Available online at www.sciencedirect.com

ScienceDirect

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

What is the prevalence of developmental prosopagnosia? An empirical assessment of different diagnostic cutoffs

Joseph DeGutis [a,b,*], *Kanisha Bahierathan* [a,b], *Katherine Barahona* [a,b], *EunMyoung Lee* [a,b], *Travis C. Evans* [a,b], *Hye Min Shin* [a], *Maruti Mishra* [c], *Jirapat Likitlersuang* [a,b] *and Jeremy B. Wilmer* [d]

[a] *Boston Attention and Learning Laboratory, VA Boston Healthcare System, Boston, MA, USA*
[b] *Department of Psychiatry, Harvard Medical School, Boston, MA, USA*
[c] *Department of Psychology, University of Richmond, Richmond, VA, USA*
[d] *Department of Psychology, Wellesley College, Wellesley, MA, USA*

ARTICLE INFO

Article history:
Received 2 May 2022
Reviewed 24 August 2022
Revised 23 October 2022
Accepted 19 December 2022
Action editor Stephan Schweinberger
Published online 4 February 2023

Keywords:
Developmental prosopagnosia
Face recognition
Diagnostic cutoffs
Population prevalence

ABSTRACT

The prevalence of developmental prosopagnosia (DP), lifelong face recognition deficits, is widely reported to be 2–2.5%. However, DP has been diagnosed in different ways across studies, resulting in differing prevalence rates. In the current investigation, we estimated the range of DP prevalence by administering well-validated objective and subjective face recognition measures to an unselected web-based sample of 3116 18-55 year-olds and applying DP diagnostic cutoffs from the last 14 years. We found estimated prevalence rates ranged from .64–5.42% when using a z-score approach and .13–2.95% when using a percentile approach, with the most commonly used cutoffs by researchers having a prevalence rate of .93% (z-score, .45% when using percentiles). We next used multiple cluster analyses to examine whether there was a natural grouping of poorer face recognizers but failed to find consistent grouping beyond those with generally above versus below average face recognition. Lastly, we investigated whether DP studies with more relaxed diagnostic cutoffs were associated with better performance on the Cambridge Face Perception Test. In a sample of 43 studies, there was a weak nonsignificant association between greater diagnostic strictness and *better* DP face perception accuracy (Kendall's tau-b correlation, τb =.18 z-score; τb = .11 percentiles). Together, these results suggest that researchers have used more conservative DP diagnostic cutoffs than the widely reported 2–2.5% prevalence. We discuss the strengths and weaknesses of using more inclusive cutoffs, such as identifying mild and major forms of DP based on DSM-5.

Published by Elsevier Ltd.

[*] *Corresponding author*. Boston Attention and Learning Laboratory, VA Boston Healthcare System, 150 S. Huntington Ave. (182JP), Boston, MA 02130, USA.

E-mail address: degutis@wjh.harvard.edu (J. DeGutis).

https://doi.org/10.1016/j.cortex.2022.12.014
0010-9452/Published by Elsevier Ltd.

1. Introduction

Developmental prosopagnosia (DP) is a severe lifelong impairment in the ability to learn and recognize faces with otherwise normal neurological, socio-cognitive, intellectual, and visual functioning. Researchers have been aware that prosopagnosia resulting from an acute brain injury is quite rare and initially, researchers also believed DP to be a relatively rare disorder (e.g., De Haan, 1999; Jones & Tranel, 2001; McConachie, 1976). However, in the past 20 years, with the help of media coverage as well as the internet and social media, there has been an appreciation that DP is not as rare as initially thought (e.g., Bate & Tree, 2017).

A handful of larger studies have provided estimates of the prevalence of DP in adults (for a study examining the prevalence of face recognition difficulties in middle childhood, see Bennetts, Murray, Boyce, & Bate, 2017). Their diagnostic methods have differed, some using only self-report measures and semi-structured interviews (Kennerknecht et al., 2006; Kennerknecht, Yee-Ho, & Wong, 2008), one using a single objective measure (Bowles et al., 2009), and another using a combination of subjective and objective measures (Zhao et al., 2018). In the initial study reporting DP prevalence across a large sample, Kennerknecht et al. (2006) had subjects fill out a questionnaire and were subsequently asked open-ended questions about their face recognition experience throughout their lifetime during an interview. Subjects were diagnosed as prosopagnosic if they reported a set of specific symptoms, such as being unable to decide whether they know a face or not, having false negative and false positive face recognition events, and using other means of recognition (e.g., gait, voice, hairstyle, etc.). The estimated prevalence rate of prosopagnosia in their sample of 689 medical students in Germany was 2.47% (95% CI: 1.31%–3.63%) and 1.88% (95% CI: 1.05%–2.71%) in a follow-up study with 533 medical students in Hong Kong (Kennerknecht, Yee-Ho, & Wong, 2008). Though this suggests high rates of self-reported face recognition deficits, the validity of these studies has been criticized due to their failure to incorporate objective tests (e.g., Tree, 2011; Arizpe et al., 2019). Though several recent studies have shown that self-reported face recognition ability significantly predicts objective face recognition, these relationships have been in the smaller-to-moderate range (e.g., $r = .22$ in younger adults, Bowles et al., 2009; $r = .44$, Arizpe et al., 2019; $r = -.39$, Gray et al., 2019; $r = -.40$, Ventura, Livingston, & Shah, 2018). This suggests that individuals generally have some insight into their objective face recognition abilities, though self-reported face recognition alone is inadequate to diagnose prosopagnosia (see Arizpe et al., 2019 for a more in-depth discussion).

In addition to self-report, other studies have used objective face recognition measures to estimate the prevalence of DP. In a sample of 240 Australians, Bowles et al. (2009) used the Cambridge Face Memory Test (CFMT, Duchaine & Nakayama, 2006), a validated and widely used test in diagnosing prosopagnosia (e.g., Bate et al., 2014; Bate, Haslam, Tree, & Hodgson, 2008; Duchaine, Yovel, & Nakayama, 2007; Rezlescu, Pitcher, & Duchaine, 2012). They diagnosed a subset of participants as prosopagnosic whose CFMT scores were more than two standard deviations below the mean, indicative of a major impairment. Based on this cutoff, they concluded that the DP prevalence rate is at least 2%, not significantly different from the self-report-based estimates. One downside with relying solely on an objective measure is that it may not capture whether individuals experience prosopagnosia in their everyday life or if they experience distress from their face recognition deficits. Notably, a large DP study by Zhao et al. (2018) combined both subjective self-reports and objective tests to screen 9533 university students in Beijing, China. Their three-step screening process included self-report face recognition questionnaires, a semi-structured prosopagnosia interview, and a previously validated computer-based Old-New face recognition test. When comparing the total sample to those who received a DP diagnosis,[1] this resulted in a DP prevalence rate of 1.15% (95% CI: .94%–1.36%), substantially lower than estimates of studies using either one subjective or one objective measure. Though the Zhao study was the most thorough with combining self-report and objective measures, a downside to both Zhao et al. and Bowles et al. are that they relied on a single objective measure, and single measures are susceptible to effects such as fortuitous guessing and may have less reliability when compared to incorporating multiple measures (Holdnack et al., 2017).

As these studies demonstrate, the prevalence of DP is dependent on the diagnostic criteria, and currently there is no widely accepted diagnostic criteria for DP. Barton and Corrow (2016) reviewed the diagnostic criteria used in 23 recently published DP studies and found a high degree of variability, with most studies using significantly more conservative criteria than those providing initial prevalence rates of 2–2.5%, or even 1.15%. Most commonly, prosopagnosia diagnostic criteria required evidence of impairment on both subjective and multiple objective assessments. While the CFMT and the Famous Faces Memory Test (FFMT) are the most commonly used objective tests, a variety of other face recognition tests have also been used (e.g., Old-New Face Recognition Test, Duchaine & Nakayama, 2005) and some studies have additionally used face perception tests, such as the Cambridge Face Perception Test (CFPT, Duchaine, Yovel, & Nakayama, 2007). Despite most of these 23 studies citing in their introductory paragraph the prevalence of DP to be 2–2.5% (based on studies using single self-report or objective tests, Kennerknecht et al., 2006; Bowles et al., 2009), the criterion they used to diagnose DP was substantially stricter. This raises the question of what the prevalence of DP is according to recent diagnostic cutoffs and whether there are more principled approaches to determining cutoffs for DP, such as using data-driven cluster analyses in a large sample or employing criteria from the most recent version of the DSM-5 (e.g., mild *vs* major neurocognitive disorders, Sanchev et al., 2014). No studies to date have provided

[1] It should be noted that out of the 180 probable DPs in this study, only 105 chose to participate. Of these 105 individuals, 64 had confirmed DP (61%). Using this rate of 61%, we estimated that 46 of the 75 individuals who chose not to participate may have also had DP. Thus, to calculate the overall prevalence of DP in this sample, we added the DP individuals who participated (64) with the estimated number of DPs who chose not to participate (46), giving a total of 110 DPs.

empirical guidance for diagnostic cutoffs, which was the focus of the current investigation.

