CORTEX 166 (2023) 348–364

Available online at www.sciencedirect.com

ScienceDirect

Journal homepage: www.elsevier.com/locate/cortex

Research Report

Data-driven studies in face identity processing rely on the quality of the tests and data sets

Anna K. Bobak [a,*,1], Alex L. Jones [b,**,1], Zoe Hilker [a], Natalie Mestry [c], Sarah Bate [c] and Peter J.B. Hancock [a]

[a] Psychology, Faculty of Natural Sciences, University of Stirling, United Kingdom
[b] School of Psychology, Swansea University, Swansea, United Kingdom
[c] Department of Psychology, Bournemouth University, United Kingdom

ARTICLE INFO

Article history:
Received 18 August 2022
Reviewed 25 October 2022
Revised 1 February 2023
Accepted 28 May 2023
Action editor Holge Wiese
Published online 30 June 2023

Keywords:
Face identity processing (FIP)
Face perception
Face memory
Individual differences
Principal component analysis
Agglomerative clustering

ABSTRACT

There is growing interest in how data-driven approaches can help understand individual differences in face identity processing (FIP). However, researchers employ various FIP tests interchangeably, and it is unclear whether these tests 1) measure the same underlying ability/ies and processes (e.g., *confirmation* of identity match or *elimination* of identity match) 2) are reliable, 3) provide consistent performance for individuals across tests online and in laboratory. Together these factors would influence the outcomes of data-driven analyses. Here, we asked 211 participants to perform eight tests frequently reported in the literature. We used Principal Component Analysis and Agglomerative Clustering to determine *factors underpinning performance*. Importantly, we examined the *reliability* of these tests, *relationships between them*, and quantified *participant consistency across tests*. Our findings show that participants' performance can be split into two factors (called here *confirmation* and *elimination* of an identity match) and that participants cluster according to whether they are strong on one of the factors or equally on both. We found that the reliability of these tests is at best moderate, the correlations between them are weak, and that the consistency in participant performance across tests and is low. Developing reliable and valid measures of FIP and consistently scrutinising existing ones will be key for drawing meaningful conclusions from data-driven studies.

* *Corresponding author*. Psychology, Faculty of Natural Sciences, University of Stirling, Stirling FK9 4LA, United Kingdom.
** *Corresponding author*. School of Psychology, Faculty of Medicine, Health and Life Science, Swansea University, Swansea, SA2 8PP, United Kingdom.
E-mail addresses: a.k.bobak@stir.ac.uk (A.K. Bobak), alex.l.jones@swansea.ac.uk (A.L. Jones).
[1] AKB and ALJ are joint first and corresponding authors. For queries pertaining to data analysis, please contact ALJ.
https://doi.org/10.1016/j.cortex.2023.05.018

1. Introduction

Data driven approaches in face identity processing (FIP) utilize large datasets and computational techniques such as Principal Component Analysis, Factor Analysis, or clustering to gain insights into how humans process facial information. Several studies have been published using this approach pertaining to the specificity of face recognition (Hildebrandt, Wilhelm, Schmiedek, Herzmann, & Sommer, 2011; Ćepulić, Wilhelm, Sommer, & Hildebrandt, 2018), trait impressions from faces (Sutherland & Young, 2022; Vernon, Sutherland, Young, & Hartley, 2014), Social and Emotion Perception (Jones & Kramer, 2021), and image characteristics (Burton, Kramer, Ritchie, & Jenkins, 2016; Hancock, Burton, & Bruce, 1996). In Face Identity Processing (FIP) this is a relatively new approach in studies with typical and superior (Baker, Stabile, & Mondloch, 2023; Bobak, Mileva, & Hancock, 2019; Nador, Vomland, Thielgen, & Ramon, 2022; Verhallen et al., 2017) as well as impaired (Bennetts et al., 2022; DeGutis et al., 2022; Lowes, Hancock, & Bobak, 2023) perceivers.

The data-driven approach allows for more unconstrained analysis of large data sets—an algorithm splits or clusters data based on a common underlying characteristic which can offer unique insight into processes driving participants' performance. This is unlike the traditional top-down hypothesis-driven approach where any analysis is constrained by an *a priori* prediction. For example, recent studies using a range of FIP tests reported clusters of participants in typical (Baker et al., 2023) and developmental prosopagnosia (Bennetts et al., 2022; DeGutis et al., 2022) populations corresponding to different perceptual impairments (Bennetts et al., 2022; DeGutis et al., 2022) or response bias (Baker et al., 2023). However, the quality, i.e., reliability of the FIP tests and consistency of performance of participants on these tests is largely unknown because they were originally developed for studies examining either isolated effects in a specific test or small groups of typical, superior, or impaired perceivers, or both. How do these tests perform together, when administered as a large battery with large number of participants, in the laboratory and online is not known. Specifically, would these tests rank participants *consistently* in terms of their general FIP abilities? This largely depends on their internal reliability that can impact across-task consistency in most participants. Indeed, initial evidence suggests that some commonly used tests may be, in fact, sub-optimal for individual-difference analyses and across-task comparison (Fysh, Stacchi, & Ramon, 2020; Stacchi, Huguenin-Elie, Caldara, & Ramon, 2020). In this study, using 'big data' and data-driven approach, we aim to answer some of these important questions pertinent to the rapidly growing data-driven psychological research.

1.1. *Assessment of individual differences in face identity processing (FIP)*

Valid and reliable testing of face identity processing (FIP) in typical perceivers (Bate et al., 2018; Bobak, Pampoulov, & Bate, 2016; Dunn, Summersby, Towler, Davis, & White, 2020; Fysh et al., 2020; Ramon, 2021; Stacchi et al., 2020; Stantic et al., 2021), super-recognisers (e.g., Bate, Portch, & Mestry, 2021; Ramon, 2021) and developmental prosopagnosics (White et al., 2017a; 2017b, 2021) has recently attracted a lot of interest amongst researchers. FIP is typically assessed by examining face perception (minimal memory demand) and face memory (sub) processes. The evidence for this perceptual and mnemonic distinction comes from individuals with acquired and developmental prosopagnosia exhibiting problems at the early encoding (Biotti, Gray, & Cook, 2019), or later, discrimination (Fysh & Ramon, 2022; White et al., 2017a, 2017b) stage.

This distinction also exists at the high end of the FIP ability (Bate et al., 2018; Bobak, Hancock, & Bate, 2016). The so-called super-recognisers (SRs) have been described as heterogenous in their presentation (Bobak, Bennetts, Parris, Jansari, & Bate, 2016), with some excelling at face perception only (Bate et al., 2018; Bobak, Hancock, & Bate, 2016), and others excelling at both face perception and memory (Bate et al., 2018; Bobak, Dowsett, & Bate, 2016; Bobak, Hancock, & Bate, 2016; Ramon, 2021). Thus, although face recognition and face perception necessitate terminological and methodological distinction in individual differences studies (Ramon, 2018; Ramon & Gobbini, 2018), in practice this 'gold standard' has not always been adopted in literature (Belanova, Davis, & Thompson, 2018; Bobak, Pampoulov, & Bate, 2016; Phillips et al., 2018) and researchers use various tests to examine the sub-processes. For example, Super-recogniser and forensic experts to date have often undertaken the Cambridge Face memory Test (CFMT+, Russell, Duchaine, & Nakayama, 2009), Cambridge Face Perception Test (CFPT, Duchaine, Germine, & Nakayama, 2007; Russell et al., 2009), the Glasgow Face Matching Test (Burton, White, & McNeill, 2010), or a combination of these tests. Occasionally, SRs have been included in experimental groups because of their professional activities, such as being a part of a police force (Phillips et al., 2018; Robertson, Noyes, Dowsett, Jenkins, & Burton, 2016). Some early research into superior face-recognition employed just one test to select high performers (Bobak, Bennetts, et al., 2016; Bobak, Dowsett, & Bate, 2016; Bobak, Hancock, & Bate, 2016; Phillips et al., 2018).

For example, Phillips et al. (2018) set out to compare face matching performance of then leading computer algorithms with several professional groups, SRs, and students, recruited using various criteria. The expert groups, forensic and fingerprint examiners, were recruited in their workplaces (e.g., Police). The SRs were identified using *either* a CFMT + score over 90 points, *or* a GFMT score of over 90% (38 items correct or above), *or* a professional activity as a SR. Individuals fulfilling *only one of these conditions* were included in the experimental group ($N = 13$). These inclusion criteria rest on the assumption that these laboratory tests and professional activity are equivalent to one another in terms of one unique underlying FIP ability. Recent evidence from the individual differences literature suggests that this may not be the case and the *selection of the tests is critical for the results and conclusions of a study* (Bate et al., 2021; Fysh et al., 2020; Stacchi et al., 2020).

At least three recent studies with typical perceivers showed that multiple tests, even when used in combination, are hard to interpret (Bate et al., 2018; Fysh et al., 2020; Stacchi

et al., 2020). Stacchi et al. (2020) administered five FIP tests to over 200 participants in controlled laboratory conditions. The authors found that while the data from three tests (the CFMT+, the Yearbook Test – YBT, and the Facial Identity Card Sorting Test – FICST) converged in terms of participants' ranking, the two challenging matching tasks, Expertise in Facial Comparison Test (EFCT) and Person Identification Challenge (PICT; White, Phillips, Hahn, Hill, & O'Toole, 2015), did not predict performance on the other tests. Specifically, the top performers on EFCT and PICT scored in the full range of the remaining tests (above and below the medians).

In another study, Bate et al. (2018) examined the *consistency of performance* across different applied screening tests. They screened 200 people using the CFMT+ and three new applied tests of face processing: Models Memory Test (MMT), Pair Matching Test (PMT), and a facial composite (EVO-Fit) to lineup matching test. Approximately 60% of people showed consistency in their performance across all tests. Interestingly, the two tests examining face memory were weakly correlated ($r = .146$, $p = .039$) in spite of seemingly testing the same FIP sub-process (memory; see also Fysh et al., 2020 for similar conclusions). The MMT used different stimuli (faces) to the CFMT+ and included target-absent trials, where participants had to decide that a learned face is not present in an array of three faces. This is unlike the CFMT + where a target face is always present. Therefore, although in principle testing seemingly the same sub-process–face memory–the two memory tests in the Bate et al. require encoding and retrieval of different types of perceptual information to correctly solve target present and target absent trials (see Boudry, Nador, & Ramon, 2023 for a detailed discussion of the role of target prevalence in FIP tests).

Bate and colleagues suggested adopting an index score to capture average performance – a composite of all tests that informs about one's general ability (see also Lowes et al., 2023; Royer, Blais, Gosselin, Duncan, & Fiset, 2015). The use of an index face score is supported by the recent evidence suggesting that across various face processing tasks, there is an underlying common factor *f*, akin to the general intelligence *g* (Verhallen et al., 2017). As such, all face processing tasks may be underpinned by the same *f* ability and thus performance should be best predicted from it.

1.2. *Reliability and validity of assessment*

These initial results, however, do not address the implications of the reliability and validity (c.f. Mayer & Ramon, 2022) of used tests for the analyses (though DeGutis et al., 2022, note that the Cambridge Face Perception Test (CFPT) has only acceptable reliability). This is problematic, because lack of reliability can lead to uninterpretable results in individual differences.

