CORTEX 172 (2024) 159–184

Available online at www.sciencedirect.com

ScienceDirect

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

Research Report

A new way of classifying developmental prosopagnosia: Balanced Integration Score

Judith Lowes[*], *Peter J.B. Hancock and Anna K. Bobak*

Psychology, Faculty of Natural Sciences, University of Stirling, United Kingdom

ARTICLE INFO

Article history:
Received 23 January 2023
Reviewed 7 March 2023
Revised 11 April 2023
Accepted 19 December 2023
Action editor Stefan Schweinberger
Published online 17 January 2024

Keywords:
Developmental prosopagnosia
Face processing
Speed-accuracy trade off
Face recognition
Balanced Integration Score
Neurodevelopmental

ABSTRACT

Despite severe everyday problems recognising faces, some individuals with developmental prosopagnosia (DP) can achieve typical accuracy scores on laboratory face recognition tests. To address this, studies sometimes also examine response times (RTs), which tend to be longer in DPs relative to control participants. In the present study, 24 potential (according to self-report) DPs and 110 age-matched controls completed the Cambridge Face and Bicycle Memory Tests, old new faces task, and a famous faces test. We used accuracy and the Balanced Integration Score (BIS), a measure that adjusts accuracy for RTs, to classify our sample at the group and individual levels. Subjective face recognition ability was assessed using the PI20 questionnaire and semi structured interviews. Fifteen DPs showed a major impairment using BIS compared with only five using accuracy alone. Logistic regression showed that a model incorporating the BIS measures was the most sensitive for classifying DP and showed highest area under the curve (AUC). Furthermore, larger between-group effect sizes were observed for a derived global (averaged) memory measure calculated using BIS versus accuracy alone. BIS is thus an extremely sensitive novel measure for attenuating speed-accuracy trade-offs that can otherwise mask impairment measured only by accuracy in DP.

1. Introduction

Developmental prosopagnosia (DP) is a neurodevelopmental syndrome that manifests in severe face recognition problems due to the visual mechanisms for face processing having failed to develop (Duchaine & Nakayama, 2006). DP occurs despite normal vision and IQ and lack of obvious brain damage. Prevalence of DP is estimated at around 2 %–2.9 % in adults (Bowles et al., 2009; Kennerknecht et al., 2006, 2008) and between 1.2 % and 4 % in children (Bennetts et al., 2017). DP is thus as common as severe dyslexia (~2 %–4 %, European Dyslexia Association, n.d.) and more common than autism (~.6 %, World Health Organisation, 2018) despite being relatively unknown. However, recent work by DeGutis et al. (2023) demonstrates that prevalence estimates vary considerably depending on the measures and cut-offs used.

* *Corresponding author*. Psychology, Cottrell Building, Faculty of Natural Sciences, University of Stirling, FK9 4LA, Scotland, United Kingdom.
E-mail address: judith.lowes@stir.ac.uk (J. Lowes).
https://doi.org/10.1016/j.cortex.2023.12.011

1.1. *Different approaches to classifying DP*

No clinical definition of DP exists, instead it is usually classified by poor performance relative to neurotypical controls (usually 1.7 or 2 standard deviations – SDs below the control mean) on at least one laboratory (i.e., objective) test of face processing alongside personal (i.e., subjective) report. Some studies argue that DP is an identifiable disorder with qualitative, as well as quantitative differences between DPs and controls (Behrmann et al., 2007; Bobak et al., 2017; Burns et al., 2014; Towler et al., 2018), others that it may simply represent the lower end of the face recognition (FR) ability spectrum (Bowles et al., 2009; Johnen et al., 2014). More generally, many studies have noted that DP is a heterogenous condition, both in presentation and severity (Bobak et al., 2017; Corrow et al., 2016; Dobel et al., 2007; Duchaine & Nakayama, 2005; Wilcockson et al., 2020). This heterogeneity has even been observed among members of the same family (De Haan, 1999; Duchaine et al., 2007; Lee et al., 2010; Schmalzl et al., 2006). For example, Lee et al. (2010) reported in their study three members of the same family with impaired face memory, but only one additionally with impaired object recognition.

Overall, the literature to date contains many mixed findings. These could be due to genuine heterogeneity, perhaps accentuated by the small sample sizes typical in neuropsychology. However, it is possible that the competing findings may, in part, also be explained by the varied approaches taken by different research groups to classifying DP (Bate & Tree, 2017) and by the range of different tests and measures used for classification and assessment of face processing in DP (for an overview see DeGutis et al., 2023; Robotham & Starrfelt, 2018). Broadly, the literature shows that approaches to categorisation differ across studies in three main, yet interrelated, ways. Firstly, in the selection of test(s) and associated cut off levels to be used; secondly, in the selection of the outcome measure(s) used to quantify test performance and, thirdly, in the inclusion and exclusion criteria used for both the DP and typical controls groups. As we show here, the DP heterogeneity might also arise from study participants adopting different strategies in approaching the tasks, i.e., by taking longer to complete it and thus potentially masking impairments when only accuracy is taken into consideration.

1.2. *Different approaches to the selection of tests for classification*

The criteria to classify DP vary between research groups, which is problematic for study and case comparison (Corrow et al., 2016). To address this, Dalrymple and Palermo (2016) recommended that, in addition to subjective reports of face recognition difficulties, individuals should exhibit impairments on at least *two* objective tests of face processing (cf., Burns et al., 2022 who argue that subjective report alone is sufficient for classification). Notably these guidelines did not specify that these should test face recognition (as opposed to face perception) specifically. This may be problematic because several studies report that as many 50 % of DPs show no face perception impairment (e.g., Dalrymple et al., 2014), meaning that perceptual tests may therefore not always be suitable for classifying DP. Indeed, DeGutis et al. (2023) have recently proposed that *two tests of face recognition* specifically should be used for classification but this has yet to become the norm. Since the 2016 guidelines were published, other researchers have gone further, arguing that converging evidence of impairment across *multiple* tests of face processing would offer stronger evidence for classifying potential DPs (Bate & Tree, 2017), and that participants scoring more than 1 *SD* below the mean on *two or more tests* should be classified as impaired (Mishra et al., 2021; Stumps et al., 2020) since this approach is common in other areas of neuropsychology such as mild cognitive impairment (Sachdev et al., 2014). At the other end of the face recognition ability spectrum, it has also been argued that super recognizers should be assessed using converging evidence from *multiple* tests (Bobak et al., 2023). The guidelines for the best combination of tests predicting real-life face recognition are constantly evolving in both superior (Mayer & Ramon, 2023) and typical face recognition (Bobak et al., 2023).

1.3. *Different approaches to the selection of outcome measures to quantify test performance*

Impaired accuracy is the outcome measure traditionally used to classify DP. However, it has been shown that DPs can sometimes perform within typical accuracy limits when the task has extended or unlimited presentation time (Albonico et al., 2017; Dobel et al., 2007; Duchaine & Nakayama, 2004) leading to recommendations for both accuracy and response time (RT) to be considered (Fysh & Ramon, 2022; Stacchi et al., 2020), typically using Inverse Efficiency Scores (IES, Townsend & Ashby, 1983). Unfortunately, the IES is *only suitable* when mean accuracy is above around 85 %, a level not commonly found in DP research, and/or when there are no accuracy versus RT trade-offs (Bruyer & Brysbaert, 2011). Pertinently, such speed-accuracy trade-offs are observed in common face perception tests (Stacchi et al., 2020) and have been shown to differ between not only DPs and controls but also DPs and participants with acquired prosopagnosia (Behrmann et al., 2005; Fysh & Ramon, 2022).

A better single measure that combines accuracy and RT is the Balanced Integration Score (BIS, Liesefeld & Janczyk, 2019, 2022). Using drift diffusion modelling, the authors demonstrated that BIS is relatively unaffected by speed-accuracy trade-offs (SATs) and is appropriate for between-subjects designs and tasks with varying levels of mean accuracy and decision thresholds thus making it suitable for individual differences face processing research. Face processing tasks often instruct participants to respond as accurately and as fast as possible. This instruction is not neutral as it requires participants to decide whether to prioritise speed or accuracy since one may affect the other. Even when instructions are silent on this matter, or the time allowed for responses is not constrained, participants must still decide how to balance speed and accuracy. Controlling for differential speed-accuracy trade-offs is essential in a lifespan study of DP. Draheim and colleagues (2019) point out in their review of RT and individual differences that multiple studies have shown speed accuracy trade-offs differing across ability level (participants with *lower ability* are more likely to *sacrifice speed for accuracy vs* those with higher ability) and age (older adults tend

to proceed more slowly and carefully than young adults regardless of instructions).

As Liesefeld and Janczyk (2019) point out, the relative weightings applied to speed or accuracy may not only differ *between* participants, but also *within* participants e.g., an individual might prioritise speed on easier trials and accuracy on harder trials. BIS addresses these challenges by equally weighting accuracy and RT and can be thought of as accuracy adjusted for RT, thereby *controlling for differential speed-accuracy trade-offs*. BIS is calculated by subtracting a participant's standardised RT score on correct trials from their standardised accuracy score **[BIS = (Z accuracy − Z RT)]**, see Fig. 4 below. So, for example, the BIS for a hypothetical DP participant whose accuracy z score is −1.2 (less accurate than average) and RT z score is 1 (slower than average) would be −2.2 (−1.2 *minus* 1), whereas BIS for a hypothetical control participant with an accuracy Z score of −1.2 (also below average) and a RT z score of −1.2 (faster than average) would be 0 (−1.2 *minus* −1.2).

1.4. *Different approaches to defining inclusion and exclusion criteria*

How researchers choose to define study inclusion and exclusion for the DP and control groups has important implications for our ability to fully understand DP. A key question is why so many individuals who report severe problems recognising familiar faces in everyday life do not appear to meet the classification for DP (for a discussion see Burns et al., 2022; DeGutis et al., 2023). For example, a recent large-scale study of 165 adults who reported severe everyday face recognition problems (Bate et al., 2019) showed that 61.8 % of the 165 suspected DPs did not meet the commonly used diagnostic threshold of at least 2 SDs below control means on a minimum of two of the three commonly-used diagnostic tests. These were the Cambridge Face Memory Test (CFMT; Duchaine & Nakayama, 2006), the Cambridge Face Perception Test (CFPT; Duchaine et al., 2007) and a Famous Faces Test (FFT). The CFMT is the gold standard test for detecting DP, but even on this test 107/165 (65 %) participants did not meet strict cut off criteria (−2 SD). Burns et al. (2022) also report that a similar proportion of DPs (56 %) did not score more than 2 SD below controls means on the CFMT.

The studies by Bate et al. (2019) and Burns et al. (2022) are relatively unusual in assessing the performance of all the self-reported DPs. More commonly in the literature only those participants who meet strict objective criteria for DP are included in studies (for a review see DeGutis et al., 2023). Unnecessary exclusion of a high proportion of potential participants who report this rare condition is problematic, both for practical and, arguably, ethical reasons (Burns et al., 2022).

A final consideration is exclusion criteria based on face test performance. It is common practice to exclude control participants whose performance falls more than 1.7 or 2 *SD*s below control mean. However, although this approach avoids the unwelcome possibility of including potentially undiagnosed DPs in the control group, it also could lead to an overestimation of the FR abilities of the remaining control group by inflating the control group mean. This is problematic because excluding individuals who produce low test scores but report no day-to-day face recognition problems – even when prompted through a comprehensive and ecologically valid PI20 questionnaire (Shah et al., 2015) or similar instrument could, in turn, create two artificially-distinct groups. After some initial uncertainty, the PI20 has been shown to be a reliable and valid means of classifying DP (Burns et al., 2022; Gray et al., 2017; Tsantani et al., 2021).

Subjective face recognition difficulty is widely considered a pre-requisite for classification as a DP. Arguably therefore, it logically follows that an *absence* of subjective difficulty – *provided* subjective experience has been interrogated using a questionnaire – rules out the possibility of undiagnosed DP. Some low scoring individuals may not experience face recognition difficulties in day-to-day life because they have developed effective compensatory techniques or because their poor test performance simply represents natural variability in face processing or other cognitive processes such as attention – or both. In reality, there are likely to be overlapping face recognition abilities between participant groups who do, and do not, report face recognition difficulties but individuals falling within this overlapping range (exact scores will vary test by test) are currently rarely researched due to prevailing selection and classification methods.

Increasingly therefore, several research groups have begun to question whether the current approaches to classifying DP are appropriate and suggest that broader inclusion criteria might help inform our understanding of DP (Berger et al., 2022; Burns et al., 2022; Dalrymple & Palermo, 2016; DeGutis et al., 2023; Mishra et al., 2021; Stumps et al., 2020). Unless future research includes both the full range of individuals who report FR problems (potential DPs) *and* the full range of neurotypical individuals who score poorly on face processing tasks but do NOT report FR problems on the PI20 or similar questionnaire, we are unlikely to be able to satisfactorily characterise DP.

1.5. *The present study*

In this study we therefore adopt a new approach that allows a more comprehensive understanding of the nature of the underlying deficits in participants who report severe and noticeable face recognition difficulties in everyday life (for simplicity we call these DPs). Specifically, we compare the classification of a group of 24 self-reported DPs on traditional accuracy measures alongside a new classification method using the Balanced Integration Score (BIS). BIS accounts for speed accuracy trade-offs possibly masking face and object processing impairments in some DPs. The aim of this study was to identify whether accuracy or BIS yields an objective measure that better accounts for self-reported face recognition ability.

1.5.1. *Definitions*

We use the term face processing generically to refer to the overall process involved in recognising a human face. More specifically, here *face recognition* and *face memory* are used to refer to the ability to say whether a newly learned target face has been seen before. Both the Old New faces task and the CFMT test face memory. We also use (object/bicycle) memory to mean the ability to say whether a specific bicycle has previously been seen. Face *memory* tasks do not require

participants to be able to name the facial exemplar or to recall any biographical details about an individual. By contrast, when we refer to face *identification*, we mean the ability to identify a face, either by name or other biographical detail (e.g., Actor who plays Mr Bean). The Famous Faces task tests face *identification*. It also tests face *familiarity* by which we mean the ability to correctly judge (yes/no) whether the face of a personally-known celebrity looks familiar.

2. Research transparency and openness

The data presented here were used to classify participants in a wider study which investigates face perception in developmental prosopagnosia and findings from the perceptual task battery will be reported separately in a future publication. Although exploratory in nature, we decided to preregister the study after data collection had commenced, but prior to analysis, to avoid any suspicion of fishing for results or hypothesising after results were known. The preregistered hypotheses, analysis plan, study data and R analysis scripts are available via The Open Science Framework https://osf.io/qne8d/. Analysis of one of the initial screening tasks, the Matrix Reasoning Item Bank (Chierchia et al., 2019) showed that the self-reported DP group adopted a different speed-accuracy strategy to controls, preferring to proceed more slowly and carefully on this visual processing task that did not involve faces (Lowes, Hancock, & Bobak, 2024). To account for these observed speed-accuracy trade-offs, we therefore report here one additional measure that was not preregistered – the Balanced Integration Score or BIS (Liesefeld & Janczyk, 2019, 2022) and describe this further in section 3.4. We also conducted additional unregistered logistic regression analysis and regression analyses to formally assess whether BIS or our preregistered variables best classified potential DPs.

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 in the study.

Legal copyright restrictions prevent public archiving of the Cambridge Face Memory Tests, the Cambridge Bicycle Memory Test, Old New Faces and the Famous Faces Test which can be obtained from the copyright holders in the cited references (details at https://osf.io/f496b/).

3. Methods

3.1. Participants

Participants were 24 individuals aged 8–71 years (7 men; 17 women) who self-reported severe everyday face recognition difficulties (DPs) and 110 age matched controls aged 6–74 years (50 men; 60 women). All participants were living in the UK, Ireland and USA and reported normal, or corrected-to-normal, vision as well as a sufficiently good level of English to understand the instructions and participant information. Participants needed access to a laptop or computer with stable broadband.

Preregistered exclusion criteria were as follows. Individuals with any neurodevelopmental or neurological condition; learning difficulty (other than mild dyslexia) or psychiatric illness; a history of major/moderate brain injury at any time or a mild head injury or concussion during the preceding 12 months or acquired prosopagnosia (i.e., any face recognition difficulty that developed suddenly).