To help address these questions, the current study had three main objectives. Our first goal was to estimate the prevalence of DP based on the most commonly used diagnostic cutoffs of DP research studies from 2008 to 2021. We estimated the cutoffs used in 68 DP studies and applied these criteria to a large, unselected sample of 3116 web-based participants who had taken diagnostic tests for prosopagnosia: one validated self-report face recognition questionnaire (Cambridge Face Memory Questionnaire, CFMQ, Arizpe et al., 2019) and two validated objective face recognition tests (unfamiliar face learning/recognition-CFMT3, famous face recognition-FFMT, Mishra et al., 2019). Our second goal was to use these measures and our large dataset to determine if there are natural clusters of participants with low objective and subjective face recognition scores that should be regarded as DP. This could provide evidence whether DP exists on a continuum, i.e., normative view, or rather represents a more discrete cluster, i.e., pathologic view (Barton & Corrow, 2016). Lastly, we sought to investigate whether studies with more relaxed diagnostic cutoffs would be less able to capture known face-related impairments in DPs. In particular, face perception has been commonly found to be impaired in DPs at the group level (e.g., using the CFPT, Duchaine, Yovel, & Nakayama, 2007; Eimer, Gosling, & Duchaine, 2012; Mishra et al., 2021). We calculated average CFPT scores from 43 available studies and tested whether CFPT averages in DPs from each study were associated with the strictness of the diagnostic cutoff used. We conclude with a discussion about the advantages and disadvantages of adopting particular diagnostic cutoffs for DP.

2. Methods and methods

2.1. Participant recruitment

Adult participants from the United States that were 18–55 years of age completed the face recognition tasks and self-report questionnaire on testmybrain.org, a cognitive testing website accessed through search engines, social media, and news sites, where participants receive feedback on their cognitive performance compared to population norms (Fortenbaugh et al., 2015; Germine et al., 2011, 2012; Riley et al., 2017). The study included 3116 unpaid US participants (1904 females) who visited the website between January 2015 and March 2015. Previous studies have shown that the mean and variance of performance in samples from testmybrain.org are similar to in-lab samples (e.g., CFMT, Germine et al., 2012) and that individuals with very poor face recognition are not more prevalent in testmybrain.org studies compared to in-lab studies (e.g., Arizpe et al., 2019). All participants gave informed consent in accordance with guidelines set forth by the Committee on the Use of Human Subjects at Harvard University and the Wellesley College Institutional Review Board. Participants completed a voluntary demographic survey which asked questions related to age, sex (male/female, note that when the data was collected only two options were given), location, native language, education, and ethnicity. All participants received feedback on their performance relative to others at the completion of all the tasks.

2.2. Task and procedure

In this study, three assessments of face recognition, in the following order, were included in the battery for each participant: (1) Cambridge Face Memory Questionnaire (CFMQ), (2) Cambridge Face Memory Test, version 3 (CFMT3), and (3) Famous Faces Memory Test (FFMT).

The Cambridge Face Memory Questionnaire (CFMQ) is a previously validated (see Arizpe et al., 2019) 18-item questionnaire designed to measure self-assessment of one's face recognition in daily life. The CFMQ, where higher scores indicate better self-reported face recognition, has been shown to positively correlate with the CFMT ($r = .44$) and FFMT ($r = .52$). The CFMQ includes questions assessing the frequency of both positive and negative face recognition occurrences and one question assessing one's face recognition skills compared to others. These questions were developed by Drs Brad Duchaine, Ken Nakayama, and Laura Germine to screen for prosopagnosia and have been used for the past 20 years for this purpose (e.g., DeGutis et al., 2012, www.faceblind.org). We found the CFMQ was highly reliable, Cronbach's alpha = .91, similar to other face recognition self-reports (e.g., Cronbach's alpha for PI20 = .93, Shah, Gaule, Sowden, Bird, & Cook, 2015).

The Cambridge Face Memory Test (CFMT, Duchaine & Nakayama, 2006) is a widely used test of novel face recognition in which participants are required to learn and recognize six target faces in conditions of varying difficulty. Faces were presented in grayscale with no hair or other distinguishing non-facial features. The first part of the test introduced six target faces to participants where each target face was shown at three different angles for 3 s each. After learning each target face, participants were presented with a three-alternative forced-choice (AFC) task to choose the face they just studied out of three options. These three choices included the learned target face and two non-target faces presented in the same angle and lighting. Participants then simultaneously studied the six target faces shown for 20 s. Afterwards, they completed 30 forced-choice trials, each including one target and two non-target faces shown in different views and lighting conditions. Finally, participants again studied the six target faces for 20 s and completed 24 3-AFC trials. For these last 24 trials, visual noise was added to stimuli to make the task more challenging. As our experiment was publicly available online, we refrained from using the original CFMT to maintain the integrity of the original CFMT for clinical purposes. Instead, we used the CFMT3 which is identical to the original version developed by Duchaine and Nakayama (2006), except that different face stimuli are used. Instead of photographs of faces, the CFMT3 uses novel artificial faces that were generated via FaceGen software (Singular Inversions, Toronto, ON). Though some studies have found that artificial faces are more difficult to remember than real faces (Balas & Pacella, 2015), others have found similar overall recognition performance and robust face inversion effects, suggesting very similar processing as real faces (Kätsyri, 2018). With regards to the CFMT3, we found it had high internal consistency, Cronbach's alpha = .76. Additionally, in a subset of 67 individuals who took both the CFMT3 and CFMT original, we found a robust correlation, $r = .61$ ($P < .001$).

For the Famous Faces Memory Test (FFMT), one of three equivalent versions were assigned to each participant (for more details on the procedure and specific faces shown in each version, see Mishra et al., 2019). The face stimuli were drawn from a pool of 69 front-view faces of famous celebrities taken from google images advanced searches that were included in three famous face tests (FFMT1–27 faces, FFMT2–40 faces, FFMT3–26 faces), with 24 faces repeated across at least one test. We do not have legal permission to publicly archive the famous faces stimuli. Readers seeking access to these materials should contact Dr Jeremy Wilmer (jwilmer@wellesley.edu). The faces were cropped to remove extra facial features like hair, ears, and area below the jawline. The visual angle for all the face images was 5.5° × 7°. The faces belonged to people from various professions including actors/actresses, politicians, musicians, and sports personalities. In all versions, participants were shown an image of a famous face and asked, "Who is this?" If they typed in a response, they were then shown the correct answer along with their response to indicate whether they correctly identified the person. By design, misspellings of the correct name or even unique descriptions of the person were allowed and scored as correct. Participants who did not respond correctly were additionally asked to indicate whether they were familiar with the person. Trials where participants said they were unfamiliar with the person were *not* included in the overall calculation of scores (similar to other DP studies, e.g., Murray & Bate, 2020). This was done to avoid very lower scores in people who had reduced media exposure. As was done in a prior study (Wilmer et al., 2012), the total score was the number of trials for which they both (a) submitted a response and (b) it was verified that their response was a correct identification. To normalize the scores across different versions, we calculated the version-specific z-score for each participant. Because the distributions of these scores were comparable in each of the FFMT versions, we treated the versions as equivalent in our analyses (similar to Mishra et al., 2019). In this paper, we refer to all three versions singularly as the FFMT. Because we only scored trials where participants were familiar with the faces, which varied across participants, we were not able to calculate reliability of the overall test. However, recent studies have shown that famous face memory tests have sufficient reliability (e.g., .75-.80, Pozo, GermineL., Scheuer, & Strong, 2021).

2.3. Selection criteria and methods for prevalence estimation

We selected 104 peer-reviewed DP studies that were published from 2008 to 2021 by using keyword searches for developmental prosopagnosia and congenital prosopagnosia into google scholar and PubMed. Next, we identified which studies used the CFMT, FFMT, and self-report questionnaire similar to the CFMQ (e.g., Prosopagnosia Index-20, PI20, Shah et al., 2015) in their diagnostic criteria and calculated their diagnostic cutoffs for these measures. If no specific cutoff was mentioned, when individual subject data was available, we attempted to determine the cutoff score based on the least impaired individual that was deemed a prosopagnosic in the study. We were able to replicate the diagnostic criteria used in 68 out of the 104 studies. In studies that were not included, they either used tests that were not similar to our tests from testmybrain.org (e.g., Old-New face recognition test, Zhao et al., 2018) or we could not confidently determine their diagnostic cutoffs.

The subjective cutoffs used in the DP studies we selected varied. Some subjective measures were more structured, such as having abnormal performance on the Faces and Emotion Questionnaire (e.g., Freeman, Palermo, & Brock, 2015a, 2015b) or scoring certain standard deviations below the mean on the PI-20 (e.g., Shah et al., 2015). Others involved anecdotal reporting of lifelong face recognition difficulties. For studies that used a questionnaire other than the CFMQ, we generated analogous cutoffs using our CFMQ data. More precisely, for the studies that specified their strict, quantitative approach for subjective cutoffs (e.g., taking two standard deviations below the mean), we employed the same method using the CFMQ scores. For studies that involved the presence of subjective face recognition complaints, we tried to approximate their diagnostic method using the first question on the CFMQ, which asked, "Compared to my peers, I think my face recognition skills are …", Far Below Average/Below Average/Average/Above Average/Far Above Average. A recent study from our lab (Arizpe et al., 2019) showed that this single question is particularly good at screening for face recognition difficulties. We included participants who answered 'Far Below Average' or 'Below Average' on this question to be comparable with studies that used qualitative criteria for subjective cutoffs.

