracy ($r_{SB} = .447$, $p < .01$) and mean RTs ($r_{SB} = .804$, $p < .01$), the

[3] 105 males vs. 105 females, Sample 1; 105 males vs. 105 females, Sample 2.

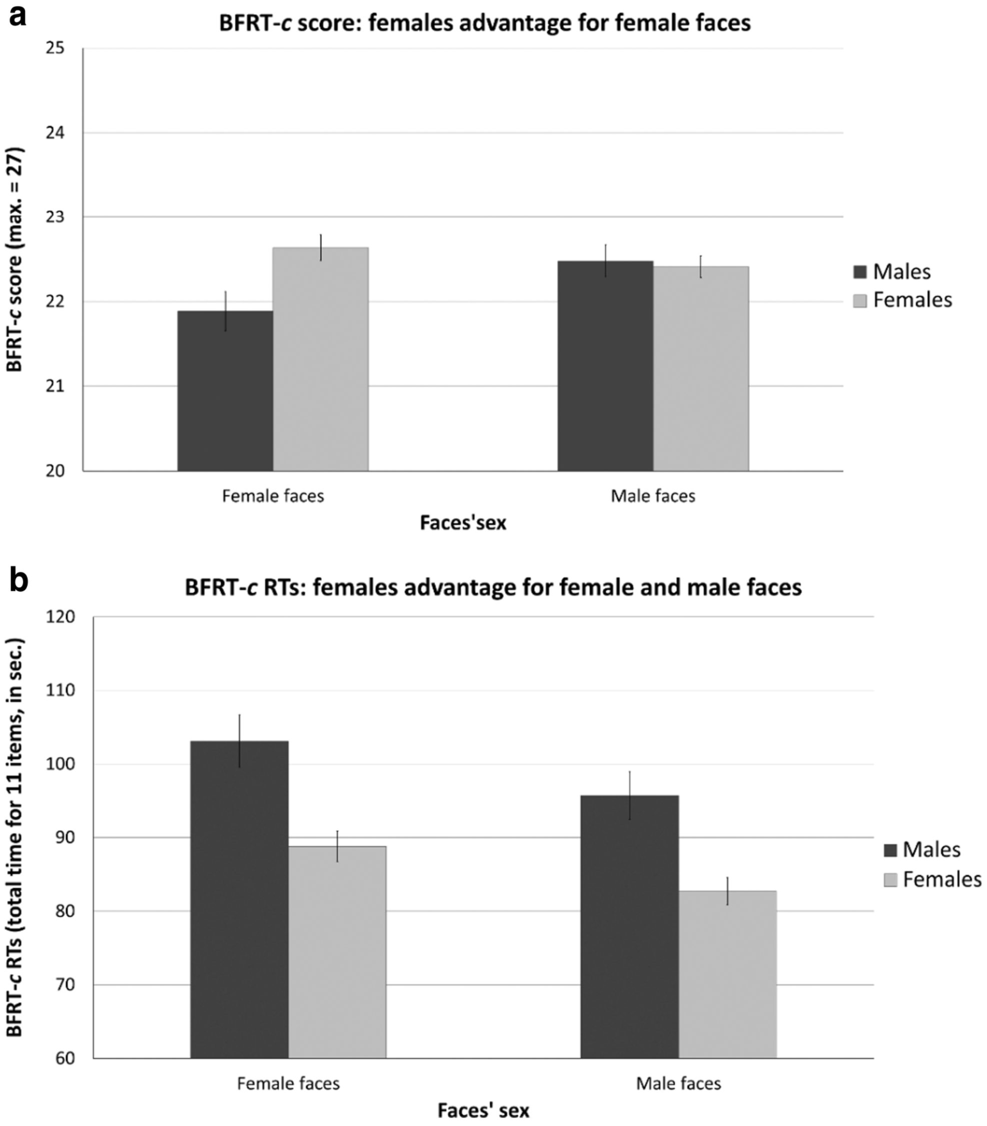

Fig. 6 Sex differences in performance on the BFRT-*c*. In terms of accuracy (**a**), females outperformed males for female faces only. In terms of RTs (**b**), females outperformed males for both female and male faces

large majority of participants had a higher score (285 out of 307 participants; see Fig. 8a) and a lower mean RT (288 out of 307; see Fig. 8b) when having to match faces across head rotation than across lighting direction differences. Only a small fraction of participants showed a slight advantage for matching faces across lighting direction changes, in terms of either accuracy (N = 5) or RTs (N = 17, with 14 participants showing an RT advantage of less than 1 s for LD items).

Item analysis

Finally, we performed an item analysis across participants. Most participants succeeded on the first six items of the BFRT-*c*. The mean score per item ranged between .96 (for Item 6: 296 participants out of 307 scored 1/1) and .99 (for Item 4: 305 participants scored 1/1). However, there were striking differences among the 16 items of the second part of the test (see Fig. 9a), with mean scores ranging from 1.83/3—that is, 61% (for Item 13: 561 correct responses out of 921 [307 participants * 3 choices])—a value still above chance level—to 2.91/3—that is, 97% (for Item 17: 895 correct responses out of 921). The distribution of participants scoring 3/3, 2/3, 1/3, or 0/3 correct on each item (see Fig. 10a) also suggests that some items are more difficult than others in this second part of the test. Indeed, whereas on some items participants almost always scored 3/3, on other items the scores were rarely 3/3.

Note that, except for Item 13 (an LD item), on which a small proportion of participants (ten out of 307) scored 0/3, null scores on the second part item were extremely rare. Unsurprisingly, the most difficult items were mostly LD items, although there were exceptions.

Table 1 The female advantage

		Females Sample 1 (FSI) (N = 105)	Females Sample 2 (FS2) (N = 105)	Males (M) (N = 105)	Test's Results	
					FSI vs. M	FS2 vs. M
Score /27	Female faces	22.59 (2.01)	22.72 (2.26)	21.88 (2.37)	Sex subj.: n.s. Sex faces: n.s. Interaction: $p < .01$	Sex subj.: n.s. Sex faces: n.s. Interaction: $p < .01$
	Male faces	22.38 (1.82)	22.47 (1.85)	22.48 (1.92)		
Score /24	Female faces	19.65 (1.93)	19.77 (2.25)	19.92 (2.32)	Sex subj.: n.s. Sex faces: n.s. Interaction: $p < .01$	Sex subj.: n.s. Sex faces: n.s. Interaction: $p < .01$
	Male faces	19.44 (1.80)	19.51 (1.82)	19.58 (1.83)		
Total RT[1]	Female faces	88.63 s (27.18)	87.72 s (31.15)	103.11 s (36.38)	Sex subj.: $p < .001$ Sex faces: $p < .001$ Interaction: n.s.	Sex subj.: $p < .01$ Sex faces: $p < .001$ Interaction: n.s.
	Male faces	81.06 s (23.86)	83.32 s (27.97)	95.71 s (33.19)		
Mean RT[a]	Female faces	9.91 s (3.25)	9.81 s (3.56)	11.59 s (4.05)	Sex subj.: $p < .001$ Sex faces: $p < .001$ Interaction: n.s.	Sex subj.: $p < .01$ Sex faces: $p < .001$ Interaction: n.s.
Mean RT Second part[b]	Male faces	9.19 s (2.84)	9.41 s (3.28)	10.84 s (3.69)		

The advantage is restricted to female faces in accuracy, but observed for both female and male faces in RTs, and observed when the 105 males were compared to an equal number of females selected with an alphabetic criterion (either the 105 first females of the alphabetic list of female participants [Female Sample 1] or the 105 last females of the alphabetic list [Female Sample 2], in each case matched in age (both *p*s > .4). [a] Calculated on 11 items (3 + 8) for each sex of faces. [b] Calculated on eight items for each sex of faces

Note that despite these discrepancies in item difficulty, the correlation between participants' performance on an individual item and participants' overall performance in the second part of the test (minus this item) was significant for each individual item, as illustrated on Fig. 11. For accuracy, correlation values ranged from .161 to .301, with no correlation value being lower than the average correlation value minus two *SD*s [0.23 − (2 × 0.04) = 0.15] or higher than the average correlation value plus two *SD*s [0.23 + (2 × 0.04) = 0.31]. Hence, we found no indication that any of the items of the BFRT-*c* should be removed from the computation of the total score on the test.

