



Published in final edited form as:

Behav Res Methods. 2019 June ; 51(3): 1102–1116. doi:10.3758/s13428-018-01195-w.

Self-Reported Face Recognition is Highly Valid, but Alone is Not Highly Discriminative of Prosopagnosia-Level Performance on Objective Assessments

Joseph M. Arizpe^{1,2,3,*}, Elyana Saad^{1,2}, Ayooluwa O. Douglas^{1,2}, Laura Germine^{1,4}, Jeremy B. Wilmer⁵, and Joseph M. DeGutis^{1,2}

¹Department of Psychiatry, Harvard Medical School, Boston, MA, USA

²Boston Attention and Learning Laboratory, VA Boston Healthcare System, Boston, MA, USA

³Science Applications International Corporation (SAIC), Fort Sam Houston, TX, USA

⁴Institute for Technology in Psychiatry, McLean Hospital, Belmont, MA, USA

⁵Department of Psychology, Wellesley College, Wellesley, MA, USA

Abstract

Severe developmental deficits in face recognition ability (Developmental Prosopagnosia, or DP) have been vigorously studied over the past decade, yet there remain many unanswered questions about their origins, nature, and social consequences. A rate-limiting factor in answering such questions is the challenge of recruiting rare DP participants. While self-reported experiences have long played a role in efforts to identify DPs, much remains unknown about how such self-reports can or should contribute to screening or diagnosis. Here, in a large, population-based web sample, we investigate the effectiveness of self-report, used on its own, as a screen to identify individuals who will ultimately fail, at a conventional cutoff, the two types of objective tests that are most commonly used to confirm DP diagnoses: the Cambridge Face Memory Test (CFMT) and the Famous Faces Memory Test (FFMT). We use a highly reliable questionnaire ($\alpha = 0.91$), the Cambridge Face Memory Questionnaire (CFMQ), and we reveal strong validity via high correlations of 0.44 with CFMT and 0.52 with FFMT. Yet cutoff analyses revealed that no CFMQ score yielded a clinical grade combination of sensitivity and positive predictive value in enough individuals to support using it alone as a DP diagnostic or screening tool. This result was replicated in an analysis of data from the widely-used PI20 questionnaire, a 20-question self-assessment of facial recognition similar in form to CFMQ. We therefore recommend that screens for DP should, wherever possible, include objective as well as subjective assessment tools.

Keywords

self-report; ability; individual differences; face recognition; screening; prosopagnosia; meta-cognition

*Corresponding author: Joseph.Arizpe.10@alumni.ucl.ac.uk, Address: Science Applications International Corporation (SAIC), 4141 Petroleum Drive, Fort Sam Houston, TX, USA.

Introduction

Over the last 20 years, there has been a growing interest in identifying and studying individuals with poor face recognition abilities, especially those with developmental prosopagnosia. Developmental prosopagnosia (DP) is characterized as a facial recognition impairment, believed to have been present throughout one's life, that is not due to some more general deficit or organicity (e.g., poor low-level vision or early-life brain trauma). It has been suggested as having a prevalence of approximately 2.5% in the general population (Bowles et al., 2009; I Kennerknecht et al., 2006; Ingo Kennerknecht, Nga, & Wong, 2008), though it should be noted that this estimate is based solely on self-report data and on a common rule of thumb for determining statistical outliers (Barton & Corrow, 2016). Researchers have been intent on developing a better understanding of the cognitive and neural mechanisms of prosopagnosia and correspondingly have sought new methods to screen the population to find, and potentially aid in the diagnosis of, research participants with prosopagnosia. Effective screening is necessary to avoid the cost (time, personnel, resources, travel) involved in having to assess a preponderance of persons without prosopagnosia in order to obtain only a small handful who do have prosopagnosia. Conversely, efficiently screening out persons with prosopagnosia from studies of normal face processing is not widely performed and would be quite useful. Additionally, there are many real-world applications for which efficient identification of individuals with poor facial recognition abilities is of value. For example, employers may want to avoid placing individuals with below-average face recognition abilities in positions that require sensitive face matching tasks as part of their job (e.g., passport officers; White, Kemp, Jenkins, Matheson, & Burton, 2014).

One potential method of screening for facial recognition deficits is to use a self-administered questionnaire analogous to that used to screen for developmental disorders such as dyslexia. Scores on the Adult Reading Questionnaire along with a Dyslexia Status question (i.e., "Do you think you are dyslexic?") have been reported to enable detection of 62.5% of those with poor reading skills (putative dyslexia), as determined by objective testing, and correct rejection of 95.04% of those without (Snowling, Dawes, Nash, & Hulme, 2012). In an attempt to efficiently identify and study individuals with poor facial recognition abilities in a similar manner, several research groups have developed self-reported facial recognition ability questionnaires or clinical interviews, the scores of which have been reported to be associated with other measures of facial recognition performance (De Heering & Maurer, 2014; I Kennerknecht et al., 2006; Ingo Kennerknecht, Plümpe, Edwards, & Raman, 2007; Palermo et al., 2017; Punit Shah, Gaule, Sowden, Bird, & Cook, 2015; Turano, Marzi, & Viggiano, 2016). Prior studies have reported moderate-to-strong correlations between self-reported face recognition abilities and objective performance, both for typical participants (Bindemann, Attard, & Johnston, 2014; Gray Bird, G., Cook, R., 2017; Livingston & Shah, 2017; Rotshtein, Geng, Driver, & Dolan, 2007; Punit Shah, Gaule, et al., 2015; Punit Shah, Sowden, Gaule, Catmur, & Bird, 2015; Turano et al., 2016; Turano & Viggiano, 2017; J. B. Wilmer et al., 2010, though see Palermo et al., 2017) as well as for those with developmental prosopagnosia (Livingston & Shah, 2017; Stollhoff, Jost, Elze, & Kennerknecht, 2011). The reported correlation magnitudes are usually in the range of about 0.35 to 0.55. This

relationship is modulated somewhat by whether the questionnaire is administered before or after the objective assessments, especially if feedback is given, thus highlighting the importance of administering questionnaires while participants are still naive to the objective tests (Bobak, Mileva, & Hancock, 2018). This evident relation between self-report and objective facial recognition ability suggests that there may be utility in using self-report as a tool for predicting whether one has prosopagnosia, as defined by objective testing.

Though such self-report questionnaires and clinical interviews for prosopagnosia seem to show promise in terms of their validity as indices of facial recognition, it is unclear whether there is a score cutoff for any such self-report measure that would be both highly predictive of as well as highly sensitive for prosopagnosia as determined by objective tests of facial recognition. Therefore, in the current study, we sought to assess the utility of the Cambridge Face Memory Questionnaire (CFMQ, a self-report measure similar to those from other research groups), to successfully screen for prosopagnosia using a large online dataset ($N = 1518$) containing a diverse sample of participants drawn from the general population. In accord with the current standard of objective measurement for face recognition ability, we also collected performance scores on a Cambridge Face Memory Test (CFMT) and a Famous Faces Memory Test (FFMT) for each participant. Because there is also publicly available normative data on the PI20 (Gray Bird, G., Cook, R., 2017; Punit Shah, Gaule, et al., 2015), another self-report questionnaire of facial recognition ability similar to the CFMQ, it was possible to likewise assess the utility of the PI20 as a screen for prosopagnosia.

We first sought to establish that the CFMQ is reliable and correlates with objective face recognition ability. Then, using receiver operating characteristic (ROC) analyses, we attempted to determine cutoff scores on the CFMQ that would be optimal for predicting prosopagnosia, as defined by the combination of participants' age- and gender-adjusted CFMT and FFMT scores being in the lowest 2.5 percentile. For these cutoffs, we calculated the rate that (i.e., proportion of) all included participants with prosopagnosia were correctly classified (i.e., sensitivity), the rate that all those without prosopagnosia were correctly classified (i.e., specificity), and the rate at which participants who were classified as having prosopagnosia did actually have prosopagnosia (i.e., positive predictive value) to assess practicality for research screening purposes (see Appendix 1). Next, in order to assess the utility of self-report to the many real-world applications for which efficient identification of individuals with below average facial recognition abilities is of value, we also sought to assess the sensitivity, specificity, and positive predictive value of the CFMQ and PI20 in classifying individuals as being below median (i.e., below the 50th percentile) at objective face recognition. Finally, though a diagnostic cutoff has previously been proposed for the PI20 (P Shah, Gaule, Sowden, Bird, & Cook, 2015), another self-report measure of facial recognition similar to the CFMQ, there are no comparable published reports of the discriminative performance of that cutoff. Therefore, we also conducted a novel analysis on data made available by Gray, Bird, & Cook (2017) to similarly determine its sensitivity, specificity, and positive predictive value for CFMT-defined developmental prosopagnosia and to compare it to the cutoffs obtained for the CFMQ.

Methods

Participant Recruitment and Inclusion Criteria

[TestMyBrain.org](https://www.testmybrain.org) hosts an online testing environment for experimental psychology. Participants in the present study were visitors to [TestMyBrain.org](https://www.testmybrain.org) who had freely opted to complete our experiment, among those experiments then available on the website, in exchange for feedback about their performance compared to the performance of the prior participants. The website attracts volunteers from the general population over a wide age range as well as of many nationalities, ethnicities, and education levels. Our study was approved by the Harvard University Institutional Review Board (protocol # F15795–122). It conforms to the principals of the Declaration of Helsinki. In the present study, only those between the ages of 18 and 40 years old were included in the analysis because of heteroscedasticity in the assessment scores across ages outside of that age range compared to those within that age range. This heteroscedasticity was controlled by this exclusion, and without doing so, a critical assumption of our linear regression and residualization (performed for equating scores across age) would have been violated. This step excluded 3,376 of the original 8,242 participants. Also, only those who identified as of European descent were included in the analyses so as to eliminate the potential influence of the “Other Race Effect” (Malpass & Kravitz, 1969; Meissner & Brigham, 2001), given that our version of the CFMT exclusively contained face stimuli of European descent and prior research suggests that the participant-stimulus ethnic match might affect CFMT scores (Bowles et al., 2009; DeGutis, Mercado, Wilmer, & Rosenblatt, 2013). The “Other Race Effect” is a robust phenomenon of reduced recognition ability for faces of different races/ethnicities from those of one’s usual experience, and so could, theoretically, introduce spuriously low scores in our assessments for those with low experience with faces of European decent. This step excluded 1,951 of the participants remaining after the age exclusion step. Those who did not report age, gender, or ethnicity were excluded from analysis. This step excluded 1,358 of the remaining participants. No exclusions were made based on level of education so as to allow generalization of our results to all education levels and because of prior evidence indicating that education and general intelligence are not associated with facial recognition ability (Peterson & Miller, 2012; Jeremy B. Wilmer, Germine, & Nakayama, 2014). Finally, among the remaining participants, those in the top 2.5% for total time to complete our version of the CFMT were excluded to remove participants who may have been distracted or taken a long break during this assessment, which could have degraded the maintenance in memory of the faces studied during the task. This step excluded 39 of the remaining participants. As the FFMT relies on long-term memory, no analogous time-based exclusion was necessary.