We enquired whether participants had previously participated in other face recognition or face training studies or had attempted any of the test battery online. Control participants who had done so were excluded (number excluded $n = 0$). DPs who had previously completed a face recognition test, but not face training, were included provided that at least three months had passed ($n = 7$, AF006, AF013, AF017, AF018, AF019 and CF008). Four of these reported prior participation in other studies but did not know which ones, two had attempted an (unspecified) online test and one child had been tested by a neuropsychologist using tasks not contained in our battery.

Participant data from any single test where responses were indicative of repeated random key press, technology failure, a clear lack of understanding of the task or failure to follow instructions were excluded. Following preliminary analysis, data from two child control participants were removed because their performance on two or more tasks suggested suboptimal effort or failure to follow instructions. Two adult controls were also excluded because of inconsistencies between their PI20 scores which were borderline indicating possible difficulty, and their follow up interview which revealed no real evidence of lifelong subjective difficulty (e.g., PI20 scores were influenced by a recent one-off failure to recognise a neighbour who was wearing a face mask). Because these two cases met neither the DP criteria (clear subjective lifelong impairment) *nor* the control criteria (no subjective impairment as measured by the PI20), we excluded them.

Participants were recruited though media coverage, social media, personal networks, and prosopagnosia support groups. Data were collected online and participants offered a £10 gift voucher to recompense them for their time.

An overview of individual DP participant accuracy performance on the tests that follow is provided in Table 1.

3.2. Materials

3.2.1. Background visual and cognitive screening

To rule out deficit in low-level vision or more general cognition as explanations for impaired face processing test performance, we screened vision and non-verbal fluid reasoning. No participant scored significantly below chance on two or more of the following sub tests of the Birmingham Object Recognition Battery (BORB, Riddoch & Humphreys, 1993): Length of line, object decision, size of circles, line orientation, and position of gap. Following initial inspection, one child participant was excluded because their BORB RT data indicated suboptimal effort. Except for one participant (CF059, see below), all remaining DP and control participants had accuracy scores within 2 SD of the gender-matched age group mean on the Matrix Reasoning Item Bank (MaRs-IB) (Chierchia et al., 2019). One DP (CF059 female, 16 years) performed significantly below average on this task, indicating a wider cognitive difficulty. A small number of controls ($n = 20$) did not

Table 1 – Individual case scores of self-reported DP participants.

ID	Age	Gender	Self-report	Old New Faces		Cambridge Face Memory Tests		Famous Faces Test		CMT difference	
			Z	*Effect size* Z CC	Estimated % of controls falling below case's score	*Effect size* Z CC	Estimated % of controls falling below case's score	*Effect size* Z CC	Estimated % of controls falling below case's score	*Effect size* Z-DCC	Probability that the standardised difference for a member of the control population would be greater than that of the case and in the same direction
CF007	8	F	−5.15	.30	61.0 %	.90	80.0 %	−1.74	5.69 %	2.10	2.80 %
CF008	10	M	−7.18	.12	54.5 %	−.08	23.1 %	**−3.33**	**.19 %**	1.19	12.9 %
CF059	16	F	−5.59	NA	NA	**−2.34**	1.66 %	−.57	29.0 %	−.43	35.4 %
CF005	17	M	−.66	−1.35	10.1 %	**−3.45**	.15 %	**−1.98**	**3.29 %**	**−2.57**	**.97 %**
AF016	23	F	−7.01	−1.35	10.1 %	−.83	21.3 %	**−2.63**	**.87 %**	−.33	37.2 %
AF002	25	F	−7.28	−.93	18.2 %	−1.54	7.45 %	.42	65.6 %	−1.14	13.7 %
AF017	29	F	−6.60	−1.74	5.22 %	−1.54	7.45 %	−.03	48.7 %	**−2.01**	**3.01 %**
AF004	31	M	−4.54	.61	72.2 %	−.93	18.8 %	**−1.87**	**4.08 %**	**−1.74**	**4.98 %**
AF007	31	F	−4.95	−.57	29.4 %	−.53	30.6 %	.34	62.8 %	−1.07	15.0 %
AF009	42	M	−5.17	.88	79.8 %	−1.22	12.4 %	−.63	27.4 %	**−1.76**	**4.75 %**
AF013	44	F	−4.73	.88	79.8 %	.92	80.9 %	.92	81.0 %	.19	42.5 %
AF075	45	F	−3.64	.29	60.8 %	−.80	22.2 %	−1.41	9.18 %	−.63	26.8 %
AF018	49	F	−5.72	.88	79.8 %	−.91	19.3 %	**−4.31**	**.02 %**	−.82	21.4 %
AF021	50	F	−3.09	.88	79.8 %	−1.55	7.39 %	**−2.04**	**2.96 %**	**−2.23**	**1.93 %**
AF001	51	F	−4.62	−.91	19.3 %	−1.45	8.61 %	−.24	40.7 %	−1.38	9.32 %
AF010	53	F	−5.17	**−9.83**	**.00 %**	**−3.04**	**.40 %**	**−2.63**	**.88 %**	**−2.48**	**1.37 %**
AF006	54	F	−6.16	**−2.70**	**.85 %**	**−2.08**	**2.85 %**	−1.59	6.80 %	−1.36	10.1 %
AF003	55	F	−6.05	.88	79.8 %	−1.55	7.39 %	−1.52	7.60 %	**−2.14**	**2.31 %**
AF060	61	F	−4.65	−.91	19.5 %	−.42	34.3 %	**−1.91**	**4.02 %**	−.85	20.6 %
AF022	64	F	−4.65	−.18	43.3 %	.04	51.7 %	.72	75.4 %	−.49	31.7 %
AF098	67	M	−2.76	**−3.13**	**.39 %**	−1.74	5.56 %	−1.45	8.88 %	−1.16	13.6 %
AF019	68	M	−7.75	**−3.87**	**.09 %**	−1.64	6.52 %	**−5.22**	**.01 %**	−1.64	6.22 %
AF099	70	M	−4.38	**−3.13**	**.39 %**	.33	62.2 %	NA	NA	−.10	46.1 %
AF008	71	F	−4.78	−.18	43.3 %	−.52	31.1 %	−1.37	10.1 %	−.91	19.0 %

ID	Age	Gender	Self-report Z	Old New Faces RT		Cambridge Face Memory Tests RT		Cambridge Bicycle Memory Test RT	
				Effect size Z CC	Estimated % of controls slower than case's mean RT	*Effect size* Z CC	Estimated % of controls slower than case's mean RT	*Effect size* Z CC	Estimated % of controls slower than case's mean RT
CF007	8	F	−5.15	.66	47.6 %	−.82	77.9 %	**1.97**	**3.90 %**
CF008	10	M	−7.18	−.26	59.7 %	−.30	61.3 %	.98	17.6 %
CF059	16	F	−5.59	NA	NA	**2.08**	**2.8 %**	.22	41.5 %
CF005	17	M	−.66	.92	19.0 %	.39	35.5 %	−.82	78.4 %
AF016	23	F	−7.01	.76	23.2 %	.69	25.4 %	−.69	74.7 %
AF002	25	F	−7.28	**3.40**	**.20 %**	1.16	13.5 %	.48	32.3 %
AF017	29	F	−6.60	1.43	9.00 %	−1.36	90.1 %	−1.14	86.1 %
AF004	31	M	−4.54	1.29	11.1 %	1.27	11.5 %	.73	24.1 %
AF007	31	F	−4.95	**2.33**	**1.70 %**	**4.14**	**.04 %**	**1.87**	**4.20 %**
AF009	42	M	−5.17	.85	21.0 %	**4.34**	**.02 %**	1.63	6.30 %

(*continued on next page*)

Table 1 – (*continued*)

ID	Age	Gender	Self-report Z	Old New Faces RT		Cambridge Face Memory Tests RT		Cambridge Bicycle Memory Test RT	
				Effect size Z CC	Estimated % of controls slower than case's mean RT	*Effect size* Z CC	Estimated % of controls slower than case's mean RT	*Effect size* Z CC	Estimated % of controls slower than case's mean RT
AF013	44	F	−4.73	1.68	5.90 %	.13	45.1 %	.18	43.1 %
AF075	45	F	−3.64	**3.35**	**.20 %**	**3.03**	**.40 %**	**2.35**	**1.60 %**
AF018	49	F	−5.72	**2.05**	**3.10 %**	**1.97**	**3.50 %**	1.73	5.30 %
AF021	50	F	−3.09	**2.26**	**2.00 %**	.48	32.2 %	−.11	54.1 %
AF001	51	F	−4.62	1.54	7.50 %	−.88	80.0 %	−1.54	92.6 %
AF010	53	F	−5.17	**2.28**	**2.00 %**	.43	34.0 %	−1.01	83.2 %
AF006	54	F	−6.16	−.27	60.3 %	−.92	81.0 %	−.59	71.6 %
AF003	55	F	−6.05	1.48	8.30 %	.70	25.2 %	.83	21.4 %
AF060	61	F	−4.65	**2.02**	**3.40 %**	−.35	63.1 %	.07	47.3 %
AF022	64	F	−4.65	.11	46.0 %	.55	29.9 %	1.58	7.10 %
AF098	67	M	−2.76	.70	25.3 %	−.72	75.2 %	.05	48.1 %
AF019	68	M	−7.75	**3.03**	**.50 %**	.08	47.0 %	.82	21.8 %
AF099	70	M	−4.38	**4.39**	**.03 %**	.10	46.1 %	−.02	50.8 %
AF008	71	F	−4.78	**5.54**	**.01 %**	.72	24.8 %	**3.04**	**.40 %**

Note. Self report = standardised score on PI20 or parental questionnaire. CMT difference = (Cambridge Face Memory accuracy z score − Cambridge Bicycle Memory accuracy z score). Mean RT is response time on correct trials. Effect sizes in bold indicate performance significantly worse than controls calculated using the SingleBayes_ES.exe computer programme for Bayesian tests of deficit (Z-CC) and the DiffBayes_ES.exe programme for Bayesian standardised difference test (Z-DCC) from Crawford et al., 2010; Crawford & Garthwaite, 2007. Alpha = .05.

Fig. 1 – Overview of recruitment and classification process.

complete the MaRs-IB but their performance on other tasks was within typical norms so they were retained (n.b. reasoning serves as a check to ensure poor test performance in DP cannot be accounted for by impaired cognitive function or difficulty following test instructions). An overview of the recruitment and classification process is shown in Fig. 1.

3.2.2. *Subjective report*

To be initially classified as a potential DP, participants had to have subjective report of lifelong difficulty with familiar face recognition. For adults this was operationalised as a score of ≥61 on the PI20 (Shah et al., 2015) as recommend by Tsantani et al. (2021). This cut-off represented roughly 2.5 standard deviations above the age group control means (and 3.6 SD above the mean in the 14–35 years age group). Inclusion and exclusion criteria were confirmed by follow up semi-structured interview which also captured details of daily face recognition experiences. For children and young people aged 6–17 years we required parental report of face recognition difficulties, classified as a score >1.7 SD above the control group mean on a questionnaire comprising 32 items: 17 items from the PI20 (Shah et al., 2015), rephrased where necessary to make it suitable for reporting a third party's ability, and 15 additional items identified as hallmarks of prosopagnosia for

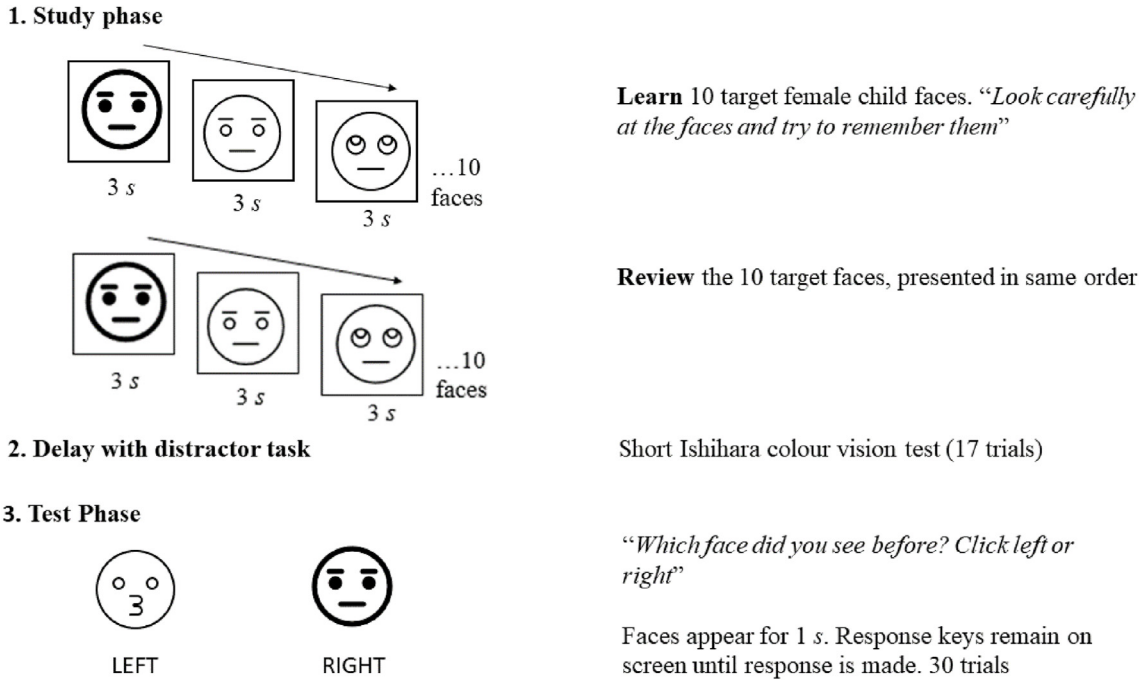

Fig. 2 – Old New Faces test design.

non-experts (Murray et al., 2018). Similar to the PI20, a higher score indicates more difficulties with face recognition. Like adults, inclusion and exclusion criteria and case histories were confirmed through screening interviews with a parent. For ease of comparison with other tasks, negative self-report z scores indicate performance worse than the control average.

3.2.3. *Face tests*

3.2.3.1. Old New Faces. The Old New Face task (see Fig. 2) is a test of delayed face identity recognition matching. Full methods are described in the original paper (Dalrymple et al., 2014), both children and adults completed the version with child faces. Since z scores were computed using the means and SD of the appropriate control age group, any own age bias is therefore controlled for because each participant is compared to the typical score for their relevant age group (Fig. 4). Briefly, participants memorise ten target child faces presented one at a time for 3 sec each. Each target was then immediately shown again in the same order for 3 sec. Both adults and children Participants were instructed to try to memorise the faces. At test, the task is to indicate which of two faces presented is the previously seen face. One target and a similar-looking distractor (age, expression, orientation) were presented simultaneously for 1 sec. Response options appeared under each face ("LEFT" or "RIGHT") and remained on screen until a response was made by mouse click. Targets appeared in randomised order three times each for a total of 30 trials alongside 30 unique distractors that were never repeated. Stimuli were presented in grayscale with hair, ears, and any obvious moles or blemishes removed.

DVs were accuracy and BIS (which uses mean RT on correct trials only). We administered an amended version of the test by introducing a distractor task between study and test. This took the form of a 17-item Ishihara colour vision test, lasting approximately two minutes. Data from one child control participant was not included in analysis of this test since their mean RT was 2.94 *SD* above the age group mean. Their other test performance was within typical norms, so they were retained as a participant.

3.2.3.2. Cambridge Face Memory Tests. Three versions of the CFMT, suitable for different ages, were administered. The CFMT is considered the gold standard test for detection of prosopagnosia. It tests viewpoint-dependent and viewpoint-independent recognition memory for newly-learned faces. Impaired performance on this test is widely used to classify DP since difficulty learning and individuating faces is the core behavioural manifestation of DP. There is a matched object recognition test and comparison of participants' standardised scores on both tests can indicate whether individuation deficits are face selective or also extend to object recognition – at least for the object class being tested. The DVs for all versions of the CFMT were accuracy and BIS. BIS is calculated using mean RT on correct trials only.