We estimated DP prevalence rates in our sample using both z-score estimates (which most studies reported) as well as percentile cutoffs calculated based on the z-scores. For instance, if a study's objective cutoff was 2 standard deviations below the mean on the CFMT, we calculated the number of participants who were in the bottom 2.275% of all CFMT scores. This percentile-based analysis was conducted to mitigate any impact that could originate from deviations from a normal distribution, since percentiles are more robust to non-normality than z-scores.

2.4. Cluster analyses

Using our large sample, we sought to determine if there was a natural cutoff for a group that performed poorly on subjective and objective face recognition tests. Prior to performing cluster analyses, we randomly split our sample into a testing dataset ($n = 1540$) and a replication dataset ($n = 1576$). Following random assignment, we normalized face processing measures separately within the testing dataset and replication dataset using a z-transformation. Prior to performing cluster analyses, we screened for multivariate outliers separately within each dataset to meet distributional assumptions. Based on a Mahalanobis distance criterion of $P \leq .001$, we removed seven multivariate outliers in the testing dataset and five multivariate outliers in the replication dataset, achieving a final sample size of 1533 and 1571, respectively.

Using R software and associated libraries (R Core Team, 2013; http://www.R-project.org/), we conducted a hierarchical cluster analysis (HCA) to determine an optimal number of clusters within the testing and replication datasets. Briefly, HCA initially assigns each participant to a unique cluster in which each cluster represents a single participant. Next, in an iterative fashion, each cluster is combined with the next most similar cluster based on the minimal multivariate distance. Clusters are iteratively combined in this manner until all data points are contained within a single cluster. Throughout this iterative process, HCA identifies multiple possible clustering solutions, which range from two clusters to $n - 1$ clusters. To compute multivariate distance between participants and/or clusters, we utilized the squared Euclidean distance between our normalized face recognition measures. To perform iterative cluster linkage, we utilized Ward's minimum variance linkage, which forms clusters that minimize the error sum of squares at each iteration (Ward, 1963). Next, we aimed to identify an optimal cluster solution in a data-driven manner using the nbClust library in *R* (Charrad, Ghazzali, Boiteau, & Niknafs, 2014). Specifically, potential cluster solutions were evaluated and compared based on 30 different criteria available (e.g., silhouette width). Though there is no accepted standard for approximating the sample size required to form a given number of clusters (k; Dolnicar, Grün, Leisch, & Schmidt, 2014), based on a conservative heuristic of 2^k (Formann, 1984), the sample size within the testing and replication datasets ($n = {\sim}1500$) was likely adequate to compare cluster solutions ranging from $k = 2$ to $k = 10$. Using this data-driven approach, the optimal cluster solution was identified among these potential cluster solutions based on performance across the previously described 30 clustering metrics.

To determine if the clustering solutions were consistent across cluster analytic approaches, we also computed a two- and three-cluster solution calculated using the k-means algorithm within the testing and replication datasets. Next, we assessed the agreement of participant assignment to each cluster between the HCA and k-means algorithms across the testing and replication datasets by calculating inter-rater reliability using Cohen's Kappa (two-cluster solution) or Cohen's weighted Kappa (three-cluster solution). Based on recently recommended guidelines (McHugh, 2012), we interpreted Kappa values < .40 to indicate no or minimal inter-rater reliability, Kappa values between .40 and .59 to indicate weak inter-rater reliability, Kappa values between .60 and .79 to indicate moderate inter-rater reliability, and .80–1.00 to indicate excellent inter-rater reliability.

2.5. *Association between Cambridge Face Perception Test and study diagnostic cutoffs*

Finally, we sought to investigate whether studies with more relaxed versus stricter diagnostic cutoffs would show differential performance on an independent face perception measure. We reviewed DP studies published in the past 14 years that administered the Cambridge Face Perception Test (CFPT, Duchaine, Germine, & Nakayama, 2007), ranked them based on the strictness of their diagnostic criteria, and compared their DPs' performance on the CFPT. The CFPT is a well-validated (e.g., Mishra et al., 2021) and widely used test of face perception used in many DP studies. The test consists of eight trials in which participants are asked to sort a set of six frontal view faces on a continuum from most to least like a target face, shown from ¾ view. We used the CFPT in this analysis because it is widely used and because DPs consistently perform worse than controls at the group level (e.g., Duchaine, Yovel, & Nakayama, 2007; Eimer et al., 2012; Mishra et al., 2021). It should be noted that though DPs perform worse on the CFPT and other face perception tests (e.g., computerized Benton, Mishra et al., 2021), they are typically not as impaired as on face memory tests, with some DPs performing within the normal range of performance on face perception tests. DP researchers have described face perception performance in DPs as a shifted distribution towards impairment (Biotti et al., 2019; Bate et al., 2019; Mishra et al., 2021) and though some researchers have distinguished apperceptive versus non-apperceptive subtypes of DPs (e.g., Biotti & Cook, 2016), there is currently limited evidence for discrete subgroups of DPs with impaired versus unimpaired face perception abilities (see Bennetts et al., 2022). To rank the strictness of diagnostic criteria of studies administering the CFPT, we applied the diagnostic criteria to our dataset of 3116 participants and used both z-score and percentile approaches. After calculating the percentages for all the studies, they were sorted from the lowest (i.e., strictest diagnostic criterion) to the highest (i.e., least strict diagnostic criterion), and Kendall's tau-b as well as a Pearson correlations were calculated to determine the relationship between the strictness of diagnostic criteria and CFPT performance.

2.6. *Sample size justification, preregistration, and inclusion/exclusion*

The sample size of the current study was based on guidelines from Naing, Winn, and Rusli (2006) for determining the sample size for prevalence studies. We used the following formula: $n = \frac{Z^2 P(1-P)}{d^2}$ where sample size is n, $Z = Z$ statistic for a level of confidence (in our case 1.96), $P =$ expected prevalence (2% from previous DP reports), and $d =$ precision (we set this at ± .5%). This gave a suggested sample size of 3012, and the sample we obtained was 3116. Note that no part of the study procedures or analyses was pre-registered prior to the research being conducted. 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.

2.7. *Data, study materials, and analysis code availability*

Data and analysis code are available at https://osf.io/5469z/.

3. Results

3.1. *Participants*

3116 volunteers (1904 females) ranging in age from 18 to 55 years (M = 30.99, SD = 10.54) performed the CFMT, CFMQ, and FFMT on testmybrain.org. Regarding the highest education attained, .6% of the participants attended middle school, 9.5% went to high school/secondary school, 28.6% attended some college/university, 26.8% held a bachelor's degree, 26.8% received had a graduate degree, and 3.3% did not indicate their level of education. There were significantly more female participants than males in the sample (overall females: 61%, overall males: 39%), similar to other studies from testmybrain.org (Germine et al., 2011).

3.2. CFMT, FFMT, and CFMQ performance and intercorrelations

We found that the overall group performance on the CFMT (M items correct = 54.26, SD = 7.39), FFMT (M z-score = −.01, SD = 1.01), and CFMQ (M score = 68.15, SD = 11.25) was very similar to previous normative samples (e.g., Arizpe et al., 2019; Germine et al., 2011; Germine et al., 2012). In terms of the distributions of scores, we found that all three measures deviated from normality and were negatively skewed, particularly the FFMT (see Supplementary Materials Table S1/S2 and Figure S1). Notably, the percentile approach we employed is robust to deviations from normality (see more on this in the discussion below). Similar to previous studies, we also observed similar moderate-to-strong correlations between these three measures: CFMT/FFMT (r = .46, p < .001), CFMT/CFMQ (r = .44, p < .001), FFMT/CFMQ (r = .51, p < .001). This suggests that the three tests all measure aspects of face recognition ability but are not so overlapping as to suggest they are measuring the exact same construct.

3.3. Prosopagnosia prevalence estimation

We were able to replicate the diagnostic cutoffs that were utilized in 68 DP studies from 2008 to 2021. As shown in Fig. 1, the diagnostic criteria varied significantly across the studies. Only one study diagnosed DP based on one objective test whereas the majority of the studies, 56%, used three tests (e.g., one subjective and two objective). The most common method to meet DP criteria was to take two standard deviations below the mean on both the CFMT and FFMT along with some subjective report of face recognition difficulties. This approach was used in 31 out of the 68 studies (46%). Other common methods included taking two standard deviations below the mean on the CFMT in combination with self-reported face recognition difficulties. This approach was used in 14 studies (21%). The third most common method, used in 4 studies (6%), focused on objective tests and incorporated the two standard deviation cutoff below the mean on both of CFMT and FFMT. The remaining studies (28%) used idiosyncratic diagnostic cutoffs that were either unique to that study or only replicated in one or two other studies.