Participants

1518 participants (941 female) were included in our analyses (mean age = 27.09, standard deviation = 6.44, age range: 18–40). All participants gave informed consent through the [TestMyBrain.org](https://www.testmybrain.org) website prior to participation.

Assessments

Three assessments of face recognition, in the following order, were included in the battery for each participant: (1) The Cambridge Face Memory Questionnaire (CFMQ), (2) The

Cambridge Face Memory Test, version 3 (CFMT3), and (3) a Famous Faces Memory Test (FFMT). Demographic information (age, gender, ethnicity) were also solicited from each participant.

The CFMQ is a 17-item self-report questionnaire of face recognition ability, with scores that range from 17 to 85, where lower scores indicate poorer self-reported ability. The questions of the CFMQ were developed by some of the same researchers who developed the original version of the CFMT. Drs. Brad Duchaine, Ken Nakayama, and Laura Germine developed the CFMQ to screen for prosopagnosia and it has been used for approximately the past decade by the Harvard Vision Lab on www.faceblind.org for this purpose. The questions were composed based upon their experience with persons with prosopagnosia. The CFMQ is similar to other self-report measures of facial recognition, including the better known 20-item Prosopagnosia Index (PI20). Similar to the PI20, each question in the CFMQ has five possible responses as a Likert scale. For the purpose of computing scores, number values of 1, 2, 3, 4, and 5 were, respectively, associated with the five possible response options. These number values were assigned such that, on one end of the scale, the option indicating highest ability had a value of 5, and, on the other end of the scale, the option indicating the lowest ability had a value of 1. One's CFMQ score was, then, simply the sum of the values corresponding to one's responses across all the 17 questions. It should be noted that questions 3, 4, 5, 6, 9, 13, 14, 15, and 17 were reverse scored (on these questions 'always' means worse face recognition performance). The full list of questions in the CFMQ are listed in Supplementary Methods and can be compared and contrasted with the list of PI20 questions published by Shah, Gaule, Sowden, Bird, and Cook (2015) or with the list of questions in the Kennerknecht screening questionnaire published by Kennerknecht, Plümpe, Edwards, and Raman (2007). Though the CFMQ has not served as the central focus of a publication before, questions from the CFMQ have been used in previously published studies (e.g., DeGutis, Cohan, Mercado, Wilmer, & Nakayama, 2012; J. B. Wilmer et al., 2010)

The CFMT3 is identical to the original version of the Cambridge Face Memory Test (CFMT) developed by Duchaine and Nakayama (Brad Duchaine & Nakayama, 2006), except that different facial stimuli are used. Instead of photographs of faces, the CFMT3 uses novel artificial faces that were generated with FaceGen software (Singular Inversions, Toronto, ON). As our experiment was widely and publicly available online, we refrained from using the original CFMT in our study so as to maintain the integrity of the original CFMT for clinical or in-lab use, for example, as part of a diagnosis of prosopagnosia. The CFMT, version 2 (CFMT2), which also uses FaceGen computer-generated faces similar to our CFMT3, has been in use on faceblind.org as a screening tool for several years. For this reason, we refrained from using the CFMT2 since it is theoretically possible that our participants could have been exposed to it before. The current study is the first published to use the CFMT3; However, another study (J. B. Wilmer et al., 2010) using a similar FaceGen-produced artificial facial stimuli CFMT variant, found a strong correlation with the original CFMT ($r(40) = 0.76$, 95% CI: 0.59–0.86). This correlation was not statistically different ($p = 0.44$, two-tailed) from the test-retest reliability ($r(387) = 0.70$, 95% CI: 0.64–0.74) reported for the original CFMT. Thus, it is likely that CFMT3 performance has a similarly high correlation with performance on the original CFMT. Additionally, in our

dataset, raw CFMT3 scores and raw FFMT scores (see below) correlate at $r(1516) = 0.47$, $p = 1.24e-85$.

One of three possible versions of the FFMT were assigned to each participant. In all versions, participants were shown an image of a famous face on each trial and asked “Who is this?”. If they typed in a response, they were then shown the correct answer along with their response and are asked to indicate whether they correctly identified the person. By design this allowed for misspellings of the correct answer or even for unique descriptions of the person, such as “Rolling Stones Singer” for “Mick Jagger”, to be scored as correct. As was done in a prior study (Jeremy B. Wilmer et al., 2012), total score was computed as the number of faces out of the total number of trials for which they both (a) submitted a response and (b) verified that their response was a correct identification. Participants who did not respond with the correct name were additionally asked to indicate whether they were familiar with the person, though this does not figure into the scoring procedure. Scoring was performed automatically. We recorded all participant responses though, so manual validation on a sub-sample of responses was done to verify that no systematic issues were present. To normalize the scores across FFMT versions, we calculated the FFMT version-specific z-score for each participant. Because the distributions of these scores from each of the FFMT versions were comparable, we treated the versions as equivalent in our analyses (see Supplementary Methods for the full list of famous faces for each version). Thus, in this paper, we refer to all three versions singularly as the FFMT.

Analyses

Identification of prosopagnosia.—We evaluated the discriminative utility of the CFMQ as a tool for predicting prosopagnosia, where participants were identified as having prosopagnosia independently based on their combination of CFMT3 and FFMT scores. More specifically, to identify participants with prosopagnosia, we first regressed out age and gender effects (such as have been reported by (Bowles et al., 2009; L. T. Germine, Duchaine, & Nakayama, 2011)) from the raw CFMT3 and FFMT scores using a third-order polynomial regression model so that all participants’ data could be compared to each other regardless of their age and gender. The resulting residualized scores were then translated to percentile scores, so that the resulting CFMT3 and FFMT scores could be averaged together to form a composite score. This manner of forming a composite score meets criteria generally preferred by prosopagnosia researchers, namely, that it involves multiple well-validated assessments reflecting both familiar (famous) and unfamiliar facial recognition ability and that the resulting score is highly interpretable in terms of performance relative to peers.

To reflect a commonly regarded prevalence for prosopagnosia (i.e., ~2.5% of the general population (I Kennerknecht et al., 2006)), the ground truth designation of prosopagnosia for this study was then defined as having a composite score within the lowest-performing 2.5 percentile among the composite scores of all of the included participants. It should be noted that this approach does not rule out prosopagnosia caused by poor low/mid-level vision, poor general memory, or general social cognitive disorders (e.g., autism), which are typically ruled out in laboratory-based studies. As there is no consensus on how prosopagnosia should be precisely defined and diagnosed, we make no strong claim that those we have identified

as having prosopagnosia for the purposes of our study would all have necessarily been identified as such by other research groups if brought into their labs. The conclusions of the present study are not contingent upon this issue as the phenomena that we report with respect to the relationship of self-report to objective performance would still hold for any reasonable set of diagnostic criteria.

Evaluation of the CFMQ as a screen for prosopagnosia.—We assessed the discriminant utility of the CFMQ as a screening test for prosopagnosia by applying receiver operating characteristic (ROC) curve analyses (Fawcett, 2006; Green & Swets, 1966; Spackman, 1989), which enabled quantification of the sensitivity, specificity, and positive predictive value associated with any given CFMQ cutoff score (Appendix 1). In the context of our study, sensitivity corresponds to the rate that (i.e., proportion of) all included participants with prosopagnosia were correctly classified as having prosopagnosia using a given CFMQ cutoff score. Specificity corresponds to the rate that all those without prosopagnosia were correctly classified as such using a given cutoff. Finally, positive predictive value corresponds to the rate at which participants who were classified with the cutoffs as having prosopagnosia did actually have prosopagnosia. Positive predictive value is implicitly determined by the combination of sensitivity, specificity, and the positive class's prevalence. Given the intended use of such CFMQ cutoff scores as practical screening criteria for prosopagnosia researchers, a viable cutoff should yield both an acceptable level of sensitivity and of positive predictive value. This is because low sensitivity would imply that many persons with prosopagnosia will be missed by the criterion and a low positive predictive value would imply that those meeting the criterion will be unlikely to actually have prosopagnosia upon further testing. Therefore, we evaluated the practical utility of the resulting cutoffs primarily by their associated sensitivities and positive predictive values.

In particular, we first determined three optimal operating points (i.e., the optimal CFMQ cutoff scores) for detection of prosopagnosia, such that each operating point was optimized according to a different criterion. The first of the three operating points was optimized on sensitivity and specificity using Youden's J index (Youden, 1950) such that the cost of loss of sensitivity was equal to the cost of loss of specificity (i.e., the operating point where sensitivity rate + specificity rate - 1 is maximized). The operating point resulting from this optimization would, in practice, be optimal in a context where sensitivity is more important than positive predictive value (i.e., where greater importance is attached to not failing to include those with prosopagnosia than to being highly confident that one meeting the cutoff indeed has prosopagnosia). The second operating point was optimized in a more standard manner such that the costs were proportional to the 2.5% base rate of prosopagnosia, where the cost of loss of specificity, therefore, greatly outweighs the cost of loss of sensitivity. Thus, compared to Youden's J index, this optimization adjusts costs to reflect that those without prosopagnosia greatly outnumber those with prosopagnosia. In practice, this second criterion attempts a balance between sensitivity and positive predictive value, and so returns the main CFMQ cutoff of interest. The third operating point was the one simply yielding the maximum positive predictive value, regardless of the associated sensitivity and specificity. Thus, the three different criteria constitute different degrees of trade-off between sensitivity and positive predictive value. We then recorded the sensitivity, specificity, and positive

predictive value at these operating points. The 95% confidence intervals for these three metrics were calculated using the Wilson score interval for binomial proportions (Wilson, 1927).