Indeed, specific tests of FIP may be *unreliable*, i.e., the results across items in a test are inconsistent, which impacts on the maximum possible strength of correlations between the tests. Using an example from Verhallen et al. (2017) where the internal reliability (α) of a holistic processing task was .53 and the reliability of the GFMT was .91, *the maximum expected correlation between these tests would be* $(.53 \times .91) = .69$. Good reliability is an inherent part of intelligence and personality research–for example, the subtests of the Wechsler Abbreviated Scale of Intelligence, Second Edition (WASI-II) have an excellent internal reliability ranging from .90 to .92 in an adult sample (McCrimmon & Smith, 2013), yet reliability has not typically been scrutinised in FIP tasks (c.f., Bowles et al., 2009; Richler, Floyd, & Gauthier, 2014), and definitely *not in every study and every sample*. Importantly for the data-driven approaches, such error variance from an unreliable test can impact the quality of PCA (Bailey, 2012) and clustering by increasing the contribution of the error variance to all the components of the PCA, spreading explained variance across more components. However, it is important to note that as performance on a test approaches the ceiling (either when a test is too easy, or the participants' FIP is superior), the reliability for such a test, or a sample may be artificially inflated.

More than only the 'pure' FIP ability may also be needed to accurately complete any given test. Participants may adopt different response strategies depending on the test difficulty, structure of the response options, and own response biases. The Bate et al. (2018) and Fysh et al. (2020) studies show this may be the case based on the modest correlations between two memory based tasks (one with a 'target absent' option, and one without). While the eyewitness literature widely acknowledges that differences between 'choosers' (i.e. those who always make a positive identification in a line-up) and 'non-choosers' exist (Wixted & Wells, 2017), and that an ideal test of eyewitness accuracy should incorporate measures of FIP ability and proclivity to choose (Baldassari, Kantner, & Lindsay, 2019), this issue is rarely acknowledged in the fundamental FIP research (c.f., Baker et al., 2023). However, a recent review by Bindemann and Burton (2021) suggested that acknowledging decision making strategies is important when assessing performance on face matching tasks, while Boudry et al. (2023) provide the first consistent examination of the role of target prevalence in FIP tests. This is particularly important given the evidence that performance on different sub-components of even simple matching tests is unrelated (Megreya & Burton, 2007).

1.3. *Differences between online and laboratory assessment*

Finally, while some individual differences data are acquired in laboratories under controlled conditions (Bennetts, Mole, & Bate, 2017; Bobak, Bennetts, et al., 2016), other studies rely on online data collection methods (Baker et al., 2023; Bate et al., 2018; Davis, Lander, Evans, & Jansari, 2016) and yet comparative analyses for these two sources of data are rarely reported in FIP literature. A thorough examination of online testing methods is long overdue given the shift to that environment over the last three years. To address these important issues we examined: 1) the directly unobservable factors that may underpin FIP performance using data-driven approaches: Principal Components Analysis (PCA) and Hierarchical Agglomerative Clustering 2) the *relationship* between and *reliability* of most commonly used FIP tests, 3) and the *consistency* in performance within participants.

2. Method

The study was approved by the University I and University II institutional Ethics Committees. Below, we report how we determined our sample size, all data exclusions, all inclusion/exclusion criteria, whether inclusion/exclusion criteria were established prior to data analysis, all manipulations, and all measures (tests) in the study. No part of the study procedures or analysis plans was preregistered prior to the research being conducted.

2.1. Participants

We obtained two samples of data on the same tasks; one in the laboratory, and one online. All participants were White. In the laboratory sample, we recruited 108 students and visitors at two UK institutions (University I, $N = 88$; University II, $N = 20$). Participants' demographic data are summarised in Table 1. Given the planned analyses (PCA and internal reliability) and the number of tests ($N = 7$), we anticipated the study needing *a minimum* of 140 participants. This estimation was based on a conservative requirement for minimum sample size being 20 times larger than the number of variables in an Exploratory Factor Analysis (Mundfrom et al., 2005). We thus opted to recruit the maximum possible number of participants given the available funding to maximise the statistical power of analyses.

The online recruitment was fully managed by University II via Prolific. Although the age of the online sample was significantly higher than the lab sample, t (208) = 5.68, $p < .001$, $d = .78$, we decided to collapse the two samples for consistency and to give more observations to the principal component and clustering analyses. We justified this decision because a) the data distributions for all tasks were largely overlapping and b) the relationship between age and face recognition ability in adults is small and linear (Susilo, Germine, & Duchaine, 2013), c) any underlying meaningful difference would be identified using the opted for data driven approach, and also that these approaches have better performance with larger datasets.

Because we were interested in individual differences we elected to remove participants only if their performance indicated inattention, repetitive button pressing, or misunderstanding of instructions, removing participants who had more than one "error" across all the tasks. We defined an error as a performance level that is indicative of a failure to attend on a task. In detail, an error was classified when a participant had a score of less than .05 on the target absent *or* less than .05 on the target present conditions of the one-in-ten task. For the PICT, KFMT, GFMT, and the upright and inverted versions of the EFCT, we classified an error as a participant scoring less than .2 on the match or mismatch conditions, *or* scoring less than .4 on the match *and* mismatch conditions. The number of errors were summed across tasks and participants excluded if they had more than one. The exact code for exclusion is available in the online materials (https://osf.io/5k9ny/). This resulted in excluding six participants from the online sample, while none were excluded from the lab sample, a removal of 5.45% of online data or 2.84% of the full sample, which is in line with exclusions in recent studies with online participants (e.g., Carragher & Hancock, 2022). Participant recruitment was restricted to the 18–35 age range with the aim of examining data relatively unaffected by cognitive ageing. All participants gave an informed consent and were reimbursed £10 for their time.

2.2. Materials and procedure

2.2.1. Tests

The paradigms and specific tests used in this study are shown in Fig. 1 and the text below. For more information, we refer readers to the specific papers that implemented these tasks (cited in this paper). These tests cannot be publicly shared by the authors due to data protection restrictions, but we include a Wiki page on how materials were obtained https://osf.io/5k9ny/wiki/Materials/.

2.2.1.1. Cambridge face memory test: long form (Russell et al., 2009). In this task, participants are asked to memorise six male faces (wearing beanie hats so the external features, i.e., hair and ears) are invisible. The test consists of 102 trials and is split into four parts. In part one, participants are presented with three views of the learned face and must subsequently pick from a lineup of three faces (containing the same image). In part two, participants must pick from lineups of three faces, but the images are novel. Parts three and four are identical, but the images are overlaid with visual noise and the faces are in varying poses and face expressions. Part four images also show hair. There are two 20 s long reviews of all learned faces after part one and after part two. All images are presented in grayscale. The maximum correct score is 102 and the chance level is at 33.3%.

2.2.1.2. The Glasgow face memory test: short (Burton et al., 2010). In this task, participants are presented with two images of faces side by side and have to decide whether they show the same person ('matched' pairs) or two different

Table 1 – Summary of the demographic data.

Sample origin	Lab	Online
N recruited	108	110
N excluded	1 (experimenter error)	6 (based on performance)
N retained (final sample)	107	104
Mean Age (*SD*)	23.4 (4.3)	27.1 (5.2)
Gender[a]	58 males, 49 females	53 males, 56 females, 1 unknown
Handedness	13 left	14 left, 5 ambidextrous

[a] The age differences between the genders were *ns* in both samples (one participant who did not disclose their gender was removed from *only this* analysis).

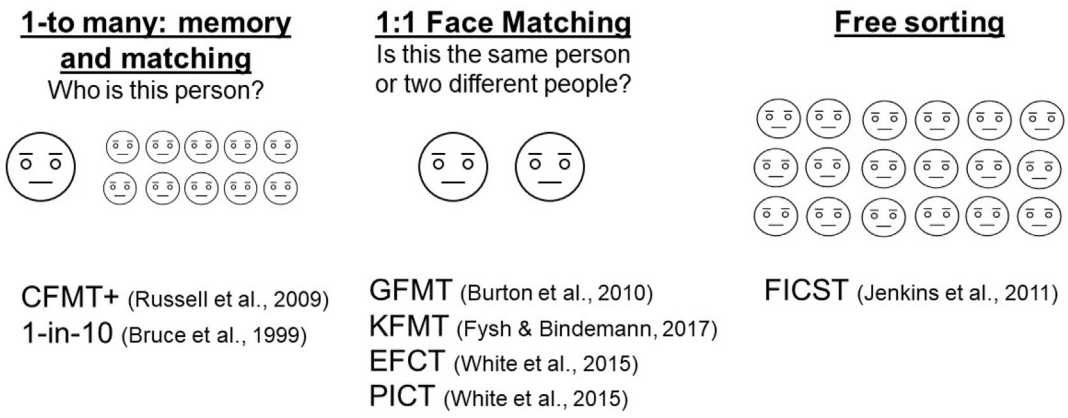

Fig. 1 – Schematic figure paradigms and specific tests participants attempted in this study.N.B. the EFCT was shown to participants in both upright and inverted versions.

people ('mismatched' pairs) and respond by pressing a keyboard key ('s' or 'k'). The images were taken on the same day by different cameras and are in grayscale to avoid matching on basic visual features such as blemishes or moles. There are 40 trials, half are matched (they show two images of the same person) and half are mismatched (they show two images of different people). There is no time limit to respond, and the chance level is at 50%.

2.2.1.3. One in 10 test (Bruce et al., 1999). In this version of the test, each trial consists of a target face extracted from a video recording displaying a male face from a 30° angle. This 'target' image is presented with a 5 × 2 array of faces below (ten images). All images are presented in colour. The target images measure 216 × 263 pixels and the arrays measure 1050 × 700 pixels. There are 80 trials and half of the trials are target present (the 'target' person appears on one of the images in the array) and half of the trials are target absent (the 'target' person does not appear in the array). All images in the array are numbered from 1 to 10 and participants have to respond with the corresponding number key if they think that the target is present (they press 0 for face number 10) or press the space bar if they think that the 'target' does not feature in the array. The chance level is at 10% in target present trials.

2.2.1.4. Expertise in facial comparison test upright (EFCT-U). This test, developed by White et al. (2015), examines face matching ability. It comprises a total of 84 trials, half of which are matched and half mismatched. The pairs are displayed for a maximum of 30 s and participants can respond during the 30 s or after images disappear. We retained the original response options which were as follows: (i) sure they are the same person; (ii) think they are the same person; (iii) do not know; (iv) think they are different people; and (v) sure they are different people. We adopted the scoring method used in previous studies (Kramer, Jones, & Gous, 2021; O'Toole et al., 2007); responses 1 and 2 were converted to 'same' judgments and 3, 4, and 5 as 'different' judgments (n.b., the EFCT as developed by White et al. (2015) intending to use AUC as the dependent measure).

