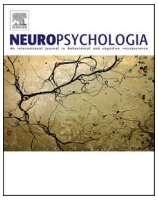
Neuropsychologia 165 (2022) 108119

Contents lists available at ScienceDirect

Neuropsychologia

journal homepage: www.elsevier.com/locate/neuropsychologia



Accurate but inefficient: Standard face identity matching tests fail to identify prosopagnosia

Matthew C. Fysh [a], Meike Ramon [b,*]

[a] School of Psychology, University of Kent, Canterbury, Kent, United Kingdom
[b] Applied Face Cognition Lab, Department of Psychology, University of Fribourg, Fribourg, Switzerland

ARTICLE INFO

Keywords:
Face cognition
Face matching
Behavioral assessment
Acquired prosopagnosia
Performance accuracy
Response times
Speed-accuracy trade-offs

ABSTRACT

In recent years, the number of face identity matching tests in circulation has grown considerably and these are being increasingly utilized to study individual differences in face cognition. Although many of these tests were designed for testing typical observers, recent studies have begun to utilize general-purpose tests for studying specific, atypical populations (e.g., super-recognizers and individuals with prosopagnosia). In this study, we examined the capacity of four tests requiring binary face-matching decisions to study individual differences between healthy observers. Uniquely, we used performance of the patient PS (Rossion, 2018), a well-documented case of acquired prosopagnosia (AP), as a benchmark. Two main findings emerged: (i) PS could exhibit typical rates of accuracy in all tests; (ii) compared to age-matched controls and when considering both accuracy and speed to account for potential trade-offs, only the KFMT — but not the EFCT, PICT or GFMT — was able to detect PS's severe impairment. These findings reflect the importance of considering both accuracy and response times to measure individual differences in face matching, and the need for comparing tests in terms of their sensitivity, when used as a measure of human cognition and brain functioning.

1. Introduction

In recent decades, interest in human face processing abilities has grown considerably. Historically, face processing has been used as a means to generate knowledge about general functioning principles of the human brain, with case studies reported for over two centuries (Quaglino and Borelli, 1867; Quaglino et al., 2003; Bodamer, 1947). Today, easy access to experimentation solutions and participants, as well as publicly available or artificially generatable face databases, has led to an explosion in tests designed to measure the perception and recognition of unfamiliar faces (for recent demonstrations, cf. e.g., Bate et al., 2018; Dunn et al., 2020; Fysh and Bindemann, 2018; Smith et al., 2021; Stantic et al., 2021; White et al., 2021). Traditionally, tools to assess these aspects of face cognition (i.e., the cognitive mechanisms underpinning processing of facial information) were developed following established neuropsychological and test theoretical procedures and with the aim of identifying impairments associated with these abilities (cf. Duchaine and Nakayama, 2006). More recently, a growing number of tests have emerged with the purported aim of understanding inter-individual differences among healthy, i.e. *normal* observers (for a review see Bate et al., 2021; Ramon et al., 2019a,b; Ramon, 2021).

However, generally speaking, limited attention has been directed towards the *de facto* appropriateness of tests for addressing specific research questions. For instance, some studies have employed tests to identify high performing individuals (e.g., Phillips et al., 2018), or probe their capacities (Tardif et al., 2019) using measures that were either developed to identify impairments (Duchaine and Nakayama, 2006) or which fail to distinguish between normal and impaired performance (White et al., 2017). Unfortunately, oftentimes little concern is directed to the specific behavioral measure, which is being considered to evaluate observers' individual levels of performance. In general, a disregard of the considerable differences in test sensitivity is detrimental to the domain of face processing.

Here, we aimed to shed light on and address this issue using an unconventional approach: Testing PS, the most extensively documented case of acquired prosopagnosia to date (for a review see e.g., Rossion, 2014), we sought to put face-matching tests that were designed to measure the abilities of healthy (i.e., neurotypical) observers to the test. Following a collision with a bus in 1992, PS sustained major brain damage which resulted in severely compromised face cognition, whilst her object recognition is unimpaired. In terms of low-level visual abilities, she presents with a small left paracentral scotoma and (slightly)

* Corresponding author. University of Fribourg, Applied Face Cognition Lab Department of Psychology, Faucigny 2, 100, Fribourg, Switzerland.
E-mail address: meike.ramon@gmail.com (M. Ramon).

https://doi.org/10.1016/j.neuropsychologia.2021.108119
Received 19 March 2021; Received in revised form 24 November 2021; Accepted 9 December 2021
Available online 15 December 2021

lower visual acuity than normal participants. For this reason, PS's willingness to participate in experimental research places her as an invaluable source of information for understanding the behavioral and cortical mechanisms that underpin the perception of faces.

PS's severe impairment at processing facial identity has been repeatedly demonstrated (Rossion et al., 2003; Ramon et al., 2017; for a review see Rossion, 2014). PS's abnormal performance on simple tasks of face identity matching provide the most striking demonstration of her profound deficit (Ramon and Rossion, 2010; Ramon et al., 2010). Our focus in this study was specifically on tests involving 2-alternative forced-choice (2AFC) paradigms, which measure simple binary decisions based on simultaneous perceptual discrimination. We reasoned that, a minimum requirement for tests to be informative for the study of individual differences in simple 2AFC face matching among healthy observers, is that they are capable of identifying and excluding drastically impaired observers. Testing four previously published tests of facial identity matching that use the same paradigm, we compare PS's performance to that of younger and older normal adults. Importantly, we also consider speed–accuracy trade-offs (Heitz, 2014; see also Luce, 1986), which are of particular importance given moderate to low test difficulty but is often neglected in tests focusing exclusively on performance accuracy.

In short, we sought to scrutinize behavioral 2AFC tests of unfamiliar face matching currently used to study (differences in) healthy cognitive functioning. Reversing the traditional approach of test development for the sake of impairment identification, we used PS's impairments to probe different tests and investigate their respective discriminatory power and informative merit.

1.1. Prosopagnosia and assessment of face cognition for clinical purposes

Individuals suffering from prosopagnosia exhibit impaired processing of facial information, leading to deficient face matching, recognition and identification (for recent reviews, see Albonico and Barton, 2019; Geskin and Behrmann, 2018). Two forms of prosopagnosia have been described (but cf. Rossion, 2018): acquired prosopagnosia (AP), which includes patients whose impairments were caused by damage to occipitotemporal regions (Rossion, 2014), and congenital or developmental prosopagnosia, which is characterized by severe lifelong difficulty with face recognition in the absence of acquired brain injury (e.g., Cattaneo et al., 2016; Geskin and Behrmann, 2018; McConachie, 1976; Rosenthal et al., 2017; Thomas et al., 2009).

In clinical settings, prosopagnosia has usually been assessed using the Benton Facial Recognition Test (BFRT; Benton and Van Allen, 1968; Benton et al., 1983; Rossion and Michael, 2018) and the Warrington Recognition Memory for Faces (RMF; Warrington, 1984; Duchaine and Weidenfeld, 2003; Rossion et al., 2003). Note that while both tests' names include the term "recognition", following the definitions adopted here and elsewhere (Fysh et al., 2020; Stacchi et al., 2020; Ramon and Gobbini, 2018; Ramon, 2018, 2021) the BFRT (Benton et al., 1983) measures *simultaneous matching* of multiple facial identities (i.e., perception), while the RMF probes *memory* for face images (i.e., recognition). In early studies, these tests were used to identify disorders of face perception and distinguish individuals with prosopagnosia from samples of control participants, and critically, enabled researchers to isolate face-specific deficits in neurologically-damaged patients from visual field defects and aphasias (Benton and Van Allen, 1968). These studies also found, however, that individuals with prosopagnosia were able to perform within normal accuracy levels when discriminating unfamiliar faces despite a profound impairment at recognizing the faces of familiar people, leading to speculation that the recognition of familiar faces and the discrimination of unfamiliar faces are dissociable (e.g., Benton and Van Allen, 1972).

However, a common limitation shared by these tests is that they may be solved via strategies unrelated to facial identity information, which are however not revealed by the measure of performance considered (i.e., accuracy). For example, impaired individuals exhibit an overreliance on external facial information and paraphernalia (e.g., Rossion et al., 2003) that can enable normal performance accuracy. For instance, patient EP obtained a normal RMF (Warrington, 1984) score by using non-facial information (Nunn et al., 2001; see also Duchaine and Weidenfeld, 2003). Second, especially in tests of face perception, impaired individuals may solve trials via time-consuming piecemeal strategies (cf. Ramon et al., 2017; Busigny et al., 2014a,b; Ramon and Rossion, 2010; Ramon et al., 2010; Ramon, 2018). This leaves open the potential for speed-accuracy trade-offs, which go undetected if response times are not measured (cf., Geskin and Behrmann, 2018). Choice of appropriate stimulus material, as well as recording informative behavioral measures such as response times (e.g., Rossion and Michel, 2018) can not only mitigate these issues, but are critical for revealing additional, or more subtle deficits (Geskin and Behrmann, 2018; Delvenne et al., 2004).