Identification of below-median face recognition performance.—A low base rate for a positive class, such as for the participants with prosopagnosia in our sample, is a disadvantage for achieving high classification performance (see Discussion). Additionally, the application of self-report measures of face recognition ability, such as the CFMQ, may potentially extend beyond screening for prosopagnosia to screening for below average facial recognition ability levels of much higher prevalence (e.g., within common professions requiring daily facial recognition for successful performance). Thus, we additionally assessed the discriminant utility of the CFMQ as a screening test simply for below-median facial recognition performance. The procedure for this assessment was identical to that for prosopagnosia screening, except that the prevalence of the positive class was set at 50% rather than at 2.5%.

Assuring our ROC/cutoffs results are not due to issues with poor performance.—Finally, it is theoretically possible that a non-negligible proportion of those in our sample who were identified as having prosopagnosia rather yielded artificially low scores on the CFMT3 and FFMT due to causes that are non-specific or unrelated to their facial recognition abilities (e.g., a general visual or cognitive impairment, distraction, lack of motivation, etc.), thereby partially obscuring the full potential of the CFMQ as a screen for prosopagnosia. Under such circumstances, the discriminant utility of the CFMQ as a screen for performance in the bottom 2.5 percentile should be lower than its ability as a screen for performance in the top 2.5 percentile, since it is not possible that scores within the top 2.5 percentile could be driven by such confounding factors. Therefore, we additionally assessed the CFMQ as a screen for those performing in the top 2.5 percentile on their CFMT3 + FFMT composite score to determine whether or not there are any indications of poor data quality for those in our sample we have identified as having prosopagnosia.

Evaluation of the PI20 as a screen for prosopagnosia.—Finally, to compare the discriminative performance of the resulting cutoffs for the CFMQ to that of another prominent self-report measure of facial recognition, we also similarly calculated the sensitivity, specificity, and positive predictive value for the cutoff score previously proposed for the PI20 (P Shah et al., 2015) from data made available from a study by Gray, Bird, & Cook (2017). Because we could not compute age- and gender-adjusted CFMT scores similar to how we had for the CFMT3 and FFMT scores in our own sample, we simply made use of the raw CFMT scores in this analysis and of the given cutoff score previously proposed.

Results

The CFMQ is reliable and correlates with objective face recognition ability

We verified that the CFMQ correlates with objective face recognition ability, as independently defined by CFMT3 and FFMT scores. In our sample, the Pearson's correlation between CFMQ scores and percentile composite scores (CFMT3 and FFMT)

was $r(1516) = .54, p = 3.39e-114$, indicating strong validity of the CFMQ as a measure of face recognition ability (Supplementary Figure 1). Also, the Cronbach's alpha for the CFMQ scores was .91, indicating high internal consistency for this measure, similar to the other self-reports of face recognition (e.g., Cronbach's alpha for PI20 = .93; P Shah et al., 2015, for Kennerknecht questionnaire = .84; Palermo et al., 2017).

Additional evidence for the strong validity of the self-report information contained in the CFMQ is that it correlated as well as or even more highly with each of the individual objective measures of facial recognition than the two objective measures correlated with each other. The Pearson correlation between CFMQ scores and age- and gender-adjusted (see methods) CFMT scores was $r(1516) = .44, p = 1.06e-73$ (Supplementary Figure 2), and the correlation between CFMQ scores and age- and gender-adjusted FFMT scores was $r(1516) = .52, p = 2.68e-106$ (Supplementary Figure 3). These two correlation values are impressive when considering that the correlation between the age- and gender-adjusted CFMT and FFMT scores was $r(1516) = .47, p = 7.34e-83$ (Supplementary Figure 4). This means that CFMQ scores were not significantly less correlated ($p = .40$, two-tailed) with CFMT performance than FFMT performance was with CFMT performance. This also means that CFMQ scores were significantly more highly correlated ($p = .048$, two-tailed) with FFMT performance than CFMT performance was with FFMT performance.

ROC analyses – Screening for Prosopagnosia

No CFMQ cutoff score produced simultaneously adequate rates for sensitivity and positive predictive value (see Table 1). Optimizing the CFMQ cutoff score by equating the cost of loss of sensitivity and specificity (i.e., with Youden's J index) returned a CFMQ cutoff score of 62 for a prosopagnosia screening criterion, and yielded acceptable sensitivity and specificity rates. Sensitivity was 89.19% (95% CI: 75.29% - 95.71%) and specificity was 68.06% (95% CI: 65.64% - 70.39%). However, this cutoff would be impractical because the prosopagnosia prevalence of 2.5% meant that the positive predictive value achieved was only 6.52% (95% CI: 4.68% - 9.02%). Thus, using this CFMQ cutoff score as a screening criterion would mean that the overwhelming majority (93.48%) of those meeting this criterion would not actually have prosopagnosia.

Optimizing the cutoff such that the costs of loss of sensitivity and of specificity reflect the 2.5% base rate of prosopagnosia yielded a CFMQ cutoff score of 34. At this cutoff, the positive predictive value was a respectable, albeit unreliable, 70% (95% CI: 39.68% - 89.22%) and specificity was 99.8% (95% CI: 99.41% - 99.93%). However, sensitivity was only 18.92% (95% CI: 9.48% - 34.21%), thus most of those with prosopagnosia were missed with this criterion. Additionally, since a score of 34 or lower was so uncommon, one would rarely encounter someone with such a low score in practice. Only 10 (i.e., 0.65%) out of the 1518 included participants had a CFMQ score this low or lower (95% CI: 0.36% - 1.21%).

Finally, optimizing the cutoff by maximizing the positive predictive value alone also returned a cutoff of 34, and so the sensitivity, specificity, and positive predictive value for this optimization were identical to those immediately above.

Figure 1 plots facial recognition ability vs. CFMQ score for all included participants. The performance level cutoff corresponding to prosopagnosia is indicated with a horizontal dotted line. The two optimized cutoffs returned by the foregoing ROC analyses are indicated with vertical dotted lines. From this plot, one can get an intuitive sense of the sensitivity for a given CFMQ cutoff score by observing the proportion of participants with prosopagnosia (i.e., those below the horizontal line) that is captured by the given cutoff (i.e., those to the left of the given horizontal line). Likewise, one can get an intuitive sense of the specificity by observing the proportion of participants without prosopagnosia (i.e., those above the horizontal line) that does not obtain to the CFMQ cutoff score (i.e., that falls to the right of the given vertical line). Finally, the positive predictive value would correspond to the proportion of those obtaining to the given CFMQ cutoff score (i.e., of those falling to the left of the given vertical line) who actually have prosopagnosia (i.e., who fall below the horizontal line).

ROC analyses – Screening for the bottom 50th Percentile

Compared to the cutoffs optimized to screen for prosopagnosics (i.e., for performers in the bottom 2.5 percentile), a more adequate combination of sensitivity and positive predictive value was possible for a cutoff optimized to screen for performers in the bottom 50th percentile (see Table 2). This provides evidence that the low base rate of participants with prosopagnosia was a critical disadvantage for achieving high classification performance with the CFMQ. Optimizing the cutoff such that the costs of loss of sensitivity and of specificity reflect the 50% base rate of scores within the bottom 50th percentile is equivalent to optimizing it by equating the cost of loss of sensitivity and specificity. In this case, the optimization yielded a CFMQ cutoff score of 66, a sensitivity of 65.35% (95% CI: 61.89% - 68.65%), a specificity of 71.41% (95% CI: 68.09% - 74.51%), and a positive predictive value of 69.57% (95% CI: 66.09% - 72.83%). The cutoff score was thus squarely above chance (i.e., 50%) in predicting whether someone was below the 50th percentile in facial recognition performance, though depending on the particular application, the positive predictive value of approximately 70% may or may not be sufficiently predictive for the purposes of a given practical application. Regardless, that either the sensitivity or the positive predictive value was lower in the two corresponding optimizations for predicting prosopagnosia indicates that the low base rate of prosopagnosia did, in practice, disadvantage the CFMQ as a screening tool for prosopagnosia.

Optimizing the cutoff by maximizing the positive predictive value alone yielded a cutoff score of 39 (the maximum cutoff score with a positive predictive value of 100%), a perfect specificity of 100% (95% CI: 99.5% - 100%), a perfect positive predictive value of 100% (95% CI: 89.57% - 100%), but a sensitivity of only 4.35% (95% CI: 3.11% - 6.04%). That the overwhelming majority of the sensitivity was lost in order to achieve maximal specificity was reflective of the fact that the distribution of CFMT3 scores for those in the bottom 50 percentiles greatly overlap that for those in the top 50 percentiles (compare those above and below the 50th percentile in facial recognition ability in the scatter plot in Figure 1).

ROC analyses – Screening for the top 2.5th Percentile

The discriminant utility of the cutoff scores optimized to screen for performers in the top 2.5 percentile of the objective assessment was not higher than those optimized for screening for prosopagnosia (see Table 3). This means that there was no indication of poor data quality for those in our sample we have identified as having prosopagnosia, and so poor data quality could not have explained the low discriminant utility of the relevant cutoff scores for prosopagnosia. Optimizing the CFMQ cutoff score using Youden's J index returns a CFMQ cutoff score of 72 for a top 2.5 percentile performance screening criterion, yields a sensitivity at 76.92% (95% CI: 61.66% - 87.35%) and a specificity at 67.68% (95% CI: 65.25% - 70.02%). However, this criterion is impractical because, as was the case with the screening of prosopagnosia under this cutoff optimization method, the prevalence of our top performers at 2.5% means that the positive predictive value achieved here is similarly only 5.91% (95% CI: 4.17% - 8.31%). The sensitivity, specificity, and positive predictive value for this optimization are comparable, if somewhat numerically lower than, the corresponding optimization for prosopagnosia screening. Thus, using this CFMQ cutoff score as a screening criterion would likewise mean that the overwhelming majority (94.09%) of those meeting this criterion will not actually perform within the top 2.5 percentile.