In terms of mean RT, the first six items ranged between 2.27 s (Item 4) and 3.68 s (Item 2). For the 16 items of the second part of the test, the range was wide (Fig. 9b), from 6.12 s (Item 17 again) to 14.66 s (Item 13 again). The items that were responded to the fastest were also associated with the highest accuracy scores (Figs. 9 and 10). However, for response accuracy, as is illustrated in Fig. 11, the correlations between participants' performance on an individual item and participants' overall performance in the second part of the test (without this item) was again significant for each individual item. The correlation values ranged from .499 to .813. Note that the lowest correlation value (Item 17) was particularly

Table 2 Cutoff values for male and female participants on the main measures of the BFRT-*c*, as determined by performance of the participants in our sample (307: 105 males, 202 females)

		MALE AND FEMALE SUBJECTS			
Global perf. (/54)	Cutoff (P5)	39-40			
	% in our sample	<39		<40	
		≈4% (12/307)		≈8% (25/307)	
Perf. second part (/48)	Cutoff (P5)	33-34			
	% in our sample	<33		<34	
		≈2.5% (8/307)		≈7.5% (23/307)	
		MALE SUBJECTS		FEMALE SUBJECTS	
Total time (s)	Cutoff (P95)	342.04		265.90	
	% in our sample	>342.04	>265.90	>342.04	>265.90
		≈5% (5/105)	≈16% (17/105)	≈1% (2/202)	≈5% (10/202)
Mean time second part (s)	Cutoff (P95)	18.38		15.55	
	% in our sample	>18.38	>15.55	>18.38	>15.55
		≈5% (5/105)	≈15% (16/105)	≈1.5% (3/202)	≈5% (10/202)

For accuracy measures, the percentage of participants in our sample performing below the cutoff score is indicated. For time-related measures, the percentages of participants in our sample performing above the cutoff and above the cutoff determined for the other group are indicated

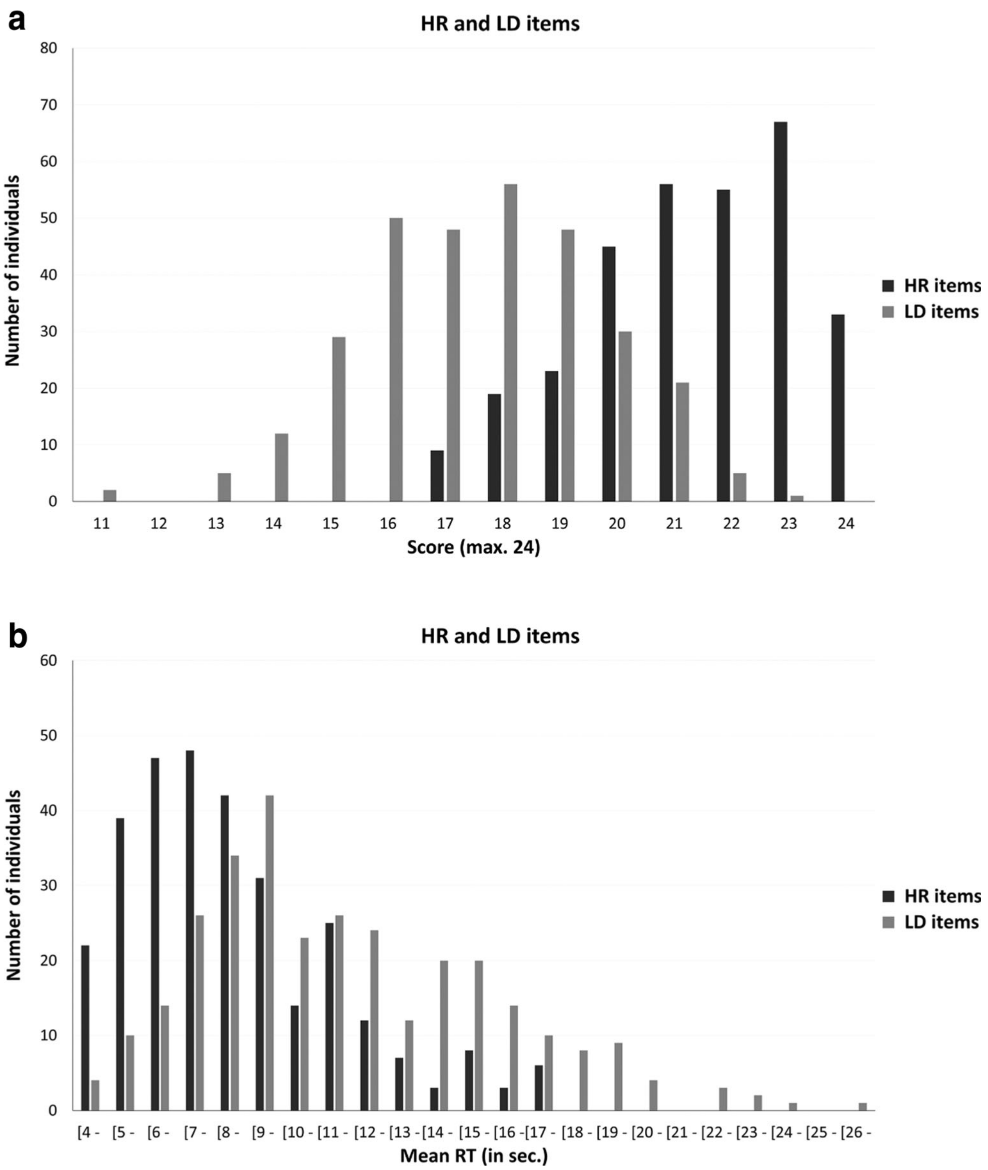

Fig. 7 Distributions of scores (**a**) and mean RTs (**b**) for head rotation (HR) items and lighting direction (LD) items reveal a clear advantage for HR items in both measures

weak—that is, weaker than the average correlation value minus two *SD*s [0.66 – (2 × 0.07) = 0.52], whereas the highest correlation value (Item 11) was higher than the average correlation value plus two *SD*s [0.66 + (2 × 0.07) = 0.8]. Note also that Item 17 also had the lowest correlation value (although within the normal range of correlation values) in terms of accuracy.