3.2.3.2.1. Cambridge Face Memory Test (adults and adolescents aged ≥14 years). A full description of the methods can be found in the original paper (Duchaine & Nakayama, 2006). Briefly, participants study six target faces each from three different viewpoints. Cropped and greyscale faces are presented for 3 sec. The test stage uses a three alternate forced choice (3 AFC) format, comprising one target and two distractor faces. In the first introduction test phase each target face is tested with three identical viewpoints. In the second test phase, after review, each target face is shown in a novel viewpoint from that learned at study and, finally, the noise test section introduces novel views of the target faces with added gaussian noise. There are 72 trials in total and chance = 33 %.

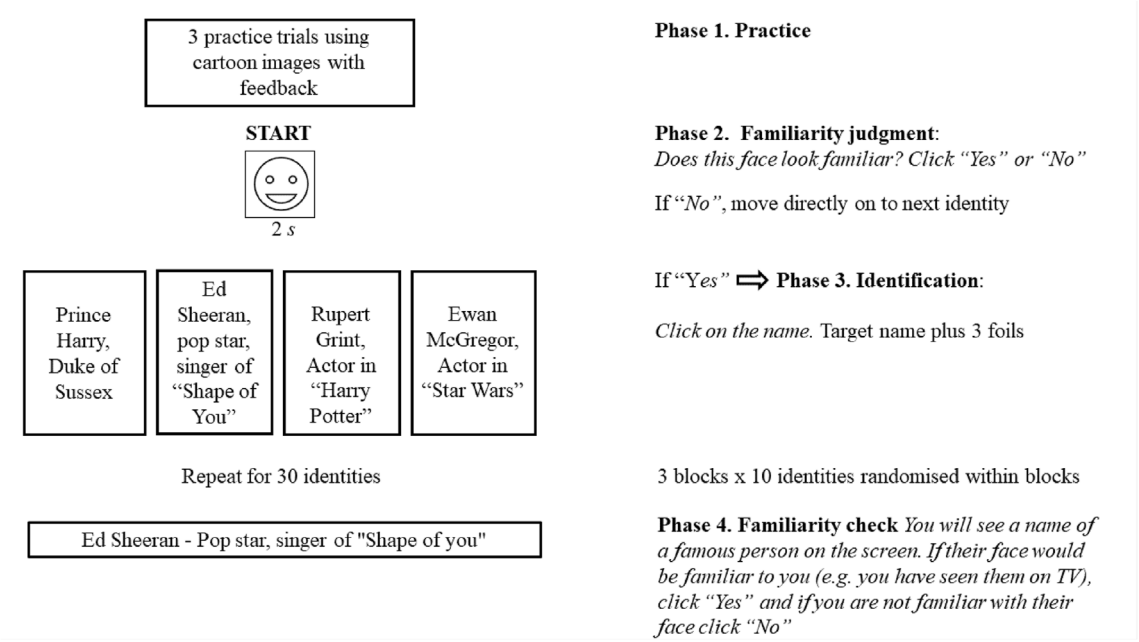

Fig. 3 – Famous Faces test design.

3.2.3.2.2. Cambridge Face Memory Test – Kids (children 10–13 years). The child version of the test (Dalrymple et al., 2014) follows a similar procedure to the CFMT but uses six child faces. There are 72 trials in total and chance = 33 %.

3.2.3.2.3. Cambridge Face Memory Test – Young Kids (children 6–9 years). Due to floor effects being observed in the CFMT-Kids in children <10 years (Dalrymple & Duchaine, 2016; Dalrymple & Palermo, 2016), a shorter version of the test was designed (Dalrymple & Palermo, 2016). Here, participants study 12 unfamiliar target faces, one at a time, from three different viewpoints. Each identity is tested three times, once from each viewpoint, using a 3AFC paradigm. Like the introduction test phase in the CFMT and CFMT-Kids, testing occurs immediately after the study phase for each identity thus minimising memory demands. There are 36 trials in total and chance = 33 %.

3.2.3.3. Cambridge Bicycle Memory Task (CBMT). The CBMT (Dalrymple et al., 2014) is matched in format to the CFMT (adult and kid versions) but uses images of bicycles rather than faces to allow a wider object agnosia to be identified. The CBMT has been used with adults as well as children as young as seven years old and does not appear to have floor effects in this age group (Bate et al., 2020), is well matched in controls to the CFMT (Biotti & Cook, 2016), and has been argued to have better diagnostic properties than the car memory test (Barton et al., 2019). So, following Bate et al. (2020), the six-target version was used with all age groups. There are 72 trials in total and chance = 33 %. As for the CFMT, the DVs were accuracy and BIS. The primary outcome measure of interest for classification purposes was the standardised CMT *difference* score which was computed as standardised face score minus standardised bicycle score.

3.2.3.4. Famous Faces Test. Difficulty identifying familiar faces is the core deficit in DP. In contrast to the CFMT which tests memory for newly learned, unfamiliar faces, the FFT measures long-term familiar face recognition memory. We were not aware of any recent published famous face tests suitable for children and young people as well as adults, so a new test was devised and administered to all age groups. First, we used social media to informally poll parents of children aged 6–16 years. From the longlist of suggested famous identities, we selected 30 identities from multiple sectors covering sport, entertainment, music, politics, and royalty and gathered facial images from the internet. Pilot testing with typically developing participants in the UK aged 6–16 years showed that the median number of facial identities that participants reported knowing was 28.5/30 indicating that the chosen identities were likely to be familiar to children and young people across our target age range.

Participants saw 30 famous faces one at a time (Fig. 3). Stimuli were presented in full colour on a black background with hair cropped but hairline and external contours retained. Identifying blemishes were removed and any jewellery blurred. Each face was presented for 2 sec and participants indicated if the face was familiar by clicking "yes" or "no". If "no", they moved immediately onto the next trial and a new identity was shown. If "yes", participants then had to click on the correct name/identity (e.g., Boris Johnson, UK Prime Minister- at the time of testing) from a choice of four (one target, three foils) before moving on to the next trial. Foils were the descriptions of other famous identities matched for gender and approximate age and, as far as practical, profession. To discourage guessing in the initial familiarity judgement, instructions stressed that it didn't matter how many identities looked familiar. The test began with three practice trials using cartoon images to familiarise participants with the task and feedback was given. No feedback was provided during the test phase. Participants could take a break after every 10 trials.

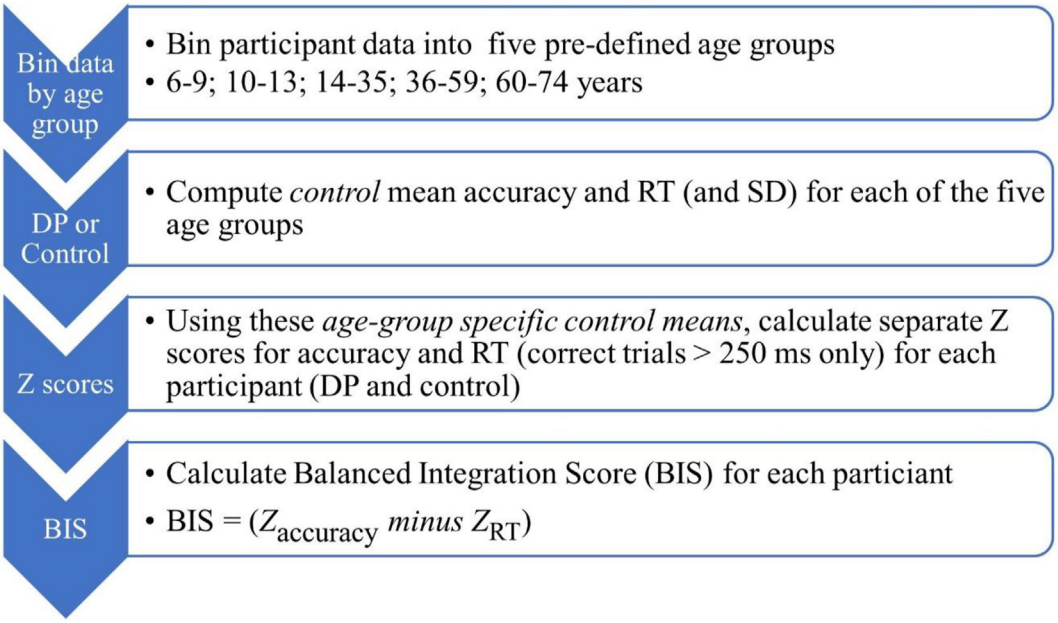

Fig. 4 – Overview of the method used to calculate z scores and BIS.

Stimuli and trial files are available from the authors on request. We do not have permission to share them publicly because images were sourced from the internet.

3.3. *Procedure*

All tests were administered using the online platform Testable (www.testable.org) with Google Chrome as the recommended browser, except the matrix reasoning screening task which was administered using Gorilla (www.gorilla.sc). Participants (or parents in the case of under 18s) were emailed a document containing written task instructions and links to each test and instructed to complete the tests in the prescribed order. Full onscreen instructions were also provided. Parents supervised children and could help with explaining tasks, but instructions stressed they must not help children with their responses. The tests reported here were completed in the following order CBMT, FFT, CFMT and then Old New Faces. Testing took place over a minimum of three self-paced sessions. Each session included only one face recognition test alongside some perceptual tasks with no memory demands. Instructions recommended a minimum break between sessions of at least 12 h for participants under 18 years and one hour for adults. In addition, participants were informed that they could stop at the end of any test to take additional breaks and parents were advised that children should take a break if they began to appear distracted or tired.

3.4. *Analysis plan*

First, we inspected the descriptive statistics and distributions for each test and compared these to published norms for controls where available. Next, test reliability was calculated using Cronbach's alpha for the sample as a whole and for the control and DP groups separately. We originally planned to include gender as a covariate, so we then checked for gender differences in all tests. Because there were no gender differences in all tests (all *p*'s > .05) and there were no significant differences in gender distribution between the DP and control groups ($X^2 = 2.14$, $p = .144$, and all *p*s > .144 in individual age groups) data were collapsed across genders for analysis.

Finally, to allow comparison across tasks, we standardised all measures by calculating z scores (Fig. 4). We centred z scores on control age group means because accuracy and RT varied on some tasks as a function of age. Age groups were predefined as 6–9, 10–13, 14–35, 36–59 and 60–74 years. The choice of age group broadly followed previous literature (e.g., Bowles et al., 2009) and was also partially driven by practical consideration as we aimed to have 20 controls in each age group. The child age groups were chosen primarily to be suitable for the CFMT_Kids and CFMT_Young Kids (7–9). When deciding which tests to administer to adolescents 14–17 we followed Bate et al. (2015) and Bennetts et al. (2017) who administered the CFMT adult version to a prosopagnosic and a super recogniser (respectively) and typical controls aged 14 & 15 years. We inspected mean control performance for 14–17 year olds and 18–35 years old and found no difference with mean scores of 58.8/72 and 58.0/72 respectively. This age group standardisation allowed us to classify DP participants' performance across the full age range since here the z scores quantify participants' performance *relative to their own age group* on any given task. We checked *within* age groups for any significant age-accuracy correlations and found none except among 6–9 year olds for Old New Faces accuracy [$r(10) = .664$, $p = .018$] and 10–13 year olds on the FFT [$r(20) = .44$, $p = .042$]. We also report unstandardised mean accuracy, z score and BIS by age group for each test.

The primary preregistered DV for the CFTM, CBMT and Old New Faces was accuracy (correct trials/total trials). This was because at the time of pre-registration we were not aware of BIS. As discussed in section 2 above, to account for the different speed accuracy trade off strategies observed between controls and DPs on the matrix reasoning screening

task (Lowes, Hancock, & Bobak, 2024)), we additionally calculated BIS (Z accuracy − Z RT) which can be thought of as accuracy adjusted for RT thereby controlling for differential speed-accuracy trade-offs.

For the FFT, three DVs were preregistered. In addition to the primary DV of *accuracy* (the proportion of identities known to the participant that were correctly identified), we also calculated the raw *number* of identities known to each participant (i.e., a familiarity check), and *familiarity* (the proportion[1] of identities known to each participant that were reported as familiar). We did not compute RTs, and therefore BIS, for the FFT since we considered reading speed to be an important potential confound.

The fifth and final preregistered measure of face memory was a global memory score computed for each participant from the mean of their standardised scores on the CFMT, Old New Faces, FFT and standardised CMT difference using pairwise deletion. We report two global memory scores, the first calculated using accuracy and the second using BIS.

3.5. Statistical analysis

To calculate group differences, we conducted Bayesian independent samples *t*-tests on the standardised scores using the default Cauchy prior with a scale of .707. We pre-registered Bayesian analysis because it enables the strength of evidence for both the alternative and null hypotheses to be compared and removes the need to correct for multiple comparisons (Gelman et al., 2012; Kruschke, 2010). For completeness, we also report Welch's *t*-tests which are recommended in independent subject designs with different experimental group sizes and/or unequal group variance (Ruxton, 2006) and are more conservative than Student's *t* tests. Data were analysed in R (R Core Team, 2021) using R Studio 2021.09.1 and the Tidyverse (Wickham et al., 2019) and jmv, version 2.3.4 (Selker et al., 2022) packages. Cronbach's alpha was calculated in SPSS 28.0.0.0 (IBM).

At the individual participant level, single case analyses were conducted using the SingleBayes_ES.exe computer programme for Bayesian tests of deficit (BTD) and the DiffBayes_ES.exe programme for Bayesian standardised difference test (BSDT) (Crawford et al., 2010; Crawford & Garthwaite, 2007). The alpha level was set at .05. These analyses were preregistered. Our sample size of DPs was determined following previous literature and our control sample size (target of 20 participants per age group) was informed by McIntosh and Rittmo's (2021) study. This recommends a minimum neuropsychological control sample of at least eight participants and that twice that number is more desirable, but that increasing control sample size above 16 does little to meaningfully increase power.

Binomial logistic regression modelling and regression analysis were used to assess which outcome measures, or combination of measures, best predicted self-reported group membership as quantified by PI20 or parental report scores. These were not a pre-registered analyses but were added in order to formally assess whether the original variables or BIS best predicted group membership.

[1] The preregistration mistakenly defined familiarity as the raw *number* of identities reported as familiar rather the *proportion*.

4. Results

The results section is structured as follows. We first report test reliability then group results for each of the four tests of face memory followed by the computed global (average) memory score. To illustrate the effect of taking RT as well as accuracy into consideration, we compare the group results calculated first using accuracy and second using BIS. Finally, we present logistic regression data showing which models best predicted group membership.

4.1. Reliability

Table 2 shows reliability (Cronbach's alpha) for the overall sample and separately for the DP and control groups. Reliabilities are rarely reported in such detail but are needed in order to calculate maximum possible correlations between tests and their suitability for individual difference studies (see Bobak et al., 2023). Reliability for most tests was excellent or good; the Old New Faces test was acceptable for controls and good for DPs.

4.2. Old New Faces

The potential DP group ($n = 23$, one self-reported DP participant did not complete this test) was, on average, less accurate than controls ($n = 87$, 23 participants did not complete this test) at judging whether a face had previously been seen or not. The results of Bayesian independent samples *t* test and Welch's *t* tests on standardised accuracy scores and BIS scores are reported in Table 8 and illustrated in Fig. 5. There was strong evidence for a true difference between groups both for accuracy ($BF_{10} = 28.8$) and BIS ($BF_{10} = 5{,}090{,}000$). In other words, the alternative hypothesis (a true BIS difference exists between groups) is more than 5 million times as likely as the null hypothesis given the data. Descriptive statistics for each age group showing unstandardised accuracy scores and BIS are shown in Table 3.

4.3. Cambridge Face Memory Tests

We analysed standardised accuracy scores to allow comparison across the different versions of the CFMT. Results are shown in Table 8 and illustrated in Fig. 6. As expected, self-reported DPs were less accurate than controls and overall

Table 2 – Test reliabilities.

Test	Cronbach's alpha		
	Full sample	Controls	DPs
Old New Faces	.79	.77	.83
CFMT	.92	.92	.86
CFMT-Kids	.89	.90	–
CFMT-Young Kids	.92	.90	–
CBMT	.93	.93	.89
Famous Faces Test	.93	.93	.93

Note. CFMT = Cambridge Face Memory Test, CBMT = Cambridge Bicycle Memory Test. Only one DP completed the CFMT-Kids and CFMT-Young Kids meaning alpha could not be calculated separately for DPs on these tests.