Optimizing the cutoff such that the costs of loss of sensitivity and of specificity reflect the 2.5% base rate of the top performers yields a CFMQ cutoff at the empirical maximum score of 85. At this cutoff, the positive predictive value is 42.86% (95% CI: 15.82% - 74.95%) and specificity is 99.73% (95% CI: 99.31% - 99.89%). However, sensitivity is only 7.69% (95% CI: 2.65% - 20.32%), thus most of the top performers were missed with this criterion. The sensitivity, specificity, and positive predictive value for this optimization are again comparable, though numerically lower than, the corresponding optimization for prosopagnosia screening. Additionally, since a score of 85 is likewise very uncommon, one will rarely encounter someone with such a high score. Only 7 (i.e., 0.46%) out of the 1518 included participants had this CFMQ score. Finally, optimizing the cutoff by maximizing the positive predictive value alone also returns a cutoff of 85, and so the sensitivity, specificity, and positive predictive value for this optimization are identical to those reported immediately above.

Low performance scores on the CFMT3 and FFMT could, theoretically, be reflective of poor data quality, though the same is not true of high performance scores. Therefore, a stronger discriminant utility of CFMQ scores for detecting our highest performers than for our lowest "prosopagnosia" performers, if it existed, could have been an indication of compromised data quality for our lowest performers. However, since the sensitivity, specificity, and positive predictive value metrics yielded by the optimizations of the screening for the top performers were comparable to, and numerically even always lower than, those corresponding to the optimizations screening for prosopagnosia, we thus find no evidence that the low metrics in the prosopagnosia optimizations are due to poor data quality for the sub-sample of our participants that had been identified as having prosopagnosia. For this reason, we can be confident that those we have identified as having prosopagnosia do indeed have face recognition impairment and are representative of the population of those with prosopagnosia.

Item Response Theory analyses

It is possible that the low discriminant utility of the CFMQ we observed at one or both of its high and/or low end could be due to lack of precision for reliably identifying exceptionally high or low self-assessed face recognition. For example, perhaps the CFMQ, which was originally developed to identify poor face recognition, is more sensitive at its low end than its high end. Another possibility is that the CFMQ's high correlation with objective measures results from a high capacity to discriminate self-assessed face recognition within the normal range of performance, but this discrimination capacity breaks down at the extremes; another way of stating this is that the CFMQ might be better at discriminating the 49th from 50th percentile of self-assessed face recognition than at discriminating the 2nd from the 3rd percentile or the 98th from the 97th percentile. Finally, it is worth checking to verify that the overall precision of the CFMQ, across its full range of values, but especially at its extremes, is reasonably comparable to that of our objective measures.

To address these three questions - (1) a potential differential precision of CFMQ at high vs. low end, (2) the breakdown of precision for CFMQ at its extremes, and (3) a potentially poor precision of CFMQ relative to objective measures - we conducted a standard item-response theory analysis that enables a direct quantification of the precision of an instrument across its full range (Jeremy B. Wilmer et al., 2012). The details of these results are included as supplemental information. The conclusions of these analyses were that (1) CFMQ is similarly precise at both high and low end, (2) CFMQ is actually more, rather than less, sensitive at its extremes than in its middle range, and (3) CFMQ has comparable, or even better, precision than our objective measures across its full range, including the extremes. Therefore, we conclude that the CFMQ has good precision across its whole range; that the high-end analysis was not limited by poor CFMQ precision in this range; and that the limited diagnostic and screening utility we observe for self-report does not appear to be a simple artifact of limited precision of self-report instruments.

Our item-response theory analysis also confirms (see supplemental information) that the sensitivity of both CFMT3 and FFMT are similar at both extremes of performance, thereby effectively ruling out the possibility that these instruments, which were developed in part to identify poor face recognition performance, are simply insensitive to high performance, thereby limiting the comparative value of our high-end ROC analysis.

Single question screening

The first question in the CFMQ is "Compared to my peers, I think my face recognition skills are...", with possible responses from 1 = "far below average" to 5 = "far above average", constituting the most general and direct self-assessment item for face recognition ability in the questionnaire. Use of the response of "far below average" on this item as a prosopagnosia screening criterion yields, arguably, the most acceptable combination of sensitivity, positive predictive value, and prevalence yet. The sensitivity was 54.05% (95% CI: 41.27% – 66.32%), the specificity was 92.1% (95% CI: 88.89% - 94.44%), the positive predictive value was 14.6% (95% CI: 9.65% – 21.48%), and the cutoff contains 9.03% (95% CI: 7.69% - 10.57%) of our included participants. Therefore, use of this single response may be relatively more practical than use of the sum score on the entire CFMQ as a screening

criterion for prosopagnosia. Nonetheless, this criterion still may not be highly useful in absolute terms since the positive predictive value suggests that most persons meeting this criterion (85.4%) will not have prosopagnosia upon further objective performance testing.

PI20 cutoff analysis on Gray, Bird, & Cook (2017)

The discriminant utility of our CFMQ cutoff scores can be compared with the discriminant utility of the PI20 cutoff score proposed by the authors of the PI20. Combining both of the datasets reported in Gray, Bird, & Cook (2017), we conducted a novel analysis of the discriminant utility of the PI20 cutoff score of 65 or greater as a screen for prosopagnosia. Our calculations yielded a total of 425 included participants, 59 of whom we identified as having prosopagnosia on the basis that they obtained a score of 44 or lower on the original version of the CFMT, which, from previously published norms (Bowles et al., 2009; B. C. Duchaine, Yovel, Butterworth, & Nakayama, 2006; Brad Duchaine & Nakayama, 2006; Bradley Duchaine, Yovel, & Nakayama, 2007; Garrido, Duchaine, & Nakayama, 2008), corresponds to performance approximately within the bottom 2.5% (no FFMT was administered). 12 participants had been classified as having prosopagnosia with the PI20 cutoff, 5 of whom actually had prosopagnosia according to their CFMT score. Thus, the positive predictive value was $5/12 = 41.67\%$, though the small number of participants classified as having prosopagnosia with the PI20 cutoff meant that the confidence interval on this rate was wide (95% CI: 19.33% - 68.05%). The sensitivity was $5/59 = 8.47\%$ (95% CI: 3.67% - 18.35%) and the specificity was $359/366 = 98.09\%$ (95% CI: 96.11% - 99.07%). Given that 12 out of the 425 participants met the PI20 cutoff, the prevalence of those with such scores was thus 2.82% (95% CI: 1.62% - 4.87%). Given the wide confidence interval on the positive predictive value, it was not possible to adequately evaluate the usefulness of the PI20 cutoff for predicting CFMT-defined prosopagnosia. However, it is clear that the sensitivity is low, indicating that most (about 91.5%) of those with prosopagnosia would not be detected.

The PI20 cutoff within the Gray, Bird, & Cook (2017) dataset thus yielded comparable sensitivity, specificity, and positive predictive value to our CFMQ cutoff of 34 within our dataset as reported above (see Table 4). Though sensitivity, specificity, and positive predictive value were all numerically lower for the PI20 cutoff than for that CFMQ cutoff, the confidence intervals for these metrics greatly overlap, indicating that they did not significantly differ between the two questionnaires. This holds true even if we identify our participants with prosopagnosia only on the basis of CFMT3 scores instead of on the basis of composite CFMT3 and FFMT scores (Gray, Bird, & Cook did not administer the FFMT to their participants). Specifically, performing the analogous cutoff optimization for CFMT3-defined prosopagnosia as for composite-score-defined prosopagnosia again yields a CFMQ cutoff score of 34, with a sensitivity of 16.22% (95% CI: 7.65% - 31.14%), specificity of 99.73% (95% CI: 99.31% - 99.89%), and positive predictive value of 60% (95% CI: 31.27% - 83.18%). The one advantage that the PI20 cutoff has over the CFMQ cutoff, though, is that the prevalence of those meeting the cutoff is higher (2.82% vs. 0.66%, respectively) and the difference is statistically significant according to a binomial test on the Gray, Bird, & Cook (2017) data where the null hypothesis is that the prevalence is equal to our value of 0.66% ($p = 0.0000070$).

Figure 2 plots the data of Gray, Bird, & Cook (2017) in a manner analogous to Figure 1. One difference to note in Figure 2 compared to Figure 1 is that those meeting the PI20 cutoff are those to the right of the vertical line, not to the left, as was the case for those meeting the CFMQ cutoff in Figure 1. Also, the x-axis is CFMT score, not the composite (CFMT3 + FFMT) percentile score we were able to calculate in our own data. Finally, the x-axis space did not need to be transformed, since CFMT scores were already normally distributed, allowing for easy visibility of the participants with CFMT-defined prosopagnosia who fell above and below the proposed PI20 cutoff score.

Discussion

We investigated the utility of a self-report measure of facial recognition ability, the Cambridge Face Memory Questionnaire (CFMQ), as a screen for discriminating between individuals with and without prosopagnosia-level performance on objective tests commonly used to confirm a diagnosis of developmental prosopagnosia. We found the CFMQ to be a highly reliable and valid index of facial recognition ability (Cronbach's $\alpha = .91$ and Pearson's $r(1516) = .54$ with objective performance). The validity of the CFMQ as a measure of facial recognition is impressive given that the reported average correlation between self-evaluated memory capability and objective recognition performance metrics in other domains of memory is only $r = .1$ (Beaudoin & Desrichard, 2011; Jeremy B. Wilmer, 2017). Nonetheless, no CFMQ cutoff score yielded a practically adequate combination of sensitivity and positive predictive value for prosopagnosia, as independently determined by performance on the objective facial recognition tests. In other words, no CFMQ cutoff score both (1) included most of the participants with prosopagnosia and (2) produced a high and reliable rate at which those participants who met the cutoff actually had prosopagnosia. That higher simultaneous rates for sensitivity and positive predictive value were possible with a CFMQ cutoff score optimized for discriminating between those above and below median performance on the objective tests indicates that the low base rate of prosopagnosia is a source of the poor discrimination between those with and without prosopagnosia. Further, that comparable rates for sensitivity and positive predictive value were produced with a CFMQ cutoff score optimized for discriminating the top 2.5 percentile on the objective tests from the rest of the participants means that there was no indication that the quality of the data for those we had identified as having prosopagnosia was poor. Finally, that comparable discriminative utility was found also for the cutoff score proposed for the PI20 suggests that the low discriminative utility of the CFMQ for prosopagnosia generalizes to other self-report measures.