Discussion

Performance and variability

We report average performance on a computerized version of the BFRT of 44.81 out of 54 (83%), a score that is remarkably close to that of the healthy participants in the original study reporting the test (45.3 or 83.9%; Benton & Van Allen, 1968), but lower than the score obtained in a recent report of Italian normative data (47.12; Albonico et al., 2017). Standard deviation (3.44; i.e., 6.37%) is smaller in our sample than reported originally (4.5; i.e., 8.33%) and this could be due to the more homogeneous population tested here in terms of age range as compared to the wide age range covered in the original study (16–65 years of age; Benton & Van Allen, 1968).

An obvious aspect of our observations on the BFRT-*c* is that the first six items, requiring finding only one face target among six faces, without any change of viewpoint or lighting (i.e., virtually identical images), should be used only as *practice trials*: The overwhelming majority of participants perform at ceiling and very rapidly on these items. As a matter of fact,

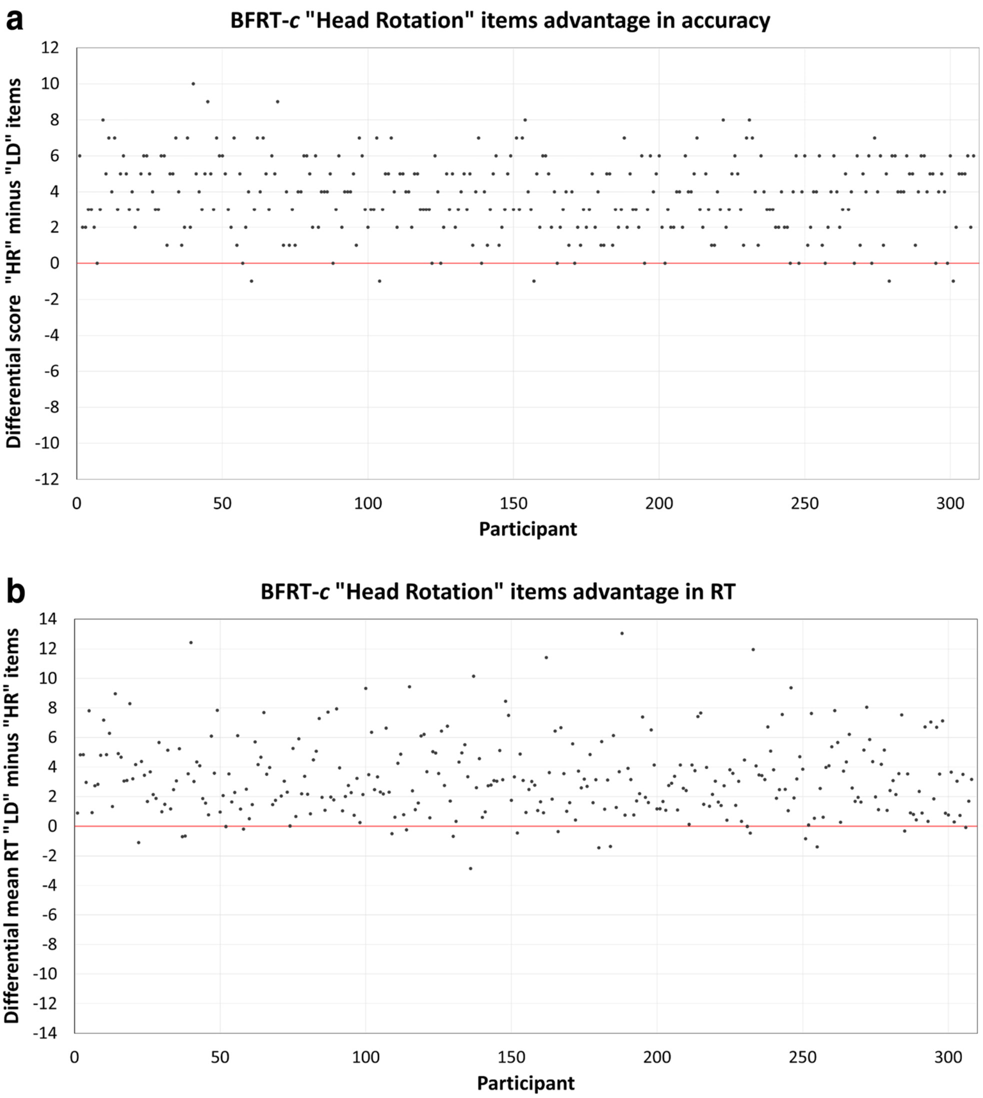

Fig. 8 The advantage for HR items, revealed by most participants scoring better (**a**) and answering faster (**b**) for HR than for LD items

the prosopagnosic patient P.S. (Rossion et al., 2003), tested recently (2015) with this version of the BFRT-*c*, obtained a score of 5/6 on these items, for a score of 35/54 in total (Liu-Shuang, Torfs, & Rossion, 2016). Even when removing these trials, performance in our sample of participants remains relatively high (38.92/48; i.e., 81%), with a low standard deviation (3.34; i.e., 7%). In future studies evaluating one's abilities at individual face matching, we thus recommend using only the second part of the BFRT-*c*.

Considering accuracy rates only, these observations would not support the view that typical human adults—here recruited simply on the basis of their self-reported absence of face identity recognition difficulties—vary substantially in their ability to individualize faces. This claim is generally based on accuracy rates in tests involving recognition memory for previously unfamiliar faces, such as in the CFMT (e.g., Duchaine & Nakayama, 2006a; see also Woodhead & Baddeley, 1981). However, several studies have also reported relatively large individual differences on unfamiliar-face-matching tasks (Bruce et al., 1999; Bruce et al., 2001; Burton et al., 2010; Duchaine et al., 2007; Megreya & Burton, 2006).

As we mentioned in the introduction, these studies relied on more complex and sometimes ambiguous tasks, involving substantial additional cognitive factors and strategies than in an individual face-matching test with simultaneous picture presentation such as the BFRT-*c*, in which the limited number of faces to match is constant and known by the participants. For instance, Duchaine et al. (2007) measured performance on a task (CFPT) requiring the ordering of morphed face series on the basis of similarity, whereas Burton et al. (2010) used a response bias-prone same/different task with very different pictures of identical faces. Other studies require finding a target face in (target-present and target-absent) arrays of ten faces similar in appearance (Bruce et al., 1999; Megreya & Burton, 2006; see also Bruce et al., 2001). As a matter