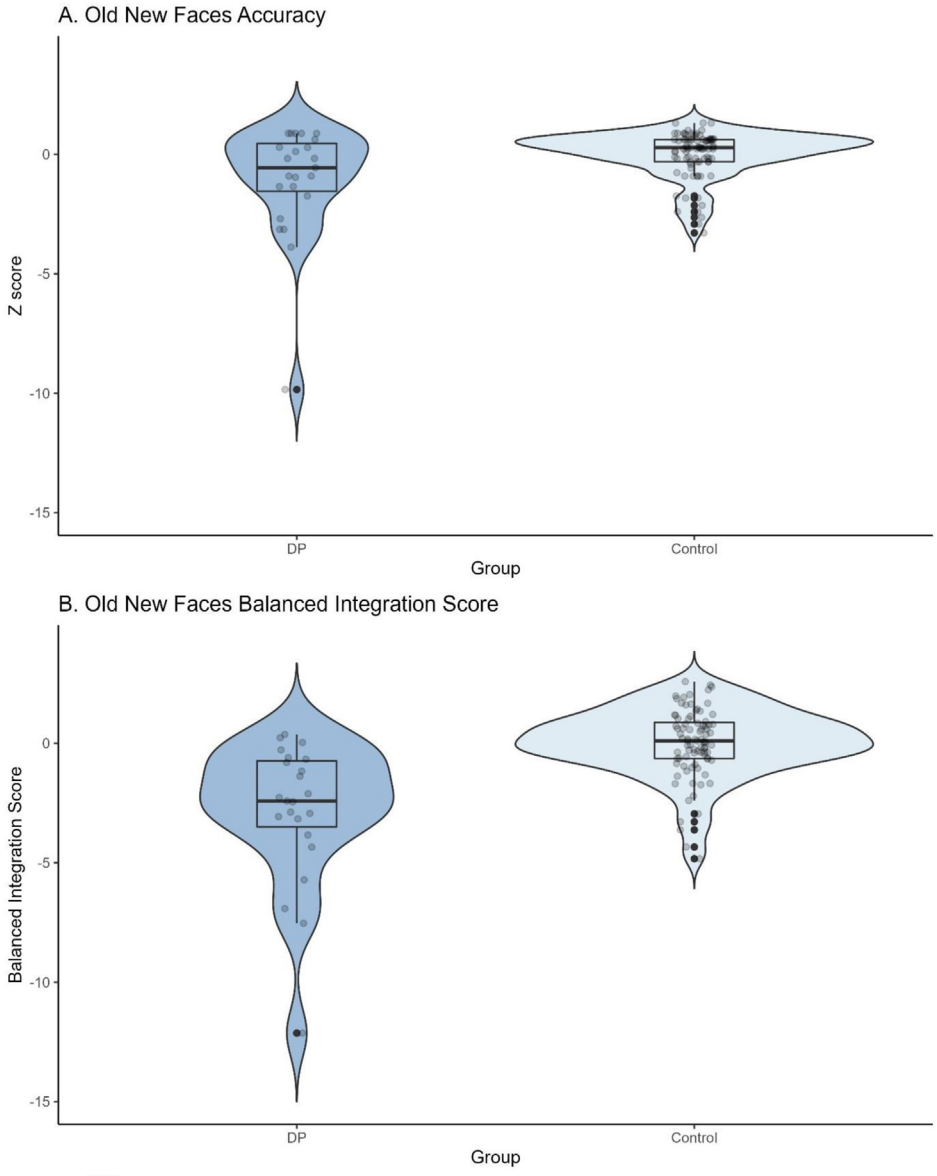

Fig. 5 – Group difference (self-reported DPs *vs* controls) in Old New Faces performance using (A) accuracy and (B) BIS ($Z_{accuracy}$ minus Z_{RT}). Each dot represents a single data point, the box shows the interquartile range (IQR) and the midline indicates the group median score. The end of each whisker line represents 1.5 × the IQR.

Table 3 – Old New Faces accuracy and BIS by age group.

Proportion correct	DP				Control		
	N	Z	*M*	*SD*	*N*	*M*	*SD*
6–9 years	1	.30	.90	–	12	.87	.09
10–13 years	1	.11	.90	–	18	.88	.15
14–35 years	6	−.89	.87	.07	21	.95	.09
36–59 years	9	−.98	.90	.20	19	.95	.06
60–74 years	6	−1.91	.86	.08	17	.94	.05
BIS							
6–9 years	1	.23		–	12	.00	1.16
10–13 years	1	.37		–	18	.00	1.49
14–35 years	6	−2.58		1.23	21	.00	1.57
36–59 years	9	−2.67		3.69	19	.00	1.65
60–74 years	6	−4.54		2.73	17	.00	1.37

Note. Chance = .5. There was only one potential DP in each of the two youngest age groups so *SD* could not be calculated separately for self-reported DPs in these age groups. Data from 23 controls is not available due to participant or technical error or participant drop out. Because BIS is a standardised measure, BIS scores for controls are 0 by definition.

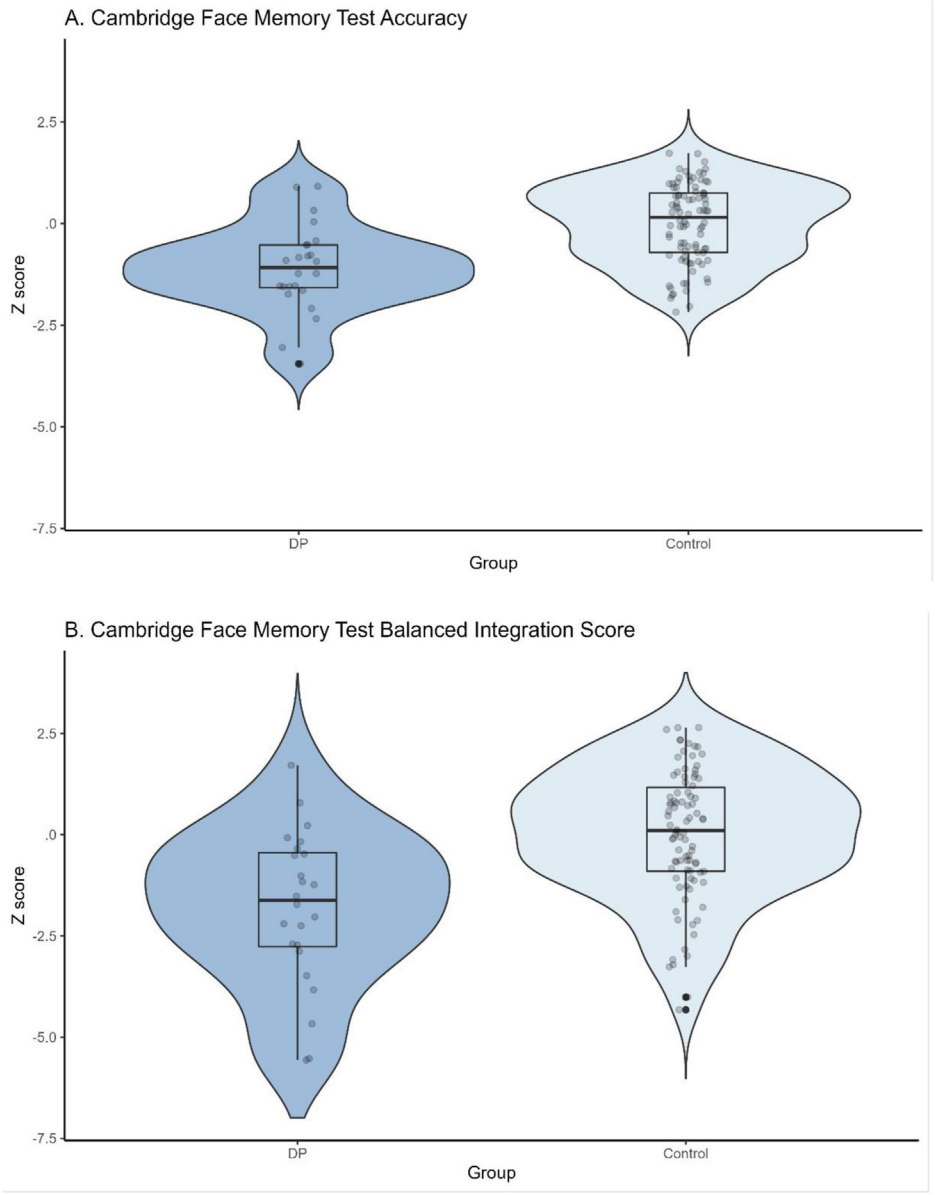

Fig. 6 – Group difference (self-reported DPs *vs* controls) in performance using (A) accuracy (correct trials/total trials) and (B) BIS ($Z_{accuracy}$ minus Z_{RT}). Each dot represents a single data point, the box shows the interquartile range (IQR) and the midline indicates the group median score. The end of each whisker line represents 1.5 × the IQR. *Note.* Participants aged 6–9 years completed the CFMT Young Kids, ages 10–13 years completed the CFMT-Kids and participants age ≥14 years completed the CFMT.

there was strong evidence of a true difference between groups for both accuracy (BF_{10} = 3,290) and BIS (BF_{10} = 2,453). Unstandardised accuracy scores and BIS are shown in Table 4.

4.4. *Difference between face (CFMT) and object (CBMT) memory accuracy*

To investigate whether any memory difficulties were face-specific, we calculated a CMT *difference* score by subtracting each participant's CBMT z score from their CFMT z score (Fig. 7). A difference score above zero indicates that a participant performed relatively better at faces than bicycles and a score below zero indicates that a participant performed relatively better at bicycles than faces. As a reminder, all z scores were centred on age group control means so this difference score is therefore a relative measure of face/bicycle memory accuracy compared to participants' own age control group. Unstandardised accuracy and BIS scores for the CBMT are shown in Table 5.

4.5. *Famous Faces Test*

Data from one potential DP were not analysed because they notified us of participant error during this task. Data from 23 controls is not available due to participant or technical error or participant drop out (i.e., some participants did not complete

Table 4 – CFMT accuracy and BIS by age group.

Proportion correct	DP				Control		
	N	Z	M	SD	N	M	SD
6–9 years	1	.90	.97	–	15	.80	.19
10–13 years	1	−.47	.65	–	17	.76	.14
14–35 years	7	−1.59	.59	.14	21	.81	.14
36–59 years	9	−1.27	.65	.14	20	.81	.13
60–74 years	6	−.65	.62	.13	17	.72	.15
BIS							
6–9 years	1	1.71		–	15	.00	1.74
10–13 years	1	−.47		–	17	.00	1.56
14–35 years	7	−2.79		1.82	21	.00	1.47
36–59 years	9	−2.31		1.92	20	.00	1.77
60–74 years	6	−.72		.74	17	.00	1.63

Note. Chance accuracy = .33. There was only one self-reported DP in each of the two youngest age groups so *SD* could not be calculated separately for DPs in these age groups. 20 controls did not complete this test. Because BIS is a standardised measure, BIS scores for controls are 0 by definition.

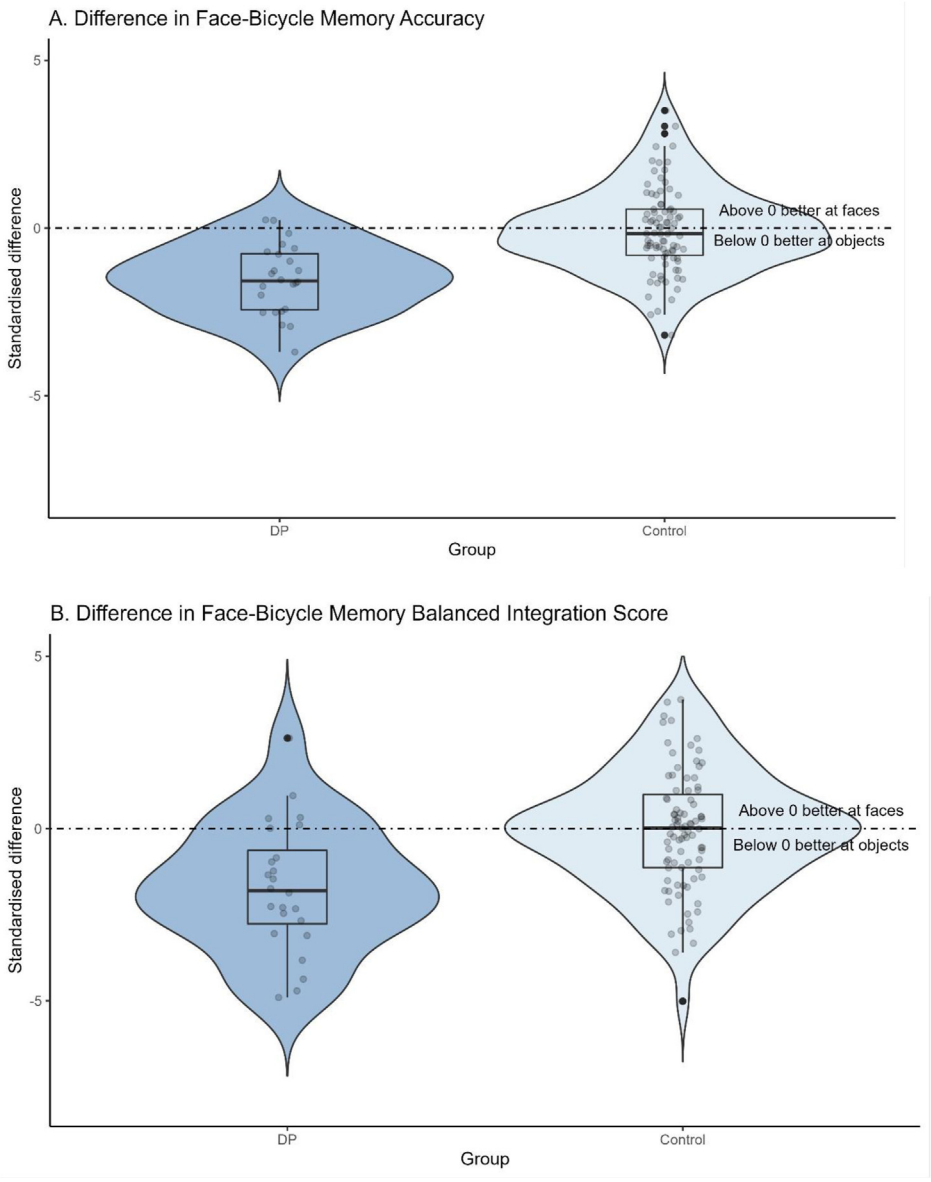

Fig. 7 – Group difference (self-reported DPs *vs* controls) in participants' relative performance on the Cambridge Face Memory Test and Cambridge Bicycle Memory Test (Z_CFMT minus Z_CBMT) using (A) accuracy and (B) BIS ($Z_{accuracy}$ minus Z_{RT}). Each dot represents a single data point, the box shows the interquartile range (IQR) and the midline indicates the group median score. The end of each whisker line represents 1.5 × the IQR.

Table 5 – Cambridge Bicycle Memory Test accuracy and BIS by age group.

Proportion correct	DP				Control		
	N	Z	M	SD	N	M	SD
6–9 years	1	1.06	.72	–	15	.57	.15
10–13 years	1	.50	.78	–	21	.69	.17
14–35 years	7	.32	.85	.19	21	.81	.12
36–59 years	9	.35	.86	.07	22	.81	.14
60–74 years	6	.56	.86	.06	18	.78	.15
BIS							
6–9 years	1	−.92		–	15	.00	1.03
10–13 years	1	−.48		–	21	.00	1.12
14–35 years	7	.28		1.93	21	.00	1.57
36–59 years	9	−.03		1.37	22	.00	1.50
60–74 years	6	−.36		1.03	18	.00	1.23

Note. Chance accuracy = .33. There was only one DP in each of the two youngest age groups so *SD* could not be calculated separately for DPs in these age groups. Because BIS is a standardised measure, BIS scores for controls are 0 by definition.

all three sessions). No floor effects were observed in any age group, 2 *SD* below the control mean accuracy was always above chance (25 %). As discussed in section 3.4, the test design meant it was not appropriate to calculate BIS for the Famous Faces Test, instead we use only accuracy. As discussed below, FFT accuracy was a strong predictor of group and so remains a useful measure.