The utility of any given screening cutoff score may depend on the context of the research and the preferences of the researcher. For example, a researcher may be seeking to maximize their sample of individuals with prosopagnosia and may have the resources to perform confirmatory diagnostic tests on many individuals even if it is expected that many of those individuals will not ultimately be found to have prosopagnosia. Such a researcher would likely value sensitivity over positive predictive value with their screening cutoff score. A researcher seeking to be more confident that an individual meeting the cutoff will actually have prosopagnosia, even if it means failing to detect many of those with prosopagnosia, will likely value positive predictive value over sensitivity. For this reason, we optimized the

CFMQ cutoff according to different criteria which either favored sensitivity, specificity, or aimed to strike a balance between the two. Cutoff optimization favoring sensitivity yielded a cutoff score with an associated sensitivity of 89.19%, but positive predictive value of only 6.52%. Cutoff optimization striking more of a balance of sensitivity and positive predictive value yielded a cutoff with an associated sensitivity of only 18.92% and positive predictive value of 70%. Though 70% is a respectable rate for a positive predictive value, it must be noted that this was not a highly reliable value (95% CI: 39.68% - 89.22%). Furthermore, those meeting that cutoff included only 0.65% of all participants. Cutoff optimization favoring positive predictive value yielded the same results as those yielded in the attempt to strike a balance between sensitivity and positive predictive value. Notwithstanding the differences in the value that researchers may place on sensitivity versus positive predictive value, no CFMQ cutoff is likely to be practical. The researcher who values sensitivity can obtain good sensitivity, but will find that more than 93% of persons meeting the cutoff optimized for sensitivity will not actually have prosopagnosia. The researcher who values positive predictive value will rarely be able to find individuals meeting the cutoff optimized for positive predictive value. Finally, the researcher seeking a balance between sensitivity and positive predictive value is left without a cutoff that is able to achieve it.

That the CFMQ is not apt for discriminating between those with and without prosopagnosia may seem inconsistent with its demonstrated reliability and validity as an index of objective facial recognition ability. However, this lack of discriminative utility stems from the low base rate of prosopagnosia (here 2.5%) along with the fact that developmental prosopagnosia constitutes the tail end of a roughly normal distribution of facial recognition ability, rather than some extreme cluster largely segregated from the rest of the distribution, when diagnosed on the basis of behavior alone (Barton & Corrow, 2016). For these reasons, an exceptionally high degree of precision in the prediction of objective performance from self-report would be required for adequate discrimination of the small number of those with prosopagnosia from the nearly 40-fold larger number of those without prosopagnosia. Though the association between self-report and objective performance is robust, it is not commensurate with the extremely high degree of precision required, meaning that the distribution of CFMQ scores for those without prosopagnosia greatly overlaps and drowns out the distribution of CFMQ scores for those with prosopagnosia. This state of affairs is plainly evident in Figure 1, where, despite the clearly robust linear relationship between CFMQ score and objective performance, there is overlap in CFMQ scores between those with and without prosopagnosia. An indication that the low base rate of prosopagnosia exacerbates the situation is our finding that when using the CFMQ to instead discriminate merely between the upper and lower 50% of facial recognition performers, higher simultaneous rates for sensitivity (65.35%) and positive predictive value (69.57%) can be achieved with an optimized cutoff score.

Our participants performed all assessments remotely through the internet, and prior evidence indicates that the quality and reliability of data collected through the web is as good as that of data collected in the laboratory (L. Germine et al., 2012). Consistent with this, the magnitude of our correlation between CFMQ scores and objective facial recognition performance fell at the high end of the range of values previously reported for associations between subjective and objective measures in the laboratory-based studies of other research

groups (Bindemann et al., 2014; Gray Bird, G., Cook, R., 2017; Livingston & Shah, 2017; Rotshtein et al., 2007; Punit Shah, Gaule, et al., 2015; Punit Shah, Sowden, et al., 2015; Stollhoff et al., 2011; Turano et al., 2016; Turano & Viggiano, 2017). Nonetheless, one could speculate that the data for the participants we had identified as having prosopagnosia may be uniquely suspect with respect to its quality, given that this subset of participants, by definition, constitutes the outliers and that such poor performance could merely be caused by factors unrelated to facial recognition ability. If this were so, the apparent utility of the CFMQ for discriminating between those with and without prosopagnosia would be spuriously reduced in our study. However, this possibility is highly unlikely, given that in using the CFMQ to attempt to discriminate the top 2.5% of facial recognition performers from the rest of our participants, we discovered comparably poor discriminative ability as we had found for prosopagnosia. Furthermore, additional Item Response Theory analyses (Supplementary Results, Supplementary Figures 1–5) of all of our assessment metrics verified that facial recognition ability estimate precisions from each of the metrics we employed were sufficient within the relevant score ranges. Thus, neither compromised quality of our participant sample nor of our assessment metrics were the cause for the CFMQ being poor at discriminating between those with and without prosopagnosia. However, even considering the hypothetical case where a significant proportion of the participants that we identified as having prosopagnosia were so due to more general vision, memory, or social cognitive deficits, the inclusion of such participants would still have negligible influence on the results and interpretation of our study. This is because all of these conditions still bear upon a person's ability to recognize faces relative to others in the population, and whether one's facial recognition deficit is domain specific or not, a comparable impairment should still be evident in the objective assessment and in self-report.

Does the limited discrimination of prosopagnosia-level performance observed here generalize broadly across self-report instruments? As previously mentioned, we found the magnitude of the correlation between the CFMQ and objective performance to be higher than almost all prior published correlations between subjective and objective measures of facial recognition. This is perhaps merely because of the steps we took to make our metric of objective performance standardized and reliable (e.g., age- and gender-adjustments, percentile scoring, and the composition of two tests) rather than because the CFMQ reflects objective performance more strongly than other self-report measures. Nonetheless, for the reasons mentioned above relating to the low base rate of prosopagnosia and the overlap in self-report scores between those with and without prosopagnosia, it seems unlikely that any of the other self-report measures currently in existence would be capable of achieving substantially better discriminative ability than the CFMQ. Consistent with this is our finding from a novel analysis of data from another self-report measure, the PI20 (Punit Shah, Gaule, et al., 2015). Using data available from a published study (Gray Bird, G., Cook, R., 2017) in which PI20 and CFMT scores were collected from a sample of 435 participants, we found that the proposed PI20 score cutoff of 65 yielded rates of sensitivity, specificity, and positive predictive value comparable, if numerically lower, to those yielded by the cutoff score of 34 for the CFMQ, even when prosopagnosia was analogously defined only on the basis of CFMT3 scores, rather than composite scores, within our participant sample. In addition to suggesting the generalizability of our findings to other self-report measures, this finding

further argues against the possibility that our dataset may in some way be contaminated by unconscious participants, since, unlike our internet-based participant sample, the participant sample of Gray, Bird and Cook were laboratory-based and so close supervision was possible during testing.

The poor performance of the CFMT and the PI20 in discriminating between those with and without prosopagnosia can be contrasted with the much better performance, for example, of the Autism-Spectrum Quotient (AQ) (Baron-Cohen, Wheelwright, Skinner, Martin, & Clubley, 2001) in discriminating between those with and without autism spectrum disorders or of the Montreal Cognitive Assessment (MoCA) in discriminating between those with and without Mild Cognitive Impairment or Alzheimer's Disease. The AQ has been reported to detect 80% of participants with high functioning autism spectrum disorders when a cutoff score of 32 or higher on the AQ was used, whereas only 2% of controls had such scores (i.e., a cutoff sensitivity of 80% and specificity of 98%, (Baron-Cohen et al., 2001). Scores on the Montreal Cognitive Assessment (MoCA) have been reported to enable a detection rate (i.e., sensitivity) of 81%, a correct rejection rate (i.e., specificity) of 77%, and positive prediction at a rate of 78% for Mild Cognitive Impairment. MoCA scores also enable a sensitivity of 88%, specificity of 98%, and positive predictive value of 98% for Alzheimer's Disease (Freitas, Simões, Alves, & Santana, 2013).

Comparisons with the AQ and the MoCA may not be appropriate, however. Critical differences between how autism spectrum disorders and dementias are screened and diagnosed compared to how prosopagnosia tends to be screened and diagnosed likely explains why the AQ and MoCA are not susceptible to the same disadvantages that self-reported facial recognition has for predicting prosopagnosia. Autism spectrum disorders are diagnosed on the basis of semi-structured observation of behavior by an examiner and the questions in the AQ are similar in content to this diagnostic assessment. Prosopagnosia, on the other hand, tends to be diagnosed not on the basis of ecologically-observed behaviors corresponding to those subjectively assessed with the CFMT or PI20, but rather on the basis of objective performance metrics of facial recognition, (e.g., the CFMT and FFMT, though see I Kennerknecht et al., 2006; Ingo Kennerknecht, Nga, et al., 2008; Ingo Kennerknecht, Pluempé, & Welling, 2008; Stollhoff et al., 2011). Alzheimer's Disease and other dementias, like prosopagnosia, are diagnosed on the basis of objective tests; however, the MoCA is itself an objective measure, not a subjective self-report measure, and so, like the AQ, is similar in nature to the relevant diagnostic tests. Thus, the AQ and MoCA are not susceptible to the reduction of discriminative performance that is likely inevitably introduced when there is a difference in kind of information (i.e., subjective/observational vs. objective performance) used between screening and diagnosis.