Applying these diagnostic cutoffs from the previous studies to our web-based sample using a z-score cutoff approach (middle panel Fig. 1), the calculated DP prevalence rates also varied considerably, ranging between .64% (95% CI: .39%–.99%) and 5.42% (95% CI: 4.65%–6.28%). The lowest rate of .64% was calculated by taking 2 SD below the mean on the FFMT and CFMQ along with 1.5 SD below the mean on the CFMT. The diagnostic criteria that involved taking two standard deviations below the mean on *either* the CFMT or the FFMT along with subjective complaints yielded the highest DP prevalence estimate of 5.42%, eight times greater than the lowest rate. The most common method of taking two standard deviations below the mean on the CFMT and FFMT with subjective reporting resulted in the prevalence estimate of .93% (95% CI: .62%–1.33%).

We found a similar pattern, though reduced prevalence, when using the corresponding percentile cutoff approach (right panel, Fig. 1). The estimated prevalence varied from .13% (95% CI: .03%–.33%) to 2.95% (95% CI: 2.39%–3.61%). For the

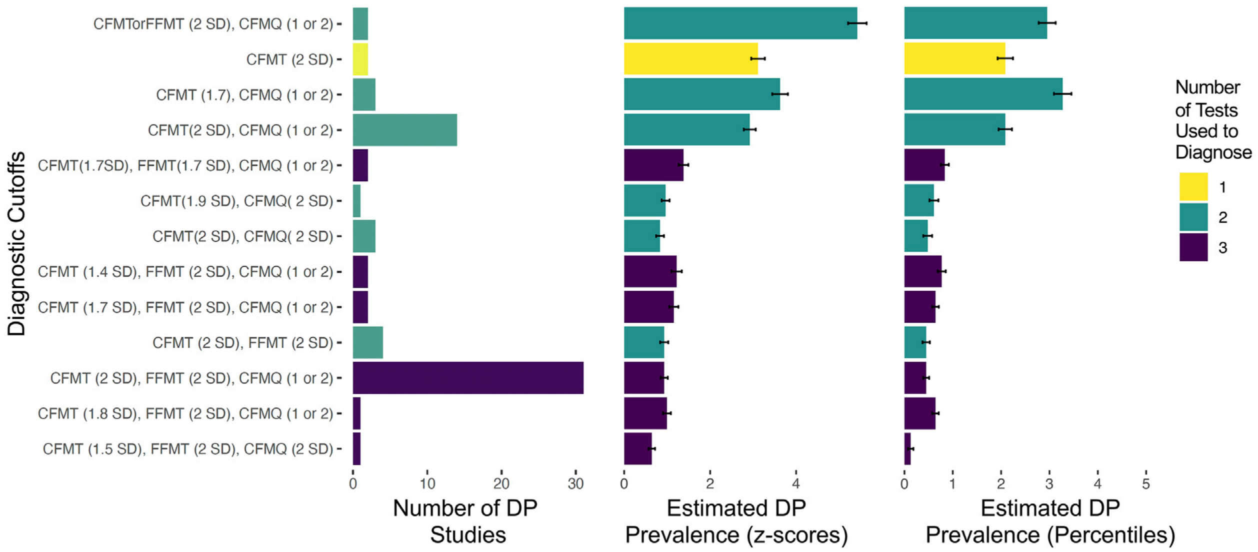

Fig. 1 – Diagnostic cutoffs of DP studies from 2008 to 2021 and the estimated prevalence rates. *Note.* CFMT = Cambridge Face Memory Test version 3, CFMQ = Cambridge Face Memory Questionnaire, FFMT = Famous Faces Memory Test, SD = standard deviations below the mean. Error bars represent 95% confidence intervals based on the sample size. CFMQ 1 or 2 indicates that participants in these studies self-reported poor face recognition, which corresponded to either 'Below Average' (2) or 'Far Below Average' (1) responses on the CFMQ item "Compared to my peers, I think my face recognition skills are... "

percentile-based estimation, the lowest rate of .13% was calculated by taking those who scored below the 2.275th percentile on the FFMT and CFMQ in combination with below the 6.68th percentile on the CFMT. The highest DP prevalence estimate of 2.95%, which is more than twenty-two times greater than the lowest rate, was based on those who scored below the 2.275th percentile on *either* the CFMT or FFMT along with self-reported face recognition deficits. The most common method of taking those below the 2.275th percentile on both the CFMT and FFMT with self-reported face recognition deficits yielded the prevalence rate of .45% (95% CI: .25%–.75%).

3.4. *Cluster analyses*

We next sought to determine if there was a more data-driven approach to identifying DPs from non-DPs. We applied cluster analyses to the testing (*n* = 1533) and replication datasets (*n* = 1571). In the testing dataset, the optimal number of clusters was identified as a two-cluster solution (favored by 10/30 metrics), which outperformed a three-cluster solution (favored by 6/30 metrics) and all other potential cluster solutions (≤2/30 metrics). In the replication dataset, the optimal number of clusters was identified as a three-cluster solution (favored by 9/30 metrics), which slightly outperformed a two-cluster solution (favored by 8/30 metrics) and all other potential cluster solutions (≤2/30 metrics). We present results for the two-cluster solution for the testing and replication datasets (see Fig. 2 and below). The three-cluster solutions can be found in the Supplementary Materials (see Figure S2).

3.4.1. *Hierarchical cluster analysis: cluster description*

In the testing dataset, the two-cluster solution was characterized by sub-groups exhibiting below-average performance (*n* = 596) or above-average performance (*n* = 937) across all face processing measures (see Fig. 2A), suggesting a unidimensional structure. In the replication dataset, the two-cluster solution was similarly characterized by subgroups exhibiting either below-average performance (*n* = 845) or above-average

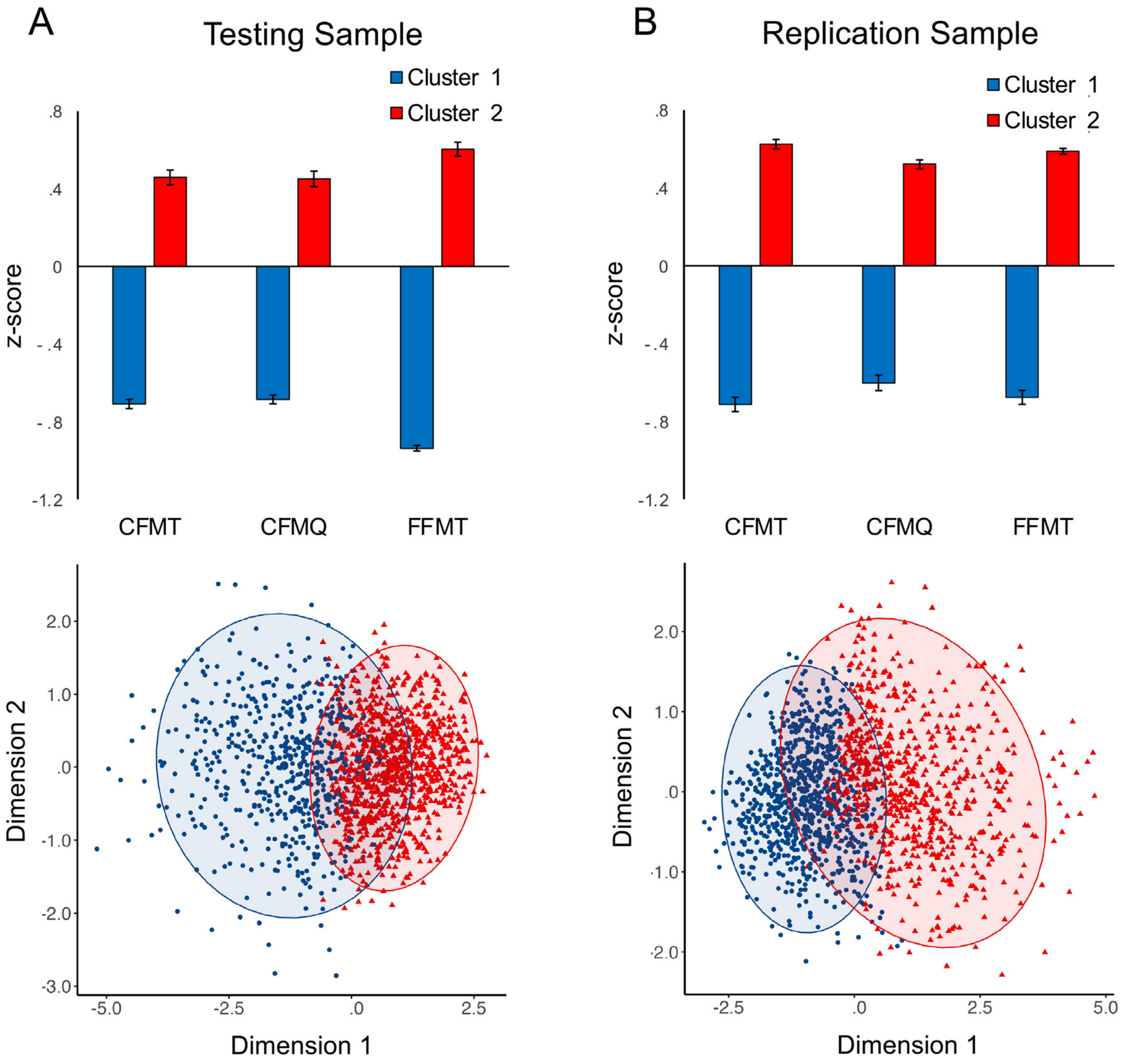

Fig. 2 – *Hierarchical Cluster Analysis 2-cluster Solution: Testing (A) and Replication (B) Samples*. Note. CFMT = Cambridge Face Memory Test version 3, CFMQ = Cambridge Face Memory Questionnaire, FFMT = Famous Faces Memory Test. Error bars represent 95% confidence intervals. Dim2 = dimension 1, Dim2 = dimension 2

performance ($n = 723$) across all face processing measures (see Fig. 2B). For the three-cluster solution, the testing dataset was again characterized by a unidimensional structure, with subgroups exhibiting slightly below average performance ($n = 440$), slightly above-average performance ($n = 937$), or below-average performance ($n = 156$) across all face processing measures (see Supplementary Figure S2A). In the replication dataset, the three-cluster solution was also similarly characterized by subgroups exhibiting slightly below average performance ($n = 522$), slightly above-average performance ($n = 848$), or below-average performance ($n = 201$) across all face processing measures (see Supplementary Figure S2B).