2.2.1.5. Kent face matching test (KFMT): short (Fysh & Bindemann, 2018). This version of the task consists of 40 identity pairs (20 males and 20 females). Twenty pairs show the same identity ('matched' pairs) and 20 show different identities ('mismatched' pairs). One image from each pair came from the Kent Unfamiliar Face Database and were taken under controlled conditions, akin to a document photograph. These images measure 1050 × 700 pixels. The other image in each pair is a student ID photograph which varies in pose, expression, and lighting. These images measure 142 × 192 pixels. Participants respond by a keypress ('*m*' or 'x'). There is no time limit and the chance level is 50%.

2.2.1.6. Expertise in facial comparison test inverted (EFCT-I). The form of this test is identical to the EFCT-U except that a new set of 84 pairs of images are presented upside down. We adopted the same scoring method as in EFCT-U. We included this test to calculate the index of holistic processing which was the aim of another analysis (see Berger, Fry, Bobak, Juliano, & DeGutis, 2022) – we report the results here for completeness but do not discuss them.

2.2.1.7. Person identification challenge test (PICT). The PICT is a test of face perception. It is exactly the same in its procedure and response options to the EFCT (White et al., 2015). The people presented in the images in the test are taken from a greater distance and with more contextual information (i.e., the background). We adopted the same scoring method as in EFCT-U.

2.2.1.8. Facial identity card sorting test (FICST). This face perception test originally adopted from Jenkins, White, Van Montfort, & Mike Burton, 2011), and recently used by Stacchi et al. (2020) involves a simultaneous card sorting procedure. The test aims to assess face matching across variable appearance. The faces are forty images of two Dutch celebrities (Chantel Janzen, Bridget Maasland) provided by Jenkins and Burton. All images measure 38 × 50 mm and are in grayscale. Participants were shown all images at once scattered on a flat surface and are instructed to sort them into as

many or as few identities they thought were in the pile. Jenkins et al. (2011) reported participants perceived on average 7.5 identities in the pile (replicated by Stacchi et al., $N = 218$, $M_{identities} = 7.6$). The experiment ended when participants informed the researcher that they had sorted the pile into the correct (in their opinion) number of identities. The number of IDs recorded was included in the analyses. We also recorded the number of errors, i.e., an image of one identity in a pile of predominantly the other identity. Because this test was not possible to conduct online at the time of testing and we collapsed data for the lab and online sample, we only report it in supplementary materials.[2]

All the tests were administered in the same (as described above) fixed order to minimize error variance. The FICST was not included in the online version of the study for practical reasons, i.e. it was not possible to implement it on Testable at the time of running the study (but see footnote 1). After completing the first four tests, participants were asked to take a 30 min long break. Testing took place in dimly lit cubicles using 19 inch monitors running 1280 × 1024 pixels resolution with refresh rate 60 Hz and 24 inch widescreen monitors running 1920 × 1080 pixels resolution with refresh rate 60 Hz at Universities I and II respectively. The laboratory testing was conducted by six Experimenters.

All online experiments were implemented in Testable and run via Prolific on participants' own devices (laptops and PCs, but not mobile phones and tablets). The devices and screen resolutions varied between participants. The consistent stimuli display size was ensured using Testable's calibration screen where participants have to press a credit-card sized card against their screen and adjust the length of a bar displayed in the browser using arrow keys on their keyboard to align with the size of the card.

3. Analytic strategy

Our three main aims to achieve in this study were to examine a) patterns of performance emerging amongst participants, b) overall reliability of the tests and the correlations between the tests c) consistency of performance within participants and We present the analysis plan for each of these aims below. All data (including not reported here reaction times) and code for analysis are available on our OSF page https://osf.io/5k9ny/. The OSF wiki further defines every file in the repository and every variable name and abbreviation in the files.

3.1. Patterns of performance: data driven approach

To examine the patterns of performance, we chose to implement Principal Components Analysis (PCA) and hierarchical agglomerative clustering, a form of cluster analysis (CA). Verhallen and colleagues reported in their paper that face processing ability is underpinned by a process akin to general intelligence (*g*) and dubbed it *f*. We sought to replicate this finding by performing PCA with all tests. As the correlations between match and mismatched performance tend to be low and these sub-components are often reported separately (e.g., Bobak et al., 2019), we split the matching tests (except for the CFMT+) by matched and mismatched performance, and the 1-in-10 test by target present and target absent performance, before z-score standardising the data and conducting the PCA.

Additionally, to investigate the presence of clusters of individuals within the dataset, we used hierarchical agglomerative clustering (Kaufman & Rousseeuw, 1990) with the retained principal components. This type of clustering groups observations together according to their similarity in Euclidean space. Briefly, hierarchical agglomerative clustering begins by considering each observation (here a participant and their scores on the respective tasks) as a single cluster. On each iteration, the two nearest clusters are merged into one larger group, until all clusters fall into a single group upon completion. Clusters were calculated according to Ward's method (Ward, 1963), which merges observations that have the smallest difference in error sums of squares between their respective clusters.

Once estimated, agglomerative clustering gives a tree-like structure that can be cut to contain a given number of clusters. To select the number of clusters to retain, we used the silhouette coefficient (Rousseeuw, 1987). This statistic compares the mean distance between an observation and all other observations within the same cluster, against the mean distance between that data point and all other points within the next-nearest cluster. We opted to use hierarchical agglomerative clustering as opposed to other methods like K-Means (Bennetts et al., 2022), because it is a deterministic algorithm that requires no random initialisation (unlike K-Means), and so is easily reproducible. In addition, it is transparent, with the estimated tree structure showing all possible cluster solutions, which is a unique advantage of the approach.

Clusters were estimated using the *agnes* method from the *cluster* package in the R programming language (Mächler, Rousseeuw, Struyf, Hubert, & Hornik, 2012). Data was Z-scored standardised before clustering, and exactly like the PCA, had all measures split by their match/nonmatch conditions (excluding the CFMT + which only has target present trials).

3.2. Reliability and correlations

The correlation strength depends, amongst other factors, on the internal reliability of tasks under scrutiny. The lower the internal reliability of these tasks, the lower the maximum expected correlation between them (relative to a perfect correlation of ±1). We calculated the maximum expected correlations (MEr) using the formula adopted by Verhallen et al. (2017), i.e., taking the square root of the product of the internal reliabilities (Cronbach's alpha) of each task $MEr = \sqrt{(\alpha 1 * \alpha 2)}$ (c.f. Nimon, Zientek, & Henson, 2012). To estimate the relationships between tasks, we employed Pearson product–moment correlations.

3.3. Consistency of the tests

Given the number of face recognition tasks in this study, we aimed to test for the consistency of performance of each

[2] The most up to date version of this task published in Ramon (2021) represents a subset of all images provided by Jenkins and colleagues and its new, online version should be requested from Prof. Dr. Ramon at the Applied Face Cognition Lab.

participant across the different tasks – more simply, to what extent does a high-scoring individual on one task also score highly on another? A statistic ideally suited to exploring this relationship is the intraclass correlation coefficient (ICC), which can summarise the extent to which participants maintain their rank orders across repeated measurements (Liljequist, Elfving, & Skavberg Roaldsen, 2019). More generally, the ICC represents a ratio of systematic variations between participants to individual and measurement noise. This is an important application, as existing literature often uses one or two tasks to classify participants as a super-recogniser or prosopagnosic, with little recourse to the reliability of FIP tests (c.f., Nador et al., 2022).

We submitted six tasks (CFMT+, GFMT, KFMT, PICT, EFCT upright, and One-in-Ten) to an ICC analysis, omitting the EFCT inverted. This is because, for practical reasons, we were interested in how consistent participants were in upright tasks. The ICC analysis can produce two values: *absolute agreement* and *consistency*. The absolute agreement tests the extent to which one may get the same score between the tasks (e.g., 75% on CFMT, GFMT etc.). The consistency ranks the individual subjects across all measures. For example ranking 1 on CFMT+ but 18 on PICT would indicate poor consistency in a subject. Conversely, ranking 1 or 60 across all tests would indicate good consistency. Although in, for example, medical sciences and repeated measurements the absolute value may be important (e.g., for repeated resting heart rate measurement in one subject), here, we were interested in scores across tasks with different ranges in performance, so we report consistency only. The analysis yields one value for the whole dataset. We used the cut-off of Koo & Li (2016) to interpret the strength of the ICC.

4. Results

4.1. Descriptive statistics

Table 2 shows summary statistics (means and SDs) for the FIP tests used in this study, split by the task origin (lab or online). The summary statistics broadly track previous literature suggesting that our data is not anomalous with respect to reported performance on these tests. However, as we lay out below, one should not be 'hung up' on descriptive statistics, given the reliability of the test and (in)consistency in performance (see Table 3).

Table 2 – Descriptive statistics across tasks and data collection origin. The CFMT+ score is the average of the raw number of trials correct, the FICST is the average of participants' FICST score: (the number of piles + number of errors) – 2 (optimal performance separating the pile to two identities without errors); all remaining tasks are the average of proportion correct responses.

Task	Lab		Online	
	Mean	SD	Mean	SD
CFMT+	69.60	11.52	67.51	11.93
EFCT inv	.68	.08	.60	.07
EFCT up	.81	.07	.74	.07
GFMT	.81	.10	.78	.15
KFMT	.64	.11	.61	.11
1-in-10	.62	.14	.54	.18
PICT	.75	.09	.61	.10
FICST	5.86	4.03	–	–

4.2. Data driven approaches

4.2.1. Principal Component Analysis

As indicated above, for the PCA, we chose to split performance, where possible, by matched and non-matched (and the target present and target absent in 1-in-10) trials. The PCA separated performance into two components explaining 50.8% of the variance in the dataset. In line with previous work indicating a lack of correlation between match and mismatch ability (e.g. Kokje, Bindemann, & Megreya, 2018; Megreya & Burton, 2007), these two components broadly reflected *match and target present* (PC1) and *non-match (nm) and target absent* (PC2) performance. The CFMT+ loaded on the PC1 (match performance). The distribution of tasks across two components and their loadings are presented in Fig. 2.[3]

4.2.2. Clustering

After fitting the hierarchical agglomerative clustering tree, we tested for the presence of clusters by examining the silhouette coefficient for candidate cluster numbers from one to ten. The highest silhouette coefficient was observed for three clusters, $SC = .16$, and the agglomerative coefficient (the dissimilarity of each observation to the first cluster it is merged with, divided by the dissimilarity of the merge in the final step) was .91, indicating a relatively good cluster solution. The cluster showed a larger group containing 60% of observations ($n = 123$), and two smaller clusters, one containing 23% ($n = 48$) and the final containing 17% ($n = 34$), see Fig. 3.

To interpret the clusters, we considered them in light of the principal component solution derived earlier. The largest cluster (60% of the sample) comprised individuals who had positive scores relatively close to zero on both components (cluster PC1 mean = .91, PC2 mean = .63), suggesting average performance on both matching and non-matching components. Cluster two, comprising around 23% of the sample, had low scores on PC1 (mean = −2.80) but positive middling scores on PC2 (mean = .87), while cluster three (17%) had negative scores close to zero on PC1 (mean = −.36), and low scores on PC2 (mean = −2.22). Given the interpretations of the PC loadings, this seems likely to indicate groups of individuals with average performance; poor match performance but normal mis-match performance, and the converse.