The raw *number* of famous faces known to participants was checked at the end of the test. At an age group level, the mean number of identities known ranged from 13.4 in young children (skewed by a higher proportion of non-UK controls, chosen to match the DP in this age group) to 24.2 in the 14–35 years age group. Considering UK control participants only, the mean number of identities known were: 16.4 (6–9 years); 21.2 (10–13 years); 25.7 (14–35 years); 23.1 (36–59 years); 20.4 (>60 years). Whilst number of known identities was not the main variable of interest and was measured in order to calculate personally familiar accuracy and familiarity scores for each participant, we noted that the DP group reported knowing significantly fewer famous faces than controls, $t(32.1) = 3.80$, $p < .001$, with extremely strong support ($BF_{10} = 133$) for a true group difference. This could be due to previously reported differences in media consumption (Dalrymple & Palermo, 2016).

4.5.1. *Famous Face Test: Identification*

As shown in Fig. 8, potential DPs ($n = 23$) on average, correctly identified (named) a lower proportion of faces than controls ($n = 100$). A Bayesian independent samples *t* test on standardised scores provided very strong evidence for a true difference between groups $BF_{10} = 204{,}436$. In other words, the alternative hypothesis (a true difference exists between groups) is more than two hundred thousand times as likely as the null hypothesis given the data (see Table 8). Unstandardised group-level accuracy descriptive statistics are shown in Table 6.

4.5.2. *Famous Face Test: Familiarity*

Similar to the primary outcome measure of famous face identification discussed above ("*choose the name that matches the face*"), the self-reported DP group was significantly less accurate at judging a known face as looking familiar versus controls

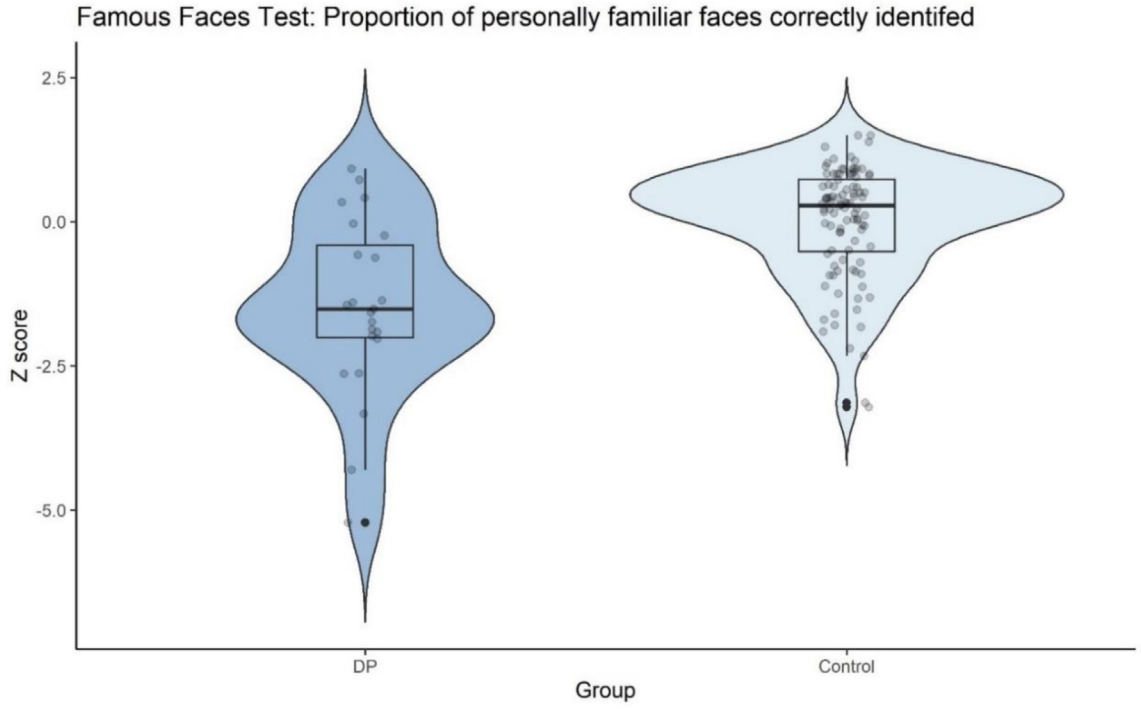

Fig. 8 – Group difference (self-reported DPs *vs* controls) in identification accuracy. Scores were calculated using only those facial identities that participants reported knowing. Each dot represents a single data point, the box shows the interquartile range (IQR) and the mid line indicates the group median score. The end of each whisker line represents 1.5 × the IQR.

Table 6 – FFT identification accuracy by age group.

Proportion correct	DP			Control		
	N	*M*	*SD*	*N*	*M*	*SD*
6–9 years	1	.33	–	15	.69	.21
10–13 years	1	.17	–	22	.75	.17
14–35 years	7	.85	.11	23	.93	.09
36–59 years	9	.70	.18	22	.89	.12
60–74 years	5	.67	.22	18	.86	.10

Note. Chance = .25 There was only one potential DP in each of the two youngest age groups so SD could not be calculated separately for DPs.

Table 7 – Global face memory BIS by age group.

Global face memory BIS	DP			Control		
	N	*M*	*SD*	*N*	*M*	*SD*
6–9 years	1	.71	–	17	−.10	1.06
10–13 years	1	−.86	–	22	−.05	.98
14–35 years	7	−2.33	.55	23	−.05	.97
36–59 years	9	−2.19	1.58	22	−.02	1.06
60–74 years	6	−1.89	1.29	18	.07	1.19

Note. There was only one potential DP in each of the two youngest age groups so SD could not be calculated separately for DPs. Mean scores were computed from CFMT BIS, CMT difference BIS, Old New Faces BIS and FFT identification accuracy.

("*does this face look familiar?*"), $t(24.9) = 4.43$, $p < .001$, $d = 1.22$. A Bayesian independent samples *t*-test also provided extremely strong evidence for a true difference between groups ($BF_{10} = 6{,}130{,}000$). Notably, an almost perfect correlation was observed between participants' familiarity scores and identification scores in both the DP [$r(21) = .997$, $p < .001$] and control [$r(98) = .983$, $p < .001$] groups. This result suggests that, at least in this cohort, a sense of familiarity was not distinct from the ability to identify a face, however the 4AFC paradigm used in the identification phase would be expected to result in a higher familiarity/identification correlation than a paradigm which required participants to generate the name themselves.

We chose a 4AFC paradigm for this task because we reasoned it would be easier for children than a standard FFT paradigm. A typical FFT requires participants to provide a name, or some other unique identifying detail, from memory, unprompted, usually by writing or typing a response. We were therefore also concerned that parental interference might become an issue with such a design since children would be likely to need parental help to type and/or spell the celebrity's name and this could result in parents answering on behalf of children. Due to our design, it was possible therefore that the expected group differences would not be observed on this somewhat easier test (see Rivolta et al., 2013). We did not find this to be the case and observed strong group differences on the FFT (see Table 8) with, as expected, the DP group being less accurate at identifying famous faces known to them than the control group (Table 6). These results suggest that our test design did not unduly assist potential DPs to identify familiar faces versus controls.

4.6. *Global face memory measures*

Because it is possible that a participant may achieve a score on one test that may be higher or lower than their true ability due to chance or to random factors such as tiredness, we computed a global memory score by averaging participants' z scores across the four face memory measures of interest (Old New Faces, CFMT, CMT Difference and FFT). Descriptive statistics for the global memory scores are provided in Table 7. As shown in Fig. 9, the self-reported DP group's global memory accuracy and global memory BIS means were both significantly lower than the respective control group means. This finding of lower DP performance versus controls, even when measured across multiple tests, suggests that the group differences observed for each individual face memory outcome measure are not due solely to noise, or chance, and is confirmed by much higher Bayes factors for global face memory versus individual tests as shown in Table 8. Table 8 also shows that statistical analysis of the group differences in accuracy provided strong support for the alternative hypothesis and that the group difference was significant.

When considering group differences averaged across multiple standardised measures of face memory, the global memory BIS (i.e., accuracy adjusted for RT) showed a larger effect size than the global memory score calculated using accuracy alone (see Table 8). We also separately calculated group differences for adults (aged 18 years and over) and found an identical pattern of results (lower portion of Table 8), and very similar effect sizes. Although face processing continues to develop throughout childhood (e.g., Pascalis et al., 2011) and declines in later life (Bowles et al., 2009), our approach of using age-matched z scores to analyse group differences in this lifespan study ensures that any individual's face processing impairment is best classified with comparison to typical controls at a similar stage of development.

4.7. *Which measures best predict group membership?*

We used binomial logistic regression to formally assess which objective face memory measure, or combination of measures, best predicted self-reported group membership as classified by cut offs on the PI20 or the parental report questionnaire. We developed five models (see Table 9). All models significantly predicted group membership, and all correctly classified over 94 % of controls. Crucially however, the models' ability to correctly classify self-reported DPs varied greatly, ranging from only 16.7 % (Model 5, with CFMT accuracy as the predictor) to 68.2 % in the best performing model (Model 2, comprising four separate predictor outcome variables (CFMT BIS, CMT difference BIS, Old New BIS, Famous Face identification accuracy)).

Crucially, Model 2 strongly predicted group membership and outperformed all other models. Model 2 correctly predicted 68.2 % of self-reported DPs and 94.9 % of controls, explaining around 44 %–57 % of the variance in subjective ratings of participants' face recognition ability.

4.8. *How does using BIS change classification of DP?*

As seen in Table 9 above, the effect sizes for mean group difference were larger using BIS than using accuracy alone. We

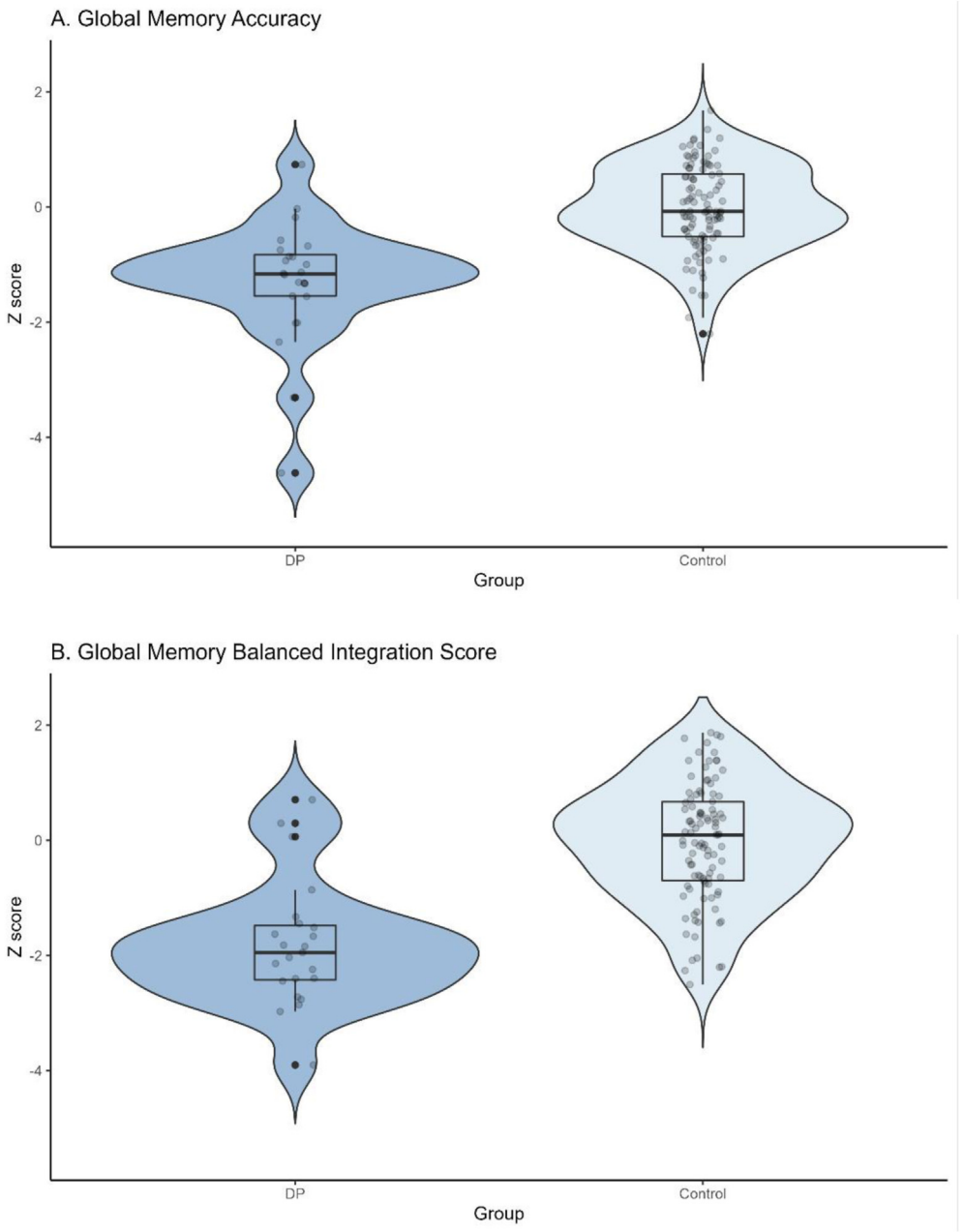

Fig. 9 – Global face memory scores showing group differences (self-reported DPs *vs* controls) in mean face memory scores using (A) accuracy and (B) BIS ($Z_{accuracy}$ minus Z_{RT}). Mean scores were computed from CFMT accuracy, CMT difference, Old New Faces accuracy and FFT identification accuracy. Each dot represents a single data point, the box shows the interquartile range (IQR) and the midline indicates the group median score. The end of each whisker line represents 1.5 × the IQR.

therefore next investigated how using BIS might change, or confirm, classification as a DP on our objective tests. Results are shown in Tables 10 and 11. As a reminder, to meet the initial classification as a potential DP, all participants firstly had to score atypically on the parental questionnaire or PI20. We then sub-classified these self-reported "DPs" into three groups using the average of four objective measures: CFMT, FFT, Old New Faces and the difference between CFMT and CBMT scores (CMT difference). "Major DPs" scored more than 1.7 SD *below* their age group control mean; "Mild DPs" scored between 1 and 1.7 SD *below* the control means and "Subjective DPs" showed atypical self-report scores but scored within 1 SD of their age group control means, showing no objective impairment despite subjective report of face recognition difficulties. One participant (CF059) did not complete the Old New Faces test but was classified as Major DP because their CFMT z score (−3.45) indicated severe impairment and their FFT score was also below average.

As shown in Table 10, the key finding of our study was that 83.3 % of the self-reported DP group showed objective face recognition deficits (20/24; 15 major, five mild) using the global memory BIS measure versus 58.3 % who showed deficits when using global memory accuracy alone (14/24: five major, nine mild). This suggests that some self-reported DPs in our sample

Table 8 – Comparing effect sizes for the difference between potential DP and control group scores before (Accuracy) and after (BIS) accounting for RT.

	Accuracy			Balanced Integration Score		
	BF_{10}	p	d	BF_{10}	p	d
Old New Faces	28.8	.042*	−.60	5,090,000	**<.001*** **	−1.24
CFMT	3,290	**<.001*** **	−1.08	2,452	**<.001*** **	−1.03
CBMT	1.25	.044*	.46	.246	.766	−.07
CMT difference	14,357	**<.001*** **	−1.27	221	**<.001*** **	−.91
FFT	204,436	**<.001*** **	−1.15	–	–	–
Global face memory	28,343,149	**<.001*** **	−1.39	4,085,749,935	**<.001*** **	−1.63
Adults aged 18 years and over only						
	Accuracy			Balanced Integration Score		
	BF_{10}	p	d	BF_{10}	p	d
Old New Faces	6.08	.06	−.60	80,900	**<.001*** **	−1.34
CFMT	153	**<.001*** **	−1.09	154	**<.001*** **	−1.06
CBMT	1.49	.011*	.60	.265	.972	−.01
CMT difference	550	**<.001*** **	−1.27	86.0	**<.001*** **	−1.04
FFT	2767	**.001** **	−1.12	–	–	–
Global face memory	57,332	**<.001*** **	−1.36	9,166,061	**<.001*** **	−1.78

Note. *$p < .05$, **$p < .01$, ***$p < .001$. p values in bold remain significant after Bonferroni correction.
CMT difference = CFMT − CBMT. RT (and consequently BIS) was not considered a relevant measure for the FFT since participants had to read names meaning RT would reflect reading speed and comprehension as well as face processing ability. Instead, we use only famous face identification accuracy (proportion of known famous faces that were correctly identified) in both versions of the global scores.

were able to achieve close to normal, or only mildly impaired, performance by trading speed for accuracy; however, once RT was considered (using BIS), their impairment became apparent. This important finding demonstrates the value of BIS as a measure for classifying DP since these participants would otherwise have been missed if classification considered accuracy alone.