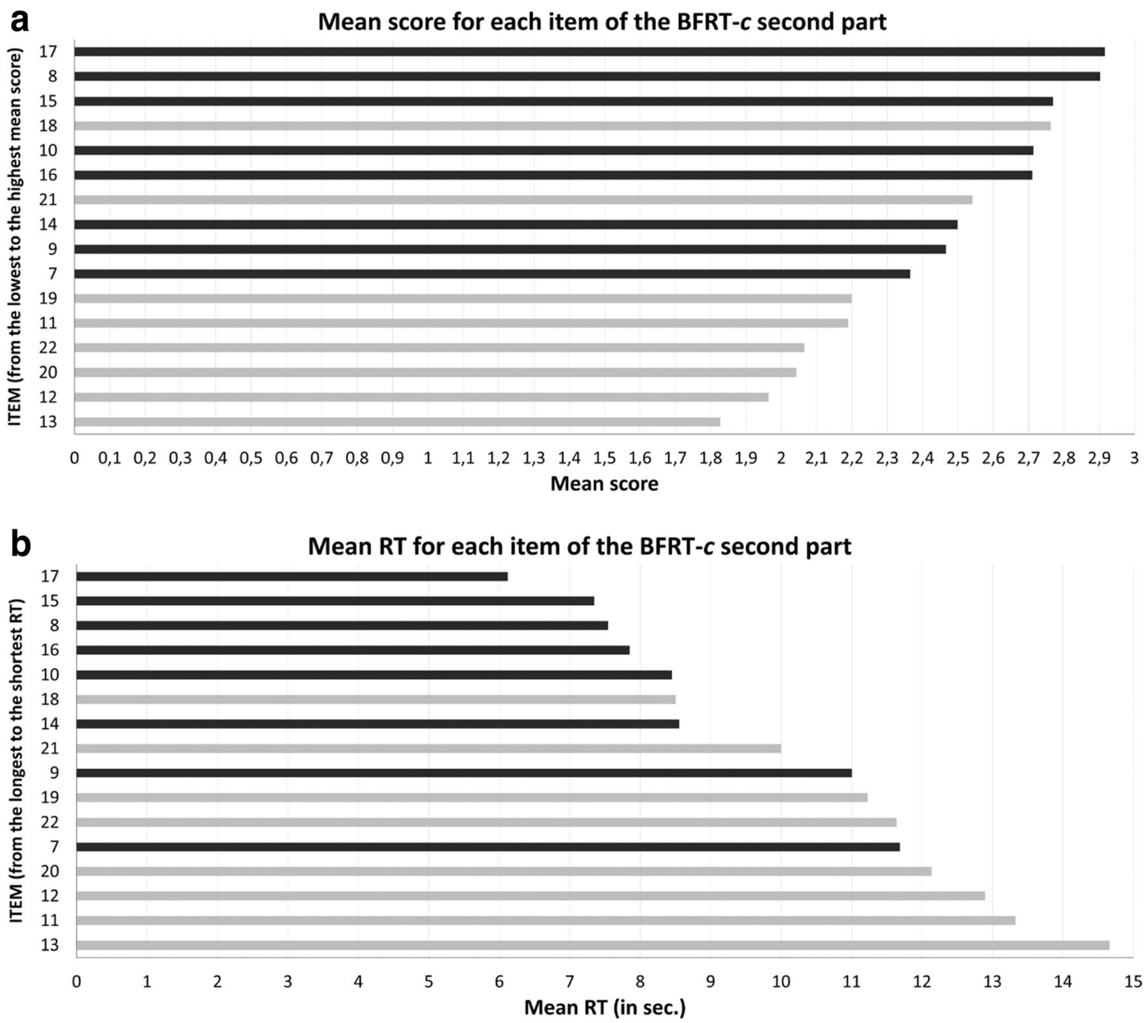

Fig. 9 Mean scores (**a**; max score = 3) and mean RTs (**b**) for each item on the second part of the BFRT-*c* (average of 307 participants). Items are ordered from the lowest mean score (Item 13: 1.83/3) to the highest mean score (Item 17: 2.91/3) and from the longest mean RT (Item 13: 14.66 s) to the shortest mean RT (Item 17: 6.12 s). LD items are represented in light gray, and HR items are represented in dark gray

of fact, RTs are usually not reported in these face-matching tasks, and are probably extremely prolonged and highly variable across participants (e.g., Özbek & Bindemann, 2011 ; see below).

Note that we do not claim here that these tasks are less appropriate than the BFRT-*c* to evaluate people's ability to match unfamiliar faces. However, we believe that interindividual variability in accuracy rates in individual face matching as measured with behavioral tasks may be overestimated, due to a variability in terms of general cognitive processes involved in many tests. Moreover, in a clinical context, an unfamiliar-face-matching test should not involve too many cognitive processes, which may be differentially affected in a patient population, and the test should be relatively fast to administer. Overall, performance in terms of percent accuracy for the BFRT-*c* is far from ceiling, yet is well above chance level, with very little variance, which makes this measure useful in a clinical context.

The importance of RT measures

It has sometimes been claimed that the traditional neuropsychological BFRT is invalid because of a lack of sensitivity to individual face-processing difficulties: Individuals presenting prosopdysgnosia would be frequently falsely diagnosed as normal (Bowles et al., 2009; Duchaine & Nakayama, 2004, 2006a) and typical participants could achieve reasonable scores even when most of the internal part of the face is masked, preventing normal processing of the face stimuli (Duchaine & Weidenfeld, 2003). In fact, as mentioned in the introduction, Benton himself noted that some patients with prosopagnosia could achieve relatively normal or borderline accuracy scores on this test (Benton, 1980, 1990; Benton & Van Allen, 1972; see, e.g., Busigny & Rossion, 2010; Delvenne et al., 2004; De Renzi et al., 1991; Gainotti, 2013; McNeil & Warrington, 1991). However, these observations do not invalidate the BFRT that, when tested in its long form, is