[3] By submitting the full dataset to the PCA we ignored the origin of the data. One advantage of PCA is that we may now find that the origin (lab/online) is correlated strongly with a component, indicating that the component may represent the source of variability in the data associated with the lab/online distinction. We tested for this here to see how much variance this may account for. The origin shows the strongest relationship with PC3 ($r = .47$), which explained around 8% of the variance. We opted to keep the first two components as they explain around 50% of the variability in the dataset, with each subsequent component after PC3 explaining less than 7% of the variance.

Table 3 – Pearson correlations, reliability (Cronbach's Alpha; α shown in the first row of each, lab and online samples), and maximum expected correlations (see the formula in the analytic strategy section). Values on the upper diagonal represent the maximum expected correlation, lower diagonal the correlation coefficient.

Origin	Task	CFMT+	EFCT Inv	EFCT Up	GFMT	KFMT	One-In- Ten	PICT
Lab	**α**	.884	.643	.662	.669	.597	.881	.554
	CFMT+	–	.754	.765	.769	.727	.882	.700
	EFCT Inv	.155[a]	–	.653	.656	.620	.753	.597
	EFCT Up	.304	.592	–	.666	.629	.764	.606
	GFMT	.474	.204	.441	–	.632	.767	.609
	KFMT	.308	.178[a]	.333	.328	–	.725	.575
	One-In -Ten	.302	.258	.440	.386	.274	–	.699
	PICT	.212	.421	.418	.285	.440	.283	–
Online	**α**	.884	.520	.690	.824	.597	.933	.517
	CFMT+	–	.678	.781	.854	.727	.909	.676
	EFCT Inv	.371	–	.599	.655	.557	.697	.519
	EFCT Up	.530	.441	–	.754	.642	.803	.597
	GFMT	.558	.484	.565	–	.702	.877	.653
	KFMT	.458	.330	.509	.428	–	.746	.555
	One-In-Ten	.540	.323	.433	.486	.363	–	.695
	PICT	.313	.343	.427	.329	.339	.201[a]	–

[a] Indicates non-significant ($p > .05$) correlations.

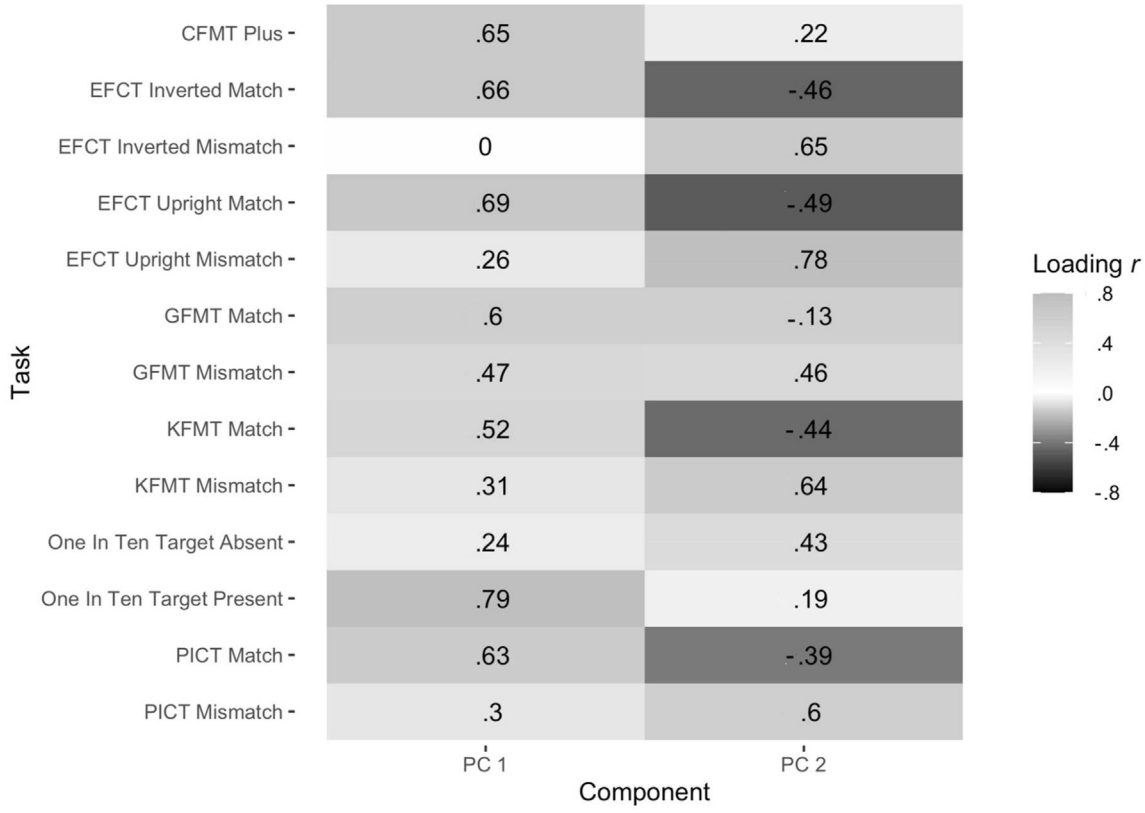

Fig. 2 – The PCA component loadings. The loadings represent the strength of the relationship of each task with the PCA1 and PCA2.

4.3. *Reliability and correlations*

We first correlated the performance of all participants on all accuracy scores in all tasks. The results are presented in Table 1. The individual *p* values are available on the OSF page. In the *FICST* task, we recorded the number of piles and mistakes, but not individual cards in these piles. It was thus impossible to calculate the reliability and the maximum expected correlations with this test.

4.4. *Consistency of performance*

We placed all tasks on the same scale for analysis, expressed as the proportion of trials correct. For the whole dataset, the

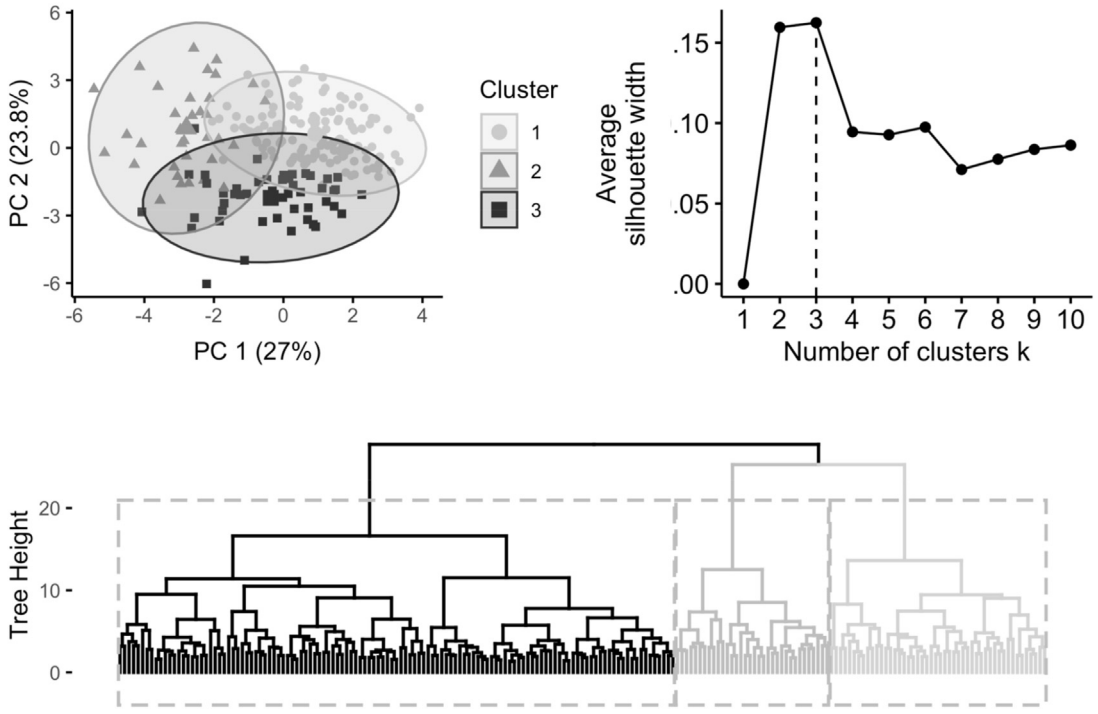

Fig. 3 – Top left: Individual data points plotted on the principal component axes and their cluster locations. Larger points represent the centroids of each cluster. Top right: The average silhouette coefficient per number of clusters. Bottom: Dendrogram of agglomerative clustering solution. Dashed lines represent cluster segregation (vertical) and height at which the dendrogram was cut (horizontal) to create the clusters.

consistency ICC was significant, ICC(C, 1) = .38 [.32, .45], F (208, 1040) = 4.70, $p < .001$. However, this estimate (.38) indicates poor consistency, falling below .50 (Koo & Li, 2016). Next, we examined the ICC in both the lab and online data to check for any divergences that may explain the result. For the lab data consistency was poor; ICC (C, 1) = .32 [.24, .41], F (104, 520) = 3.77, $p < .001$. For the online data, the consistency was marginally higher; ICC(C, 1) = .38 [.29, .47], F (103, 515) = 4.71, $p < .001$. We further opted to visualise these data to allow for cross-study comparisons, using methods described by Fysh et al. (2020).

First, we computed the rank sum for each participant across all tests by ranking performance on each individual test, and summing across these per participant. Ranking this summed ranking allowed us to extract the highest 2.5% (those with the lowest rank scores, i.e. better performance) the middle 2.5%, and the and lowest 2.5%, giving a view across the range of abilities. Radar plots (see Fig. 4) were created that mapped the relative performance abilities of these individuals (five per category) illustrating the variability at the top, middle, and bottom of the sample.

5. General discussion

In this study, we examined a large dataset ($N = 211$) with online and laboratory samples of young adults, using data driven approaches (Principal Component Analysis and Hierarchical Agglomerative Clustering) to determine patterns of performance across tasks. We found that the two main principal components reflected accuracy on matched and mismatched (Target Present and Target Absent) trials, respectively. The clustering further corroborated these findings by splitting participants into three clusters, reflecting similar performance on both Principal Components (bias-free responding), higher performance on PC1 (biased towards 'match' responses) or higher performance on PC2 (biased towards non-match response). We also showed that the internal reliability is mostly low in commonly used tests, the correlations between FIP tests are small to moderate, and the consistency of performance within participants (across tests) is poor (see also Nador et al., 2022). Interestingly these seemed irrespective of the origin of the data, i.e., the online data is not any more unreliable or inconsistent than data collected in the laboratory.