3.4.2. *Cluster consistency between hierarchical and k-means approaches*

To examine the robustness and reliability of our HCA findings, we next performed k-means cluster analyses for two- and three-cluster solutions and found a very similar pattern of results in both the testing and replication datasets (see Supplementary Materials Figures S3 and S4). For the two-cluster solution, we observed moderate-to-strong inter-rater reliability between the HCA and k-means algorithms for the testing dataset ($\kappa = .83$, 95% CI = .80–.86, $P < .001$) and the replication dataset ($\kappa = .69$, 95% CI = .66–.73, $P < .001$). For the three-cluster solution, we observed slightly reduced inter-rater reliability between the HCA and k-means algorithms across for the testing dataset ($\kappa = .38$, 95% CI = .32–.44, $P < .001$, there was a discrepancy in assigning participants between the ‘average’ versus ‘above average’ clusters, $\kappa = .02$) and a higher correspondence in the replication dataset ($\kappa = .80$, 95% CI = .78–.82, $P < .001$). Together, this shows that HCA results largely generalized to the k-means approach and that neither method identified clusters of individuals with poorer face recognition that could be considered in the prosopagnosic range of performance.

3.4.3. *Post-Hoc analysis in individuals with subjective face recognition deficits*

Because individuals with below-average self-reported face recognition are those more likely to seek out prosopagnosia researchers or visit prosopagnosia websites (e.g., www.faceblind.org, www.troublewithfaces.org), we also sought to determine if this particular subset of individuals had defined clusters or subgroups. We performed cluster analyses in individuals reporting “below average” or “far below average” face recognition compared to their peers ($n = 927$, based on a single item in the CFMQ, see Methods). Using HCA, we found that the optimal number of clusters was identified as a two-cluster solution (10/30 metrics), which outperformed a three-cluster solution (1/30 metrics). Similar to cluster analyses of the entire sample, the two clusters represented overall high ($n = 437$) and low face recognition abilities ($n = 488$) and failed to identify a cluster close to what would be considered prosopagnosic performance (see Supplementary Figure S5).

3.5. *CFPT performance comparison across diagnostic criteria*

We finally analyzed face perception performance between DP studies using different diagnostic criteria to see if the strictness of the cutoffs employed was associated with face perception abilities. For this analysis, studies that explicitly used the CFPT in the screening process and studies that did not administer or report individual-level CFPT results were excluded, which resulted in a total of 43 studies included. As can be seen in Fig. 3, the studies overlapped considerably in their CFPT performance.

After ranking these studies from the most to least strict diagnostic criteria, we calculated Kendall's tau-b and Pearson correlations (using both z-score and percentile approaches applied to our unselected web sample, see Supplementary Figure S6) to determine the relationship between the strictness of diagnostic criteria and CFPT performance of the DPs. For the z-score approach, there was nonsignificant association between CFPT and cutoff strictness (Kendall's tau-b correlation, $\tau b = .18$, $p = .125$; Pearson $r = .17$, $p = .267$), with stricter studies having numerically *better* CFPT scores. We found a similar pattern when using a percentile approach to calculating prevalence, with a nonsignificant association between CFPT and cutoff strictness (Kendall's tau-b correlation, $\tau b = .11$, $p = .339$; Pearson $r = .28$, $p = .067$), with stricter studies again having numerically *better* CFPT scores. These results clearly do not support the assertion that stricter diagnostic cutoffs allow one to better capture known face-related impairments in DPs.

4. Discussion

The current investigation illustrates the range of diagnostic criteria that DP studies have employed over the last 14 years and the associated DP prevalence rates. Applying these differing criteria to our sample of 3116 unselected web participants, we found estimated DP prevalence rates ranged from .64 to 5.42% when using a z-score approach and .13–2.95% when using a percentile approach, with the most commonly used cutoffs by researchers having a prevalence rate of .93% (z-score) and .45% (percentile). These estimates are considerably lower than the 2–2.5% prevalence commonly reported in the media and in introduction sections of many DP publications. These variable estimates of the prevalence of DP bring up the issue of whether there is a more data-driven approach to estimating the prevalence of DP. We addressed this in the current study by applying cluster analyses to our large dataset as well as a subset of individuals with self-reported below average face recognition. In both cases, we found unidimensional clusters based on better versus worse face recognition ability, but no clusters that identified those with close to prosopagnosia-level performance. This provides support for DP existing on a continuum rather than representing a discrete group. Finally, we examined whether the use of more relaxed versus stricter DP cutoffs in studies affected group-level face perception performance on the CFPT. We found a weak and nonsignificant correlations between cutoff strictness and CFPT performance, suggesting that more relaxed versus stricter criteria are likely not capturing mechanistically distinct populations of DPs. These findings have important theoretical and practical implications for how DP is diagnosed, and we conclude with recommendations for future studies.

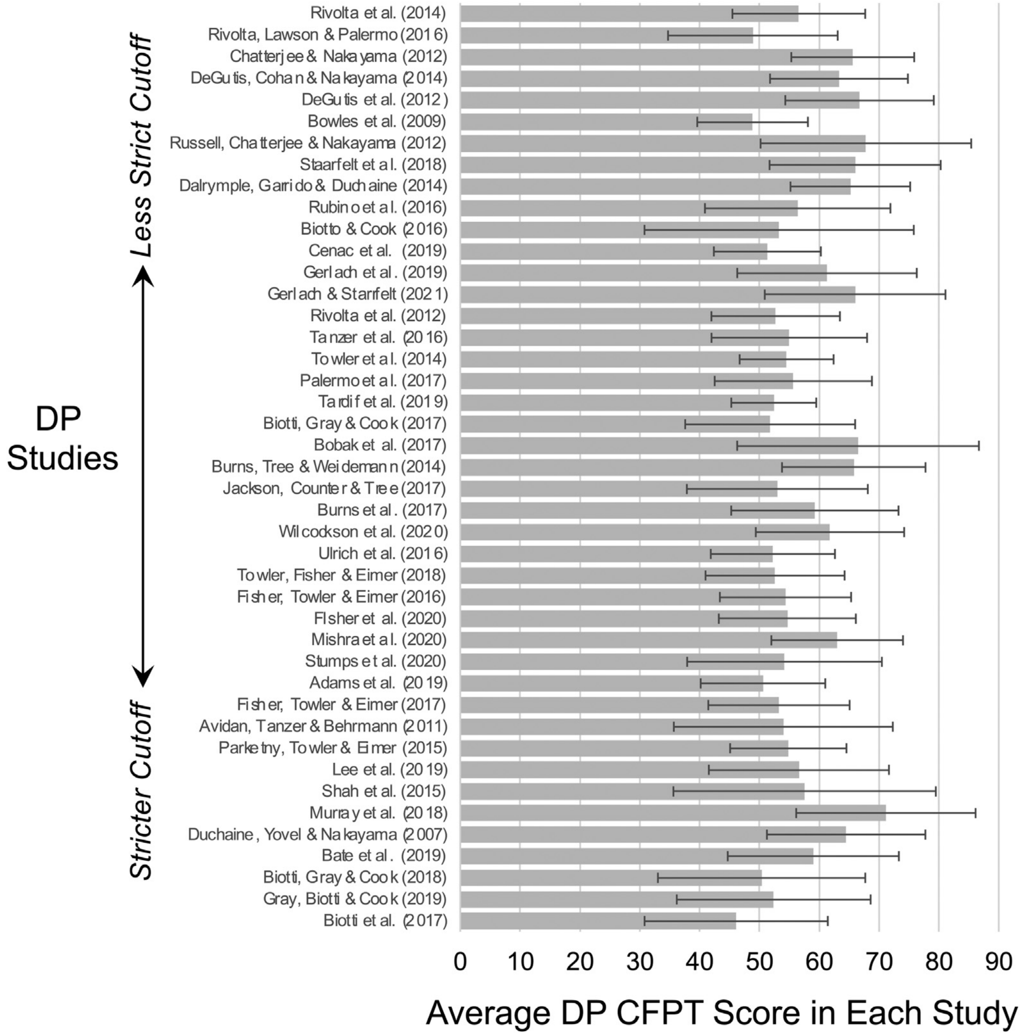

Fig. 3 – Cutoff strictness of developmental prosopagnosia studies and relationship to cambridge face perception test scores. Note. CFPT = Cambridge Face Perception Test. Error bars represent the standard deviation of the developmental prosopagnosia group. Note that higher scores on the CFPT indicate poorer performance.