Three self-reported DP participants showed no objective impairment regardless of the measure used, we refer to these as having "subjective" DP since no objective impairment was observed. Additionally, one potential child DP showed a mild impairment ($Z = -1.32$) when considering global accuracy alone but was not considered objectively impaired once RT was accounted for (BIS) although their global memory BIS was still below average ($Z = -.86$). In summary, many more of the participants with atypically high PI20/parental questionnaire scores were classified as DP using BIS than using classical measures. Here we classify a major impairment as $z \leq -1.7$ but, for comparison, using a stricter ≤ -2 SD cut off, our data show that 12 participants would be classified as DP using the BIS measures versus only five when using accuracy alone.

Finally, we compared how classification using individual test scores, rather than the global face memory score shown in Table 10, would differ using accuracy versus BIS and these are reported in Table 11.

Table 9 – Logistic regression models predicting self-reported group membership (DP or control).

Model	Predictors	Overall model				AUC	% correctly classified	
		X^2	p	Adj R^2 McF[a]	Adj R^2 N[b]		DP	Control
1	**Accuracy**: CFMT, CMT Difference, Old New Faces, FFT	43.6	**<.001*** **	.41	.54	.90	59.1	97.4
2	**BIS**: Old New Faces, CFMT, CMT Difference; FFT accuracy	46.7	**<.001*** **	.44	.57	.90	68.2	94.9
3	Global memory **accuracy**	37.5	**<.001*** **	.31	.41	.87	29.2	95.1
4	Global memory **BIS**	45.6	**<.001*** **	.38	.50	.88	54.2	94.1
5	CFMT **accuracy**	20.3	**<.001*** **	.17	.25	.77	16.7	96.7

Model 2 Coefficients – group

Predictor	Estimate	SE	Z	p	Odds ratio	95 % Confidence Interval	
						Lower	Upper
Intercept	−3.03	.61	−4.96	<.001	.05	.01	.16
Old New Faces BIS	−.63	.21	−3.00	**.003** **	.53	.35	.80
CFMT BIS	−.08	.28	−.28	.783	.93	.53	1.61
FFT	−.91	.31	−2.91	**.004** **	.40	.22	.74
CMT Difference BIS	−.38	.28	−1.38	.168	.68	.39	1.18

Note. Estimates represent the log odds of "group = DP" vs. "group = Control". P values in bold are significant * p< .05, ** p < .01, *** p<.001
[a] Adj R^2 McF = McFaddens R^2.
[b] Adj R^2 N = Nagerkerke's R^2.

Table 10 – Global memory accuracy and BIS classification for all self-reported DP participants.

ID	Self report z score	Global memory accuracy z score	Global memory BIS	DP classification (global memory accuracy)	DP classification (global memory BIS)	Change using BIS vs accuracy
AF019	-7.75	-3.31	-3.90	Major	Major	=
AF002	-7.28	-.93	-2.24	Subjective	Major	↑
CF008	-7.18	-1.32	-.86	Mild	Subjective	↓
AF016	-7.01	-1.32	-2.03	Mild	Major	↑
AF017	-6.60	-1.55	-1.51	Mild	Mild	=
AF006	-6.16	-2.01	-1.63	Major	Mild	↓
AF003	-6.05	-1.15	-1.66	Mild	Mild	=
AF018	-5.72	-1.33	-2.40	Mild	Major	↑
CF059	-5.59	-1.54	-2.85	Mild	Major	↑
AF009	-5.17	-.75	-2.72	Subjective	Major	↑
AF010	-5.17	-4.62	-5.65	Major	Major	=
CF007	-5.15	-.18	.71	Subjective	Subjective	=
AF007	-4.95	-.57	-2.76	Subjective	Major	↑
AF008	-4.78	-.86	-1.84	Subjective	Major	↑
AF013	-4.73	.74	.30	Subjective	Subjective	=
AF060	-4.65	-1.13	-1.44	Mild	Mild	=
AF022	-4.65	-.03	.06	Subjective	Subjective	=
AF001	-4.62	-.996	-1.33	Subjective	Mild	↑
AF004	-4.54	-1.17	-1.95	Mild	Major	↑
AF099	-4.38	-.86	-2.40	Subjective	Major	↑
AF075	-3.64	-.68	-2.44	Subjective	Major	↑
AF021	-3.09	-1.31	-2.14	Mild	Major	↑
AF098	-2.76	-2.01	-1.82	Major	Major	=
CF005	-.67	-2.34	-2.97	Major	Major	=

Note. Major DP: $z \leq -1.7$ shown in red, Mild DP: $-1 \geq z \geq -1.69$ shown in yellow and Subjective DP: $z > -1$. Change column on far right indicates whether the DP classification severity rating changed when classifying using BIS versus accuracy; ↑ indicates rating increased, = indicates unchanged, ↓ indicates rating decreased.

5. Discussion

Twenty-four participants with self-reported face recognition problems (DPs) and 110 age matched controls completed four tests measuring different aspects of objective face recognition ability alongside a self-report questionnaire. To examine the best way of best identifying DP, we used both traditional accuracy measures (proportion correct) and a novel integrated measure, the Balanced Integration Score (BIS), which adjusts accuracy to take account of RT while controlling for speed-accuracy trade-offs. At an individual participant level, 15 individuals who self-reported face recognition difficulties were classified as having a major face recognition impairment using BIS, but only four were classified as having a major impairment when using accuracy alone. Overall, observed between-group effect sizes for computed global (averaged) memory scores were also larger using BIS than using accuracy alone (see Table 8).

5.1. *Is self-report a useful initial classification measure for DP?*

Before comparing how the objective face memory measures were able to detect and classify DP, we first checked that self-report – which we used to make a preliminary classification of participants into potential DPs or likely controls – was in fact a valid basis on which to make this preliminary classification. The extent to which self-report is a valid indication of face recognition ability has been much debated. Furthermore, direct comparisons are difficult to make across studies using different self-report questionnaires. Nonetheless, previous studies have reported that self-reported and objective face recognition measures are, at best, modestly correlated in naïve typical perceivers (Bobak et al., 2019; Matsuyoshi & Watanabe, 2021; Palermo et al., 2017) and in individuals who were aware that they met the diagnostic criteria for DP before completing the self-report questionnaire (Murray & Bate,

Table 11 – Showing how individual DP case performance compares using accuracy and BIS.

ID	Age	Self report	Old New Faces	FFT	CFMT	CBMT	CMT Difference	Global face memory	Face memory measures	
									No. impaired (of 4)	Severe impairment on ≥ 2 measures?
Accuracy										
AF019	68	-7.75	-3.88	-5.22	-1.64	.84	-2.48	-3.31	4	yes
AF002	25	-7.28	-.96	.41	-1.53	.11	-1.64	-.93	2	no
CF008	10	-7.18	.11	-3.33	-.78	.50	-1.27	-1.32	2	no
AF016	23	-7.01	-1.35	-2.63	-.83	-.35	-.48	-1.32	2	no
AF017	29	-6.60	-1.74	-.04	-1.53	1.36	-2.89	-1.55	3	yes
AF006	54	-6.16	-2.70	-1.58	-2.09	-.41	-1.67	-2.01	4	yes
AF003	55	-6.05	.88	-1.52	-1.55	.87	-2.42	-1.15	3	no
AF018	49	-5.72	.88	-4.30	-.91	.08	-.99	-1.33	1	no
CF059	16	-5.59	NA	-.58	-3.45	-2.84	-.60	-1.54	1 of 3	no*
AF009	42	-5.17	.88	-.63	-1.23	.77	-2.00	-.74	2	no
AF010	53	-5.17	-9.84	-2.63	-3.05	-.12	-2.93	-4.62	4	yes
CF007	8	-5.15	.30	-1.74	.90	1.06	-.16	-.18	1	no
AF007	31	-4.95	-.56	.34	-.53	1.02	-1.54	-.57	1	no
AF008	71	-4.78	-.17	-1.37	-.52	.84	-1.36	-.85	2	no
AF013	44	-4.73	.88	.92	.92	.67	.24	.74	0	no
AF022	64	-4.65	-.17	.72	.04	.75	-.70	-.03	0	no
AF060	61	-4.65	-.92	-1.92	-.42	.84	-1.27	-1.13	2	no
AF001	51	-4.62	-.91	-.24	-1.23	.38	-1.60	-1.00	2	no
AF004	31	-4.54	.62	-1.87	-.93	1.58	-2.51	-1.17	2	yes
AF099	70	-4.38	-3.14	NA	.33	.09	.23	-.86	1 of 3	no*
AF075	45	-3.64	.28	-1.40	-.80	-.02	-.78	-.67	1	no
AF021	50	-3.09	.88	-2.03	-1.55	.97	-2.52	-1.31	3	yes
AF098	67	-2.76	-3.14	-1.45	-1.74	.00	-1.74	-2.01	4	yes
CF005	17	-.66	-1.35	-1.98	-2.34	1.36	-3.70	-2.34	4	yes
Proportion impaired			.35	.65	.50	.04	.67	.58		.33
Balanced Integration Scores										
AF019	68	-7.75	-6.92	-5.22	-1.72	.02	-1.74	-3.90	4	yes
AF002	25	-7.28	-4.35	.41	-2.70	-.37	-2.33	-2.24	3	yes
CF008	10	-7.18	.37	-3.33	-.47	-.48	.01	-.86	1	no
AF016	23	-7.01	-2.11	-2.63	-1.52	.35	-1.87	-2.03	4	yes
AF017	29	-6.60	-3.17	-.04	-.17	2.50	-2.67	-1.51	2	yes
AF006	54	-6.16	-2.42	-1.58	-1.16	.18	-1.34	-1.63	4	no
AF003	55	-6.05	-.60	-1.52	-2.25	.04	-2.29	-1.66	3	yes
AF018	49	-5.72	-1.17	-4.30	-2.88	-1.65	-1.23	-2.40	4	yes
CF059	16	-5.59	NA	-.58	-5.53	-3.07	-2.46	-2.85	2 of 3	yes
AF009	42	-5.17	.03	-.63	-5.56	-.85	-4.71	-2.72	2	yes
AF010	53	-5.17	-12.13	-2.63	-3.48	.89	-4.37	-5.65	4	yes
CF007	8	-5.15	.23	-1.74	1.71	-.91	2.63	.71	1	no
AF007	31	-4.95	-2.88	.34	-4.67	-.85	-3.82	-2.76	3	yes
AF008	71	-4.78	-5.72	-1.37	-1.23	-2.19	.96	-1.84	3	yes
AF013	44	-4.73	-.80	.92	.79	.49	.29	.30	0	no
AF022	64	-4.65	-.28	.72	-.51	-.83	.32	.06	0	no
AF060	61	-4.65	-2.94	-1.92	-.07	.77	-.85	-1.44	2	yes
AF001	51	-4.62	-2.45	-.24	-.35	1.92	-2.26	-1.33	2	yes
AF004	31	-4.54	-.67	-1.87	-2.20	.85	-3.05	-1.95	3	yes
AF099	70	-4.38	-7.53	NA	.22	.11	.11	-2.40	1 of 3	no*
AF075	45	-3.64	-3.07	-1.40	-3.83	-2.37	-1.46	-2.44	4	yes
AF021	50	-3.09	-1.38	-2.03	-2.03	1.08	-3.11	-2.14	4	yes
AF098	67	-2.76	-3.84	-1.45	-1.02	-.05	-.97	-1.82	3	no
CF005	17	-.66	-2.27	-1.98	-2.73	2.18	-4.90	-2.97	4	yes
Proportion impaired			.70	.65	.67	.17	.67	.83		.71

Note. FFT = Famous Faces Test, CFMT = Cambridge Face Memory Tests, CBMT = Cambridge Bicycle memory Test, CMT difference = CFMT – CBMT, Global face memory = standardised mean average. BIS was not calculated for the FFT so the results are the same as for accuracy alone but are shown again in the second table for ease of comparison. Major impairment $z \leq -1.7$ shown in red, mild impairment $-1 \geq z \geq -1.69$ shown in yellow. All scores are standardised and centred on age-matched control means. The two farthest right columns show firstly the number of independent face processing measures on which a participant was impaired (mild or major) and secondly whether the participant showed an impairment of at least 2 SD below control means on at least two of the four face memory measures, * indicates that participant completed three rather than four tasks.

2019). By contrast, studies of individuals with DP, mainly using the PI20 (Shah et al., 2015), found that DPs *do* have insight into their own face processing difficulties (Burns et al., 2022; Gray et al., 2017; Livingston & Shah, 2018; Tsantani et al., 2021; Ventura et al., 2018), or at least the fact that their face recognition ability was poor relative to others (Palermo et al., 2017). Our data also support the use of self-report measures for classifying DP. All the potential DP participants except one made contact with our lab because they (or a parent) believed they struggled with face recognition or had a family history of DP, but, unlike the DP participants in the Murray and Bate (2019) study, participants in the present study were *not* told prior to completion of the self-report questionnaire whether they met the 'diagnostic' threshold for DP. They were arguably therefore less likely to be influenced by a DP 'diagnosis' when subjectively rating their face recognition ability. However, four individuals reporting poor face recognition ability had previously participated in other face recognition studies which may have provided insight into their ability.

It has been reported that women rated their prosopagnosia symptoms as more severe than men (Murray & Bate, 2019) but we found no support for this. Among adult DPs (n = 20; 15 female, 5 male) who showed both objective and subjective impairment, we found no evidence of gender differences in PI20 scores ($d = .32$, $p = .604$), and anecdotal evidence for the null hypothesis ($BF_{10} = .52$). It is possible that there is an interaction between gender and status (naive; informed) that should be explored further by researchers who disclose status to participants prior to administering a self-report questionnaire but this was not a relevant issue in our study.

Regression analysis showed that, overall, the standardised subjective rating score was a significant predictor of global face memory BIS, $F(1,124) = 57.7$, $p < .001$ explaining around 31 % of the variance in objective scores (adjusted $R^2 = .31$). When considering adults and children separately, unsurprisingly adults' subjective rating of their *own* face recognition ability was a better predictor of group than parental report of *their child's* face recognition ability, explaining around 40 % and 6 % respectively of the variance in global face memory BIS. Binomial logistic regression analysis showed that parental report was nevertheless a significant predictor of group suggesting that parents did, on average, have insight into their children's face memory ability in binary terms, i.e., whether it was much worse than average or not. However, parents' ability to predict more precisely their child's ability *relative* to their age group, as measured by global face memory BIS z score, was only just above chance ($p = .044$).

One factor that may have limited parental ability to accurately judge their child's face recognition ability was the parent's own face recognition ability. Face recognition is a highly heritable ability (Wilmer et al., 2010) and many reported cases of DP have a known family history (De Haan, 1999; Duchaine et al., 2007; Gruter et al., 2008; Lee et al., 2010; Schmalzl et al., 2006). In our study, the face recognition ability rating for three of the four child DP candidates was provided by a parent who themselves reported having difficulty recognising faces. Thus, it is likely that these parents lacked an accurate reference point for judging typical, and consequently atypical, face recognition ability.