5.1. *Patterns of performance: Principal Component Analysis and clustering*

Our first aim was to investigate the patterns of performance in our sample. Verhallen et al. (2017) proposed that FIP ability is underpinned by a single process *f* akin to general intelligence (*g*). If this were true, we would expect the tasks to load strongly on one component. However, it has previously been reported that the correlations between match and mismatched trials in simultaneous matching tasks is low (Bobak et al., 2019) and thus match vs. mismatch performance may represent

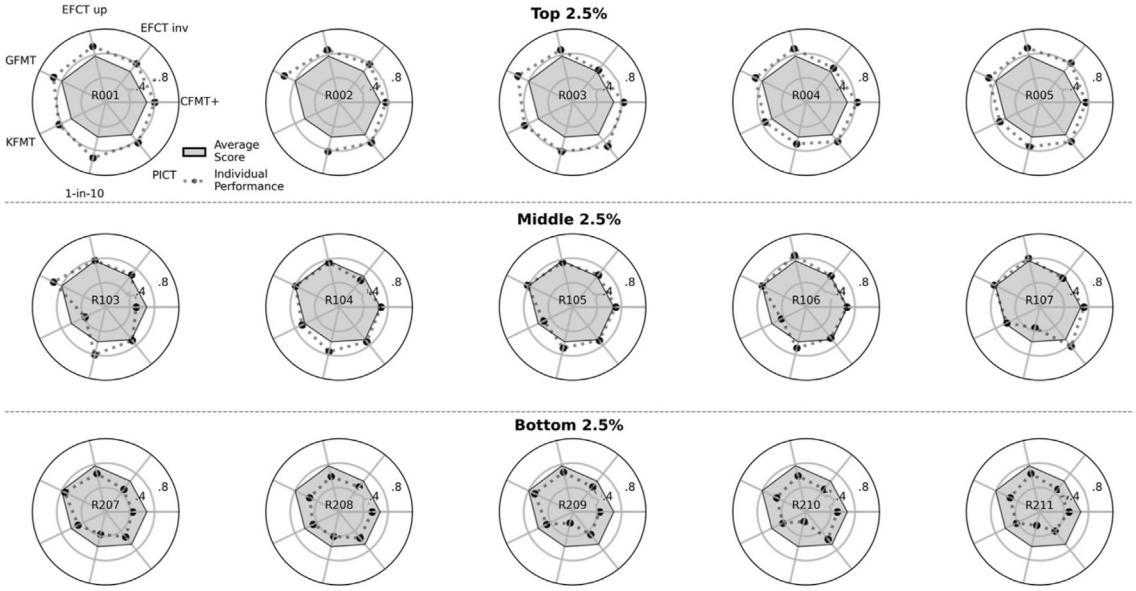

Fig. 4 – Radar plots of individual score profiles on each task at the top, middle, and bottom levels of performance of the dataset. The shaded area represents the average score in the dataset on each task, and the dots and dashed lines represent the individual participants' profile on that task. The rank sum – the rank of the sum across individual task rankings—is displayed centrally for the participant. We then took the rank scores of each participant on each variable, and computed a cosine similarity matrix between all possible pairs of scores. This approach is a geometric measure of the angle between participants' ranks on each variable, and functions much like a correlation coefficient. Those with a cosine similarity of one will have the same pattern of rank performance across each task, while those with zero will be maximally divergent. We sorted this similarity matrix by the rank sum across tasks (the variable used to select participants in Fig. 4), and visualised it as a heatmap in Fig. 5. This heatmap uses the cosine similarity between all possible pairs of participants' scores, producing a performance similarity matrix, and orders this by the rank sum variable. This illustrates the relationships among individuals according to their overall performance across each task.

different cognitive processes (see also, Berger et al., 2022) and/or decisional strategies or biases (Baker et al., 2023; Baldassari et al., 2019).

Indeed, the first two components (PC1 and PC2) of the unrotated solution explained nearly 51% of the variance in the dataset. For all the simultaneous matching tasks and the 1-in-10, the 'match' and 'target-present' sub-tests loaded strongly on PC1, the 'mismatch' and 'target-absent' sub-test loaded strongly on PC2 suggesting two main sub-processes underpinning participants' performance, we call these confirmation and elimination (of an identity match), respectively. The CFMT+ loaded more strongly on PC1, most likely reflecting the structure of this task where participants always have to choose a target on each trial.

This finding is consistent with the low reliability of most of the matching tasks. One of the reasons for insufficient internal reliability of a test is that it measures more than one construct (or here: sub-process; Cook & Beckman, 2006). This appears to be the case in most of the FIP tasks used here and suggests that the match and mismatch (or TP and TA in the 1-in-10 task) trials tap somewhat into different cognitive sub-processes. One possible explanation could be that 'match' trials are treated as a 'confirmation' task (i.e. "they look similar – are they both face A?") whereas 'mismatch' trials are treated as a elimination task (i.e. "they look quite different – is face A also a face B?"). Similarly for the 1-in-10 task, the TP trials may reflect detection (of a match amongst 10 foils) and confirmation, whether TA trials may involve a process of elimination.

Do participants differ with respect to how they approach these tasks? The clustering of the principal components further suggested that participants have somewhat different response preferences. While the largest cluster consisted of participants who scored averagely on both components (broadly corresponding to bias-free performance), two clusters indicated groups of people who are more prone to declare a pair of images or a lineup a 'mismatch' (*Cluster 2*) or a 'match' (*Cluster 3*), suggesting that routinely examining patterns of responding by separate analysis of different types of trials has a merit.

Additionally, we found limited evidence for the notion of *f* (a single general face-factor) proposed by Verhallen et al. (2017). This discrepancy may not be a result of perceptual processes, but instead reflect decisional strategies of participants and different tests used in ours and Verhallen et al.'s studies. Our study constitutes a conceptual, rather than direct replication of Verhallen et al.'s results (Wilson, Harris, & Wixted, 2020) and thus is not directly comparable.

Our results can also be conceptualised as participants having a consistent bias towards saying match or mismatch across tests. A signal detection analysis, which does separate performance into sensitivity and bias, shows that, indeed, bias scores are correlated across our tests (see supplementary analysis). The two interpretations are equivalent: someone who is better at match trials than mismatch trials will have an

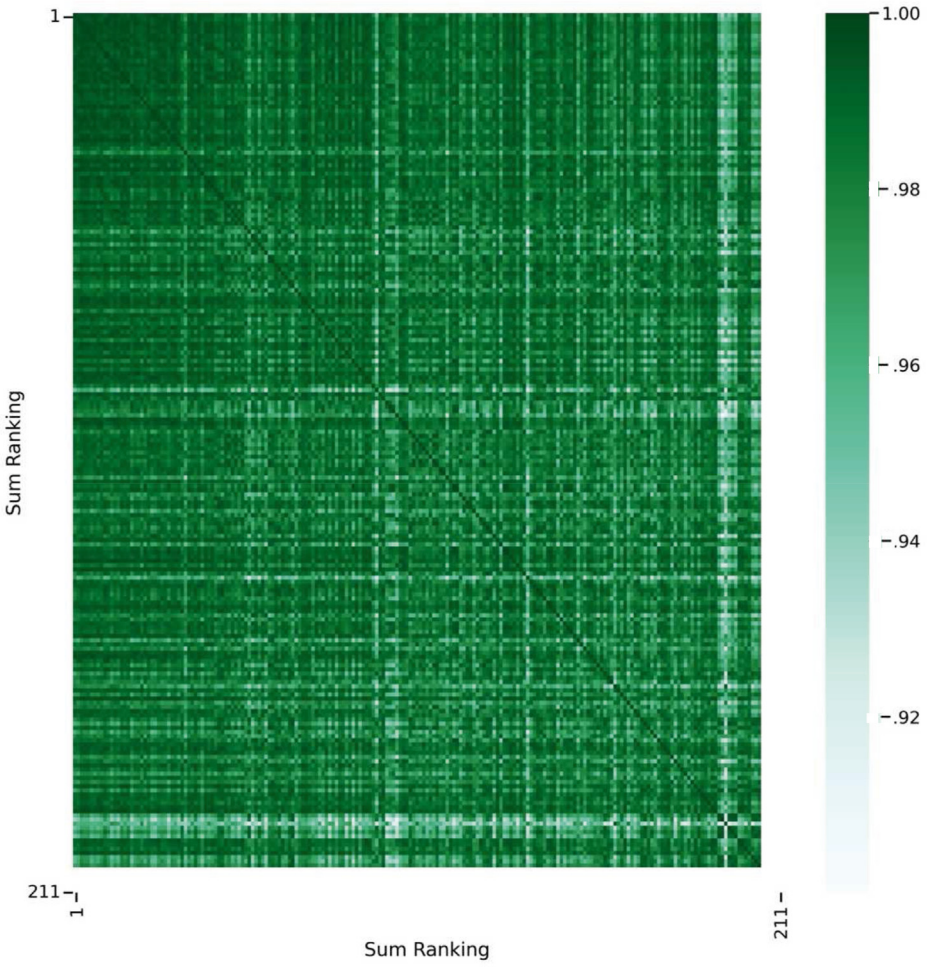

Fig. 5 – The cosine similarity matrix between all pairings of individuals, ordered by their rank sum. Darker areas of the figure indicate participants with higher similarity in their performance across tasks to one another, while lighter areas indicate greater divergences. The figure suggests those with a general higher performance across tasks, as indicated by their rank sum, are more similar to each other.

overall bias towards saying match (see Hancock, 2023 for further discussion). It is possible, however, that there are different underlying causes. Some participants may simply be more cautious about declaring a match, leading directly to differences in bias (e.g. Baker et al., 2023). There may also be distinct abilities. For example, Berger et al. (2022) found that match score and eye-processing ability correlated, while mismatch scores correlated with holistic processing.

It is also worth noting that although some researchers use signal detection analysis (e.g., Baker et al., 2023), it is arguably not an appropriate approach for matching tasks. Burton and Bindemann (2020) point out that mismatched trials cannot be conceptualised as 'noise' and matched trials as 'signal' and that detecting a mismatched trial is as much of a signal as stating that two images match (for example in passport control settings). We agree with this approach.

5.2. *Reliability of the FIP tests*

Internal reliability is a proxy for how consistent a measure is, i.e. it informs us whether items in a test or a questionnaire measure the same construct. It also offers an approximation to other forms of reliability, for instance test-retest when obtaining multiple time-point measurements is impractical (e.g., testing large assessment batteries (Cook & Beckman, 2006). The 'acceptable' value of alpha differs depending on the scale application (Tavakol & Dennick, 2011). It can be deemed as satisfactory at level .7 or .8 when comparing groups, but be only acceptable at .9 or .95 for clinical applications (Bland & Altman, 1997).

Here, we showed that only two tests, the CFMT+ and the 1-in-10, consistently met the more lenient, satisfactory, criterion of $\alpha = .7$ (see Fig. 6).