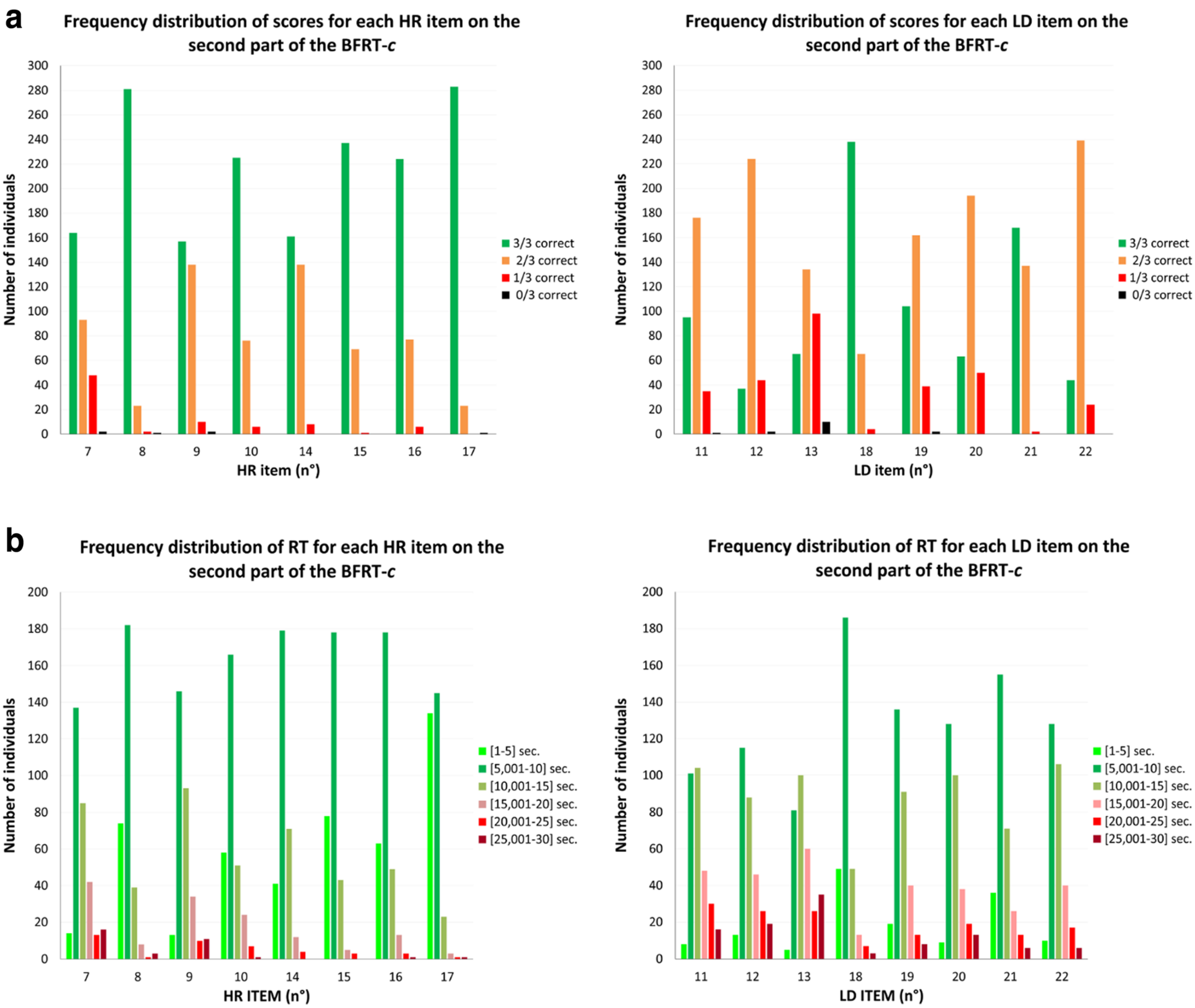

Fig. 10 (**a**) Numbers of individuals scoring 3/3, 2/3, 1/3, and 0/3 on each item of the second part of the BFRT-*c*. (**a**) Numbers of individuals responding within 5, 10, 15, 20, 25, and 30 s on each item (head rotation items [HR]: left panels; lighting direction [LD] items: right panels)

failed by many brain damaged patients with prosopagnosia (e.g., Busigny, Graf, Mayer, Rossion, 2010; Delvenne et al., 2004; Eimer & McCarthy, 1999; Levine & Calvanio, 1989; Rossion et al., 2003). Groups of patients with posterior right hemisphere brain-damage have also generally shown lower scores for the BFRT than patients with left hemisphere posterior brain damage (Benton & Van Allen, 1968, 1972; Trahan, 1997; Tranel, Vianna, Manzel, Damasio, & Grabowski, 2009; Young et al., 1993), which is consistent with the well-known right hemispheric dominance for face perception and individual face recognition (e.g., Gilbert & Bakan, 1973; Hecaen & Angelergues, 1962; Jonas et al., 2016; Sergent, Ohta, & MacDonald, 1992). Moreover, the BFRT is associated with a large face inversion effect in the normal population (i.e., about 20%) but a lack of inversion effect in prosopagnosia (Busigny & Rossion, 2010; McNeil & Warrington, 1991), indicating that performance on this test is not based merely on simple image matching, but also captures human expertise at individualizing faces.

Most importantly however, criticisms of the BFRT essentially reflect the need to go beyond the mere accuracy score—that is, to add RT measures, as in the present validation of the computerized version of the test. Here, we found that participants took about 10 s by item on average for the second part of the test, which is less than 3 min for 16 items with three faces to match by item. Interestingly, variability in RTs (i.e., *SD* of 33.8% of the mean) was much higher than for percent accuracy, with some of the participants being particularly slow. This suggests that a fair proportion of individuals can indeed achieve a normal score on the BFRT(-*c*), at the expense of abnormally prolonged RTs, potentially revealing difficulties in individual face matching. Moreover, the prosopagnosic patient PS, who scored 30/48 in the BFRT-*c* (second part), the lowest performance of all the sample of participants of the

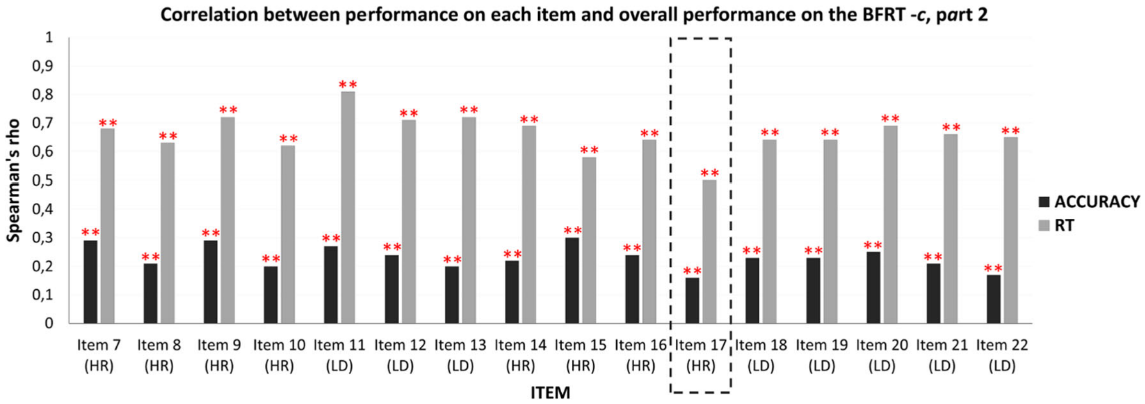

Fig. 11 For each item of the second part of the BFRT-*c*, Spearman's rho values for the correlations between the 307 participants' scores on this item (0/3, 1/3, 2/3, or 3/3) and participants' overall performance on the second part of the BFRT-*c* minus this item (max 45: i.e., 48 minus 3), as well as for the correlations between the 307 participants' RTs (continuous variable) on this item and their mean RTs on the second part of the BFRT-*c* without this item. **All Spearman's rho values are significant ($p < .01$). Item 17 (dashed rectangle) has the lowest correlation value for accuracy (although still within the normal range of correlation values), and also is the only item for which the Spearman's rho value for RT is lower than the average Spearman's rho value for RT minus 2 *SD*s

present study but one, was also extremely slow, taking 46.55 s on average by item to perform this second part of the BFRT-*c*. This 4.5-fold increase relative to control participants cannot simply be accounted for by the age difference between PS and the participants of the present study (see Liu-Shuang et al., 2016, Table 1) or by a general slowing down on perceptual-matching tasks for this patient (see, e.g., Busigny et al., 2010) and thus undoubtedly reflects her prosopagnosic impairment.

In summary, reporting RTs for the BFRT-*c* appears critical to evaluate performance. This is also supported by the fact that, in line with accuracy rates, RT measures supported an advantage of female participants, and of head rotation over lighting changes, in line with accuracy and the literature on face recognition (see below). Two caveats should be mentioned, however. First, since the participant has to find three targets for each item, it would be virtually impossible to report RTs for correct trials only, unless RTs could be normalized in some way for the number of items correctly recognized in each panel. Second, since the test is not performed under strict time constraints, one can never exclude that prolonged RTs are due to excessive care and control rather than a genuine impairment in individual face matching. In this context, the two variables, percent accuracy and RTs, should always be taken into account for a full evaluation.

As a matter of fact, patients with prosopagnosia with extremely prolonged RTs on the classical BFRT do not appear to take much advantage of this prolonged inspection to achieve very high scores: if they are not impaired in terms of accuracy rates, they nevertheless also generally perform in the lower range. This issue is particularly well illustrated in a study that Young et al. (1993) conducted on patients with posterior brain damage. Their two patients with right posterior brain lesions had borderline scores on the BFRT (40 and 38/54), but at the expense of extremely prolonged RTs (65.3 and 50.4 s/item, respectively; for 23.4 s/item in their control group). Only one patient (out of five) with left posterior brain damage was also significantly slowed down for the BFRT (40 s/item). However, this patient's accuracy was very high (48/54), in fact significantly above the control mean, prompting the authors to suggest that his increased time was due to his taking extra care with this task, rather than impaired performance per se. Hence, consideration of both accuracy rates and RTs appears necessary to evaluate performance on this test.