5.2. DP performance on Old New Faces

In line with previous work (Dalrymple et al., 2014), we found that Old New Faces was a useful test for classifying DP. Although the test could be criticised as being more of an image memory than a face recognition test since the images used at study and test are identical, Dalrymple et al. (2014) reported that all adult DPs ($n = 16$) showed impaired accuracy on the Old New Faces test. By contrast, 0/16 were impaired at a matched old/new houses test and only 1/16 was impaired in a matched old/new horses test. Among children, 4/6 DPs were significantly impaired versus controls on the Old New Faces test but – similar to adults – 0/6 were impaired on the matched object task, in this case an old/new flowers task. By contrast, the authors observed that 10/16 adult DPs were *unimpaired* on the Cambridge Face Perception Test accuracy (CFPT, Duchaine et al., 2007). Together, these results suggest firstly that the Old New Faces test is useful for identifying DP and that the impairment detected by the task is face specific. Secondly, the results suggest that the presence of a memory demand in the Old New Faces task – even when using the same image at study and test – produces different patterns of impairment compared with the CFPT which is also a test of face matching but without memory demands. Overall, Dalrymple et al. (2014) show that the Old New Faces task is a useful source of converging evidence of face recognition difficulties when the CFPT is not suitable. Our data support these findings, namely that the DP group, on average, produced significantly lower accuracy and BIS scores than the control group. Additionally, on the Old New Faces task we found stronger Bayesian evidence and larger effect sizes for a group difference for BIS versus accuracy alone.

5.3. DP performance on Famous Faces Test

Despite not being able to calculate BIS for FFT, since reading speed would have confounded RT, we nevertheless found this novel FFT to be a useful measure for classifying DP using identification accuracy. This measure produced the strongest accuracy effect size ($d = 1.15$, $p < .001$) of the four individual face memory tests we administered. Additionally, Bayesian analysis showed extremely high evidence for a group difference ($BF_{10} = 204{,}436$), again the highest of any single test (Table 8). To check that group differences were not driven by a small number of individuals with extreme scores, we conducted individual case analysis on all accuracy measures (Tables 1 and 11). On the FFT, 19/23 potential DPs produced identification accuracy scores below the age group mean, and of these, 9/23 scored significantly below mean control accuracy. A previous large-scale study (Bate et al., 2019) found evidence of a dissociation between memory for familiar faces (FFT) and memory for newly-learned faces in 63 of 165 individuals and suggested that long term memory for familiar faces (as indexed by a FFT) may be selectively impaired in DP. Our data support the use of an FFT as one of several measures to classify DP, even though BIS could not be calculated.

5.4. *Do accuracy or BIS measures best predict group membership?*

The logistic regression model that included three BIS face memory measures plus FFT accuracy (Model 2, see Table 9) was *the most sensitive* and classified the most self-reported DP participants (68.2 %) as being objectively impaired. Importantly, Model 2's *highest sensitivity* did not come at the cost of reduced *specificity* as the ability of this model to classify controls was very similar (94.9 % *vs* 96.7 % for Models 2 and 5 respectively). In comparison, CFMT accuracy (Model 5), which is the measure traditionally used to diagnose DP, classified only 16.7 % of our self-reported DP group. Although AUC was only slightly lower than this for Model 3 (AUC = .87, global face memory accuracy) and Model 4 (AUC = .88, global face memory BIS), these models were much less sensitive than Model 2, correctly classifying only 29.2 % and 54.2 % of self-reported DPs respectively.

One DP participant produced very low scores on three of the four tests. We therefore checked to ensure that this potential outlier was not unduly influencing results by repeating the logistic regression modelling with this participant removed. Although exact values changed, the pattern of results was very similar and Model 2 remained the best model for classifying DP.

We also reran these analyses on a sub-sample of only the self-reported DPs who showed mild or major impairment on the global face memory accuracy score, that is without the ten participants classed as "subjective DPs" and found the same pattern of results (see supplementary materials sections 2 and 3). Again, Model 2 with BIS as the outcome measure remained the best model (AUC = .98, $p < .001$). As would be expected from this approach (i.e., excluding the potential DPs who achieved typical accuracy due to atypical RT) the group effect size differences when comparing accuracy and BIS were smaller than when we analysed the full sample of potential DPs. This is because in this alternative approach, only the self-reported DPs with impaired accuracy are included in the analysis and so it logically follows that the utility of accuracy as a classification measure would improve. Nevertheless, the fact that BIS measures – even among this sub-sample – were more sensitive than accuracy measures for classifying DP further strengthens the value of BIS.

Our results confirm the importance of considering RT as well as accuracy when classifying DP and these findings are in line with recent studies (Fysh & Ramon, 2022; Stacchi et al., 2020) and a large literature review by Geskin and Behrmann (2018).

However, it is possible that RT (and therefore BIS) is more useful in some tasks and paradigms than others. Our data show that for the CFMT, a test specifically designed to detect DP, BIS added little or no additional information compared to accuracy alone. Effect sizes for both measures were similar, and indeed slighter larger for accuracy. Others have also questioned whether RT always adds value over accuracy. A recent study (DeGutis et al., 2022) investigating both accuracy and response times on a face matching task, the Benton Face Recognition Test (BFRT-c, Rossion & Michel, 2018) reported that RT alone did not reliably predict group membership. The BFRT-c is an updated version of the original face matching test (Benton et al., 1983) that was specifically designed to emphasise speed as well as accuracy. However, there are two important differences between the BFRT-c and the tasks used here. Firstly, despite its name, the BFRT is a perceptual task since it involves no memory demands. Secondly, the BFRT-c requires participants to click up to *three* faces from a choice of six meaning that motor control is likely to influence RT more than it would in our tasks where participants had to make a single response. Finally, as the authors explain, the BFRT-c design means that it is not possible to analyse RT on correct trials only as is common practice in other face cognition tasks. Notably RT from correct trials *only* is the measure used to calculate BIS (Liesefeld & Janczyk, 2019). Despite this, some theoretical papers have incorrectly used RT on *all* trials when comparing BIS with other integrated measures of speed accuracy which could lead to confusion about how to calculate BIS (for a fuller explanation see Liesefeld & Janczyk, 2022). BIS may therefore not be informative on tasks where RT measures include both incorrect and correct trials.

Our findings support the use of BIS as an *integrated* measure that adjusts accuracy to account for RT. However, we make no claims about the use of RT as a sole measure which was the question DeGutis et al. (2022) investigated. As the authors correctly caution, if researchers wish to use RT *instead* of accuracy on a given task, RT should first be validated as a measure. In the present study we were instead interested in whether accuracy and RT *together* might be more informative than accuracy alone. Our data show that accuracy and RT together (BIS) explained more of the variance in self-reported face recognition ability than accuracy alone and, additionally, showed greater sensitivity for classification of DP than accuracy alone. A third benefit of BIS was that for the global face memory score and Old New Faces task, the observed effect sizes were larger for BIS than accuracy (on the CFMT there was little difference). This is an important consideration in neuropsychology research where effect sizes are typically modest and sample sizes often small, thereby limiting power (McIntosh & Rittmo, 2021). Using BIS thus provides a practical solution to ensuring that no impaired participants are omitted from analyses. This increased power to detect differences between different populations will allow for better understanding of deficits underpinning DP and provide pathways to effective training. Better classification will also ensure that neuropsychological research is more ethical. The British Psychological Society Code of Human Research Ethics calls for maximising the benefit to participants (point 2.4) from inception to dissemination (Oates et al., 2021). Thus by improving the methods of studying DP, our work contributes to that principle of ethical research.

In addition to offering a practical solution, BIS, as an integrated measure of both accuracy and RT, is a more ecologically valid approach to identifying DP than accuracy alone. This is because in typical social interactions, the amount of time an individual with DP has to make a correct identification is effectively the time it takes the person they are interacting with to recognise the DP. It is time limited. If the DP has not recognised the face before the person greets or speaks to them, then this will appear to be a failure of recognition (even

if – given much longer to study the face – the DP might have been able to identify the person). Response latency and accuracy are therefore *both* important elements with regard to ecological validity.

5.5. *Use of global scores and choice of cut off*

In their editorial of a special issue on DP, Bate and Tree (2017) argued that seeking converging evidence of impairment across *multiple* tests provides more compelling evidence of a true deficit as well as mitigating the risk that unimpaired controls may be misclassified as DP. The same argument is made for super recogniser research (Bate et al., 2018; Ramon, 2021). We therefore administered four separate tests and used these scores to compute a global face memory score (see Table 7). Global scores showed large differences between DPs and controls and thus appear useful. However, logistic regression showed the global face memory scores produced slightly lower AUCs than Models 1 and 2 which used non-averaged scores from the individual tests. Our results therefore suggest that although global effect sizes are larger than those from individual tasks and therefore will increase the power to detect group level differences, the four *individual* BIS measures (CFMT BIS, CMT Difference BIS, Old New Faces BIS and FFT accuracy) best predicted group membership. This could be because averaging results can attenuate the insights provided by multiple individual test scores. Researchers may need to decide whether they wish to prioritise sensitivity (classification), or ability to detect group difference (effect sizes). Individual BIS measures were slightly better for classification purposes and global measures showed larger effect sizes which may be an important consideration when sample sizes are small.

Finally, we also compared the patterns of impairment across the four independent face memory measures (Table 11) at the individual case level. Using this alternative, and more commonly used approach, results once again supported the overall conclusion that BIS is a more sensitive measure for classifying DP than accuracy alone. Data showed that more than twice as many (71 %) of the self-reported DP participants showed severe objective impairment (<−2 *SD*) on at least two individual face memory BIS measures compared with only 33 % who were severely impaired on at least two accuracy measures. BIS therefore appears equally valuable whether using a global (averaged) score or multiple independent measures of face memory. Notably, although we used more liberal cut offs of −1 *SD* to classify mild and −1.7 *SD* for major impairment to classify potential DPs, *all* participants who were classified as impaired (mild *or* major) on global face memory BIS also showed *severe* impairment (<−2 *SD*) on two individual BIS measures suggesting that a more liberal cut off is justified provided multiple objective tests are administered.

5.6. *Online versus lab-based data*

We compared test results with previous literature and found that our control group mean accuracy data was broadly very similar, almost always within 1 SD, to published lab-based studies testing similar age groups, thus giving us confidence in our results. However, one important difference we wish to highlight was that we observed a greater score variability, and thus larger standard deviations, compared with previous literature. For example, CFMT accuracy in both the 14–25 and 36–59 years control groups was 81 %, almost identical to previously reported scores of 80.4 % (Duchaine & Nakayama, 2006). By contrast, the standard deviations in these control groups were 14 % and 13 % respectively versus 11 % in the original Duchaine and Nakayama (2006) study. Among control participants aged 60–74 years, mean accuracy data (72 % ± 15 %) was again very similar to published age group norms for 60–69 year olds of 70.14 % ± 12 % (Bowles et al., 2009) but standard deviations were once again higher. We also observed similar patterns in the child data.

This finding is highly relevant for classification because is DP is 'diagnosed' or classified using the mean and standard deviations of the control group. For online research, we therefore caution against using "standard cut-offs" on popular tests such as the CFMT as these may not be valid for online data collection. Instead, online specific norms should be used. Further, our data support the need to use age group norms (Bowles et al., 2009) since our data also showed that accuracy and standard deviations both varied by age group resulting in different cut offs for mild and major impairment in each age group. This finding is also in line with results from the large sample of 165 DPs (Bate et al., 2019).

6. Conclusion

In conclusion, our key finding is that, whether using multiple individual scores or an averaged global score, and regardless of whether the cut off applied was −1.7 SD or −2 SD, BIS was a more sensitive measure of difference between self-reported DPs and controls than accuracy alone. Furthermore, BIS better predicted group membership compared with accuracy alone. Our data show that using four measures (Old New Faces BIS, CFMT BIS, CMT difference BIS and FFT identification accuracy) alongside the PI20 to classify DP *considerably improved sensitivity* (captured more DPs) with *no reduction in specificity* (did not decrease the proportion of controls correctly classified) compared with traditional accuracy measures. These results have important applied value for researchers who must identify and classify DP. Using measures that show strong effect sizes can increase power, an important consideration in research into rare conditions such as DP where large sample sizes are difficult to achieve. Improved classification will also allow a better understanding of the underpinnings of DP and avoid unnecessary exclusion of participants whose impairments may be masked by speed-accuracy trade-offs.

Data availability statement

Data and code are available at https://osf.io/f496b/.

CRediT author statement:

Judith Lowes: Conceptualisation, Data curation, Formal analysis, Writing - Original draft preparation, Visualisation,

Investigation, Funding acquisition, Project administration. **Anna Bobak**: Conceptualisation, Supervision, Writing - Reviewing and Editing, Funding Acquisition. **Peter Hancock**: Conceptualisation, Supervision; Writing - Reviewing and Editing.

Funding

This work was supported by an Economic and Social Research Council studentship (grant number ES/P000681/1) and grant from the British Psychological Society Cognitive Section to Judith Lowes and a Leverhulme Trust Early Career Fellowship (grant number ECF-2019-416) to Anna K. Bobak. **Role of the funding source**: Funders had no influence on the study.

Open practices

The study in this article earned Open Data and Preregistered badges for transparent practices. The data used in this study are available at: https://osf.io/f496b/ and preregistered study at: https://osf.io/qne8d/.

Declaration of competing interest

No competing interests.

Acknowledgements

Thanks to Dr Gabriele Chierchia, Dr Kirsten Dalrymple, and Professor Brad Duchaine for kindly sharing their materials and online experimental set ups with us and to Professor Jane Riddoch for permission to create an online version of the Birmingham Object Recognition Battery. We would also like to thank undergraduate student Mhairi Webster for helping to develop the Famous Faces Test and all the participants and parents for taking part. We also thank the funders (ESRC and British Psychological Society Cognitive Section to Judith Lowes and the Leverhulme Trust ECF-2019-416 to Anna Bobak) for funding this research.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cortex.2023.12.011.

REFERENCES

Albonico, A., Malaspina, M., & Daini, R. (2017). Italian normative data and validation of two neuropsychological tests of face recognition: Benton Facial Recognition Test and Cambridge Face Memory Test. *Neurological Sciences, 38*(9), 1637–1643. https://doi.org/10.1007/s10072-017-3030-6. APA PsycInfo.

Barton, J. J. S., Albonico, A., Susilo, T., Duchaine, B., & Corrow, S. L. (2019). Object recognition in acquired and developmental prosopagnosia. *Cognitive Neuropsychology, 36*(1–2), 54–84. https://doi.org/10.1080/02643294.2019.1593821

Bate, S., Adams, A., & Bennetts, R. J. (2020). Guess who? Facial identity discrimination training improves face memory in typically developing children. *Journal of Experimental Psychology: General, 149*(5), 901–913. https://doi.org/10.1037/xge0000689

Bate, S., Bennetts, R. J., Gregory, N., Tree, J. J., Murray, E., Adams, A., Bobak, A. K., Penton, T., Yang, T., & Banissy, M. J. (2019). Objective patterns of face recognition deficits in 165 adults with self-reported developmental prosopagnosia. *Brain Sciences, 9*(6). https://doi.org/10.3390/brainsci9060133, 133–133.

Bate, S., Bennetts, R., Mole, J. A., Ainge, J. A., Gregory, N. J., Bobak, A. K., & Bussunt, A. (2015). Rehabilitation of face-processing skills in an adolescent with prosopagnosia: Evaluation of an online perceptual training programme. *Neuropsychological rehabilitation, 25*(5), 733–762.

Bate, S., Frowd, C., Bennetts, R., Hasshim, N., Murray, E., Bobak, A. K., Wills, H., & Richards, S. (2018). Applied screening tests for the detection of superior face recognition. *Cognitive Research: Principles and Implications, 3*(1), 22. https://doi.org/10.1186/s41235-018-0116-5

Bate, S., & Tree, J. J. (2017). The definition and diagnosis of developmental prosopagnosia. *The Quarterly Journal of Experimental Psychology, 70*(2), 193–200. https://doi.org/10.1080/17470218.2016.1195414

Behrmann, M., Avidan, G., Gao, F., & Black, S. (2007). Structural imaging reveals anatomical alterations in inferotemporal cortex in congenital prosopagnosia. *Cerebral Cortex, 17*(10), 2354–2363. https://doi.org/10.1093/cercor/bhl144. APA PsycInfo.