1.2. Assessment of face cognition for non-clinical purposes

As noted, the number of face cognition tests continues to increase. These tests vary in difficulty, stimulus properties, and task demands (Burton et al., 2010; Bate et al., 2018; Dunn et al., 2020; Fysh and Bindemann, 2018; see Box 1 for a discussion of general criteria for test selection). Despite these differences, however, such tests are typically designed to assess face processing ability in the normal population. As a consequence, it is possible that these tests may be of limited value for studying non-normal participants.

For example, in the short version of the Glasgow Face Matching Test (GFMT; Burton et al., 2010) it is not possible to clear two standard deviations above the mean (Bate et al., 2018), meaning that the GFMT is unlikely to be able to reliably distinguish exceptionally high performers (i.e. "super-recognizers") from normal observers. Nonetheless, recent work has utilized the GFMT for this very purpose (see, Phillips et al., 2018). Conversely, it was demonstrated in another study that a sample of six participants with developmental prosopagnosia were also capable of achieving normal performance accuracy in this test (White et al., 2017). These aspects of the GFMT indicate that performance accuracy on this test should not be used to distinguish atypical observers from the population. Conversely, other tests, such as the Crowds Matching or Models Memory Test (Bate et al., 2018), or the 10-item version of the Yearbook Test (YBT-10; Fysh et al., 2020) may be extremely difficult, and therefore would not separate clinically impaired observers from typical participants.

At the same time, tests that *were* designed to study face identity processing in specific settings or populations, such as algorithms (White et al., 2015), SRs (Russell et al., 2009), and individuals with prosopagnosia (Duchaine and Nakayama, 2006) are being increasingly utilized for more general purposes (e.g., Balsdon et al., 2018; Phillips et al., 2018; White et al., 2015). A key problem with this approach, however, is that tests which were designed to measure the proficiency of face recognition algorithms - for example - might reliably distinguish between different qualities of algorithms, but may not be calibrated to achieve this same goal with humans. Due to the near-instantaneous speed with which algorithms process faces, response times are often not considered when evaluating performance in these specific tasks. Likewise, the type of stimuli used in tests that were designed to measure specific populations may also lack generalizability to faces in the real world (Russell et al., 2009; White et al., 2015; Ramon et al., 2019b; Ramon, 2021). Consequently, tests of this kind may also fail to capture the full capacity and robustness of human face cognition (Ramon et al., 2019b; Ramon and Gobbini, 2018; Ramon, 2021).

1.3. Comparing a patient with acquired prosopagnosia to healthy controls on two difficult tests of face perception: 1-to-many identity matching

To illustrate the importance of the above points, we recently assessed the performance of patient PS on two challenging tasks of face

Box 1
Practical solutions for assessment of face cognition via currently available tests

A variety of tools have been developed to assess face perception and recognition using unfamiliar facial identities. *Face perception* is typically assessed via simultaneous or 1-to-*n* matching or discrimination tasks in which observers e.g. arrange stimuli according to their perceived similarity (e.g., Duchaine et al., 2007, distinguish between simultaneously presented faces (Ramon and Rossion, 2010), or determine which of *n* probes corresponds to a target stimulus (cf. Benton et al., 1983; Ramon et al., 2010; Stacchi et al., 2020; Fysh et al., 2020). Assessment of *face recognition*, on the other hand, typically involves distinguishing between experimentally learned and novel facial identities (cf. Duchaine and Nakayama, 2006; Bate et al., 2018; Warrington, 1984).

Which of the many tests that are currently in circulation should one choose for a given purpose? This question is not trivial. First and foremost, one needs to ensure selection of a test probing the same *sub-process* (i.e., face perception or recognition; see above). Beyond this, at least three additional important aspects require careful consideration: stimuli, design and analyses.

Stimulus material

Carefully look at the type of stimulus material a test involves. As shown in Fig. 2, even when involving the same basic setup (here: 2-alternative forced-choice (2AFC) identity matching) tests may use extremely different stimuli. Choose the test that uses the type of stimuli that are most relevant for your specific goal.

For example, do you want realistic images and welcome natural variability, or does your research question require ensuring that all stimuli are spatially aligned, devoid of color and contour information? Alternatively, you can search for more appropriate material in publicly available databases or create your own stimuli. Bear in mind that stimulus examples portrayed in papers may provide an impression that might not necessarily be representative of the entire stimulus set. The value of good stimulus material cannot be overstated.

Test characteristics: design and measures

The choice of stimuli has a large impact on the difficulty of a test. In the context of 2AFC matching tasks, pairing highly distinctive facial identities on a trial basis, will result in decreased task difficulty. Different measures can be implemented to address this issue. On the one extreme, stimulus similarity can be quantified objectively, which has the advantage of being able to reproduce the setup with other stimulus material. On other hand, simple criteria can be adopted, e.g. pairing identities that have the same gender, age, hair and eye color, and (if present) external paraphernalia (cf. Fig. 2).

Additional considerations should be made regarding the effects that different procedures have on measured performance. For instance 2AFC matching tests may involve different stimulus presentation modes and durations. Consider possible scenarios in which the identity of two stimuli A and B has to be judged as same/different: (i) A and B are briefly presented, side by side, for 1000 ms; (ii) A is presented for 1000 ms and B remains on screen until a decision is reached; (iii) A and B are presented simultaneously until a response is provided. The latter design (iii) was employed in the face matching tests we used here. The lack of constraints regarding presentation duration can result in ceiling effects for performance accuracy (i.e, with most participants achieving high scores). In this scenario, accuracy scores have little or no informative value (cf. GFMT), and measuring and analysis of response times as additional information become ever more important. Note that response times also provide additional information about potential differences in difficulty *across* trials, and relatedly the consistency of a given test.

Finally, different scenarios require different test design and procedural approaches. For instance, if the goal is to distinguish between average and high performers or Super-Recognizers (Ramon, 2021), 2AFC matching tasks with unlimited presentation duration are inappropriate. In the context of such simple binary decisions, speeded procedures (e.g. Ramon et al., 2011, 2018, 2019a) become more interesting. In short, there are several test-theoretical and setting-specific considerations that should be considered when making choices about tests to use.

Performance estimation and analyses

There are several ways to estimate participants' performance, and the degree of elaboration should be linked to the intended goal. For example, in cases where deficits are well-established or obvious, such as in PS, individual performance averages as considered here in the context of simple 2AFC tasks might be sufficient. Of course, in a task that can be performed above chance by the severely impaired patient PS, normal individual participants may choose to emphasize speed over accuracy. Therefore, to account for speed-accuracy trade-offs that must be considered *at the individual* (not group) *level*, we calculated composite performance scores for each participant (cf. Fig. 3). Crucially, individuals' performances *across independent tests* should never be estimated by simple raw score summation (see Nador and Ramon, 2021).

Again, if differences between typical observers are of interest, analyses should be conducted at a higher level of granularity – zooming in on each individual observer's behavioral performance. This allows to examine potential variations of performance across e.g. stimulus repetitions, or trials of an experiment (Nador et al., 2021a, 2021b; Ramon et al., 2011, 2018). Going one step further, if the goal is to characterize differences between typical and high-performing Super-Recognizers (Ramon, 2021), the analytical focus has yet again to change. For instance, in some scenarios, inter-observer differences in ability may be reflected more accurately via measured differences in performance *consistency* (e.g., Nador et al., 2021). It is clear that such questions may require the development of novel analytical procedures (e.g., Nador et al., 2021b).

perception. The first of these was the YBT-10 (Fysh et al., 2020), an extremely difficult test requiring (per gender) matching of five young target identities among ten probes that are 25 years older, for which mean accuracy is 3.71/10 (SD = 1.99) (Fysh et al., 2020). The second task was the Facial Identity Card Sorting Task (FICST; Jenkins et al., 2011; Stacchi et al., 2020), which requires observers to correctly sort 40 intermixed images of two people into identity piles. The FICST is divided into two phases – in Phase 1, observers complete the task without any instruction regarding the number of identities present among the cards. In Phase 2, observers repeat the task, but with the additional information that only two identities are present, thus greatly reducing the difficulty of the task (see, Andrews et al., 2015; Stacchi et al., 2020).

Our aim in employing the FICST and YBT-10 was to demonstrate that for difficult tests of face matching, PS could indeed achieve normal accuracy relative to neurotypical controls. In such documentedly difficult tests (Stacchi et al., 2020; Fysh et al., 2020; Ramon, 2021), her profound impairment is revealed by performance under conditions that decrease task difficulty and enhance performance for typical observers. Compared to a control group of age-matched and younger observers, PS exhibited within normal range performance on the YBT-10 (2/10).