If all forms of self-report, by themselves, are of limited value in screening for prosopagnosia, then what alternative screening approaches could be practicable for researchers seeking to identify and recruit persons with prosopagnosia? Our study may serve as a proof of concept for an approach of remote *objective* performance testing via the internet. Though the current gold standard assessments of facial recognition performance (e.g., the CFMT original and FFMT) may be reserved for laboratory-based confirmation of prosopagnosia, collecting data from alternative forms of such objective performance

assessments online and then screening based on performance scores could serve as an efficient, economical, and reliable means of locating persons highly likely to have prosopagnosia. Indeed, alternative forms of the CFMT, along with norms for performance within the general population, already exist which may be suitable for such purposes (e.g., the CFMT3 and norms from the current study, or the CFMT2 and norms as in (L. Germine, Duchaine, & Nakayama, 2013; L. T. Germine et al., 2011)).

Though self-report information, by itself, is of limited use for predicting prosopagnosia as independently defined by objective tests, this does not preclude it as an integral aspect of the very definition of prosopagnosia, and thus, as essential in the context of diagnosis. Because self-report often reflects aspects of everyday functioning and disability not fully captured by objective tests (B. C. Duchaine & Weidenfeld, 2003; Stollhoff et al., 2011), self-report could potentially be regarded either as the primary diagnostic information or, at least, as on par with objective measures. This possibility of using self-report as primary diagnostic information for prosopagnosia is similar to how self-report, in the form of extended structured clinical interviews, is the primary or co-primary diagnostic information for disorders of which subjective distress and/or functional disability are, by definition, essential characteristics. Developmental disorders such as autism, dyslexia, and attention deficit hyperactivity disorder, as well as psychiatric disorders such as depression, obsessive-compulsive disorder, and posttraumatic stress disorder are examples of such disorders. Therefore, some prosopagnosia researchers treat subjective report as the primary information for a putative diagnosis (e.g., I Kennerknecht et al., 2006; Ingo Kennerknecht, Pluempfe, et al., 2008; Ingo Kennerknecht et al., 2007; Stollhoff et al., 2011), though it should be noted that prosopagnosia researchers using this approach are currently in the minority.

Alternatively, self-report could instead be used as a prerequisite for the objective tests of facial recognition that are then subsequently used as the primary information to confirm the diagnosis of prosopagnosia. Such a step-wise approach has precedent in screening for cognitive decline, where objective measures are collected only if self- or informant-reported complaints exist (Crooks, Buckwalter, Petitti, Brody, & Yep, 2005; Winblad et al., 2004). In line with this approach, most labs studying prosopagnosia have employed some form of self-report or self-identification as a screen to select individuals for further laboratory testing where objective measures are then collected as the ground truth information to perform the putative diagnosis.

Presently, no consensus exists within the research community for how self-report information should be utilized in the diagnosis of prosopagnosia. However, our data reveal that this unresolved detail is not without significant consequence. For the purposes of our study, we identified persons with prosopagnosia within our sample entirely on the basis of their performance on objective tests typically utilized in the confirmation of a diagnosis of prosopagnosia. It is evident that many of the participants that we identified as having prosopagnosia instead subjectively perceive their facial recognition ability as within the normal, or even high, range compared to others (see Figure 1). To meet a “true” diagnosis, should those even without insight into their exceptionally poor facial recognition ability be considered as having prosopagnosia or should some form of subjective complaint instead be required? Likewise, for those whose performance on objective measures is within the normal

range, but who nonetheless present with complaints that their ability is exceptionally impaired with realworld consequences, how much, if any, provision should be made for diagnosing such individuals with prosopagnosia? The data of the present study cannot resolve this issue of diagnostic convention, but do highlight some of what is at stake in any convention that may be adopted.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Appendix 1: ROC analysis definitions in the context of detecting prosopagnosia with the CFMQ:

Sensitivity –

The rate at which participants *with prosopagnosia* were correctly classified when using a given CFMQ cutoff score. A low sensitivity rate would imply that a high proportion of persons with prosopagnosia would not be detected on the basis of their CFMQ score alone.

Specificity –

The rate at which participants *without prosopagnosia* were correctly classified when using a given CFMQ cutoff score. A low specificity rate would imply that a high proportion of persons without prosopagnosia would be falsely labeled as having prosopagnosia on the basis of their CFMQ score alone.

Positive predictive value –

The rate at which those classified as having prosopagnosia *actually had prosopagnosia*. High positive predictive value does not imply high sensitivity, or vice versa, so it is possible for a high proportion of persons with prosopagnosia to fail to be detected even if a high proportion of those classified as having prosopagnosia actually have prosopagnosia.

References

- Baron-Cohen S, Wheelwright S, Skinner R, Martin J, & Clubley E (2001). The autism-spectrum quotient (AQ): evidence from Asperger syndrome/high-functioning autism, males and females, scientists and mathematicians. *J Autism Dev Disord*, 31(1), 5–17. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/11439754> [PubMed: 11439754]
- Barton JJS, & Corrow SL (2016). The problem of being bad at faces. *Neuropsychologia*. 10.1016/j.neuropsychologia.2016.06.008
- Beaudoin M, & Desrichard O (2011). Are Memory Self-Efficacy and Memory Performance Related? A Meta-Analysis. *Psychological Bulletin*, 137(2), 211–241. 10.1037/a0022106 [PubMed: 21244133]
- Bindemann M, Attard J, & Johnston RA (2014). Perceived ability and actual recognition accuracy for unfamiliar and famous faces. *Cogent Psychology*, 1(1). 10.1080/23311908.2014.986903
- Bobak AK, Mileva VR, & Hancock PJ (2018). Facing the facts: Naive participants have only moderate insight into their face recognition and face perception abilities. *Quarterly Journal of Experimental Psychology* 10.1177/1747021818776145

- Bowles DC, McKone E, Dawel A, Duchaine B, Palermo R, Schmalzl L, ... Yovel G (2009). Diagnosing prosopagnosia: Effects of ageing, sex, and participant-stimulus ethnic match on the cambridge face memory test and cambridge face perception test. *Cognitive Neuropsychology*, 26(5), 423–455. 10.1080/02643290903343149 [PubMed: 19921582]
- Crooks VC, Buckwalter JG, Petitti DB, Brody KK, & Yep RL (2005). Self-reported severe memory problems as a screen for cognitive impairment and dementia. *Dementia*, 4(4), 539–551. 10.1177/1471301205058310
- De Heering A, & Maurer D (2014). Face memory deficits in patients deprived of early visual input by bilateral congenital cataracts. *Developmental Psychobiology*, 56(1), 96–108. 10.1002/dev.21094 [PubMed: 23192566]
- DeGutis J, Cohan S, Mercado RJ, Wilmer J, & Nakayama K (2012). Holistic processing of the mouth but not the eyes in developmental prosopagnosia. *Cognitive Neuropsychology*, 29(5–6), 419–446. 10.1080/02643294.2012.754745 [PubMed: 23428080]
- DeGutis J, Mercado RJ, Wilmer J, & Rosenblatt A (2013). Individual Differences in Holistic Processing Predict the Own-Race Advantage in Recognition Memory. *PLoS ONE*, 8(4). 10.1371/journal.pone.0058253
- Duchaine BC, & Weidenfeld A (2003). An evaluation of two commonly used tests of unfamiliar face recognition. *Neuropsychologia*, 41(6), 713–720. 10.1016/S0028-3932(02)00222-1 [PubMed: 12591028]
- Duchaine BC, Yovel G, Butterworth EJ, & Nakayama K (2006). Prosopagnosia as an impairment to face-specific mechanisms: Elimination of the alternative hypotheses in a developmental case. *Cognitive Neuropsychology*, 23(5), 714–747. 10.1080/02643290500441296 [PubMed: 21049351]
- 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. 10.1016/j.neuropsychologia.2005.07.001 [PubMed: 16169565]
- Duchaine B, Yovel G, & Nakayama K (2007). No global processing deficit in the Navon task in 14 developmental prosopagnosics. *Social Cognitive and Affective Neuroscience*, 2(2), 104–113. 10.1093/scan/nsm003 [PubMed: 18985129]
- Fawcett T (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), 861–874. 10.1016/j.patrec.2005.10.010
- Freitas S, Simões MR, Alves L, & Santana I (2013). Montreal cognitive assessment: Validation study for mild cognitive impairment and alzheimer disease. *Alzheimer Disease and Associated Disorders*, 27(1), 37–43. 10.1097/WAD.0b013e3182420bfe [PubMed: 22193353]
- Garrido L, Duchaine B, & Nakayama K (2008). Face detection in normal and prosopagnosic individuals. *Journal of Neuropsychology*, 2(Pt 1), 119–140. 10.1348/174866407X246843 [PubMed: 19334308]
- Germine L, Duchaine B, & Nakayama K (2013). Face Recognition Memory Data Across the Lifespan N approx 60,000. Figshare. 10.6084/m9.figshare.767320.v1
- Germine L, Nakayama K, Duchaine BC, Chabris CF, Chatterjee G, & Wilmer JB (2012). Is the Web as good as the lab? Comparable performance from Web and lab in cognitive/perceptual experiments. *Psychonomic Bulletin and Review*, 19(5), 847–857. 10.3758/s13423-012-0296-9 [PubMed: 22829343]
- Germine LT, Duchaine B, & Nakayama K (2011). Where cognitive development and aging meet: Face learning ability peaks after age 30. *Cognition*, 118(2), 201–210. 10.1016/j.cognition.2010.11.002 [PubMed: 21130422]
- Gray Bird G, Cook R, H. L (2017). Robust associations between the 20-Item Prosopagnosia Index and the Cambridge Face Memory Test in the general population. *Royal Society Open Science*. Retrieved from [internal-pdf://228.60.152.105/Gray, Bird & Cook \(in press\).pdf](https://doi.org/10.1098/rsos.170200)
- Green DM, & Swets JA (1966). Signal detection theory and psychophysics. *Society*, 1, 521 10.1901/jeab.1969.12-475
- 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). *Am J Med Genet A*, 140(15), 1617–1622. 10.1002/ajmg.a.31343 [PubMed: 16817175]