For the last decade or so, the prevalence of DP has been reported in academic research papers and in the media to be 2–2.5%. In this study, we found that the prevalence of DP based on the most common cutoffs used across 31 of 68 research studies from 2008 to 2021 was .93% (z-score) and .45% (percentiles) but also that there was considerable variability. In studies using one diagnostic test, the DP prevalence rate was as high as 3.11% (z-score) and 2.09% (percentiles) whereas with three diagnostic tests, it was as low as .64% (z-score) and .13% (percentiles). This variability highlights the lack of diagnostic agreement amongst DP researchers and shows that there is a conservative bias towards a more rigorous criterion, where a DP identified in one study would be able to meet most of the existing criteria that researchers use. Though these conservative criteria could potentially identify more differences between DPs and controls, one downside of this approach is that it may make recruiting and screening DPs very burdensome and time-consuming, resulting in smaller sample sizes and less power to discover DP versus control group differences. Even recent DP studies still use quite small samples (e.g., $N = 10$, Gerlach & Starrfelt, 2021; $N = 13$, Haeger et al., 2021), making them more susceptible to potential sampling biases and more challenging to replicate. An overly conservative approach may also dissuade researchers from performing DP studies due to the burden of recruiting rare participants. Further, selecting only the most impaired DPs would make it more difficult to identify behavioral and biological markers that differentiate “pure” DP cases from borderline DP cases, if such markers exist.

In our DP prevalence estimates, it is notable that we found a sizeable difference between higher estimated prevalence rates based on z-score cutoffs versus lower prevalence rates based on a percentile approach, begging the question of what the most accurate estimation is. Because the distributions of the CFMT, CFMQ, and especially the FFMT deviated from normality and were skewed towards lower scores (see Supplementary Figure S1), the z-score cutoff analysis likely overestimated the prevalence of DP compared to if the tests were more normally distributed. Since the percentile approach is robust to deviations from normality, this approach may represent a better theoretical estimate of the DP prevalence. However, if the goal is to determine the prevalence of DP based on the measures and methods that researchers typically use (our CFMT3, FFMT, and CFMQ measures are very similar to most DP studies), then we suggest that our z-score cutoff results may better reflect the population prevalence rates of DP as is typically studied by DP researchers. Thus, our estimate of the population prevalence of DP based on the most common practices employed by DP researchers is 1 out of 108 individuals, or .93%.

To better understand the impact of studies using different face recognition cutoffs for DP, we analyzed whether stricter cutoffs could allow researchers to better capture face matching deficits commonly reported in developmental prosopagnosia (see Mishra et al., 2021). We compared DPs' face perception performance on the CFPT across 43 studies, none of which used the CFPT in diagnosing DPs. If stricter diagnostic criteria were associated with worse CFPT performance, it would support that DPs diagnosed with stricter criteria could be mechanistically distinct (in terms of their face perception abilities) from DPs diagnosed with looser criteria. Notably, our results revealed weak and non-significant correlations in the *opposite* direction, with more strictly diagnosed prosopagnosics having numerically *better* face perception performance. This finding provides preliminary support for the assertion that using more relaxed diagnostic criteria does not appreciably change the nature of the disorder being studied, though it would be useful to replicate these findings with other, potentially more sensitive face perception tests (e.g., computerized Benton, Mishra et al., 2021; Murray, Bennetts, Tree, & Bate, 2021) as well as other behavioral (e.g., face recollection *vs* familiarity abilities, Stumps, Saad, Rothlein, Verfaellie, & DeGutis, 2020) and neural measures (e.g., fMRI/EEG). A beneficial implication of this finding is that previous DP results using looser diagnostic criteria would likely generalize to DPs identified using stricter diagnostic criteria.

The current study also investigated whether there are natural cutoffs for identifying prosopagnosics when using subjective and objective diagnostic face recognition measures (CFMQ and CFMT/FFMT). Performing hierarchical and k-means cluster analyses on separate testing ($n = 1533$) and replication samples ($n = 1571$) consistently identified either two or three clusters of individuals with generally below- versus generally above-average subjective and objective face recognition abilities (as well as an ‘average’ group in the three-cluster solution). This suggests that there is not a discrete cluster of prosopagnosic individuals that emerge when taking this data-driven approach amongst an unselected sample. We additionally performed cluster analyses within just those individuals with self-reported below average/far below average face recognition abilities, who may often be referred to prosopagnosia websites (e.g., faceblind.org) or prosopagnosia researchers. Again, clusters emerged of those with generally average versus generally below average subjective and objective face recognition abilities, though far from prosopagnosia performance levels. Together, these results, along with a visual inspection of the data, suggest that face recognition performance is graded and that face recognition difficulties lie on a continuous spectrum rather than representing a discrete population, supporting the normative rather than pathologic view of DP (Corrow et al., 2016). This is similar to several other developmental and neurological disorders, including autism (Lord, Elsabbagh, Baird, & Veenstra-Vanderweele, 2018), multiple sclerosis (Vollmer, Nair, Williams, & Alvarez, 2021), and Alzheimer's Disease (Hampel et al., 2021).

The continuous nature of face recognition performance that the cluster analyses revealed is consistent with studies showing that DPs and typically developed participants are qualitatively similar. For example, Abudarham, Bate, Duchaine, and Yovel (2021) found that DPs, controls, and super recognizers used similar facial features for successful face recognition. Together, this advocates for more generally using an individual differences approach rather than a categorical/diagnostic approach to study face recognition ability. However, behavioral and fMRI evidence suggests there are qualitative differences in DP versus control face processing (Tian et al., 2020) and that associations found within the more general population can break down at the DP end of the continuum (e.g., association between social-cognitive and face recognition abilities, Barton and Corrow, 2016; Fry et al., 2022). Additionally, just because face recognition is graded does not mean that all aspects of face processing that contribute to face recognition are. Other measures such as holistic processing (Bennetts et al., 2022) or preferential fixation location (Pertzov et al., 2020) may reveal more distinct DP versus control differences or distinctions within DPs. For example, with regards to face perception ability, a recent study of 37 DPs by Bennetts et al. (2022) found DP subgroups with similar face perception deficits but either intact or deficient holistic face processing ability. Additional studies using better-characterized samples of DPs and controls would be useful to establish whether there are discrete differences between DPs and controls and if there are perceptual subtypes of DPs or rather a graded continuum of perceptual ability (e.g., shifted distribution model of DP perceptual deficits, Biotti et al., 2019).

Together, the current findings have important implications for diagnosing DP. Because our cluster analyses demonstrated that face recognition, particularly objective performance, is on a continuum, this suggests that validated methods used to diagnose other continuous neurocognitive disorders (e.g., dementia) could be applied to DP. One standard, validated approach that is currently used to diagnose continuously distributed neurocognitive disorders is from the latest edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-5, Sachdev et al., 2014). Based on poor reliability associated with using a single measure used in diagnosis (Holdnack et al., 2017), the DSM-5 recommends that at least two objective validated measures within a domain (in the case of DP, two face recognition measures) are impaired (z-score < -2 for major neurocognitive disorder) to receive a diagnosis. It also suggests that there should be subjective evidence of impairment. This criterion of self-reported face recognition deficits and z-score < -2 on two or more face recognition tests is consistent with previous recommendations (Dalrymple et al., 2014) and has been the most common method used to diagnose DP in the last 14 years (see Fig. 1) and we suggest this would be a useful standard for the field moving forward. When using this criterion, we estimate the prevalence of prosopagnosia in the population to be .93% (z-score approach) or .45% (percentile approach).

The DSM-5 also differentiates major from mild neurocognitive disorders, which may be a useful distinction for DP research going forward. Mild neurocognitive disorder is defined as performance worse than one standard deviation below the normative mean on multiple tests whereas major neurocognitive disorder requires z-scores < -2 (Sachdev et al., 2014). Based on this and the fact that we found no significant DP performance differences on the CFPT based on diagnostic criteria, it could be fruitful for future studies to include mild prosopagnosics with subjective face recognition complaints. When applying the DSM-5 mild neurocognitive criterion to our large web-based sample using the z-score approach, we found the prosopagnosia prevalence was 3.08%, with 2.15% having mild prosopagnosia and .93% having major prosopagnosia (with percentiles, the prosopagnosia prevalence was 3.27%, with 2.82% having mild prosopagnosia and .45% having major prosopagnosia). Thus, including mild prosopagnosics could improve recruitment efforts and allow for appreciably larger prosopagnosia study sample sizes. These larger sample sizes have the potential to better characterize individual differences amongst prosopagnosics and could help discover mechanistic differences between prosopagnosics that could further refine diagnostic cutoffs (e.g., identify a "true" cut-off if one exists). Further, larger DP sample sizes could improve the replicability and generalizability of DP findings. A downside to including mild prosopagnosics would be, if those participants dominated the sample, it could potentially obscure important prosopagnosic versus control differences. For this reason, we suggest that if researchers include mild prosopagnosics they also include an equal or greater number of major prosopagnosics as well. Further, it would be important to perform all key analyses with only major prosopagnosics in addition to the larger sample of mild and major prosopagnosics.