This finding limits inferences we can make from these widely used tests (and possibly many others widely used in the FIP field). Reliability is an important (but not sufficient) component of validity of a measurement. Unreliable tests are equally problematic for estimating one's FIP ability, as would be a thermometer giving three consecutive body temperature readings of 36 °C, 34 °C, and 37.5 °C for diagnosing hypothermia. It is thus unsurprising that large FIP screening attempts with multiple tests show limited evidence of generalizability from one test to another, often with procedurally similar tasks (Fysh et al., 2020; Stacchi et al., 2020). Using a composite score

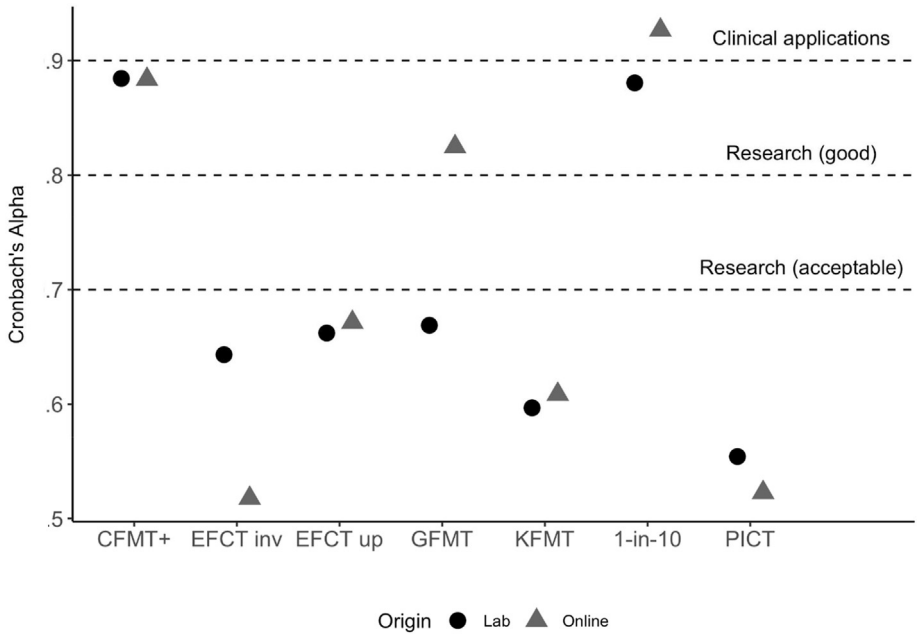

Fig. 6 – Reliability of laboratory and online tests.

(Royer et al., 2015) does not entirely remedy the issue. The sources of this measurement error and reliability can be either in the variability of the items (i.e., images) or of the participants, potentially limiting the generalizability of the results to *one specific stimuli set* and *one specific participant sample* (Shavelson, Webb, & Rowley, 1992). Imagine an instance where one of the four simultaneous matching tasks would be used in a study to estimate one's face matching ability (e.g., Phillips et al., 2018). If another researcher would run the same study with another task, they might arrive at very different conclusions pertaining to their sample, despite testing, in principle, the same ability with a somewhat different stimuli set and, sometimes, procedure.

The solution to this problem might lie in the analytic approach, specifically the use of linear mixed models where participants and stimuli ate treated as random factors. Studies in FIP research are typically analyzed using aggregated responses (e.g., average accuracy on task A for a sample of subjects). This method treats the stimuli and participants as a fixed effect and assumes that the stimuli and participants are representative of, in the instance of FIP research, all faces, and all people. To make results more generalizable over stimuli and participants, FIP researchers may wish to adopt linear mixed modelling, an approach widely used in e.g., psycholinguistics (DeBruine & Barr, 2021; Yarkoni, 2020).

5.3. *Relationships between the FIP tests*

All upright tasks correlated with one another significantly, but the strength of these correlations varied from weak to moderate, in line with previous studies (Balsdon, Summersby, Kemp, & White, 2018; Dunn et al., 2020; McCaffery, Robertson, Young, & Burton, 2018). Interestingly, even the correlation between tests that were procedurally identical, i.e. EFCT-up and PICT, and KFMT and GFMT, were moderate ($r = .418$) and weak ($r = .328$), respectively. This is lower than in the previous comparison of these matching tasks. For example, Stacchi et al. (2020) reported a correlation of .78 between PICT and EFCT but they administered the task with binary response options (2AFC paradigm, see also Noyes, Hill, & O'Toole, 2018; Phillips et al., 2018) suggesting that *even minor changes to response options* may have an impact on the outcomes of a test. This is important, because while most studies use binary response options, some opt for a 5- (White et al., 2015) or even 6-point (Carragher & Hancock, 2022) Likert-like scales and it is unclear how exactly these procedural differences impact participant behaviour. Future work would benefit from a systematic investigation of response options in FIP tasks.

Fysh and Bindemann (2018) reported a moderate correlation of .45 between the GFMT and the KFMT. While our lab results may represent a simple regression to the mean, it is peculiar that the shared variance of two tests differing merely by the face images used as stimuli and participant samples, e.g., the GFMT and the KFMT, vary from 10.8% to 20.2%, depending on the study. Clearly, the contribution of error variance to these tasks is high and conclusions drawn from studies employing the tests ought to be treated with caution. It is particularly interesting that PICT, KFMT, EFCT, and GFMT are almost procedurally identical (bar response options in the EFCT and PICT) and yet the *individual performance ranks vary* depending on the test used, which may result in different conclusions about participant's face matching (or perception) ability. This inconsistency may be driven by stimulus properties, pairings, and, in this study, response options (binary vs. 1–5 confidence scale). We chose to retain the original format of response options for the EFCT and PICT given that this is how these tests were administered in several studies (O'Toole et al., 2007; White et al., 2015). Although this format could be partially responsible for the weaker correlations with tests using two forced response choices, the correlation between PICT (confidence response) and KFMT (binary) is actually the strongest.

Indeed, the variability in stimuli properties between the tests is considerable: the GFMT images were taken on the same day, with two different cameras, the KFMT images were taken months apart, with different devices and at varying distances, while the EFCT and PICT images were taken at varying distances, in different lighting and over a period of two years (Phillips et al., 2011). Although this variability may be overcome to some extent at the upper end of the FIP ability continuum, i.e., the super-recognizers (Ramon, 2021), this does not seem to be the case for most participants. We show in Figs. 4 and 5, that indeed, the top 2.5% of the whole sample appear to be more consistent in their performance than the remaining participants which is unsurprising – to make it to the top of the ranking, one needs to be consistently good. The problem remains with estimating the ability of the middle of the distribution.

The internal reliability of the tests also has implications for the interpretation of correlations between them. Although a maximum absolute value of a correlation is in theory ±1, in practice this can only be achieved if the internal reliability (α) is also one. This is not the case with any of the FIP tests in this study. As such, the maximum expected correlations are not perfect and their interpretation may need adjustment (see Table 2). In their meta-analysis of over 700 individual differences studies, Gignac and Szodari (Gignac & Szodorai, 2016) recommended adjusted coefficients of .10, .20, and .30 as small, medium, and large effect sizes, respectively. This was based on the average correlation strength reported in individual difference studies. Following these guidelines, most of the correlations in our study could be interpreted as strong. What we find important, however, is how little shared variance between these tests these analyses explain and thus this recommended adjustment may not be appropriate. Where estimation of individual's ability is important (e.g., when selecting for occupational roles within national security settings), such estimation clearly must not rely on one test and should be carried out using valid a reliable tests of FIP (Ramon, 2021). While high performers can overcome these test-specific challenges and score consistently high (Nador, Zoia, Pachai, & Ramon, 2021; Nador, Zoia, Pachai, & Ramon, 2021, 2022), low performers tend to struggle with most FIP tasks, but estimating the ability of the 'middle' part of the FIP continuum posits a challenge that researchers should strive to address. It is also possible that high reliability and correlations between measures could simply reflect high performance.

Construction of an instrument to measure FIP is a difficult task: different images of the same person need to be sufficiently different to be uncertain, while images of different people need to be similar enough. Typical performance should be around 70% to allow for a range of abilities to be tested without hitting floor or ceiling. Participants will use any cues available, such as hairlines, ears or spots to discern a match. Such techniques are valid in a forensic setting but hardly typical of everyday face recognition and so need to be minimised in a test of FIP. Early tests were assembled largely by eye from a set of available images. Individual pairs varied in difficulty but not in any very systematic way. Recently there have been attempts to produce more psychometrically valid tests, with items of graded difficulty (Stantic et al., 2021; White, Guilbert, Varela, Jenkins, & Burton, 2021). In principle, such tests should show greater consistency across observers. Unfortunately, neither was available at the time of our testing.

5.4. *Consistency of performance*

Our last aim was to quantify the consistency of performance across the tests for all subjects. Specifically, if a participant ranks first or second on one test, would they also rank in the upper range on the remaining face processing tasks? We already suspected this might not be the case given the weak–moderate correlations in our sample and the findings of previous studies. Others have reported a lack of consistency in performance of participants unselected for FIP ability using data visualization and examining the top and bottom 5% of the accuracy distribution for various tasks (Stacchi et al., 2020; Fysh et al., 2020), but Ramon (2021) reported that super-recognizers are consistently better than participants unselected for FIP ability across three challenging test assessing perception and memory for facial identity.

Indeed, in our combined data set, as well as lab and online data analyzed separately yielded ICC values below .40 suggesting that there was little consistency in performance across tests in our sample (see Fig. 7).

This lack of consistency may not always be noticeable. Large datasets are often collected as neurotypical comparison data for neurodivergent populations with severe processing deficits (e.g., some cases of developmental prosopagnosia), irrespective of the task, so the inconsistent performance of neurotypical participants is not of primary interest, but still affects the quality of the data and conclusions. These results also raise questions about the convergent validity of new FIP tests and their relationship with other tests measuring the same and different sub-processes, i.e. face perception and face memory. Although despite until now often arbitrary approaches to test development, we almost always see positive (albeit weak-to-moderate) correlations between them (Balsdon et al., 2018; Dunn et al., 2020; McCaffery et al., 2018), it is unclear what these correlations represent. The possible candidates are shared processes in these often interchangeably used tasks, their reliability, or other variance, such as procedural similarity. It is, indeed, possible that large proportion of the variance is owed to procedural similarity of the tests. For example, in our study, the correlation between the EFCT *inverted* and PICT (procedurally identical tests, but with different stimuli) supersedes most other correlations between upright matching tasks. Some of the correlations may also reflect good internal reliability, such as those with the CFMT+.

Recently, several labs have devised new tests for typical perceivers, SRs, and prosopagnosic participants (Stantic et al., 2021; White et al., 2021) or batteries of tests (Bate et al., 2018; Dunn et al., 2020; Ramon, 2021), and two of these labs recommended using multiple converging tests to examine face processing ability (Bate et al., 2018; Ramon, 2021). Our results show that even with this more conservative approach the conclusions that can be drawn from an assessment battery might be limited. Ranking 5 on one test, but 25 on another, extremely similar test limits the generalizability of

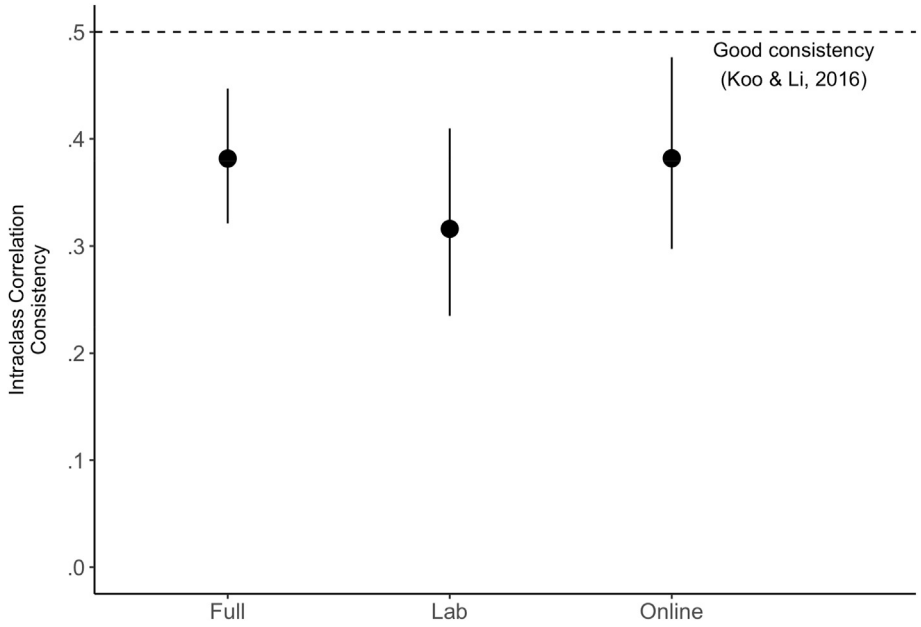

Fig. 7 – Intraclass correlation coefficients for all, lab, and online samples relative to the recommended cut -off. Error bars represent 95% confidence intervals.

conclusions in FIP literature. We recommend that *reliability and consistency are routinely examined and reported in all studies* (rather than cited from the original publication) and, whenever possible, assessment is carried out using multiple (reliable) tests of FIP (see also Ramon, 2021).