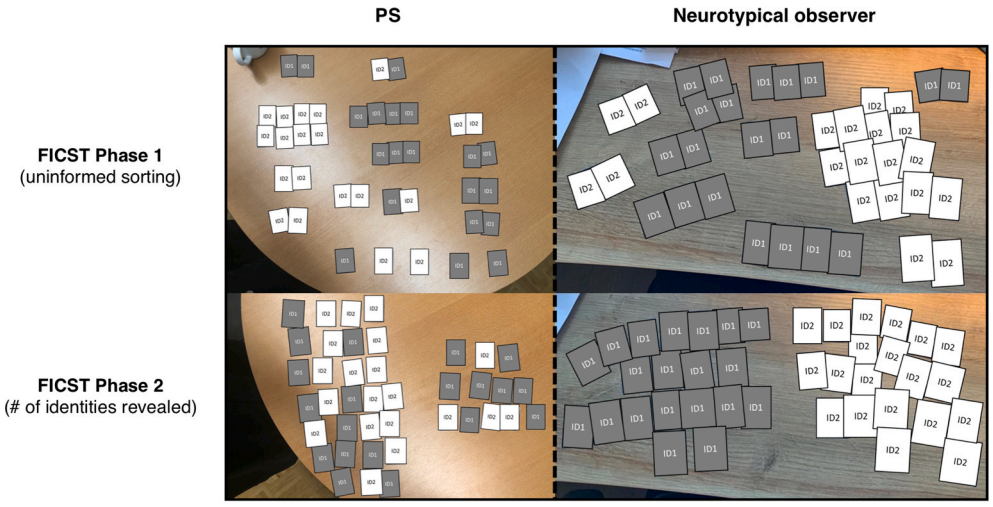


Fig. 1. Phases 1 and 2 of the FICST in PS versus a control participant. The left column represents PS′ performance in Phase 1 (upper) and Phase 2 (lower) of the FICST, in which it can be seen that even when aware that only two identities are present, she still struggled nonetheless to group identity cards accordingly. Conversely, the right column represents a randomly-sampled control participant at Phase 1 (upper) and Phase 2 (lower), who perceived too many identities in Phase 1, but who then achieved a perfect score in Phase 2 once aware of the fact that only two identities were present.

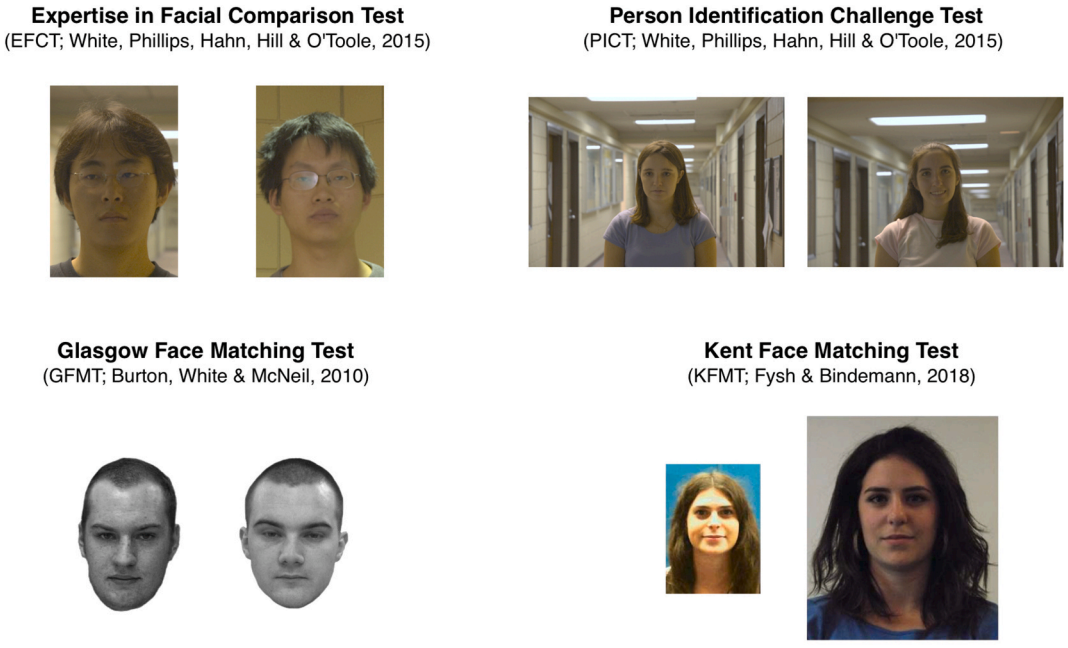


Fig. 2. Examples of stimuli from each test used in this study. All tests represent 2-alternative forced-choice face matching tasks, requiring same/different identity decisions for a face pair presented on each trial (see Methods for procedural details). Displayed for each test are examples of *mismatch trials*.

Likewise, in the difficult Phase 1 of the FICST, PS was comparable to age-matched and non-age-matched controls. Interestingly, her profound impairment became strikingly evident in the much easier Phase 2 of the task (cf. Busigny et al., 2010a,b). Neurotypical observers' performance improves when they are informed that the 40 images present only two identities. Contrariwise, PS's performance remained relatively stable, due to numerous inclusion errors among the two identity groups formed, as shown in Fig. 1. (For the full details of PS's and controls' performance in these tasks, see Supplementary Material.)

These data demonstrate that tests which are calibrated to be difficult (i.e., the YBT-10, and Phase 1 of the FICST) are unlikely to differentiate between poor and normal ability ranges. That is, based on accuracy *alone*. Although response times were not formally recorded for the YBT-10 or the FICST, PS nonetheless took an abnormally long time in completing both of these tasks relative to controls, as one would expect. Indeed, her profound deficit has been repeatedly revealed via abnormally prolonged RTs in the context of comparatively simpler simultaneous matching tasks with fewer choice options (e.g., Ramon et al., 2010; Ramon and Rossion, 2010; Orban de Xivry et al., 2008; Busigny et al., 2010a,b; Ramon et al., 2017). In the neuropsychological literature it is well established that normal levels of performance accuracy may be achieved at the cost of abnormally prolonged response times (e.g., Delvenne et al., 2004; Geskin and Behrmann, 2018), also captured via composite measures integrating both accuracy and speed (c.f. Bruyer and Brysbaert, 2011; Townsend and Ashby, 1978).

1.4. Testing the tests: which can or fail to identify an acquired prosopagnosic patient's deficit?

Thus, whilst challenging tasks may be effective for separating high-performing observers from those exhibiting typical face recognition ability, these might fail to distinguish abnormally low from normal performance, due to a restricted range at the lower end of the performance scale. However, easier tests have the opposite problem: if baseline accuracy is already high, then it might be statistically impossible to demonstrate accuracy that exceeds the mean by more than two standard

Table 1
Normative samples' and the patient PS's demographics.

Test	PS	Younger controls			Older controls		
	Age	*n*	Mean age ±SD	Age range	*n*	Mean age ±SD	Age range
KFMT[1]	69	60[a]	20.25 ± 3.55	18–41	5	73.20 ± 5.54	70–83
GFMT[2]	69	60[a]	20.25 ± 3.55	18–41	5	73.20 ± 5.54	70–83
EFCT[3]	70	175[b]	27.74 ± 10.73	18–58	6[b]	70.50 ± 8.22	59–80
PICT[3]	70	175[b]	27.74 ± 10.73	18–58	6[b]	70.50 ± 8.22	59–80

[1]Fysh and Bindemann (2018); [2]Burton et al. (2010); [3]White et al. (2015). Data were sourced from: [a]Fysh and Bindemann (2018); [b]Stacchi et al. (2020).

deviations. Consequently, tasks for which normal performance is already high may be capable of identifying clinical impairment in face matching, but not superior performance. These considerations for easy tests are particularly pertinent when performance accuracy alone is measured. Disregarding response times as an important source of information can lead to a failure to detect impairments as described above.

Following these considerations, in the present study we adopted an unconventional approach: We investigated which tests and measures of face perception should *not* be used to describe inter-individual differences among *healthy* individuals. We reasoned that simple 2AFC tests are insufficiently sensitive to this end if PS can achieve normal levels of performance accuracy despite her deficit at facial identity processing. Given PS's established record of impaired performance in unfamiliar face matching tasks that have been designed to probe her perceptual deficits (e.g., Ramon and Rossion, 2010; Ramon et al., 2010), she represents an ideal benchmark against which these tests might be evaluated.

Using four commonly used tests involving 2AFC simultaneous face matching/discrimination, we sought to identify which of these would detect PS's perceptual impairments. The tests used were the GFMT (Burton et al., 2010), the short version of the Kent Face Matching Test (KFMT; Fysh and Bindemann, 2018), the Expertise in Facial Comparison Test (EFCT; White et al., 2015), and the Person Identification Challenge Test (PICT; White et al., 2015). Following previous procedures that consider both performance accuracy *and* speed (e.g., Busigny et al., 2010a,b; Ramon et al., 2010; Ramon et al., 2017), we compared PS's performance to that of smaller groups of age-matched controls, as well as previously reported large and heterogeneous observer samples (Fysh and Bindemann, 2018; Stacchi et al., 2020; see Methods).