A recent study by Burns, Gaunt, Kidane, Hunter, and Pulford (2022) also suggests that looser criteria should be employed when diagnosing DP. They studied 61 individuals with self-reported lifelong difficulty with faces with either impaired CFMT scores (z-score < -2, $n = 27$, so-called 'Classical DPs') or unimpaired CFMT scores ($n = 34$, so-called 'Excluded DPs' because they are routinely excluded from DP studies). They found that the excluded group showed many deficits on objective face processing measures, though smaller in magnitude than classical DPs, and argue that self-reported face recognition difficulties, as measured by the PI20, should be used as the sole criteria to diagnose DP. Though our findings agree with the sentiment of having more inclusive DP diagnostic criteria, there are problems with solely relying on self-report for a diagnosis, including biases in self-report and lack of insight into one's face recognition abilities (e.g., 10–18 and 51–70-year-olds as well as males in general overestimate their self-reported face recognition abilities in comparison with 19–50-year-olds and females, DeGutis et al., 2023). Additionally, since Burns et al. (2022) showed that many of their excluded sample have face recognition deficits, several of these individuals would likely meet the DSM-5 criteria for mild prosopagnosia. Though we do not believe that those with self-reported face recognition deficits and normal objective face recognition performance (z-score > -1 across several tests) should be classified as developmental prosopagnosics, it would be important to study these individuals and understand the source of their self-report versus objective discrepancies (e.g., social anxiety impairing face recognition in the real world though not during lab testing) and work with these individuals to develop more sensitive face recognition tests that better reflect their self-reported difficulties.

There are several limitations with the current study. First, in estimating the prevalence of prosopagnosia in our web sample based on the cutoffs of published studies, we relied on our CFMQ, CFMT3, and FFMT measures, but about one third of the studies that we reviewed did not employ similar measures. Given that the CFMT and FFMT are the most commonly and traditionally used DP diagnostic tests, it is unlikely that these prevalence estimations differ from studies that used other diagnostic tests, yet there still may be some variance. Additionally, although we used the CFMT3 in place of the original CFMT (Duchaine & Nakayama, 2006) as to not widely distribute the original CFMT, there may be subtle differences between the CFMT3 and original which could affect prevalence rate, such as the use of artificial faces in the CFMT3 (though Kätsyri, 2018, suggests that artificial faces are processed similarly to real faces). Another limitation is that participants recruited via testmybrain.org tend to be younger, more educated, and female than a fully representative sample and testmybrain.org could have attracted more individuals with poor face recognition abilities interested in seeing if they have a deficit, which would potentially inflate the DP prevalence (though the similar Mean and SD of the tasks compared to the lab suggests this is not a widespread issue). Replicating these findings in a sample more representative of the general population would be useful. Another limitation is that the CFPT has complex instructions and may have less-than-ideal reliability (e.g., Controls $\alpha = .74$, DPs $\alpha = .79$, Mishra et al., 2021; Controls $\alpha = .67$, Rezlescu et al., 2012; Controls $\alpha = .74$, Bowles et al., 2009), suggesting that alternative face perception measures could have been more ideal. Additionally, importantly, a diagnosis of prosopagnosia requires ruling out other factors that could cause face recognition deficits (e.g., poor low-level vision, see Corrow et al., 2016; Dalrymple et al., 2014), which we were unable to assess in our large online sample. Thus, our estimates of prosopagnosia prevalence rates are likely slightly higher than had these individuals been screened out. A final limitation of the current study is that the results are specific to developmental prosopagnosia and do not generalize to the much rarer disorder of acquired prosopagnosia, which often presents with more severe face processing deficits (e.g., Barton, Albonico, Susilo, Duchaine, & Corrow, 2019). It would be important for future investigations to systematically review the diagnostic cutoffs used in acquired prosopagnosia studies and estimate the prevalence of prosopagnosia amongst individuals with acquired brain injury.

In sum, the current study reviewed the different approaches used to diagnose DP over the last 14 years and calculated corresponding prevalence rates in a large, unselected web-based sample. Our results highlight that the most common DP diagnostic cutoffs used have been substantially more conservative (e.g., .93% prevalence when using a z-score approach) than the widely reported DP prevalence rate of 2–2.5%. Using cluster analyses, we also found that there is a continuous distribution of face recognition abilities with no natural demarcation for a DP cutoff. Additionally, we found that face perception performance was very similar across DP studies with looser and stricter diagnostic cutoffs. Considering these findings, we suggest that DP researchers adopt standardized neurocognitive disorder cutoffs from DSM-5 to identify major (self-report + at least 2 validated face recognition tests z-score < -2) and mild (self-report + at least 2 validated face recognition tests z-score < -1) forms of prosopagnosia until more mechanistically grounded cutoffs can be identified.

Funding

This study was funded by the National Eye Institute grants RO1EY026057, and R21EY031000 awarded to JD.

Ethical statement

Participants gave informed consent in accordance with guidelines set forth by the Committee on the Use of Human Subjects at Harvard University and the Wellesley College Institutional Review Board.

Author contributions

JD was responsible for study conceptualization and design. JW set up and supervised participant recruitment and data collection. KanB, KB, EML, HMS, and MM conducted the review of the previous developmental prosopagnosia literature and performed behavioral data analyses with supervision from JD. TE and JL conducted the cluster analyses. JD, EML, and TE drafted the manuscript and KanB and TE created the figures. KanB, KB, EML, TE, MM, and JW provided critical feedback and edits to the manuscript. JD acquired funding for the study. All authors approved the final version of the manuscript.

Open practices

The study in this article earned Open Data badge for transparent practices. The data and analysis code for the study are available at: https://osf.io/5469z/

Declaration of competing interest

All authors declare no competing interests.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cortex.2022.12.014.

REFERENCES

Abudarham, N., Bate, S., Duchaine, B., & Yovel, G. (2021). Developmental prosopagnosics and super recognizers rely on the same facial features used by individuals with normal face recognition abilities for face identification. *Neuropsychologia, 160*, Article 107963.

Arizpe, J. M., Saad, E., Douglas, A. O., Germine, L., Wilmer, J. B., & DeGutis, J. M. (2019). Self-reported face recognition is highly valid, but alone is not highly discriminative of prosopagnosia-level performance on objective assessments. *Behavior Research Methods, 51*(3), 1102–1116. https://doi.org/10.3758/s13428-018-01195-w

Balas, B., & Pacella, J. (2015). Artificial faces are harder to remember. *Computers in human behavior, 52*, 331–337.

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

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

Bate, S., Bennetts, R. J., Gregory, N., Tree, J. J., Murray, E., Adams, A., et al. (2019a). Objective patterns of face recognition deficits in 165 adults with self-reported developmental prosopagnosia. *Brain Sciences, 9*(6), 133. https://doi.org/10.3390/brainsci9060133

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

Bate, S., Cook, S. J., Duchaine, B., Tree, J. J., Burns, E. J., & Hodgson, T. L. (2014). Intranasal Inhalation of oxytocin improves face processing in developmental prosopagnosia. *Cortex; a Journal Devoted To the Study of the Nervous System and Behavior, 50*, 55–63. https://doi.org/10.1016/j.cortex.2013.08.006

Bate, S., Haslam, C., Tree, J. J., & Hodgson, T. L. (2008). Evidence of an eye movement-based memory effect in congenital prosopagnosia. *Cortex; a Journal Devoted To the Study of the Nervous System and Behavior, 44*(7), 806–819. https://doi.org/10.1016/j.cortex.2007.02.004

Bate, S., & Tree, J. J. (2017). The definition and diagnosis of developmental prosopagnosia. In *Quarterly journal of experimental psychology*. London, England: SAGE Publications Sage UK. https://doi.org/10.1080/17470218.2016.1195414.

Bennetts, R. J., Gregory, N. J., Tree, J., Luft, C. D. B., Banissy, M. J., Murray, E., … Bate, S. (2022). Face specific inversion effects provide evidence for two subtypes of developmental prosopagnosia. *Neuropsychologia, 174*, 108332.

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.

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

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

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*, 1–24.

Charrad, M., Ghazzali, N., Boiteau, V., & Niknafs, A. (2014). NbClust: an R package for determining the relevant number of clusters in a data set. *Journal of statistical software, 61*, 1–36.

Corrow, J. C., Corrow, S. L., Lee, E., Pancaroglu, R., Burles, F., Duchaine, B., et al. (2016). Getting lost: Topographic skills in acquired and developmental prosopagnosia. *Cortex; a Journal*

Devoted To the Study of the Nervous System and Behavior, 76, 89–103. https://doi.org/10.1016/j.cortex.2016.01.003

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

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

DeGutis, J., Cohan, S., Mercado, R. J., Wilmer, J., & Nakayama, K. (2012b). Holistic processing of the mouth but not the eyes in developmental prosopagnosia. *Cognitive Neuropsychology, 29*(5–6), 419–446. https://doi.org/10.1080/02643294.2012.754745

DeGutis, J., Yosef, B., Lee, E. A., Saad, E., Arizpe, J., Song, J. S., … Esterman, M. (2023). The rise and fall of face recognition awareness across the life span. *Journal of Experimental Psychology: Human Perception and Performance, 49*(1), 22–33.