6. Conclusion

This study administered seven commonly used tasks of face processing (face memory and face perception) to 211 typical perceivers in the laboratory and online. Data driven analyses (PCA and clustering) revealed that performance splits into two components, 'confirmation' PC1 and 'elimination' PC2, explaining approximately 51% of variance in performance, suggesting that most tasks measure more than one subprocess. Agglomerative clustering of these principal components further grouped participants into three clusters reflecting different response strategies (or biases), one with similar levels of confirmation and elimination performance, one with higher elimination than confirmation performance, and the converse.

To our knowledge, this is the first study to *systematically assess the internal reliability* of the tests commonly used in neuropsychological and individual differences studies (c.f., Nador et al., 2022). The reliability varied from acceptable (CFMT+ and 1-in-10) to poor (PICT, KFMT). Extending previous work (Bate et al., 2018; Fysh et al., 2020; Stacchi et al., 2020), our results revealed small to medium correlations between tests and thus large individual differences between them, even when operating within the same paradigm (simultaneous matching tasks). The intraclass correlation coefficient (ICC) further confirmed low consistency in performance (i.e. large variability in ranking on different tests between participants) in our sample.

We make several recommendations for future studies in FIP. Firstly, researchers may re-think either the use of some of the commonly employed tests, or adjust the analyses to account for stimuli as random factors. Secondly, studies in FIP (both with neuropsychological and/or super-recogniser populations and with typical perceivers) should *routinely report internal reliability* of the used tests for *each attempt and each sample*. Finally, newly developed tasks of face processing should be examined for convergent validity with other established and ecologically valid measures (e.g., Ramon, 2021), even if the only difference are the images used in the task. Until we improve the FIP tests, the inconsistent findings and reports of heterogeneity in special populations and typical perceivers are not only unsurprising but expected and *cannot be attributed to individual differences only*.

In sum, while the 'big data' individual differences approach (Bate et al., 2018; Bennetts et al., 2022; DeGutis et al., 2022; Fysh et al., 2020; Stacchi et al., 2020) can help researchers discover processes driving human performance, there are limitations how robust this approach can be given the available data. This is akin to automated face recognition algorithms and their limitations. Often, an algorithms' 'test' performance does not only depend on the algorithm itself but is reliant on the quality and diversity of the 'training set'. The same is true in the FIP data. The conclusions of data-driven analyses are going to be as meaningful as the quality of the data sets allows.

CRediT author statement

Conceptualization, methodology, funding acquisition: AKB, SB, PJBH. Investigation: AB, SB, NM, ZH. Data curation, Formal analysis: ALJ, PJBH. Writing – original draft: AKB, ALJ. Writing-review and editing: AKB, ALJ, PJBH.

Open practices

The data used in this study are available at: https://osf.io/5k9ny/

Acknowledgments

Anna K. Bobak was funded by the Leverhulme Early Career Fellowship, grant number ECF-2019-416; Peter J.B. Hancock was funded by the EPSRC, grant number EP/N007743/1.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cortex.2023.05.018.

REFERENCES

Bailey, S. (2012). Principal component analysis with noisy and/or missing data. *Publications of the Astronomical Society of the Pacific, 124*(919), 1015. https://doi.org/10.1086/668105

Baker, K. A., Stabile, V. J., & Mondloch, C. J. (2023). Stable individual differences in unfamiliar face identification: Evidence from simultaneous and sequential matching tasks. *Cognition, 232*, Article 105333. https://doi.org/10.1016/j.cognition.2022.105333

Baldassari, M. J., Kantner, J., & Lindsay, D. S. (2019). The importance of decision bias for predicting eyewitness lineup choices: Toward a Lineup Skills Test. *Cognitive Research: Principles and Implications, 4*(1), 2. https://doi.org/10.1186/s41235-018-0150-3

Balsdon, T., Summersby, S., Kemp, R. I., & White, D. (2018). Improving face identification with specialist teams. *Cognitive Research: Principles and Implications, 3*(1), 25. https://doi.org/10.1186/s41235-018-0114-7

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

Bate, S., Portch, E., & Mestry, N. (2021). When two fields collide: Identifying "super-recognisers" for neuropsychological and forensic face recognition research. *The Quarterly Journal of Experimental Psychology: QJEP, 74*(12), 2154–2164. https://doi.org/10.1177/17470218211027695

Belanova, E., Davis, J. P., & Thompson, T. (2018). Cognitive and neural markers of super-recognisers' face processing superiority and enhanced cross-age effect. *Cortex; a Journal Devoted To the Study of the Nervous System and Behavior, 108*, 92–111. https://doi.org/10.1016/j.cortex.2018.07.008

Bennetts, R. J., Gregory, N. J., Tree, J., Di Bernardi Luft, C., Banissy, M. J., Murray, E., et al. (2022). Face specific inversion effects provide evidence for two subtypes of developmental prosopagnosia. *Neuropsychologia, 174*, Article 108332. https://doi.org/10.1016/j.neuropsychologia.2022.108332

Bennetts, R. J., Mole, J., & Bate, S. (2017). Super-recognition in development: A case study of an adolescent with extraordinary face recognition skills. *Cognitive Neuropsychology, 34*(6), 357–376. https://doi.org/10.1080/02643294.2017.1402755

Berger, A., Fry, R., Bobak, A. K., Juliano, A., & DeGutis, J. (2022). Distinct abilities associated with matching same identity faces versus discriminating different faces: Evidence from individual differences in prosopagnosics and controls. *The Quarterly Journal of Experimental Psychology: QJEP*. https://doi.org/10.1177/17470218221076817, 17470218221076816.

Biotti, F., Gray, K. L. H., & Cook, R. (2019). Is developmental prosopagnosia best characterised as an apperceptive or mnemonic condition? *Neuropsychologia, 124*, 285–298. https://doi.org/10.1016/j.neuropsychologia.2018.11.014

Bland, J. M., & Altman, D. G. (1997). Statistics notes: Cronbach's alpha. *Bmj: British Medical Journal, 314*(7080), 572. https://doi.org/10.1136/bmj.314.7080.572

Bobak, A. K., Bennetts, R. J., Parris, B. A., Jansari, A., & Bate, S. (2016). An in-depth cognitive examination of individuals with superior face recognition skills. *Cortex, 82*, 48–62. https://doi.org/10.1016/j.cortex.2016.05.003

Bobak, A. K., Dowsett, A. J., & Bate, S. (2016). Solving the border control problem: Evidence of enhanced face matching in individuals with extraordinary face recognition skills. *Plos One, 11*(2), Article e0148148. https://doi.org/10.1371/journal.pone.0148148

Bobak, A. K., Hancock, P. J. B., & Bate, S. (2016). Super-recognisers in action: Evidence from face-matching and face memory tasks. *Applied Cognitive Psychology, 30*(1), 81–91. https://doi.org/10.1002/acp.3170

Bobak, A. K., Mileva, V. R., & Hancock, P. J. B. (2019). A grey area: How does image hue affect unfamiliar face matching? *Cognitive Research: Principles and Implications, 4*. https://doi.org/10.1186/s41235-019-0174-3

Bobak, A. K., Pampoulov, P., & Bate, S. (2016). Detecting superior face recognition skills in a large sample of young British adults. *Frontiers in Psychology, 7*. https://doi.org/10.3389/fpsyg.2016.01378

Boudry, L., Nador, J. D., & Ramon, M. (2023). Determinants of face recognition: The role of target prevalence and similarity. *PsyArXiv*. https://doi.org/10.31234/osf.io/ux3yc

Bowles, D. C., McKone, E., Dawel, A., Duchaine, B., Palermo, R., Schmalzl, L., et al. (2009). Diagnosing prosopagnosia: Effects of ageing, sex, and participant–stimulus ethnic match on the Cambridge face memory test and Cambridge face perception test. *Cognitive Neuropsychology, 26*(5), 423–455. https://doi.org/10.1080/02643290903343149

Bruce, V., Henderson, Z., Greenwood, K., B, J., Mike, A., & Miller, P. (1999). Verification of face identities from images captured on video. *Journal of Experimental Psychology. Applied, 5*(4), 339–360. https://doi.org/10.1037/1076-898X.5.4.339

Burton, A. M., Kramer, R. S. S., Ritchie, K. L., & Jenkins, R. (2016). Identity from variation: Representations of faces derived from multiple instances. *Cognitive Science, 40*(1), 202–223. https://doi.org/10.1111/cogs.12231

Burton, A. M., White, D., & McNeill, A. (2010). The Glasgow face matching test. *Behavior Research Methods, 42*(1), 286–291. https://doi.org/10.3758/BRM.42.1.286

Carragher, D. J., & Hancock, P. J. B. (2022). Simulated automated facial recognition systems as decision-aids in forensic face matching tasks. *Journal of Experimental Psychology: General, No Pagination Specified-No Pagination Specified*. https://doi.org/10.1037/xge0001310

Ćepulić, D.-B., Wilhelm, O., Sommer, W., & Hildebrandt, A. (2018). All categories are equal, but some categories are more equal than others: The psychometric structure of object and face cognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 44*, 1254–1268. https://doi.org/10.1037/xlm0000511