2. Methods

All procedures reported here were approved by the local research Ethics Committee of the University of Fribourg (approval number 473); upon publication of the manuscript, all data are publicly available via an Open Science Framework repository (osf.io/wj5bp).

2.1. Participants

Patient PS. PS's lesions and deficits have been previously described in over 30 papers published by the Face Categorization Lab and its collaborators since the first report by Rossion et al. (2003). To summarize briefly, PS suffered traumatic brain injury after she was hit by a bus in London in 1992, resulting in a severe closed-head injury. She presents with lesions in the left mid-ventral (mainly fusiform gyrus), and in the right inferior occipital regions, as well as minor damage to the left posterior cerebellum and the right middle temporal gyrus (Rossion et al., 2003; Sorger et al., 2007). After successful neuropsychological rehabilitation, PS was able to return to her profession as a kindergarten teacher (c.f. Ramon et al., 2017); her severe and longstanding prosopagnosia is unaccompanied by object processing deficits (c.f. Busigny et al., 2010a,b; Busigny and Rossion, 2010). PS was 69 when she completed the GFMT (Burton et al., 2010), KFMT (Fysh and Bindemann, 2018), YBT-10 (Fysh et al., 2020), and the FICST (Jenkins et al., 2011; Stacchi et al., 2020), and 70 years old when she completed the PICT and the EFCT (White et al., 2015).

Control observers. Control data from healthy volunteers was taken from previous studies (EFCT and PICT data from Stacchi et al. (2020); and GFMT and KFMT data from Fysh and Bindemann (2018)). From the EFCT and PICT datasets we created two sub-samples: (i) larger cohorts of younger observers, (ii) smaller cohorts of older observers (ranging between ± 10 years of PS's age at the time of testing). Since the KFMT and the GFMT did not contain any older controls, we tested 6 older observers (3 females, 3 males) for each of the EFCT and the PICT, and 5 older observers (1 female, 4 males) for the KFMT and the GFMT. Table 1 summarizes the demographics of each test cohort.

2.2. Tests of face processing

Fig. 2 provides example stimuli from each test that was employed. Note that all tests were administered with (virtually) no time constraints (but 30s duration in the EFCT and PICT; see below). For all tests, participants' response times (RTs) were recorded. This follows previous procedures, which aim to measure PS's abilities without pressure, while at the same time considering the potential for well-established speed-accuracy trade-offs.

Kent Face Matching Test (short version) (KFMT). This test also measures simultaneous face matching/discrimination. Participants are presented with 40 image pairs (20 males, 20 females) depicting two Caucasian faces and are required to indicate whether these depict the same person, or two different people. Half of trials present the same identity (i.e., an identity match) and the remaining 20 trials depict two different identities (i.e., an identity mismatch). Each trial features one high-quality digital face photograph of a person in frontal pose whilst bearing a neutral expression, which is paired with a non-controlled ambient student ID photograph which was acquired a minimum of three months earlier. For full details of the KFMT, see Fysh and Bindemann (2018).

Glasgow Face Matching Test (short version) (GFMT). The GFMT requires observers to match 40 pairs of cropped greyscale images of well-lit frontal faces displaying neutral expressions, half of which depict the same identity, and the remaining 20 trials depict two different people. In this task, identity match trials comprise two images of the same person that were acquired on the same day within a single session. For full details of the GFMT, see Burton et al. (2010).

Expertise in Facial Comparison Test (EFCT) and Person Identification Challenge Test (PICT). Both tests involve simultaneous unfamiliar face matching/discrimination and were originally reported by White et al. (2015). Each test includes 84 and 40 trials, respectively, consisting of presentation of pairs of full-frontal colored portraits, half of which depict the same identity. Making use of normative data provided by Stacchi et al. (2020), we applied the same procedures as these authors did. Specifically, on each trial (maximum presentation duration of 30s) participants provided a binary (i.e., same/different) response on each trial, and both accuracy and response times were recorded. The images used in both tests were taken across different days (see Phillips et al., 2012), in different locations and consequently depict different background information and involve differences in lighting. The difference between the EFCT and PICT concerns the distance from which the images were taken. As images used in the PICT were taken from a greater distance, its stimuli include more body and environmental cues.

2.3. Analyses

The three following measures were generated from these data: response accuracy, response times, and inverse efficiency (IE) scores.

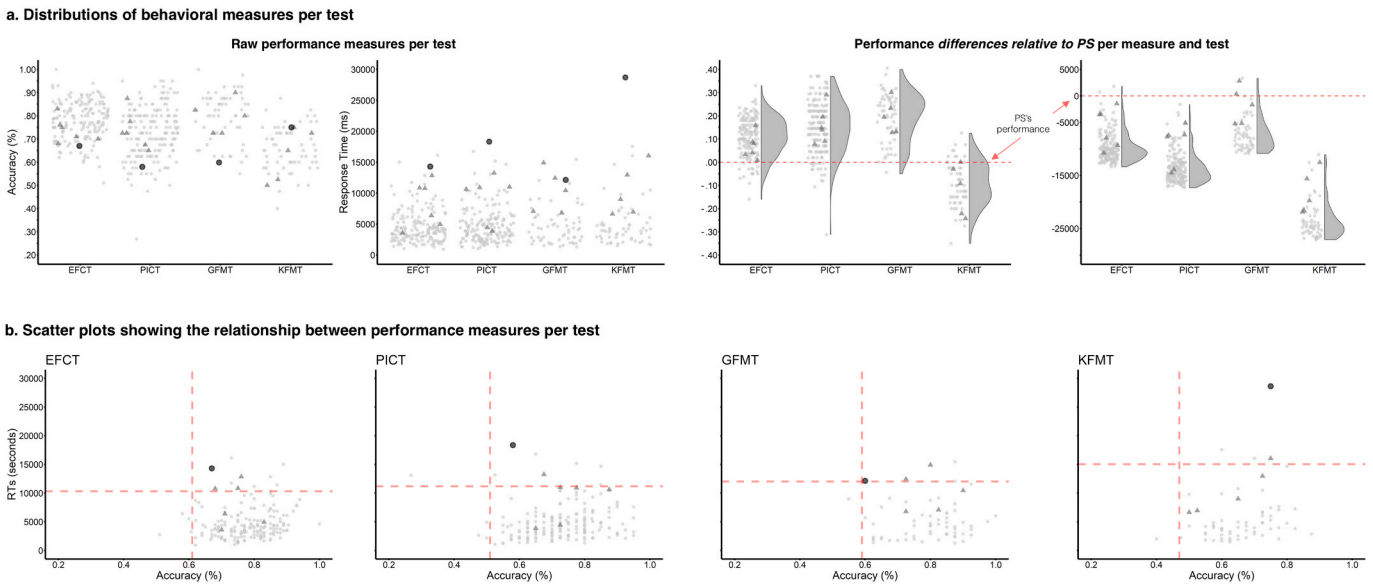


Fig. 3. Individual observers' performances per test and measure reported. a. Each observer's data is plotted per measure (accuracy, response times) and per test. The two left panels plot individual raw performance; the two right panels plot the same data standardized to PS's performance. **b.** Scatterplots illustrate the relationship between measures for each observer and test, crosshairs indicate 2 standard deviations below (for accuracy) and above (for RTs) the mean performance. Dark circles, triangles and light grey circles represent the data from PS, age-matched controls, and young controls, respectively.

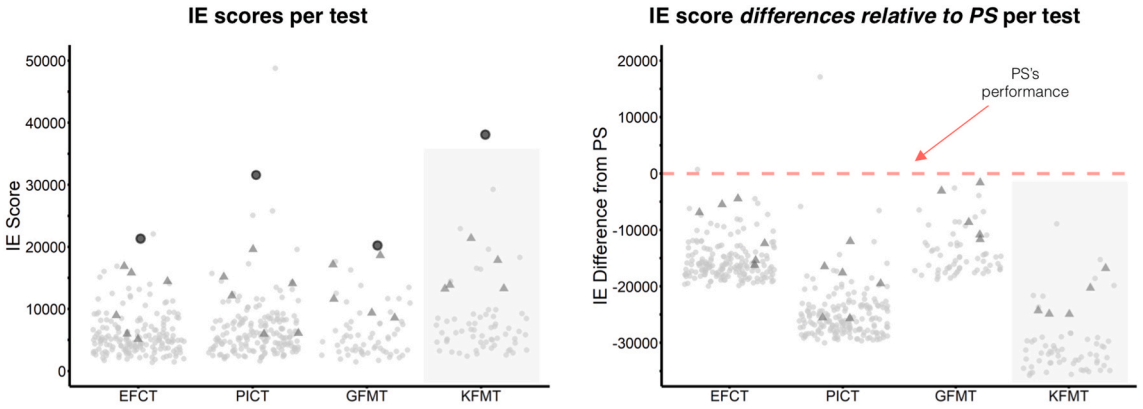


Fig. 4. Distributions of individual observers' inverse efficiency (IE) scores per test relative to the patient PS. Inferential statistical analyses revealed that PS's IE score was deficient only for the KFMT.