Dolnicar, S., Grün, B., Leisch, F., & Schmidt, K. (2014). Required sample sizes for data-driven market segmentation analyses in tourism. *Journal of Travel Research, 53*(3), 296–306.

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., & 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

Duchaine, B., Yovel, G., & Nakayama, K. (2007). No global processing deficit in the Navon task in 14 developmental prosopagnosics. *[Social Cognitive and Affective Neuroscience Electronic Resource], 2*(2), 104–113. https://doi.org/10.1093/scan/nsm003

Eimer, M., Gosling, A., & Duchaine, B. (2012). Electrophysiological markers of covert face recognition in developmental prosopagnosia. *Brain: a Journal of Neurology, 135*(2), 542–554. https://doi.org/10.1093/brain/awr347

Formann, A. K. (1984). *Die latent-class-analyse: Einführung in Theorie und Anwendung*. Beltz.

Fortenbaugh, F. C., Degutis, J., Germine, L., Wilmer, J. B., Grosso, M., Russo, K., et al. (2015). Sustained attention across the life span in a sample of 10,000: Dissociating ability and strategy. *Psychological Science, 26*(9), 1497–1510. https://doi.org/10.1177/0956797615594896

Freeman, P., Palermo, R., & Brock, J. (2015a). *Faces and emotion questionnaire*. Figshare. Advance Online Publication.

Freeman, P., Palermo, R., & Brock, J. (2015b). *Faces and emotion questionnaire*.

Fry, R., Li, X., Evans, T. C., Esterman, M., Tanaka, J., & DeGutis, J. (2022). Investigating the influence of autism spectrum traits on face processing mechanisms in developmental prosopagnosia. *Journal of Autism and Developmental Disorders, Sep, 29*, 1–22.

Gerlach, C., & Starrfelt, R. (2021). Patterns of perceptual performance in developmental prosopagnosia: An in-depth case series. *Cognitive Neuropsychology, 38*(1), 27–49. https://doi.org/10.1080/02643294.2020.1869709

Germine, L. T., Duchaine, B., & Nakayama, K. (2011). Where cognitive development and aging meet: Face learning ability peaks after age 30. *Cognition, 118*(2), 201–210. https://doi.org/10.1016/j.cognition.2010.11.002

Germine, L., Nakayama, K., Duchaine, B. C., Chabris, C. F., Chatterjee, G., & Wilmer, J. B. (2012). Is the web as good as the lab? Comparable performance from web and lab in cognitive/perceptual experiments. *Psychonomic Bulletin & Review, 19*(5), 847–857. https://doi.org/10.3758/s13423-012-0296-9

Gray, K. L. H., Biotti, F., & Cook, R. (2019). Evaluating object recognition ability in developmental prosopagnosia using the Cambridge Car Memory Test. Cognitive. *Neuropsychology, 36*(1–2), 89–96. https://doi.org/10.1080/02643294.2019.1604503

Haeger, A., Pouzat, C., Luecken, V., N'diaye, K., Elger, C., Kennerknecht, I., … Dinkelacker, V. (2021). Face processing in developmental prosopagnosia: Altered neural representations in the fusiform face area. *Frontiers in Behavioral Neuroscience, 15*.

Hampel, H., Cummings, J., Blennow, K., Gao, P., Jack, C. R., & Vergallo, A. (2021). Developing the ATX (N) classification for use across the Alzheimer disease continuum. *Nature Reviews Neurology, 17*(9), 580–589.

Holdnack, J. A., Tulsky, D. S., Brooks, B. L., Slotkin, J., Gershon, R., Heinemann, A. W., et al. (2017). Interpreting patterns of low scores on the NIH toolbox cognition battery. *Archives of Clinical Neuropsychology, 32*(5), 574–584. https://doi.org/10.1093/arclin/acx032

Jones, R. D., & Tranel, D. (2001). Severe developmental prosopagnosia in a child with superior intellect. *Journal of Clinical and Experimental Neuropsychology, 23*(3), 265–273. https://doi.org/10.1076/jcen.23.3.265.1183

Kätsyri, J. (2018). Those virtual people all look the same to me: Computer-rendered faces elicit a higher false alarm rate than real human faces in a recognition memory task. *Frontiers in psychology, 9*, 1362.

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

Kennerknecht, I., Nga, Y. H., & Wong, V. C. N. (2008). Prevalence of hereditary prosopagnosia (HPA) in Hong Kong Chinese population. *American Journal of Medical Genetics, Part A, 146*(22), 2863–2870. https://doi.org/10.1002/ajmg.a.32552

Lord, C., Elsabbagh, M., Baird, G., & Veenstra-Vanderweele, J. (2018). Autism spectrum disorder. *Lancet, 392*(10146), 508–520.

McConachie, H. R. (1976). Developmental prosopagnosia. A single case report. *Cortex; a Journal Devoted To the Study of the Nervous System and Behavior, 12*(1), 76–82. https://doi.org/10.1016/S0010-9452(76)80033-0

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia medica, 22*(3), 276–282.

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.

Mishra, M. V., Likitlersuang, J., Wilmer, J., Cohan, S., Germine, L., & DeGutis, J. M. (2019). Gender differences in familiar face recognition and the influence of sociocultural gender inequality. *Scientific reports, 9*(1), 1–12.

Murray, E., & Bate, S. (2020). Diagnosing developmental prosopagnosia: Repeat assessment using the cambridge face memory test. *Royal Society Open Science, 7*(9), Article 200884. https://doi.org/10.1098/rsos.200884

Murray, E., Bennetts, R., Tree, J., & Bate, S. (2021). An update of the Benton facial recognition test. *Behavior Research Methods*, 1–16.

Naing, L., Winn, T. B. N. R., & Rusli, B. N. (2006). Practical issues in calculating the sample size for prevalence studies. *Archives of orofacial Sciences, 1*, 9–14.

Pertzov, Y., Krill, D., Weiss, N., Lesinger, K., & Avidan, G. (2020). Rapid forgetting of faces in congenital prosopagnosia. *Cortex; a Journal Devoted To the Study of the Nervous System and Behavior, 129*, 119–132. https://doi.org/10.1016/j.cortex.2020.04.007

Pozo, E., Germine, L. T., Scheuer, L., & Strong, R. W. (2021). *Evaluating the reliability and validity of the famous faces doppelgangers test, a novel measure of familiar face recognition*, Article 10731911221087746. Assessment.

Rezlescu, C., Pitcher, D., & Duchaine, B. (2012). Acquired prosopagnosia with spared within-class object recognition but impaired recognition of degraded basic-level objects. *Cognitive Neuropsychology, 29*(4), 325–347. https://doi.org/10.1080/02643294.2012.749223

Riley, E., Okabe, H., Germine, L., Wilmer, J., Esterman, M., & DeGutis, J. (2017). Erratum: Gender differences in sustained attentional control relate to gender inequality across countries, 2016 *Plos One, 11*(11), Article e0165100. https://doi.org/10.1371/journal.pone.0170876. DOI: 10.1371/journal.pone.0165100). PLoS ONE, 12(1), e0165100.

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

Shah, P., Gaule, A., Sowden, S., Bird, G., & Cook, R. (2015). The 20-item prosopagnosia index (PI20): A self-report instrument for identifying developmental prosopagnosia. *Royal Society Open Science, 2*(6), Article 140343. https://doi.org/10.1098/rsos.140343

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

Tian, X., Wang, R., Zhao, Y., Zhen, Z., Song, Y., & Liu, J. (2020). Multi-item discriminability pattern to faces in developmental prosopagnosia reveals distinct mechanisms of face processing. *Cerebral Cortex, 30*(5), 2986–2996.

Tree, J. J. (2011). Mental imagery in congenital prosopagnosia: A reply to grüter et al. *Cortex; a Journal Devoted To the Study of the Nervous System and Behavior, 47*(4), 514–518. https://doi.org/10.1016/j.cortex.2010.11.005

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. *Quarterly Journal of Experimental Psychology, 71*(12), 2677–2679.

Vollmer, T. L., Nair, K. V., Williams, I. M., & Alvarez, E. (2021). Multiple sclerosis phenotypes as a continuum: The role of neurologic reserve. *Neurology: Clinical Practice, 11*(4), 342–351.

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

Wilmer, J. B., Germine, L., Chabris, C. F., Chatterjee, G., Gerbasi, M., & Nakayama, K. (2012). Capturing specific abilities as a window into human individuality: The example of face recognition. *Cognitive Neuropsychology, 29*(5–6), 360–392. https://doi.org/10.1080/02643294.2012.753433

Zhao, Y., Zhen, Z., Liu, X., Song, Y., & Liu, J. (2018). The neural network for face recognition: Insights from an fMRI study on developmental prosopagnosia. *Neuroimage, 169*, 151–161. https://doi.org/10.1016/j.neuroimage.2017.12.023

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