- Kennerknecht I, Nga YH, & Wong VCN (2008). Prevalence of hereditary prosopagnosia (HPA) in Hong Kong Chinese population. *American Journal of Medical Genetics, Part A*, 146(22), 2863–2870. 10.1002/ajmg.a.32552
- Kennerknecht I, Pluemp N, & Welling B (2008). Congenital prosopagnosia- a common hereditary cognitive dysfunction in humans. *Frontiers in Bioscience*, 13, 3150–3158. <https://doi.org/2916> [pii] [PubMed: 17981784]
- Kennerknecht I, Plümpe N, Edwards S, & Raman R (2007). Hereditary prosopagnosia (HPA): The first report outside the Caucasian population. *Journal of Human Genetics*, 52(3), 230–236. 10.1007/s10038-006-0101-6 [PubMed: 17186317]
- Livingston LA, & Shah P (2017). People with and without prosopagnosia have insight into their face recognition ability. *Quarterly Journal of Experimental Psychology*, pp. 1–3. 10.1080/17470218.2017.1310911
- Malpass RS, & Kravitz J (1969). Recognition for faces of own and other race. *Journal of Personality and Social Psychology*, 13(4), 330–334. 10.1037/h0028434 [PubMed: 5359231]
- Meissner CA, & Brigham JC (2001). Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychology, Public Policy, and Law*, 7(1), 3–35. 10.1037//1076-8971.7.1.3
- Palermo R, Rossion B, Rhodes G, Laguesse R, Tez T, Hall B, ... McKone E (2017). Do people have insight into their face recognition abilities? *Q J Exp Psychol (Hove)*, 70(2), 218–233. 10.1080/17470218.2016.1161058 [PubMed: 26935244]
- Peterson E, & Miller SF (2012). The eyes test as a measure of individual differences: How much of the variance reflects verbal IQ? *Frontiers in Psychology*, 3(JUL). 10.3389/fpsyg.2012.00220
- Rotshtein P, Geng JJ, Driver J, & Dolan RJ (2007). Role of features and second-order spatial relations in face discrimination, face recognition, and individual face skills: behavioral and functional magnetic resonance imaging data. *Journal of Cognitive Neuroscience*, 19(9), 1435–1452. 10.1162/jocn.2007.19.9.1435 [PubMed: 17714006]
- 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), 140343 10.1098/rsos.140343 [PubMed: 26543567]
- 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. *R Soc Open Sci*, 2(6), 140343 10.1098/rsos.140343 [PubMed: 26543567]
- Shah P, Sowden S, Gaule A, Catmur C, & Bird G (2015). The 20 item prosopagnosia index (PI20): relationship with the Glasgow face-matching test. *Royal Society Open Science*, 2(11), 150305 10.1098/rsos.150305 [PubMed: 26715995]
- Snowling M, Dawes P, Nash H, & Hulme C (2012). Validity of a protocol for adult selfreport of dyslexia and related difficulties. *Dyslexia*, 18(1), 1–15. 10.1002/dys.1432 [PubMed: 22271419]
- Spackman KA (1989). Signal detection theory: valuable tools for evaluating inductive learning. In *Proceedings of the Sixth International Workshop on Machine Learning* (pp. 160–163). 10.1016/B978-1-55860-036-2.50047-3
- Stollhoff R, Jost J, Elze T, & Kennerknecht I (2011). Deficits in long-term recognition memory reveal dissociated subtypes in congenital prosopagnosia. *PLoS ONE*, 6(1). 10.1371/journal.pone.0015702
- Turano MT, Marzi T, & Viggiano MP (2016). Individual differences in face processing captured by ERPs. *International Journal of Psychophysiology*, 101, 1–8. 10.1016/j.ijpsycho.2015.12.009 [PubMed: 26747730]
- Turano MT, & Viggiano MP (2017). The relationship between face recognition ability and socioemotional functioning throughout adulthood. *Aging, Neuropsychology, and Cognition*, 24(6), 613–630. 10.1080/13825585.2016.1244247
- White D, Kemp RI, Jenkins R, Matheson M, & Burton AM (2014). Passport officers' errors in face matching. *PLoS One*, 9(8), e103510 10.1371/journal.pone.0103510 [PubMed: 25133682]
- Wilmer JB (2017). Individual Differences in Face Recognition: A Decade of Discovery. *Current Directions in Psychological Science*, 26(3), 225–230. 10.1177/0963721417710693
- Wilmer JB, Germine L, Chabris CF, 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. 10.1080/02643294.2012.753433 [PubMed: 23428079]

Wilmer JB, Germine L, Chabris CF, Chatterjee G, Williams M, Loken E, ... Duchaine B (2010). Human face recognition ability is specific and highly heritable. *Proceedings of the National Academy of Sciences*, 107(11), 5238–5241. 10.1073/pnas.0913053107

Wilmer JB, Germine LT, & Nakayama K (2014). Face recognition: a model specific ability. *Frontiers in Human Neuroscience*, 8 10.3389/fnhum.2014.00769

Wilson EB (1927). Probable Inference, the Law of Succession, and Statistical Inference. *Journal of the American Statistical Association*, 22(158), 209–212. 10.1080/01621459.1927.10502953

Winblad B, Palmer K, Kivipelto M, Jelic V, Fratiglioni L, Wahlund L-O, ... Petersen RC (2004). Mild cognitive impairment--beyond controversies, towards a consensus: report of the International Working Group on Mild Cognitive Impairment. *Journal of Internal Medicine*, 256(3), 240–246. 10.1111/j.1365-2796.2004.01380.x [PubMed: 15324367]

Youden WJ (1950). Index for rating diagnostic tests. *Cancer*, 3(1), 32–35. 10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3 [PubMed: 15405679]

CFMQ Cutoff Score Performance

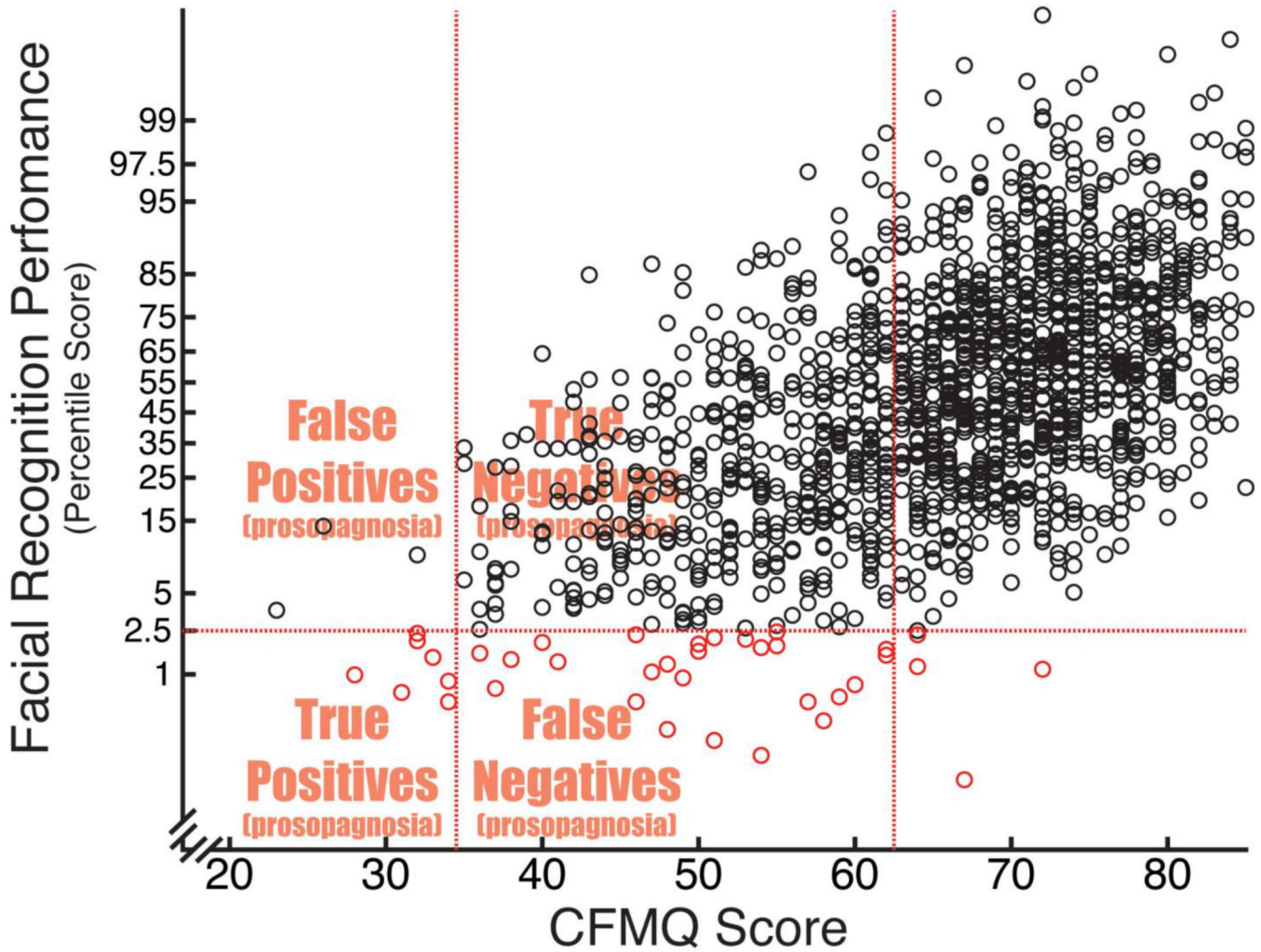


Figure 1. Facial recognition ability vs. CFMQ score. The performance level cutoff corresponding to prosopagnosia is indicated with a horizontal dotted line, such that those participants below the line were those with prosopagnosia. The two optimized CFMQ cutoffs are indicated with vertical dotted lines, such that those participants to the left of each given line were those obtaining to the given cutoff. The x-axis is in a z-transformed space to allow for ease of viewing the participants with prosopagnosia above or below the indicated cutoff scores. The sensitivity, specificity, positive predictive value, and CFMQ score prevalence rates already described can be intuited from this graph. Sensitivity for a given CFMQ cutoff score is the proportion of participants with prosopagnosia (i.e., of those below the horizontal line) that is captured by the given cutoff (i.e., those to the left of the given vertical line). Specificity is the proportion of participants without prosopagnosia (i.e., those above the horizontal line) that does not obtain to the CFMQ cutoff score (i.e., that falls to the right of the given vertical line). Positive predictive value is the proportion of those obtaining to the given CFMQ cutoff

score (i.e., of those falling to the left of the given vertical line) who actually have prosopagnosia (i.e., who fall below the horizontal line).