In general terms, this latter example serves as a useful reminder that a RT measure for this test, or any face-matching test, does not reflect the speed at which individual faces are *perceptually matched*, but the speed at which a cascade of perceptual, attentional, decisional and motor processes are accomplished. More specifically, the test requires a visual analysis of the faces, selective and spatial attention, visual search, comparison of visual representations, discrimination, selection of items and motor responses. The accuracy rates variable obtained for the BFRT also depends on the integrity of these processes, which contaminate any other explicit face-matching/recognition task, such as the CFMT for instance. For this reason, when using behavioral measures at least, isolating one's ability to recognize individual faces can only be accomplished by combining performance across a variety of tests with different task requirements and stimuli, or comparison of the performance to control stimuli such as inverted faces, arguably the most well-matched control stimuli for upright faces. In the future, purer and more stable measures of individual face *perception* in neuropsychology may require the development of alternative tests measuring this process under time constraints and without explicit behavioral responses, such as with fast periodic visual stimulation in electroencephalography (Liu-Shuang et al., 2016; Xu, Liu-Shuang, Rossion, & Tanaka, 2017).

Sex differences

We found that females outperformed males on the BFRT-*c*, a finding that is generally in agreement with the literature. Several studies showed that females perform better than males on face-processing tasks (e.g., Bowles et al., 2009; Heisz, Pottruff, & Shore, 2013; McBain, Norton, & Chen, 2009; Megreya, Bindemann, & Havard, 2011; Rehnman & Herlitz, 2007; Sommer, Hildebrandt, Kunina-Habenicht, Schacht, & Wilhelm, 2013), although this female advantage was not found in all studies, in particular in the original report of the BFRT (Benton & Van Allen, 1968; see also Albonico et al., 2017). This advantage is also observed sometimes only in specific conditions (see Weirich, Hoffmann, Meißner, Heinz, & Bengner, 2011), and could be specific for female faces (the so-called "female own-sex bias"; see, e.g., Lewin & Herlitz, 2002; Lovén, Herlitz, & Rehnman, 2011; McKelvie, Standing, St Jean, & Law, 1993; Megreya et al., 2011) as observed here for accuracy rates. Whether this higher female performance, in terms of both accuracy rates and (irrespective of face sex) correct RTs, reflects a higher level of expertise specifically at individual face matching or recognizing different types of materials (not only faces) in episodic memory (see, e.g., Herlitz, Nilsson, & Bäckman, 1997; Herlitz & Rehnman, 2008; Lewin, Wolgers, & Herlitz, 2001), or yet other factors involved in this explicit face-matching task (i.e., attention, decision making, association of the faces with semantic information, etc.), remains unknown and could be clarified in the future by using implicit measures. In any case, our data suggest that regarding time-related measures, cutoff scores (P95) to use with the BFRT-*c* in a diagnostic approach should be determined separately for males and females.

Reliability and item variability

In our study, the split-half reliability of the BFRT-*c* in accuracy rates was moderate, as in previous reports (Albonico et al., 2017; Christensen, Riley, Heffernan, Love, & Santa Maria, 2002). However, we believe that this was merely due to the small number of items (eight in each subpart of the test, for three responses in each item), and to the high variance in difficulty of the different items of the test (see Fig. 7). Thus, it seems that the reliability of the BFRT-*c* would be estimated better by a measure of test–retest variability of accuracy rates and RTs on the exact same items—for instance, at several weeks' or months' interval. Another factor that may come into play is that, contrary to other tests such as the CFMT or the CFPT (see Bowles et al., 2009), variance on the BFRT-*c* is low in the normal population, so reliability based on between-participants variability may be lower than for other tests. What is important to keep in mind here is that the reliability of a face-processing test is determined by many factors other than the process of interest. For example, if performance on a given test depends heavily on general memory or decision-making processes, the test might be highly reliable for reasons that have nothing to do with specifically how faces are perceived. In this context, the high split-half reliability value observed for RTs on the BFRT-*c* is interesting, but it should also be taken with caution: It may be partly due to general factors such as simple processing speed (Schretlen, Pearlson, Anthony, & Yates, 2001).

The difficulty of unfamiliar face matching across variable lighting conditions

In the present study, we also found that matching faces across lighting changes was far more difficult than across head orientation changes. This new observation on the BFRT-*c* is in line with previous studies that have reported difficulties with unfamiliar-face-matching tasks with pictures under the same head orientation but with substantial lighting changes (Bruce et al., 1999; Bruce et al., 2001; Jenkins et al., 2011; Megreya & Burton, 2006, 2008; Menon, White, & Kemp, 2015). Indeed, illumination parameters can change the overall magnitude of light intensity reflected back from an object, as well as the pattern of shading and shadows visible in an image (Tarr & Bülthoff, 1998). Both shading and shadows provide important cues about the three-dimensional shape of a face, and varying the direction of illumination can thus result in larger image differences than varying the head orientation of a face (O'Toole, Jiang, Roark, & Abdi, 2006; Tarr & Bülthoff, 1998). In fact, differences in lighting direction result in larger image differences than varying the identity (Adini, Moses, & Ullman, 1995). Hence, the faces of two different people shown under the same conditions of light and view look more similar in many ways—the same features are visible and receive direct illumination—and are more similar according to objective measures (Adini, Moses, & Ullman, 1995) than two pictures of a single person taken under different conditions (Hill & Bruce, 1996). Thus, this difference in lighting direction makes the images of the same face particularly difficult to match, but also prevent discriminating similar distractors, because differences on the same idiosyncratic features cannot be compared across changes of lighting direction.

An important and unresolved issue is whether these changes of lighting are particularly elevated in the BFRT(-*c*), in which pictures often miss important features, for instance of half of the face, or may not reflect the range of variations encountered in real life and that our visual system has to cope with. In any case, we are not aware of another test that has these two kinds of manipulations separated (see Fig. 1). Our data strongly suggest also reporting separate scores on the BFRT-*c* for the two types of trials. Interestingly, the prosopagnosic patient PS also showed an advantage for HR items like normal participants, but this advantage was quite small as compared to the advantage observed in healthy participants (HR items score: 16/24, mean RT: 45.7 s; LD items score: 14/24, mean RT: 47.4 s; Liu-Shuang et al., 2016). Hence, her impairment was relatively more marked for HR than LD trials overall.

Conclusions

In conclusion, we provide a normative dataset on a computerized version of the BFRT, the BFRT-*c*, which can be used to assess individuals' ability at face matching. The BFRT is not associated with a large variability across individuals in accuracy rates and yet performance is well above chance and below ceiling, making it a good clinical test to use in complement with other measures of individual face recognition. However, our data also indicate that the time taken to perform the test—which is much more variable across individuals—must be considered in the evaluation. Our normative dataset also supports previous findings that females outperform male participants at individual face-matching tasks, and that performance is much more affected by lighting changes than head rotations. Overall, this normative dataset supports the BFRT-*c* as a key component in a battery of tests identifying clinical impairments in individual face recognition, such as observed in prosopagnosia.