The first two measures are reported descriptively in the following Results section; IE scores (obtained via dividing per observer mean correct response times by the proportion of correct responses; Townsend and Ashby, 1978) were subject to inferential analysis. PS's IE scores were compared to those of age-matched and non-age-matched controls using modified *t*-tests for single-case analyses (Singlims_ES.exe program; Crawford et al., 2010), which provide point and interval estimates of effect sizes for the comparison of a case and the control group. Additionally, it provides an estimate of the percentage of the population that would perform below the studied patient.

3. Results

3.1. Description of observers' accuracy and response times (RTs)

First, individual averages per performance measure (i.e., mean percentage correct and mean RTs) were calculated for PS and the control groups (Fig. 3a, left). Per test and performance measure, we also subtracted PS' score from each subject, to obtain and visualize the difference between PS and the two groups (Fig. 3a, right). Additionally, we plotted the relationship between accuracy scores and RTs per observer and test (Fig. 3b). These descriptive analyses demonstrate that relative to two control cohorts, and across all tests, PS exhibited (i) performance accuracy within the normal range, and (ii) prolonged RTs. This is particularly evident for the KFMT and the PICT, on which she was slower to respond than any of the controls. Similarly, only two observers were slower than PS on the EFCT, and likewise, only three took longer to respond than PS on the GFMT.

3.2. Analyses of observers' inverse efficiency (IE)

To consider potential speed-accuracy trade-offs at the individual observer level, we combined response accuracy and RTs into IE scores as a composite measure of overall behavioral proficiency (performance) based on which inferential statistics were then performed. Computed by dividing each subject's mean correct RT by their accuracy score (Bruyer and Brysbaert, 2011; Townsend and Ashby, 1978), higher scores reflect poorer efficiency in the task; individual observers' IE scores are plotted in Fig. 4. As with the other measures, we computed the difference between PS and controls by subtracting PS's IE score from that of each

subject.

To formally examine IE score differences between PS and controls, *t*-tests modified for single case comparisons were performed (Crawford et al., 2010) using Holmes' sequential Bonferroni correction for multiple comparisons. In comparison to age-matched controls, PS demonstrated poorer performance on the KFMT, $t\,(4) = 5.63$, $p_{(\text{two-tailed})} = 0.005$, but not on the PICT, $t\,(5) = 3.37$, $p_{(\text{two-tailed})} = 0.020$. PS's performance was also comparable to that of age-matched controls on the EFCT and the GFMT, $t\,(5) = 1.81$, $p_{(\text{two-tailed})} = 0.129$ and $t\,(4) = 1.69$, $p_{(\text{two-tailed})} = 0.166$, respectively. The analogous comparisons for the younger control group revealed that PS displayed poorer performance than the younger control subjects for all tests, all $ts \geq 3.80$, all $ps < .001$.

4. Discussion

In this study, we sought to put four previously used tests of face perception to the test. The tests considered — the KFMT, the GFMT, the EFCT and the PICT (Burton et al., 2010; Fysh and Bindemann, 2018; White et al., 2015) — all require 2-alternative forced-choice (2AFC) decisions of facial identity matching among two simultaneously presented faces (see Fig. 2). Unconventionally, we probed these tests by comparing neurotypical observers' performance to that of PS, the likely most extensively documented case of acquired prosopagnosia to date (cf. Rossion, 2014: Ramon et al., 2017). Three main findings emerged.

Firstly, across all tests, PS exhibited patterns of response accuracy that comfortably fell within normal range. Secondly and in conjunction with previous work (e.g., Busigny and Rossion, 2010; Schiltz et al., 2006; Ramon and Rossion, 2010; Ramon et al., 2017), PS took considerably longer to correctly discriminate faces than younger controls across all four tests, as well as older controls on three of the four tests. This aligns with previous evidence from cases of acquired and developmental prosopagnosia exhibiting abnormally prolonged RTs (Behrmann et al., 2005; Bate et al., 2009, 2019; Behrmann et al., 2005), due to time-consuming, analytical/feature-based strategies, which can enable normal levels of performance accuracy (e.g., Avidan et al., 2011; Busigny et al., 2010a,b; Duchaine and Nakayama, 2006; Palermo et al., 2017). Thirdly, and finally, compared to age-matched controls PS's inverse efficiency performance was deficient for the KFMT, despite her numerically poorer accuracy on the GFMT, EFCT, and the PICT than her age-matched control counterparts. In the following we discuss the implications of these findings for present and future work on face cognition in neurotypical observers.

4.1. Accuracy and response times provide complementary information

Taken together, reports of profoundly impaired observers achieving performance accuracy scores that fall within normal range as reported here and previously (White et al., 2017) emphasize the importance of generally considering both response accuracy *and* response speed in parallel. This applies both when (i) examining differences between a/typical populations, and (ii) seeking to determine individual levels of processing proficiency. As discussed in Box 1, this is particularly crucial when simple binary (e.g., same/different) decisions are required, or stimuli are fairly easy to discriminate. However, it also applies to more difficult testing scenarios, e.g. when more complex 1-to-*n* decisions are required (cf. Stacchi et al., 2020; Fysh et al., 2020) and/or given higher inter-stimulus similarity (e.g. Ramon et al., 2010; Ramon and Van Belle, 2016).

Assessment and reporting of response times alongside accuracy measures has been practiced for many decades in experimental psychology and neuropsychology (see, e.g., Ellis et al., 1990; Marotta et al., 2002). Nonetheless, several more recent studies aiming to evaluate individuals presumed to occupy the high performing end of the face processing continuum have neglected this measure (e.g., Phillips et al., 2018; Robertson et al., 2016; White et al., 2015). This is particularly critical as some of these studies report findings that are practically relevant in applied settings. For instance, professionals have been reported to show relatively greater face matching performance, which could easily be accounted for by having been permitted substantially more time than controls to e.g. study to-be-matched faces (White et al., 2015) or perform binary face discriminations on EFCT trials (e.g., forensic examiners had up to 3 months for 20 2AFC trials, contrary to control and allegedly superior observers; Phillips et al., 2018).

Here, we demonstrated the need to consider both accuracy and response times when examining individual differences in face cognition via a case of acquired prosopagnosia. Critically, even when adopting this approach, for three frequently used tests PS exhibited performance that was indistinguishable from age-matched controls. Moving forward, the need to consider potential speed-accuracy trade-offs should finally transfer into research among typical populations. Ultimately, tests in which impaired individuals such as PS can achieve "normal" performance even when both accuracy and response times are considered, may be limited in their capacity to describe individual differences in healthy observers (e.g., Robertson et al., 2016), or to benchmark performance of algorithms (e.g., White et al., 2015; Phillips et al., 2018).

4.2. Tests differ in their ability to identify highly deficient performance

Analyses of IE scores revealed that, across all tasks PS performed abnormally low compared to younger controls. Interestingly, however, relative to older controls, PS's IE was abnormal only for the KFMT. For the remaining three tests – i.e., the PICT, GFMT and the EFCT – PS's composite performance was not distinguishable from that of age-matched controls. These findings highlight critical differences between the KFMT and the remaining tests: the KFMT's careful selection of identities portrayed in mis/match trials demonstrably leads to substantially higher task difficulty. Given this stimulus material related difficulty, the KFMT — despite requiring only binary same/different decisions — is able to detect PS's impairments when both performance measures are considered (see Box 1).[1] Additionally, the distributions of individual observers' IE scores across all tests further indicate that the PICT, GFMT and EFCT involve identity pairs that are more easily discriminated, and therefore could not detect her profound deficit and abnormal processing strategies.

The present findings also emphasize that several of the currently employed face-matching tests lack the sensitivity to discriminate between individuals at the extreme lower end of the face recognition continuum. This is not necessarily a surprising revelation in and of itself – none of the tests employed here were ever designed with the specific intention of distinguishing individuals with prosopagnosia from the general population (see, Burton et al., 2010; Fysh and Bindemann, 2018; White et al., 2015). This speaks to the broader issue of selecting the most appropriate test for assessing a given population. If the intention is to study individual differences across the entire continuum of ability in face matching, from observers with prosopagnosia all the way up to super-recognizers (Ramon, 2021), then it is essential to select a test that was designed with this purpose in mind (see, e.g., Stantic et al., 2021) and critically establish that the selected test in fact meets its intended purpose.

[1] At first sight, the reason for this pattern may not be immediately evident. However, it is possible that PS's feature-based matching strategy was well-suited for some trials of the KFMT, given that there are some identities that can be classified based on the presence of skin blemishes (i.e. moles), which typical observers are relatively more prone to overlooking (Towler et al., 2017, 2021). Relative to the GFMT, EFCT, and PICT, it would appear that the KFMT can be more efficiently resolved via a piecemeal matching strategy, which incidentally, also appears to be utilized by professionals (e.g., Towler et al., 2019; for a review see Moreton, 2021).