Behrmann, M., Avidan, G., Marotta, J., & Kimchi, R. (2005). Detailed exploration of face-related processing in congenital prosopagnosia: 1. Behavioral findings. *Journal of Cognitive Neuroscience, 17*, 1130–1149. https://doi.org/10.1162/0898929054475154

Bennetts, R. J., Murray, E., Boyce, T., & Bate, S. (2017). Prevalence of face recognition deficits in middle childhood. *Quarterly Journal of Experimental Psychology, 70*(2), 234–258. https://doi.org/10.1080/17470218.2016.1167924

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

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

Biotti, F., & Cook, R. (2016). Impaired perception of facial emotion in developmental prosopagnosia. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior, 81*, 126–136. https://doi.org/10.1016/j.cortex.2016.04.008

Bobak, A. K., Jones, A. L., Hilker, Z., Mestry, N., Bate, S., & Hancock, P. J. B. (2023). Data-driven studies in face identity processing rely on the quality of the tests and data sets. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior*. https://doi.org/10.1016/j.cortex.2023.05.018

Bobak, A. K., Mileva, V. R., & Hancock, P. J. B. (2019). Facing the facts: Naive participants have only moderate insight into their face recognition and face perception abilities. *Quarterly Journal of Experimental Psychology, 72*(4), 872–881. https://doi.org/10.1177/1747021818776145

Bobak, A. K., Parris, B. A., Gregory, N. J., Bennetts, R. J., & Bate, S. (2017). Eye-movement strategies in developmental

prosopagnosia and 'super' face recognition. *Quarterly Journal of Experimental Psychology, 70*(2), 201–217. https://doi.org/10.1080/17470218.2016.1161059

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

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)? *Psychologica Belgica, 51*(1), Article 1. https://doi.org/10.5334/pb-51-1-5

Burns, E. J., Gaunt, E., Kidane, B., Hunter, L., & Pulford, J. (2022). A new approach to diagnosing and researching developmental prosopagnosia: Excluded cases are impaired too. *Behavior Research Methods*. https://doi.org/10.3758/s13428-022-02017-w

Burns, E. J., Tree, J. J., & Weidemann, C. T. (2014). Recognition memory in developmental prosopagnosia: Electrophysiological evidence for abnormal routes to face recognition. *Frontiers in Human Neuroscience, 8*. https://doi.org/10.3389/fnhum.2014.00622. APA PsycInfo.

Chierchia, G., Fuhrmann, D., Knoll, L. J., Pi-Sunyer, B. P., Sakhardande, A. L., & Blakemore, S.-J. (2019). The Matrix Reasoning Item Bank (MaRs-IB): Novel, open-access abstract reasoning items for adolescents and adults. *Royal Society Open Science, 6*(10), Article 190232. https://doi.org/10.1098/rsos.190232

Corrow, S. L., Dalrymple, K. A., & Barton, J. J. (2016). Prosopagnosia: Current perspectives. *Eye and Brain, 8*, 165. https://doi.org/10.2147/EB.S92838

Crawford, J. R., & Garthwaite, P. H. (2007). Comparison of a single case to a control or normative sample in neuropsychology: Development of a Bayesian approach. *Cognitive Neuropsychology, 31*.

Crawford, J. R., Garthwaite, P. H., & Porter, S. (2010). Point and interval estimates of effect sizes for the case-controls design in neuropsychology: Rationale, methods, implementations, and proposed reporting standards. *Cognitive Neuropsychology, 27*(3), 245–260. https://doi.org/10.1080/02643294.2010.513967

Dalrymple, K. A., & Duchaine, B. (2016). Impaired face detection may explain some but not all cases of developmental prosopagnosia. *Developmental Science, 19*(3), 440–451. https://doi.org/10.1111/desc.12311

Dalrymple, K. A., Garrido, L., & Duchaine, B. (2014). Dissociation between face perception and face memory in adults, but not children, with developmental prosopagnosia. *Developmental Cognitive Neuroscience, 10*, 10–20. https://doi.org/10.1016/j.dcn.2014.07.003 (Journal Article).

Dalrymple, K. A., & Palermo, R. (2016). Guidelines for studying developmental prosopagnosia in adults and children. *Wiley Interdisciplinary Reviews-Cognitive Science, 7*(1), 73–87. https://doi.org/10.1002/wcs.1374

De Haan, E. H. F. (1999). A familial factor in the development of face recognition deficits. *Journal of Clinical and Experimental Neuropsychology, 21*(3), 312–315. https://doi.org/10.1076/jcen.21.3.312.917

DeGutis, J., Bahierathan, K., Barahona, K., Lee, E., Evans, T. C., Shin, H. M., Mishra, M., Likitlersuang, J., & Wilmer, J. B. (2023). What is the prevalence of developmental prosopagnosia? An empirical assessment of different diagnostic cutoffs. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior, 161*, 51–64. https://doi.org/10.1016/j.cortex.2022.12.014

DeGutis, J., Li, X., Yosef, B., & Mishra, M. V. (2022). Not so fast! Response times in the computerized Benton Face Recognition Test may not reflect face recognition ability. *Cognitive Neuropsychology, 39*(3–4), 155–169. https://doi.org/10.1080/02643294.2022.2114824

Dobel, C., Bölte, J., Aicher, M., & Schweinberger, S. R. (2007). Prosopagnosia without apparent cause: Overview and diagnosis of six cases. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior, 43*(6), 718–733. https://doi.org/10.1016/S0010-9452(08)70501-X. APA PsycInfo.

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

Duchaine, B. C., & Nakayama, K. (2004). Developmental prosopagnosia and the Benton Facial Recognition Test. *Neurology, 62*(7), 1219–1220. https://doi.org/10.1212/01.WNL.0000118297.03161.B3. APA PsycInfo.

Duchaine, B., & Nakayama, K. (2005). Dissociations of face and object recognition in developmental prosopagnosia. *Journal of Cognitive Neuroscience, 17*(2), 249–261. https://doi.org/10.1162/0898929053124857

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

Draheim, C., Mashburn, C. A., Martin, J. D., & Engle, R. W. (2019). Reaction time in differential and developmental research: A review and commentary on the problems and alternatives. *Psychol Bull, 145*(5), 508–535. https://doi.org/10.1037/bul0000192. Epub 2019 Mar 21. PMID: 30896187.

European Dyslexia Association. (n.d.). *What is dyslexia – European Dyslexia Association*. What Is Dyslexia. Retrieved 10 January 2022, from https://eda-info.eu/what-is-dyslexia/.

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

Gelman, A., Hill, J., & Yajima, M. (2012). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness, 5*(2), 189–211. https://doi.org/10.1080/19345747.2011.618213

Geskin, J., & Behrmann, M. (2018). Congenital prosopagnosia without object agnosia? A literature review. *Cognitive Neuropsychology, 35*(1–2), 4–54. https://doi.org/10.1080/02643294.2017.1392295

Gray, K. L. H., Bird, G., & Cook, R. (2017). Robust associations between the 20-item prosopagnosia index and the Cambridge Face Memory Test in the general population. *Royal Society Open Science, 4*(3). https://doi.org/10.1098/rsos.160923, 160923–160923.

Gruter, T., Gruter, M., & Carbon, C.-C. (2008). Neural and genetic foundations of face recognition and prosopagnosia. *Journal of Neuropsychology, 2*(1, SI), 79–97. https://doi.org/10.1348/174866407X231001

Johnen, A., Schmukle, S. C., Hüttenbrink, J., Kischka, C., Kennerknecht, I., & Dobel, C. (2014). A family at risk: Congenital prosopagnosia, poor face recognition and visuoperceptual deficits within one family. *Neuropsychologia, 58*, 52–63. https://doi.org/10.1016/j.neuropsychologia.2014.03.013. APA PsycInfo.

Kennerknecht, I., Grueter, T., Welling, B., Wentzek, S., Horst, J., Edwards, S., & Grueter, M. (2006). First report of prevalence of non-syndromic hereditary prosopagnosia (HPA). *American Journal of Medical Genetics. Part A, 140A*(15), 1617–1622. https://doi.org/10.1002/ajmg.a.31343

Kennerknecht, I., Ho, N., & Wong, V. (2008). Prevalence of hereditary prosopagnosia (HPA) in Hong Kong Chinese population. *American Journal of Medical Genetics. Part A, 146A*, 2863–2870. https://doi.org/10.1002/ajmg.a.32552

Kruschke, J. K. (2010). What to believe: Bayesian methods for data analysis. *Trends in Cognitive Sciences, 14*(7), 293–300. https://doi.org/10.1016/j.tics.2010.05.001

Lee, Y., Duchaine, B., Wilson, H. R., & Nakayama, K. (2010). Three cases of developmental prosopagnosia from one family: Detailed neuropsychological and psychophysical investigation of face processing. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior, 46*(8), 949–964. https://doi.org/10.1016/j.cortex.2009.07.012

Liesefeld, H. R., & Janczyk, M. (2019). Combining speed and accuracy to control for speed-accuracy trade-offs(?). *Behavior Research Methods, 51*(1), 40–60. https://doi.org/10.3758/s13428-018-1076-x

Liesefeld, H. R., & Janczyk, M. (2022). Same same but different: Subtle but consequential differences between two measures to linearly integrate speed and accuracy (LISAS vs. BIS). *Behavior Research Methods*. https://doi.org/10.3758/s13428-022-01843-2

Livingston, L. A., & Shah, P. (2018). People with and without prosopagnosia have insight into their face recognition ability. *Quarterly Journal of Experimental Psychology, 71*(5), 1260–1262. https://doi.org/10.1080/17470218.2017.1310911

Lowes, J., Hancock, P. J., & Bobak, A. K. (2024). *Evidence for different visual processing strategy for non-face stimuli in developmental prosopagnosia*. *PsyArXiv*. https://doi.org/10.31234/osf.io/ky7cs.

Matsuyoshi, D., & Watanabe, K. (2021). People have modest, not good, insight into their face recognition ability: A comparison between self-report questionnaires. *Psychological Research, 85*(4), 1713–1723. https://doi.org/10.1007/s00426-020-01355-8

Mayer, M., & Ramon, M. (2023). Improving forensic perpetrator identification with Super-Recognizers. *Proceedings of the National Academy of Sciences of the United States of America, 120*(20), Article e2220580120. https://doi.org/10.1073/pnas.2220580120

McIntosh, R. D., & Rittmo, J.Ö. (2021). Power calculations in single-case neuropsychology: A practical primer. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior, 135*, 146–158. https://doi.org/10.1016/j.cortex.2020.11.005

Mishra, M. V., Fry, R. M., Saad, E., Arizpe, J. M., Ohashi, Y.-G. B., & DeGutis, J. M. (2021). Comparing the sensitivity of face matching assessments to detect face perception impairments. *Neuropsychologia, 163*, Article 108067. https://doi.org/10.1016/j.neuropsychologia.2021.108067

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

Murray, E., Hills, P. J., Bennetts, R. J., & Bate, S. (2018). Identifying hallmark symptoms of developmental prosopagnosia for non-experts. *Scientific Reports, 8*. https://doi.org/10.1038/s41598-018-20089-7 (Journal Article), 1690–1690.

Oates, J., Carpenter, D., Fisher, M., Goodson, S., Hannah, B., Kwiatkowski, R., Prutton, K., Reeves, D., & Wainwright, T. (2021). BPS Code of Human Research Ethics (p. bpsrep.2021.inf180). *British Psychological Society*. https://doi.org/10.53841/bpsrep.2021.inf180

Palermo, R., Rossion, B., Rhodes, G., Laguesse, R., Tez, T., Hall, B., Albonico, A., Malaspina, M., Daini, R., Irons, J., Al-Janabi, S., Taylor, L. C., Rivolta, D., & McKone, E. (2017). Do people have insight into their face recognition abilities? *Quarterly Journal of Experimental Psychology (2006), 70*(2), 218–233. https://doi.org/10.1080/17470218.2016.1161058

Pascalis, O., de Viviés, X. de M., Anzures, G., Quinn, P. C., Slater, A. M., Tanaka, J. W., & Lee, K. (2011). Development of face processing. *Wiley Interdisciplinary Reviews. Cognitive Science, 2*(6), 666–675. https://doi.org/10.1002/wcs.146

R Core Team. (2021). *R: The R project for statistical computing* [Computer software]. R Foundation for Statistical Computing https://www.r-project.org/.

Ramon, M. (2021). Super-recognizers—A novel diagnostic framework, 70 cases, and guidelines for future work. *Neuropsychologia, 158*. https://doi.org/10.1016/j.neuropsychologia.2021.107809. APA PsycInfo.

Riddoch, M. J., & Humphreys, G. W. (1993). *Birmingham Object Recognition Battery*. Lawrence Erlbaum Associates.

Rivolta, D., Palermo, R., & Schmalzl, L. (2013). What is overt and what is covert in congenital prosopagnosia? *Neuropsychology Review, 23*(2), 111–116. https://doi.org/10.1007/s11065-012-9223-0

Robotham, R. J., & Starrfelt, R. (2018). Tests of whole upright face processing in prosopagnosia: A literature review. *Neuropsychologia, 121*, 106–121. https://doi.org/10.1016/j.neuropsychologia.2018.10.018 (Journal Article).

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

Ruxton, G. D. (2006). The unequal variance t-test is an underused alternative to Student's t-test and the Mann–Whitney U test. *Behavioral Ecology, 17*(4), 688–690. https://doi.org/10.1093/beheco/ark016

Sachdev, P. S., Blacker, D., Blazer, D. G., Ganguli, M., Jeste, D. V., Paulsen, J. S., & Petersen, R. C. (2014). Classifying neurocognitive disorders: The DSM-5 approach. *Nature Reviews Neurology, 10*(11), 634–642. https://doi.org/10.1038/nrneurol.2014.181

Schmalzl, L., Palermo, R., & Coltheart, M. (2006). Congenital prosopagnosia: A family study. *Australian Journal of Psychology, 58* (Journal Article), 89–89.

Selker, R., Love, J., Dropmann, D., & Moreno, V. (2022). *jmv: The 'jamovi' Analyses (2.3.4)* [Computer software] https://CRAN.R-project.org/package=jmv.

Shah, P., Gaule, A., Gaigg, S. B., Bird, G., & Cook, R. (2015). Probing short-term face memory in developmental prosopagnosia. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior, 64*, 115–122. https://doi.org/10.1016/j.cortex.2014.10.006

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

Stumps, A., Saad, E., Rothlein, D., Verfaellie, M., & DeGutis, J. (2020). Characterizing developmental prosopagnosia beyond face perception: Impaired recollection but intact familiarity recognition. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior, 130*, 64–77. https://doi.org/10.1016/j.cortex.2020.04.016

Towler, J., Fisher, K., & Eimer, M. (2018). Holistic face perception is impaired in developmental prosopagnosia. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior, 108*, 112–126. https://doi.org/10.1016/j.cortex.2018.07.019 (Journal Article).

Townsend, J. T., & Ashby, F. G. (1983). *Stochastic modelling of elementary psychological processes*. New York, NY: Cambridge University Press.

Tsantani, M., Vestner, T., & Cook, R. (2021). The Twenty Item Prosopagnosia Index (PI20) provides meaningful evidence of face recognition impairment. *Royal Society Open Science, 8*(11), Article 202062. https://doi.org/10.1098/rsos.202062

Ventura, P., Livingston, L. A., & Shah, P. (2018). Adults have moderate-to-good insight into their face recognition ability: Further validation of the 20-item Prosopagnosia Index in a Portuguese sample. *The Quarterly Journal of Experimental Psychology, 71*(12), 2677–2679. https://doi.org/10.1177/1747021818765652. APA PsycInfo.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., … Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software, 4*(43), 1686. https://doi.org/10.21105/joss.01686

Wilcockson, T. D. W., Burns, E. J., Xia, B., Tree, J., & Crawford, T. J. (2020). Atypically heterogeneous vertical first fixations to faces in a case series of people with developmental prosopagnosia. *Visual Cognition, 28*(4), 311–323. https://doi.org/10.1080/13506285.2020.1797968

Wilmer, J. B., Germine, L., Chabris, C. F., Chatterjee, G., Williams, M., Loken, E., Nakayama, K., & Duchaine, B. (2010). Human face recognition ability is specific and highly heritable. *Proceedings of the National Academy of Sciences of the United States of America, 107*(11), 5238–5241. https://doi.org/10.1073/pnas.0913053107

World Health Organisation. (2018). *Autistic spectrum disorders* (2 April 2018). Retrieved from the World Health Organisation web site: https://www.who.int/news-room/fact-sheets/detail/autism-spectrum-disorders.