Author note This work was supported by the Belgian National Research Fund (FNRS) and an ERC grant facessvep 284025. We thank the members of the Face Categorization Lab for help in data collection, Talia Retter for editing the manuscript, Pernille Hemmer and three anonymous reviewers for helpful comments on a previous version of the manuscript.

References

Adini, Y., Moses, Y., & Ullman, S. (1995). *Face recognition: The problem of compensating for changes in illumination direction* (Report No. CS93-21). Rehovot, Israel: Weizmann Institute of Science.

Albonico, A., Malaspina, M., & Daini, R. (2017). Italian normative data and validation of two neuropsychological tests of face recognition: Benton Facial Recognition Test and Cambridge Face Memory Test. *Neurological Sciences*, *38*, 1637–1643.

Benton, A. L. (1980). The neuropsychology of face recognition. *American Psychologist*, *35*, 176–186.

Benton, A. L. (1990). Facial recognition. *Cortex*, *26*, 491–499.

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). New York, NY: Oxford University Press.

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. L., & Van Allen, M. W. (1972). Prosopagnosia and facial discrimination. *Journal of Neurological Sciences*, *15*, 167–172.

Bodamer, J. (1947). Die Prosopagnosie. *Archiv fur Psychiatrie und Zeitschrift fur Neurologie*, *179*, 6–54.

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

Bruce, V., Henderson, Z., Greenwood, K., Hancock, P. J. B., Burton, A. M., & Miller, P. (1999). Verification of face identities from images captured on video. *Journal of Experimental Psychology: Applied*, *5*, 339–360.

Bruce, V., Henderson, Z., Newman, C., & Burton, A. M. (2001). Matching identities of familiar and unfamiliar faces caught on CCTV images. *Journal of Experimental Psychology: Applied*, *7*, 207–218.

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.

Burton, A. M., White, D., & McNeill, A. (2010). The Glasgow Face Matching Test. *Behavior Research Methods*, *42*, 286–291.

Busigny, T., Graf, M., Mayer, E., Rossion, B. (2010). Acquired prosopagnosia as a face-specific disorder: Ruling out the general visual similarity account. *Neuropsychologia*, *48*, 2051–2067.

Busigny, T., Mayer, E., & Rossion, B. (2013). Prosopagnosia. In O. Godefroy (Ed.), *The behavioral and cognitive neurology of stroke* (2nd ed., pp. 315–334). Cambridge, UK: Cambridge University Press.

Busigny, T., & Rossion, B. (2010). Acquired prosopagnosia abolishes the face inversion effect. *Cortex*, *46*, 965–981.

Caharel, S., *d'*Arripe, O., Ramon, M., Jacques, C., & Rossion, B. (2009). Early adaptation to unfamiliar faces across viewpoint changes in the right hemisphere: Evidence from the N170 ERP component. *Neuropsychologia*, *47*, 639–643.

Christensen, K. J., Riley, B. E., Heffernan, K. A., Love, S. B., & Santa Maria, M. E. M. (2002). Facial Recognition Test in the elderly: Norms, reliability and premorbid estimation. *Clinical Neuropsychology*, *16*, 51–56. https://doi.org/10.1076/clin.16.1.51.8332

Davidoff, J., & Landis, T. (1990). Recognition of unfamiliar faces in prosopagnosia. *Neuropsychologia*, *28*, 1143–1161.

De Renzi, E., Faglioni, P., Grossi, D., & Nichelli, P. (1991). Apperceptive and associative forms of prosopagnosia. *Cortex*, *27*, 213–221.

De Renzi, E., Faglioni, P., & Spinnler, H. (1968). The performance of patients with unilateral brain damage on face recognition tasks. *Cortex*, *4*, 17–34. https://doi.org/10.1016/S0010-9452(68)80010-3

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.

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

Duchaine, B., & Nakayama, K. (2004). Developmental prosopagnosia and the Benton Facial Recognition Test. *Neurology*, *62*, 1219–1220.

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

Duchaine, B. C., & Nakayama, K. (2006b). Developmental prosopagnosia: A window to content-specific face processing. *Current Opinion in Neurobiology*, *16*, 166–173.

Duchaine, B. C., & Weidenfeld, A. (2003). An evaluation of two commonly used tests of unfamiliar face recognition. *Neuropsychologia*, *41*, 713–720.

Eimer, M., & McCarthy, R. A. (1999). Prosopagnosia and structural encoding of faces: Evidence from event-related potentials. *NeuroReport*, *10*, 255–259.

Estudillo, A. J., & Bindemann, M. (2014). Generalization across view in face memory and face matching. *i-Perception*, *5*, 589–601.

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

Gainotti, G. (2013). Is the right anterior temporal variant of prosopagnosia a form of "associative prosopagnosia" or a form of "multimodal person recognition disorder"? *Neuropsychology Review*, *23*, 99–110.

Garrido, L., Duchaine, B. C., & Nakayama, K. (2008). Face detection in normal and prosopagnosic individuals. *Journal of Neuropsychology*, *2*, 119–140.

Gilbert, C., & Bakan, P. (1973). Visual asymmetry in perception of faces. *Neuropsychologia*, *11*, 355–362.

Hecaen, H., & Angelergues, R. (1962). Agnosia for faces (prosopagnosia). *Archives of Neurology*, *7*, 92–100.

Heisz, J. J., Pottruff, M. M., & Shore, D. I. (2013). Females scan more than males: A potential mechanism for sex differences in recognition memory. *Psychological Science*, *24*, 1157–1163.

Herlitz, A., Nilsson, L.-G., & Bäckman, L. (1997). Gender differences in episodic memory. *Memory & Cognition*, *25*, 801–811.

Herlitz, A., & Rehnman, J. (2008). Sex differences in episodic memory. *Current Directions in Psychological Science*, *17*, 52–56.

Herzmann, G., Danthiir, V., Schacht, A., Sommer, W., & Wilhelm, O. (2008). Toward a comprehensive test battery for face processing: Assessment of the tasks. *Behavior Research Methods*, *40*, 840–857.

Hill, H., & Bruce, V. (1996). Effects of lighting on the perception of facial surfaces. *Journal of Experimental Psychology: Human Perception and Performance*, *22*, 986–1004.

Hsiao, J. H., & Cottrell, G. (2008). Two fixations suffice in face recognition. *Psychological Science*, *19*, 998–1006.

Jacques, C., d'Arripe, O., & Rossion, B. (2007). The time course of the inversion effect during individual face discrimination. *Journal of Vision*, *7*(8), 3:1–9. https://doi.org/10.1167/7.8.3

Jansari, A., Miller, S., Pearce, L., Cobb, S., Sagiv, N., Williams, A. L., … 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

Jenkins, R., White, D., Van Montfort, X., & Burton, A. M. (2011). Variability in photos of the same face. *Cognition*, *121*, 313–323.

Jonas, J., Jacques, C., Liu-Shuang, J., Brissart, H., Colnat-Coulbois, S., Maillard, L., & Rossion, B. (2016). A face-selective ventral occipito-temporal map of the human brain with intracerebral potentials. *Proceedings of the National Academy of Sciences*, *113*, E4088–E4097.

Levine, D. N., & Calvanio, R. (1989). Prosopagnosia: A defect in visual configural processing. *Brain and Cognition*, *10*, 149–170.

Lewin, C., & Herlitz, A. (2002). Sex differences in face recognition: Women's faces make the difference. *Brain and Cognition*, *50*, 121–128.