4.3. Evaluating tests - in defense of the single case and representative small n

There are several potential aspects to consider when interpreting these findings. First, our data are based on the performance of a single case study of acquired prosopagnosia (i.e., PS). Because acquired prosopagnosia is, by definition, acquired typically through brain damage or illness, one cannot expect a sample of patients presenting with acquired prosopagnosia to display homogenous performance profiles. For example, other cases that have been described as patients with acquired prosopagnosia reportedly presented with incomplete visual fields and/or object recognition deficits (e.g., Rezlescu et al., 2014; Barton et al., 2002). We argue that the specificity of PS's deficit — the *de facto* conceptual prerequisite for the diagnosis of prosopagnosia (Bodamer, 1947; Rossion, 2018) — makes her particularly valuable and extremely well-suited for evaluating face-matching tests.

Secondly, as pointed out by one reviewer, it is worth discussing PS's age-matched control samples. These comprised five participants who completed the GFMT and the KFMT, and six different participants who completed the EFCT and the PICT. Our aim here was not to formally compare the tests' sensitivity (this would have required a within-observer design). Moreover, these sample sizes are not unlike those employed in studies of a similar nature (see, e.g., Humphreys et al., 2007; Ramon et al., 2017; Rezlescu et al., 2014). Interestingly, while our age-matched control participants were also generally slower than their younger counterparts in the four tasks employed, this did not translate into inverse efficiency, which contrasts with findings for PS. Finally, we contend that small samples should not be treated as a *per se* limitation. Indeed, a highly sensitive test (with optimized stimuli, procedures, sufficient number of trials, etc.; cf. Box 1) would be able to provide an accurate, and representative description of a given individual (cf. Smith and Little, 2018). So, while it is possible that given larger samples we might have found differences between PS and controls for the EFCT, GFMT, and PICT, this would not make a stronger case for the sensitivity of these tests. The important point is that with comparable control samples, the EFCT, GFMT, and PICT failed to identify PS — despite claims that they are sensitive and difficult (e.g. White et al., 2015; Phillips et al., 2018).

Finally, the group-dependent results that emerged here are noteworthy. Compared to older controls, only the most challenging of the tests probed, the KFMT, was able to detect PS's impairment. Note that we report and display the available, previously reported younger control groups' data alongside those of older controls and PS, despite one reviewer's comment that "… *the young controls play no role in the paper's conclusions*". We respectfully disagree with their opinion, as the differences between groups emphasize the widely acknowledged but oftentimes neglected general need for using *appropriate* control samples in scientific research (Barrett, 2020; Broesch et al., 2020; Gurven and Lieberman, 2020; Henrich et al., 2010; Laajaj et al., 2019; Masuda et al., 2020; Nielsen et al., 2017; Rad et al., 2018). That is, regardless of whether an assessment of normal or potentially atypical (inferior *or* superior) performance is intended, individual performance should be assessed within the context of *representative* cohorts. Achieving representation, inclusion and diversity has never been easier, and should become the standard in psychology to achieve progress in our understanding of cognition and brain functioning.

5. Conclusion

In this study we sought to emphasize the importance of utilizing multiple, complementary performance measures when studying face matching both within, and between different populations. We showed that accuracy measures alone are insufficient to discriminate an individual with acquired prosopagnosia from control subjects. Yet when combined with response times in the form of inverse efficiency scores, differences between populations may emerge that were not initially apparent. That is, given accurate decisions, the time taken to submit these requires additional attention. Combined, both measures provide insight into observers' proficiency and test sensitivity.

In a field where the number of face-matching tests and stimulus databases is increasing rapidly (see, Bate et al., 2021), these findings speak to the importance of not necessarily developing additional tests *per se*. Indeed, rather than simply creating more tests or generating more data, we need to find and use *the most appropriate ones* (Ramon, 2021). Box 1 provides practical advice on how to make the most of existing methods and data. Judging the appropriateness of a given test requires practitioners and researchers to critically assess tests' procedures and limitations, as well as the correspondence of these tests to the real-world settings or research question that they seek to measure (Ramon et al., 2019b; Ramon, 2021). Our data highlight the need to consider multiple performance measures to fully exploit the informative value of face-matching tests. Moreover, our findings reiterate the expressed need for increased diversity and representation in control samples.

Finally, irrespective of the cohort studied — but especially at the extremes of the continuum of ability — multiple tests to assess face cognition skill should be administered. As our results demonstrate: tests using the same procedure to measure the same sub-process can differ substantially. Thus, the goal of describing behavior — from impairments to superior skill — should aim for a *reliable* characterization, which requires accumulation of evidence and transparent reporting (Rossion, 2014; Ramon, 2021).

CRediT author statement

MCF: Data curation; Formal analysis; Investigation; Methodology; Resources / Software; Visualization; Writing: Writing – original draft, Writing – review & editing. **MR**: Conceptualization; Data curation; Funding acquisition; Investigation; Methodology; Project administration; Resources / Software; Supervision; Visualization; Writing: Writing – original draft,Writing – review & editing.

Declaration of competing interest

The authors declare no competing interests.

Acknowledgements

We thank PS and all volunteers for their participation. MR is supported by a Swiss National Science Foundation PRIMA (Promoting Women in Academia) grant (PR00P1_179872).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.neuropsychologia.2021.108119.

References

Albonico, A., Barton, J., 2019. Progress in perceptual research: the case of prosopagnosia, F1000Research, 8. https://doi.org/10.12688/f1000research.18492.1.

Andrews, S., Jenkins, R., Cursiter, H., Burton, A.M., 2015. Telling faces together: learning new faces through exposure to multiple instances. Q. J. Exp. Psychol. 68 (10), 2041–2050. https://doi.org/10.1080/17470218.2014.1003949.

Avidan, G., Tanzer, M., Behrmann, M., 2011. Impaired holistic processing in congenital prosopagnosia. Neuropsychologia 49 (9), 2541–2552. https://doi.org/10.1016/j.neuropsychologia.2011.05.002.

Balsdon, T., Summersby, S., Kemp, R.I., White, D., 2018. Improving face identification with specialist teams. Cognit. Res.: Princip. Imp. 3 (1), 1–13. https://doi.org/10.1186/s41235-018-0114-7.

Barrett, H.C., 2020. Towards a cognitive science of the human: cross-cultural approaches and their urgency. Trends Cognit. Sci. 24 (8), 620–638. https://doi.org/10.1016/j.tics.2020.05.007.

Bate, S., Haslam, C., Jansari, A., Hodgson, T.L., 2009. Covert face recognition relies on affective valence in congenital prosopagnosia. Cogn. Neuropsychol. 26 (4), 391–411. https://doi.org/10.1080/02643290903175004.