Cook, D. A., & Beckman, T. J. (2006). Current concepts in validity and reliability for psychometric instruments: Theory and

application. *The American Journal of Medicine, 119*(2), 166. https://doi.org/10.1016/j.amjmed.2005.10.036. e7-166.e16.
Davis, J. P., Lander, K., Evans, R., & Jansari, A. (2016). Investigating predictors of superior face recognition ability in police super-recognisers. *Applied Cognitive Psychology, 30*(6), 827–840. https://doi.org/10.1002/acp.3260
DeBruine, L. M., & Barr, D. J. (2021). Understanding mixed-effects models through data simulation. *Advances in Methods and Practices in Psychological Science, 4*(1), Article 2515245920965119. https://doi.org/10.1177/2515245920965119
DeGutis, J., Bahierathan, K., Barahona, K., Lee, E., Evans, T. C., Shin, H. M., et al. (2022). What is the prevalence of prosopagnosia? An empirical assessment of different diagnostic cutoffs. *PsyArXiv*. https://doi.org/10.31234/osf.io/qtjus
Duchaine, B., Germine, L., & Nakayama, K. (2007). Family resemblance: Ten family members with prosopagnosia and within-class object agnosia. *Cognitive Neuropsychology, 24*(4), 419–430. https://doi.org/10.1080/02643290701380491
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), Article e0241747. https://doi.org/10.1371/journal.pone.0241747
Fysh, M. C., & Bindemann, M. (2018). The kent face matching test. *British Journal of Psychology, 109*(2), 219–231. https://doi.org/10.1111/bjop.12260
Fysh, M. C., & Ramon, M. (2022). Accurate but inefficient: Standard face identity matching tests fail to identify prosopagnosia. *Neuropsychologia, 165*, Article 108119. https://doi.org/10.1016/j.neuropsychologia.2021.108119
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. Royal Society Open Science, 7(9), 200233. https://doi.org/10.1098/rsos.200233.
Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences, 102*, 74–78. https://doi.org/10.1016/j.paid.2016.06.069
Hancock, P. J. (2023). Bias and sensitivity in face matching: The independence of match and mismatch trials. *PsyArXiv*. https://doi.org/10.31234/osf.io/f2a9j
Hancock, P. J. B., Burton, A. M., & Bruce, V. (1996). Face processing: Human perception and principal components analysis. *Memory & Cognition, 24*(1), 26–40. https://doi.org/10.3758/BF03197270
Hildebrandt, A., Wilhelm, O., Schmiedek, F., Herzmann, G., & Sommer, W. (2011). On the specificity of face cognition compared with general cognitive functioning across adult age. *Psychology and Aging, 26*, 701–715. https://doi.org/10.1037/a0023056
Jenkins, R., White, D., Van Montfort, X., & Mike Burton, A. (2011). Variability in photos of the same face. *Cognition, 121*(3), 313–323. https://doi.org/10.1016/j.cognition.2011.08.001
Jones, A. L., & Kramer, R. S. S. (2021). Facial first impressions form two clusters representing approach-avoidance. *Cognitive Psychology, 126*, Article 101387. https://doi.org/10.1016/j.cogpsych.2021.101387
Kokje, E., Bindemann, M., & Megreya, A. M. (2018). Cross-race correlations in the abilities to match unfamiliar faces. *Acta Psychologica, 185*, 13–21. https://doi.org/10.1016/j.actpsy.2018.01.006
Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012
Kramer, R. S. S., Jones, A. L., & Gous, G. (2021). Individual differences in face and voice matching abilities: The relationship between accuracy and consistency. *Applied Cognitive Psychology, 35*(1), 192–202. https://doi.org/10.1002/acp.3754
Liljequist, D., Elfving, B., & Skavberg Roaldsen, K. (2019). Intraclass correlation – a discussion and demonstration of basic features. *Plos One, 14*(7), Article 0219854. https://doi.org/10.1371/journal.pone.0219854
Lowes, J., Hancock, P. J., & Bobak, A. K. (2023). Balanced integration score: A new way of classifying developmental prosopagnosia. *PsyArXiv*. https://doi.org/10.31234/osf.io/g85k7
Mächler, M., Rousseeuw, P., Struyf, A., Hubert, M., & Hornik, K. (2012). Cluster: Cluster analysis basics and extensions. In *R packages*. Vol. 1.
Mayer, M., & Ramon, M. (2022). Improving forensic perpetrator identification with Super-Recognizers. *PsyArXiv*. https://doi.org/10.31234/osf.io/9zq7j
McCaffery, J. M., Robertson, D. J., Young, A. W., & Burton, A. M. (2018). Individual differences in face identity processing. *Cognitive Research: Principles and Implications, 3*(1), 21. https://doi.org/10.1186/s41235-018-0112-9
McCrimmon, A. W., & Smith, A. D. (2013). Review of wechsler abbreviated scale of intelligence, second edition (WASI-II). *Journal of Psychoeducational Assessment, 31*(3), 337–341. https://doi.org/10.1177/0734282912467756
Megreya, A. M., & Burton, A. M. (2007). Hits and false positives in face matching: A familiarity-based dissociation. *Perception & Psychophysics, 69*(7), 1175–1184. https://doi.org/10.3758/BF03193954
Nador, J. D., Vomland, M., Thielgen, M. M., & Ramon, M. (2022). Face recognition in police officers: Who fits the bill? *Forensic Science International: Reports, 5*, Article 100267. https://doi.org/10.1016/j.fsir.2022.100267
Nador, J. D., Zoia, M., Pachai, M. V., & Ramon, M. (2021). Psychophysical profiles in super-recognizers. *Scientific Reports, 11*(1), Article 13184. https://doi.org/10.1038/s41598-021-92549-6
Nimon, K., Zientek, L. R., & Henson, R. K. (2012). The assumption of a reliable instrument and other pitfalls to avoid when considering the reliability of data. *Frontiers in Psychology, 3*, 102. https://doi.org/10.3389/fpsyg.2012.00102
Noyes, E., Hill, M. Q., & O'Toole, A. J. (2018). Face recognition ability does not predict person identification performance: Using individual data in the interpretation of group results. *Cognitive Research: Principles and Implications, 3*(1), 23. https://doi.org/10.1186/s41235-018-0117-4
O'Toole, A. J., Jonathon Phillips, P., Jiang, F., Ayyad, J., Penard, N., & Abdi, H. (2007). Face recognition algorithms surpass humans matching faces over changes in illumination. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 29*(9), 1642–1646. https://doi.org/10.1109/TPAMI.2007.1107
Phillips, P. J., Beveridge, J. R., Draper, B. A., Givens, G., O'Toole, A. J., Bolme, D. S., et al. (2011). An introduction to the good, the bad, amp; the ugly face recognition challenge problem. In *2011 IEEE International Conference on Automatic Face Gesture Recognition (FG)* (pp. 346–353). https://doi.org/10.1109/FG.2011.5771424
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. *Proceedings of the National Academy of Sciences, 115*(24), 6171–6176. https://doi.org/10.1073/pnas.1721355115
Ramon, M. (2018). The power of how—lessons learned from neuropsychology and face processing. *Cognitive Neuropsychology, 35*(1–2), 83–86. https://doi.org/10.1080/02643294.2017.1414777
Ramon, M. (2021). Super-Recognizers – a novel diagnostic framework, 70 cases, and guidelines for future work. *Neuropsychologia, 158*, Article 107809. https://doi.org/10.1016/j.neuropsychologia.2021.107809

Ramon, M., & Gobbini, M. I. (2018). Familiarity matters: A review on prioritized processing of personally familiar faces. *Visual Cognition, 26*(3), 179–195. https://doi.org/10.1080/13506285.2017.1405134

Richler, J. J., Floyd, R. J., & Gauthier, I. (2014). The vanderbilt holistic face processing test: A short and reliable measure of holistic face processing. *Journal of Vision, 14*(11). https://doi.org/10.1167/14.11.10

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), Article e0150036. https://doi.org/10.1371/journal.pone.0150036

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Royer, J., Blais, C., Gosselin, F., Duncan, J., & Fiset, D. (2015). When less is more: Impact of face processing ability on recognition of visually degraded faces. *Journal of Experimental Psychology: Human Perception and Performance, 41*(5), 1179–1183. https://doi.org/10.1037/xhp0000095

Russell, R., Duchaine, B., & Nakayama, K. (2009). Super-recognizers: People with extraordinary face recognition ability. *Psychonomic Bulletin & Review, 16*(2), 252–257. https://doi.org/10.3758/PBR.16.2.252

Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1992). Generalizability theory. In *Methodological issues & strategies in clinical research* (pp. 233–256). American Psychological Association. https://doi.org/10.1037/10109-051.

Stacchi, L., Huguenin-Elie, E., Caldara, R., & Ramon, M. (2020). Normative data for two challenging tests of face matching under ecological conditions. *Cognitive Research: Principles and Implications, 5*(1), 8. 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. *Behavior Research Methods*. https://doi.org/10.3758/s13428-021-01609-2

Susilo, T., Germine, L., & Duchaine, B. (2013). Face recognition ability matures late: Evidence from individual differences in young adults. *Journal of Experimental Psychology: Human Perception and Performance, 39*(5), 1212–1217. https://doi.org/10.1037/a0033469

Sutherland, C. A. M., & Young, A. W. (2022). Understanding trait impressions from faces. *British Journal of Psychology, 113*(4), 1056–1078. https://doi.org/10.1111/bjop.12583

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education, 2*, 53–55. https://doi.org/10.5116/ijme.4dfb.8dfd

Verhallen, R. J., Bosten, J. M., Goodbourn, P. T., Lawrance-Owen, A. J., Bargary, G., & Mollon, J. D. (2017). General and specific factors in the processing of faces. *Vision Research, 141*, 217–227. https://doi.org/10.1016/j.visres.2016.12.014

Vernon, R. J. W., Sutherland, C. A. M., Young, A. W., & Hartley, T. (2014). Modeling first impressions from highly variable facial images. *Proceedings of the National Academy of Sciences, 111*(32), E3353–E3361. https://doi.org/10.1073/pnas.1409860111

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *The Journal of the Acoustical Society of America, 58*(301), 236–244. https://doi.org/10.1080/01621459.1963.10500845

White, D., Guilbert, D., Varela, V. P. L., Jenkins, R., & Burton, A. M. (2021). GFMT2: A psychometric measure of face matching ability. *Behavior Research Methods*. https://doi.org/10.3758/s13428-021-01638-x

White, D., Phillips, P. J., Hahn, C. A., Hill, M., & O'Toole, A. J. (2015). Perceptual expertise in forensic facial image comparison. *Plastic and Reconstructive Surgery, 282*(1814), Article 20151292. https://doi.org/10.1098/rspb.2015.1292

White, D., Rivolta, D., Burton, A. M., Al-Janabi, S., & Palermo, R. (2017a). Face matching impairment in developmental prosopagnosia. *The Quarterly Journal of Experimental Psychology: QJEP, 70*(2), 287–297. https://doi.org/10.1080/17470218.2016.1173076

White, D., Rivolta, D., Burton, A. M., Al-Janabi, S., & Palermo, R. (2017b). Face matching impairment in developmental prosopagnosia. *The Quarterly Journal of Experimental Psychology: QJEP, 70*(2), 287–297. https://doi.org/10.1080/17470218.2016.1173076

Wilson, B. M., Harris, C. R., & Wixted, J. T. (2020). Science is not a signal detection problem. *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.1914237117

Wixted, J. T., & Wells, G. L. (2017). The relationship between eyewitness confidence and identification accuracy: A new synthesis. *Psychological Science in the Public Interest, 18*(1), 10–65. https://doi.org/10.1177/1529100616686966

Yarkoni, T. (2020). The generalizability crisis. *The Behavioral and Brain Sciences, 1–37*. https://doi.org/10.1017/S0140525X20001685