Lewin, C., Wolgers, G., & Herlitz, A. (2001). Sex differences favoring women in verbal but not in visuospatial episodic memory. *Neuropsychology*, *15*, 165–173.

Liu-Shuang, J., Torfs, K., & Rossion, B. (2016). An objective electrophysiological marker of face individualisation impairment in acquired prosopagnosia with fast periodic visual stimulation. *Neuropsychologia*, *83*, 100–113.

Logan, A. J., Wilkinson, F., Wilson, H. R., Gordon, G. E., & Loffler, G. (2016). The Caledonian face test: A new test of face discrimination. *Vision Research*, *119*, 29–41.

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.

McBain, R., Norton, D., & Chen, Y. (2009). Females excel at basic face perception. *Acta Psychologica*, *130*, 168–173. https://doi.org/10.1016/j.actpsy.2008.12.005

McKelvie, S. J., Standing, L., St Jean, D., & Law, J. (1993). Gender differences in recognition memory for faces and cars: Evidence for the interest hypothesis. *Bulletin of the Psychonomic Society*, *31*, 447–448.

McNeil, J. E., & Warrington, E. K. (1991). Prosopagnosia: A reclassification. *Quarterly Journal of Experimental Psychology*, *43A*, 267–287.

Megreya, A. M., Bindemann, M., & Havard, C. (2011). Sex differences in unfamiliar face identification: Evidence from matching tasks. *Acta Psychologica*, *137*, 83–89. https://doi.org/10.1016/j.actpsy.2011.03.003

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

Megreya, A. M., & Burton, A. M. (2008). Matching faces to photographs: Poor performance in eyewitness memory (without the memory). *Journal of Experimental Psychology: Applied*, *14*, 364–372.

Menon, N., White, D., & Kemp, R. I. (2015). Identity-level representations affect unfamiliar face matching performance in sequential but not simultaneous tasks. *Quarterly Journal of Experimental Psychology*, *68*, 1777–1793. https://doi.org/10.1080/17470218.2014.990468

Milner, B. (1968). Visual recognition and recall after right temporal-lobe excision in man. *Neuropsychologia*, *6*, 191–209.

O'Toole, A. J., Jiang, F., Roark, D., & Abdi, H. (2006). Predicting human performance for face recognition. In R. Chellappa & W. Zhao (Eds.), *Face processing: Advanced models and methods* (pp. 293–320). San Diego, CA: Academic Press.

Özbek, M., & Bindemann, M. (2011). Exploring the time course of face matching: Temporal constraints impair unfamiliar face identification under temporally unconstrained viewing. *Vision Research*, *51*, 2145–2155. https://doi.org/10.1016/j.visres.2011.08.009

Palermo, R., Rossion, B., Rhodes, G., Laguesse, R., Tez, T., Hall, B., … McKone, E. (2017). Do people have insight into their face recognition abilities? *Quarterly Journal of Experimental Psychology*, *70*, 218–233. https://doi.org/10.1080/17470218.2016.1161058

Prieto, E. A., Van Belle, G., Liu, J., Norcia, A. M., & Rossion, B. (2013) The 6 Hz fundamental stimulation frequency rate for individual face discrimination in the right occipito–temporal cortex. *Neuropsychologia*, *51*, 2863–2975.

Quaglino, A., & Borelli, G. (1867). Emiplegia sinistra con amaurosi-guarigione-perdita totale della percezione dei colouri e della memoria della configurazione degli oggetti. *Giornale d'Oftalmologia Italiano*, *10*, 106–117.

Ramon, M., Busigny, T., Gosselin, F., & Rossion, B. (2016). All new kids on the block? Impaired holistic processing of personally familiar faces in a kindergarten teacher with acquired prosopagnosia. *Visual Cognition*, *24*, 321–355. https://doi.org/10.1080/13506285.2016.1273985

Rehnman, J., & Herlitz, A. (2007). Women recognize more faces than men do. *Acta Psychologica*, *124*, 344–355.

Retter, T. L., & Rossion, B. (2015). Global shape information increases but color information decreases the composite face effect. *Perception*, *44*, 511–528. https://doi.org/10.1068/p7826

Rossion, B. (in press). Prosopdysgnosia? What could it tell us about the neural organization of face and object recognition? *Cognitive Neuropsychology*.

Rossion, B., Caldara, R., Seghier, M., Schuller, A.-M., 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*, 2381–2395.

Rossion, B., & Jacques, C. (2011). The N170: Understanding the time-course of face perception in the human brain. In S. Luck & E. Kappenman (Eds.), *The Oxford handbook of ERP components* (pp. 115–142). Oxford, UK: Oxford University Press.

Rossion, B. (2014). Understanding face perception by means of prosopagnosia and neuroimaging. *Frontiers in Bioscience (Elite Ed.)*, *6*, 308–317.

Schretlen, D. J., Pearlson, G. D., Anthony, J. C., & Yates, K. O. (2001). Determinants of Benton Facial Recognition Test performance in normal adults. *Neuropsychology*, *15*, 405–410.

Sergent, J. (1984). An investigation into component and configural processes underlying face perception. *British Journal of Psychology*, *75*, 221–242.

Sergent, J., Ohta, S., & MacDonald, B. (1992). Functional neuroanatomy of face and object processing. A positron emission tomography study. *Brain*, *115*, 15–36.

Sommer, W., Hildebrandt, A., Kunina-Habenicht, O., Schacht, A., & Wilhelm, O. (2013). Sex differences in face cognition. *Acta Psychologica*, *142*, 62–73. https://doi.org/10.1016/j.actpsy.2012.11.001

Tarr, M., & Bülthoff, H. (1998). Image-based object recognition in man, monkey and machine. *Cognition*, *67*, 1–20.

Trahan, D. (1997). Relationship between facial discrimination and visual neglect in patients with unilateral vascular lesions. *Archives of Clinical Neuropsychology*, *12*, 57–62.

Tranel, D., Vianna, E., Manzel, K., Damasio, H., & Grabowski, T. (2009). Neuroanatomical correlate of teh Benton Facial Recognition Test and Judgment of Line Orientation Test. *Journal of Clinical and Experimental Neuropsychology*, *31*, 219–233. https://doi.org/10.1080/13803390802317542

Tzavaras, A., Hécaen, H., & Le Bras, H. (1970). Le problème de la spécificité du déficit de la reconnaissance du visage humain lors des lésions hémisphériques unilatérales. *Neuropsychologia*, *8*, 403–416.

Weirich, S., Hoffmann, F., Meißner, L., Heinz, A., & Bengner, T. (2011). Sex influence on face recognition memory moderated by presentation duration and reencoding. *Neuropsychology*, *6*, 806–813. https://doi.org/10.1037/a0025633

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

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

Woodhead, M. M., & Baddeley, A. D. (1981). Individual differences and memory for faces, pictures, and words. *Memory & Cognition*, *9*, 368–370.

Xu, B., Liu-Shuang, J., Rossion, B., & Tanaka, J. (2017). Individual differences in face identity processing with fast periodic visual stimulation. *Journal of Cognitive Neuroscience*, *29*, 1368–1377. https://doi.org/10.1162/jocn_a_01126

Young, A. W., & Burton, A. M. (2018). Are we face experts? *Trends in Cognitive Sciences, 22*, 100–110.

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.