Bate, S., Frowd, C., Bennetts, R., Hasshim, N., Murray, E., Bobak, A.K., Richards, S., 2018. Applied screening tests for the detection of superior face recognition. Cognit. Res.: Princip. Imp. 3 (1), 1–19. https://doi.org/10.1186/s41235-018-0116-5.
Bate, S., Mestry, N., Portch, E., 2021. Individual differences between observers in face matching. In: Bindemann, M. (Ed.), Forensic Face Matching: Research and Practice. Oxford University Press, UK, pp. 115–143. https://doi.org/10.1093/oso/9780198837749.003.0011.
Barton, J.J., Press, D.Z., Keenan, J.P., O'Connor, M., 2002. Lesions of the fusiform face area impair perception of facial configuration in prosopagnosia. Neurology 58 (1), 71–78. https://doi.org/10.1212/WNL.58.1.71.
Behrmann, M., Avidan, G., Marotta, J.J., Kimchi, R., 2005. Detailed exploration of face-related processing in congenital prosopagnosia: 1. Behavioral findings. J. Cognit. Neurosci. 17 (7), 1130–1149. https://doi.org/10.1162/0898929054475154.
Benton, A.L., Van Allen, M.W., 1968. Impairment in facial recognition in patients with cerebral disease. Trans. Am. Neurol. Assoc. 93, 38–42.
Benton, A.L., Van Allen, M.W., 1972. Prosopagnosia and facial discrimination. J. Neurol. Sci. 15 (2), 167–172. https://doi.org/10.1016/0022-510x(72)90004-4.
Benton, A.L., Sivan, A.B., Hamsher, K., Vareny, N.R., Spreen, O., 1983. Facial recognition: stimulus and multiple choice pictures. In: Benton, A.L., Sivan, A.B., Hamsher, K.D.S., Varney, N.R., Spreen, O. (Eds.), Contributions to Neuropsychological Assessment. Oxford University Press, New York, pp. 30–40.
Bodamer, J., 1947. Die prosop-agnosie. Archiv für Psychiatrie und Nervenkrankheiten 179 (1), 6–53. https://doi.org/10.1007/BF00352849.
Broesch, T., Crittenden, A.N., Beheim, B.A., Blackwell, A.D., Bunce, J.A., Colleran, H., Mulder, M.B., 2020. Navigating cross-cultural research: methodological and ethical considerations. Proc. Royal Soc. 287 (1935), 20201245. https://doi.org/10.1098/rspb.2020.1245.
Bruyer, R., Brysbaert, M., 2011. Combining speed and accuracy in cognitive psychology: is the inverse efficiency score (IES) a better dependent variable than the mean reaction time (RT) and the percentage of errors (PE)? Psychol. Belg. 51 (1), 5–13.
Burton, A.M., White, D., McNeill, A., 2010. The Glasgow face matching test. Behav. Res. Methods 42 (1), 286–291. https://doi.org/10.3758/BRM.42.1.286.
Busigny, T., Graf, M., Mayer, E., Rossion, B., 2010a. Acquired prosopagnosia as a face-specific disorder: ruling out the general visual similarity account. Neuropsychologia 48 (7), 2051–2067. https://doi.org/10.1016/j.neuropsychologia.2010.03.026.
Busigny, T., Joubert, S., Felician, O., Ceccaldi, M., Rossion, B., 2010b. Holistic perception of the individual face is specific and necessary: evidence from an extensive case study of acquired prosopagnosia. Neuropsychologia 48 (14), 4057–4092. https://doi.org/10.1016/j.neuropsychologia.2010.09.017.
Busigny, T., Rossion, B., 2010. Acquired prosopagnosia abolishes the face inversion effect. Cortex 46 (8), 965–981. https://doi.org/10.1016/j.cortex.2009.07.004.
Busigny, T., Prairial, C., Nootens, J., Kindt, V., Engels, S., Verplancke, S., et al., 2014a. CELEB: une batterie d'évaluation de la reconnaissance des visages célèbres et de l'accès aux noms propres. Rev. Neuropsychol. 6 (1), 69–81. https://doi.org/10.3917/rne.061.0069.
Busigny, T., Van Belle, G., Jemel, B., Hosein, A., Joubert, S., Rossion, B., 2014b. Face-specific impairment in holistic perception following focal lesion of the right anterior temporal lobe. Neuropsychologia 56, 312–333. https://doi.org/10.1016/j.neuropsychologia.2014.01.018.
Cattaneo, Z., Daini, R., Malaspina, M., Manai, F., Lillo, M., Fermi, V., et al., 2016. Congenital prosopagnosia is associated with a genetic variation in the oxytocin receptor (OXTR) gene: an exploratory study. Neuroscience 339, 162–173. https://doi.org/10.1016/j.neuroscience.2016.09.040.
Crawford, J.R., Garthwaite, P.H., Wood, L.T., 2010. Inferential methods for comparing two single cases. Cogn. Neuropsychol. 27 (5), 377–400. https://doi.org/10.1080/02643294.2011.559158.
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 (5), 597–612. https://doi.org/10.1016/j.neuropsychologia.2003.10.008.
Duchaine, B.C., 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. Cogn. Neuropsychol. 24 (4), 419–430. https://doi.org/10.1080/02643290701380491.
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.
Dunn, J.D., Summersby, S., Towler, A., Davis, J.P., White, D., 2020. UNSW Face Test: a screening tool for super-recognizers. PLoS One 15 (11), e0241747. https://doi.org/10.1371/journal.pone.0241747.
Ellis, A.W., Young, A.W., Flude, B.M., 1990. Repetition priming and face processing: priming occurs within the system that responds to the identity of a face. Q. J. Exp. Psyhol. Sec. 42 (3), 495–512. https://doi.org/10.1080/14640749008401234.
Fysh, M.C., Bindemann, M., 2018. The kent face matching test. Br. J. Psychol. 109 (2), 219–231. https://doi.org/10.1111/bjop.12260.
Fysh, M.C., Stacchi, L., Ramon, M., 2020. Differences between and within individuals, and subprocesses of face cognition: implications for theory, research and personnel selection. R. Soc. Open Sci. 7 (9), 200233. https://doi.org/10.1098/rsos.200233.
Geskin, J., Behrmann, M., 2018. Congenital prosopagnosia without object agnosia? A literature review. Cogn. Neuropsychol. 35 (1–2), 4–54. https://doi.org/10.1080/02643294.2017.1392295.
Gurven, M.D., Lieberman, D.E., 2020. WEIRD bodies: mismatch, medicine and missing diversity. Evol. Hum. Behav. 41 (5), 330–340. https://doi.org/10.1016/j.evolhumbehav.2020.04.001.
Heitz, R.P., 2014. The speed-accuracy tradeoff: history, physiology, methodology, and behavior. Front. Neurosci. 8, 150. https://doi.org/10.3389/fnins.2014.00150.
Henrich, J., Heine, S.J., Norenzayan, A., 2010. Most people are not WEIRD. Nature 466 (7302), 29. https://doi.org/10.1038/466029a.
Humphreys, K., Avidan, G., Behrmann, M., 2007. A detailed investigation of facial expression processing in congenital prosopagnosia as compared to acquired prosopagnosia. Exp. Brain Res. 176, 356–373. https://doi.org/10.1007/s00221-006-0621-5.
Jenkins, R., White, D., Van Montfort, X., Burton, A.M., 2011. Variability in photos of the same face. Cognition 121 (3), 313–323. https://doi.org/10.1016/j.cognition.2011.08.001.
Laajaj, R., Macours, K., Hernandez, D.A.P., Arias, O., Gosling, S.D., Potter, J., Vakis, R., 2019. Challenges to capture the big five personality traits in non-WEIRD populations. Sci. Adv. 5 (7) https://doi.org/10.1126/sciadv.aaw5226 eaaw5226.
Luce, D., 1986. Response Times: Their Role in Inferring Elementary Mental Organization. Oxford University Press, New York, NY.
Marotta, J.J., McKeeff, T.J., Behrmann, M., 2002. The effects of rotation and inversion on face processing in prosopagnosia. Cogn. Neuropsychol. 19 (1), 31–47. https://doi.org/10.1080/02643290143000079.
Masuda, T., Batdorj, B., Senzaki, S., 2020. Culture and attention: future directions to expand research beyond the geographical regions of WEIRD cultures. Front. Psychol. 11, 1394. https://doi.org/10.3389/fpsyg.2020.01394.
McConachie, H.R., 1976. Developmental prosopagnosia. A single case report. Cortex 12 (1), 76–82. https://doi.org/10.1016/S0010-9452(76)80033-0.
Moreton, R., 2021. Forensic face matching: procedures and application. In: Bindemann, M. (Ed.), Forensic Face Matching: Research and Practice. Oxford University Press. https://doi.org/10.1093/oso/9780198837749.003.0007.
Nador, J.D., Ramon, M., 2021. Harnessing fast periodic visual stimulation to study face cognition: sub-processes, brain-behavior relationships, and objectivity. Eur. J. Neurosci. https://doi.org/10.1111/ejn.15115.
Nador, J.D., Zoia, M., Pachai, M.V., Ramon, M., 2021a. Psychophysical profiles in super-recognizers. Sci. Rep. 11 (1), 1–11. https://doi.org/10.1038/s41598-021-92549-6.
Nador, J.D., Alsheimer, T.A., Gay, A., Ramon, M., 2021b. Image or identity? Only Super-Recognizers'(memor) ability is consistently viewpoint-invariant. Swiss Psychol. Open: Off. J. Swiss Psychol. Soc. 1 (1), 2. https://doi.org/10.5334/spo.28.
Nielsen, M., Haun, D., Kärtner, J., Legare, C.H., 2017. The persistent sampling bias in developmental psychology: a call to action. J. Exp. Child Psychol. 162, 31–38. https://doi.org/10.1016/j.jecp.2017.04.017.
Nunn, J.A., Postma, P., Pearson, R., 2001. Developmental prosopagnosia: should it be taken at face value? Neurocase 7 (1), 15–27. https://doi.org/10.1093/neucas/7.1.15.
Orban de Xivry, J.J., Ramon, M., Lefevre, P., Rossion, B., 2008. Reduced fixation on the upper area of personally familiar faces following acquired prosopagnosia. J. Neuropsychol. 2 (1), 245–268. https://doi.org/10.1348/174866407X260199.
Palermo, R., Rossion, B., Rhodes, G., Laguesse, R., Tez, T., Hall, B., et al., 2017. Do people have insight into their face recognition abilities? Q. J. Exp. Psychol. 70 (2), 218–233. https://doi.org/10.1080/17470218.2016.1161058.
Phillips, P.J., Beveridge, J.R., Draper, B.A., Givens, G., O'Toole, A.J., Bolme, D., et al., 2012. The good, the bad, and the ugly face challenge problem. Image Vis Comput. 30 (3), 177–185. https://doi.org/10.1016/j.imavis.2012.01.004.
Phillips, P.J., Yates, A.N., Hu, Y., Hahn, C.A., Noyes, E., Jackson, K., et al., 2018. Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. Proc. Natl. Acad. Sci. Unit. States Am. 115 (24), 6171–6176. https://doi.org/10.1073/pnas.1721355115.
Quaglino, A., Borelli, G., 1867. Emiplegia sinistra con amaurosi–guarigione–perdita totale della percezione dei colori e della memoria della configurazione degli oggetti. Giornale d'Oftalmologia Italiano 10, 106–117.
Quaglino, A., Borelli, G.B., Della Sala, S., Young, A.W., 2003. Quaglino's 1867 case of prosopagnosia. Cortex 39 (3), 533–540. https://doi.org/10.1016/s0010-9452(08)70263-6.
Rad, M.S., Martingano, A.J., Ginges, J., 2018. Toward a psychology of Homo sapiens: making psychological science more representative of the human population. Proc. Natl. Acad. Sci. Unit. States Am. 115 (45), 11401–11405. https://doi.org/10.1073/pnas.1721165115.
Ramon, M., Busigny, T., Gosselin, F., Rossion, B., 2017. All new kids on the block? Impaired holistic processing of personally familiar faces in a kindergarten teacher with acquired prosopagnosia. Vis. Cognit. 24, 321–355. https://doi.org/10.1080/13506285.2016.1273985.
Ramon, M., Rossion, B., 2010. Impaired processing of relative distances between features and of the eye region in acquired prosopagnosia–two sides of the same holistic coin? Cortex 46 (3), 374–389. https://doi.org/10.1016/j.cortex.2009.06.001.
Ramon, M., 2018. The power of how-lessons learned from neuropsychology and face processing. Cogn. Neuropsychol. 35, 83–86. https://doi.org/10.1080/02643294.2017.1414777.
Ramon, M., Busigny, T., Rossion, B., 2010. Impaired holistic processing of unfamiliar individual faces in acquired prosopagnosia. Neuropsychologia 48 (4), 933–944. https://doi.org/10.1016/j.neuropsychologia.2009.11.014.
Ramon, M., Caharel, S., Rossion, B., 2011. The speed of recognition of personally familiar faces. Perception 40 (4), 437–449. https://doi.org/10.1068/p6794.
Ramon, M., Gobbini, M.I., 2018. Familiarity matters: a review on prioritized processing of personally familiar faces. Vis. Cognit. 26 (3), 179–195. https://doi.org/10.1080/13506285.2017.1405134.
Ramon, M., Sokhn, N., Lao, J., Caldara, R., 2018. Decisional space determines saccadic reaction times in healthy observers and acquired prosopagnosia. Cogn. Neuropsychol. 35, 304–313. https://doi.org/10.1080/02643294.2018.1469482.

Ramon, M., Sokhn, N., Caldara, R., 2019a. Decisional space modulates visual categorization - evidence from saccadic reaction times. Cognition 186, 42–49. https://doi.org/10.1016/j.cognition.2019.01.019.
Ramon, M., Bobak, A.K., White, D., 2019b. Super-recognizers: from the lab to the world and back again. Br. J. Psychol. 110 (3), 461–479. https://doi.org/10.1111/bjop.12368.
Ramon, M., 2021. Super-Recognizers–a novel diagnostic framework, 70 cases, and guidelines for future work. Neuropsychologia, 107809. https://doi.org/10.1016/j.neuropsychologia.2021.107809.
Ramon, M., Van Belle, G., 2016. Real-life experience with personally familiar faces enhances discrimination based on global information. PeerJ e1465. https://doi.org/10.7717/peerj.1465.
Rezlescu, C., Barton, J.J.S., Pitcher, D., Duchaine, B., 2014. Normal acquisition of expertise with greebles in two cases of acquired prosopagnosia. Proc. Natl. Acad. Sci. 111 (14), 5123–5128. https://doi.org/10.1073/pnas.1317125111.
Robertson, D.J., Noyes, E, Dowsett, A.J., Jenkins, R., Burton, A.M., 2016. Face recognition by metropolitan police super-recognisers. PloS One 11 (2), e0150036. https://doi.org/10.1371/journal.pone.0150036.
Rosenthal, G., Tanzer, M., Simony, E., Hasson, U., Behrmann, M., Avidan, G., 2017. Altered topology of neural circuits in congenital prosopagnosia. Elife 6, e25069. https://doi.org/10.7554/eLife.25069.
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. https://doi.org/10.1093/brain/awg241.
Rossion, B., 2014. Understanding face perception by means of prosopagnosia and neuroimaging. Front. Biosci. 6 (258), e307.
Rossion, B., 2018. Prosopdysgnosia? What could it tell us about the neural organization of face and object recognition? Cogn. Neuropsychol. 35 (1–2), 98–101. https://doi.org/10.1080/02643294.2017.1414778.
Rossion, B., Michel, C., 2018. Normative accuracy and response time data for the computerized Benton Facial Recognition Test (BFRT-c). Behav. Res. Methods 50 (6), 2442–2460. https://doi.org/10.3758/s13428-018-1023-x.
Russell, R., Duchaine, B., Nakayama, K., 2009. Super-recognizers: people with extraordinary face recognition ability. Psychon. Bull. Rev. 16 (2), 252–257. https://doi.org/10.3758/PBR.16.2.252.
Schiltz, C., Sorger, B., Caldara, R., Ahmed, F., Mayer, E., Goebel, R., Rossion, B., 2006. Impaired face discrimination in acquired prosopagnosia is associated with abnormal response to individual faces in the right middle fusiform gyrus. Cerebr. Cortex 16 (4), 574–586. https://doi.org/10.1093/cercor/bhj005.
Smith, H.M.J., Andrews, S., Baguley, T.S., Colloff, M.F., Davis, J.P., White, D., Rockey, J. C., Flowe, H.D., 2021. Performance of typical and superior face recognizers on a novel interactive face matching procedure. Brit. J. Psychol. 112 (4), 964–991. https://doi.org/10.1111/bjop.12499.
Smith, P.L., Little, D.R., 2018. Small is beautiful: In defense of the small-*N* design. Psycho. Bull. Rev. 25, 2083–2101. https://doi.org/10.3758/s13423-018-1451-8.
Sorger, B., Goebel, R., Schiltz, C., Rossion, B., 2007. Understanding the functional neuroanatomy of acquired prosopagnosia. Neuroimage 35 (2), 836–852. https://doi.org/10.1016/j.neuroimage.2006.09.051.
Stacchi, L., Huguenin-Elie, E., Caldara, R., Ramon, M., 2020. Normative data for two challenging tests of face matching under ecological conditions. Cognit. Res.: Princip. Imp. 5 (1), 1–17. https://doi.org/10.1186/s41235-019-0205-0.
Stantic, M., Brewer, R., Duchaine, B., Banissy, M.J., Bate, S., Susilo, T., et al., 2021. The Oxford Face Matching Test: a non-biased test of the full range of individual differences in face perception. Behav. Res. Methods 1–16. https://doi.org/10.3758/s13428-021-01609-2.
Tardif, J., Morin Duchesne, X., Cohan, S., Royer, J., Blais, C., Fiset, D., et al., 2019. Use of face information varies systematically from developmental prosopagnosics to super-recognizers. Psychol. Sci. 30 (2), 300–308. https://doi.org/10.1177/0956797618811338.
Thomas, C., Avidan, G., Humphreys, K., Jung, K.J., Gao, F., Behrmann, M., 2009. Reduced structural connectivity in ventral visual cortex in congenital prosopagnosia. Nat. Neurosci. 12 (1), 29–31. https://doi.org/10.1038/nn.2224.
Towler, A., White, D., Kemp, R.I., 2017. Evaluating the feature comparison strategy for forensic face identification. J. Exp. Psychol. Appl. 23 (1), 47. https://doi.org/10.1037/xap0000108.
Towler, A., Kemp, R.I., Burton, A.M., Dunn, J.D., Wayne, T., Moreton, R., White, D., 2019. Do professional facial image comparison training courses work? PLoS One 14 (2), e0211037. https://doi.org/10.1371/journal.pone.0211037.
Towler, A., Keshwa, M., Ton, B., Kemp, R.I., White, D., 2021. Diagnostic feature training improves face matching accuracy. J. Exp. Psychol. Learn. Mem. Cognit. https://doi.org/10.1037/xlm0000972.
Townsend, J.T., Ashby, F.G., 1978. Methods of modeling capacity in simple processing systems. Cognit. Theory 3, 199.
Warrington, E.K., 1984. Recognition Memory Test. NFER- NELSON, Windsor (UK).
White, D., Phillips, P.J., Hahn, C.A., Hill, M., O'Toole, A.J., 2015. Perceptual expertise in forensic facial image comparison. Proc. Biol. Sci. 282 (1814), 20151292. https://doi.org/10.1098/rspb.2015.1292.
White, D., Rivolta, D., Burton, A.M., Al-Janabi, S., Palermo, R., 2017. Face matching impairment in developmental prosopagnosia. Q. J. Exp. Psychol. 70 (2), 287–297. https://doi.org/10.1080/17470218.2016.1173076.
White, D., Guilbert, D., Varela, V.P., Jenkins, R., Burton, A.M., 2021. GFMT2: a psychometric measure of face matching ability. Behav. Res. Methods 1–9. https://doi.org/10.3758/s13428-021-01638-x.