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

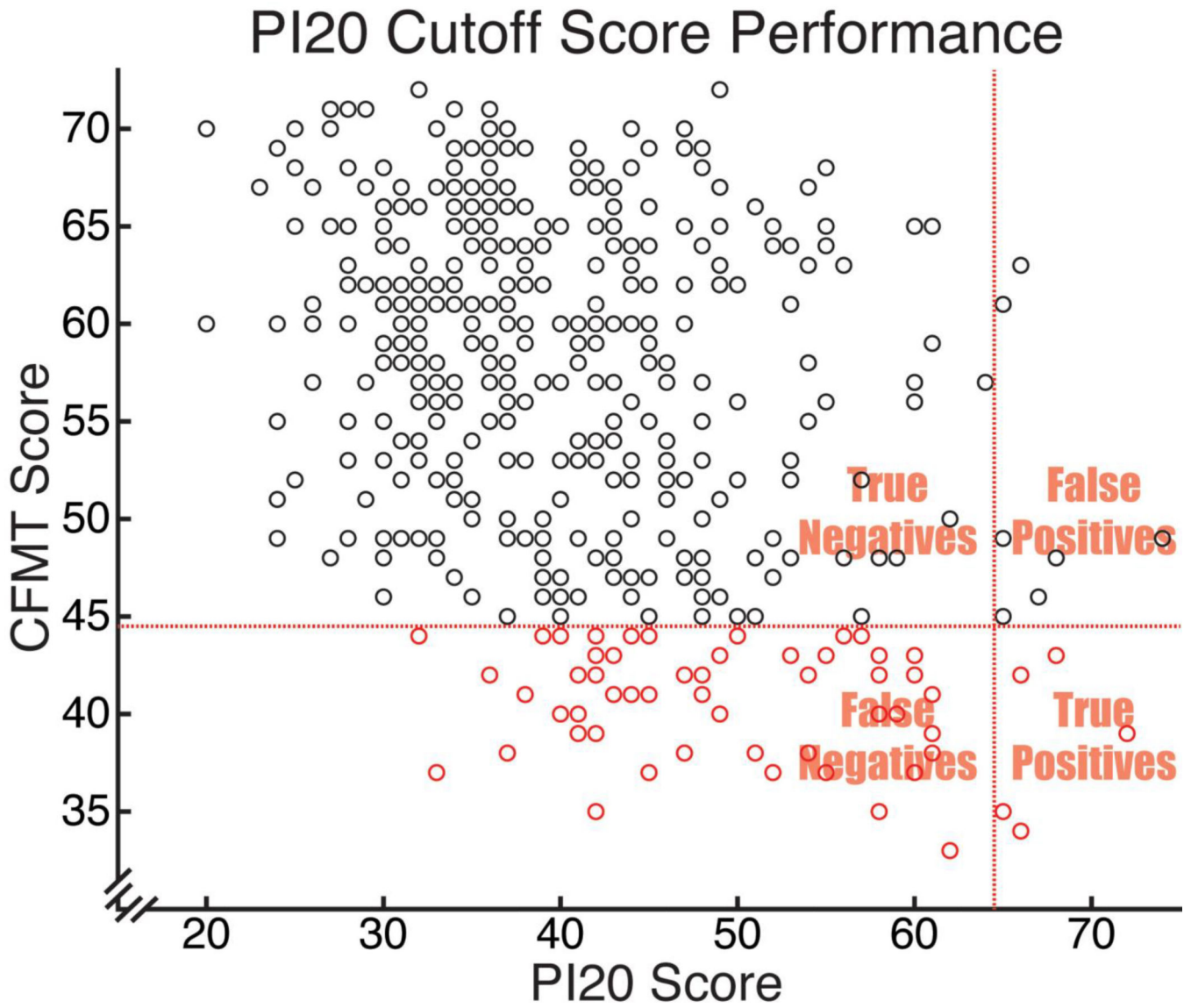


Figure 2. CFMT Score vs. PI20 score. The CFMT score corresponding to prosopagnosia is indicated with a horizontal dotted line, such that those participants below the line were those with CFMT-defined prosopagnosia. The proposed PI20 cutoff is indicated with a vertical dotted line, such that those participants to the right of the line were those obtaining to the cutoff. The sensitivity, specificity, positive predictive value, and CFMQ score prevalence rates already described can be intuited from this graph. Sensitivity for the PI20 cutoff score is the proportion of participants with CFMT-defined prosopagnosia (i.e., of those below the horizontal line) that is captured by the cutoff (i.e., those to the right of the vertical line). Specificity is the proportion of participants without prosopagnosia (i.e., those above the horizontal line) that does not obtain to the PI20 cutoff score (i.e., that falls to the left of the horizontal line). Positive predictive value is the proportion of those obtaining to the PI20

cutoff score (i.e., of those falling to the right of the vertical line) who actually have CFMT-defined prosopagnosia (i.e., who fall below the horizontal line).

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 1

ROC analyses of CFMQ score screening for prosopagnosia

| ROC Analyses - Screening for Prosopagnosia | | | | | | | |
|---|---------------------|--------------------|-----------------|--------------------|-----------------|------------|-----------------|
| Optimization type | Cutoff score | Sensitivity | 95% CI | Specificity | 95% CI | PPV | 95% CI |
| Youden's J | 62 | 89.19% | 75.29% – 95.71% | 68.06% | 65.64% – 70.39% | 6.52% | 4.68% – 9.02% |
| Standard | 34 | 18.92% | 9.48% – 34.21% | 99.80% | 99.41% – 99.93% | 70% | 39.68% – 89.22% |
| Maximum PPV | 34 | 18.92% | 9.48% – 34.21% | 99.80% | 99.41% – 99.93% | 70% | 39.68% – 89.22% |

Note. PPV = Positive Predictive Value

Youden's J returns the operating point where sensitivity rate + specificity rate -1 is maximized. Thus, with this optimization, the cost of a loss in sensitivity rate is equal to the cost of a loss in specificity rate.

The standard optimization is similar to Youden's J, except that costs are adjusted to reflect that those without prosopagnosia greatly outnumber those who with prosopagnosia. The cost of loss of specificity, therefore, greatly outweighs the cost of loss of sensitivity. In practice, this second criterion attempts a balance between sensitivity and positive predictive value, and so returns the main CFMQ cutoff of interest.

Table 2

ROC analyses of CFMQ score screening for the bottom 50th percentile of facial recognition ability.

| ROC Analyses - Screening for 50%-ile | | | | | | | |
|---|---------------------|--------------------|-----------------|--------------------|-----------------|------------|-----------------|
| Optimization type | Cutoff score | Sensitivity | 95% CI | Specificity | 95% CI | PPV | 95% CI |
| Youden's J | 66 | 65.35% | 61.89% – 68.65% | 71.41% | 68.09% – 74.51% | 69.57% | 66.09% – 72.83% |
| Standard | 66 | 65.35% | 61.89% – 68.65% | 71.41% | 68.09% – 74.51% | 69.57% | 66.09% – 72.83% |
| Maximum PPV | 39 | 4.35% | 3.11% – 6.04% | 100% | 99.5% – 100% | 100% | 89.57% – 100% |

Note. PPV = Positive Predictive Value

Youden's J returns the operating point where sensitivity rate + specificity rate -1 is maximized. Thus, with this optimization, the cost of a loss in sensitivity rate is equal to the cost of a loss in specificity rate.

The standard optimization is similar to Youden's J, except that costs are adjusted to reflect that those without prosopagnosia greatly outnumber those who with prosopagnosia. The cost of loss of specificity, therefore, greatly outweighs the cost of loss of sensitivity. In practice, this second criterion attempts a balance between sensitivity and positive predictive value, and so returns the main CFMQ cutoff of interest.

Table 3

ROC analyses of CFMQ score screening for the top 2.5th percentile of facial recognition ability.

ROC Analyses - Screening for top 2.5%-ile

| Optimization type | Cutoff score | Sensitivity | 95% CI | Specificity | 95% CI | PPV | 95% CI |
|-------------------|--------------|-------------|-----------------|-------------|-----------------|--------|-----------------|
| Youden's J | 72 | 76.92% | 61.66% – 87.35% | 67.68% | 65.25% – 70.02% | 5.91% | 4.17% – 8.31% |
| Standard | 85 | 7.69% | 2.65% – 20.32% | 99.73% | 99.31% – 99.89% | 42.86% | 15.82% – 74.95% |
| Maximum PPV | 85 | 7.69% | 2.65% – 20.32% | 99.73% | 99.31% – 99.89% | 42.86% | 15.82% – 74.95% |

Note. PPV = Positive Predictive Value

Youden's J returns the operating point where sensitivity rate + specificity rate -1 is maximized. Thus, with this optimization, the cost of a loss in sensitivity rate is equal to the cost of a loss in specificity rate.

The standard optimization is similar to Youden's J, except that costs are adjusted to reflect that those without prosopagnosia greatly outnumber those who with prosopagnosia. The cost of loss of specificity, therefore, greatly outweighs the cost of loss of sensitivity. In practice, this second criterion attempts a balance between sensitivity and positive predictive value, and so returns the main CFMQ cutoff of interest.

Table 4

ROC Analyses of PI20 score screening for CFMT-defined prosopagnosia.

| ROC Analyses - PI20 Screening for Prosopagnosia | | | | | | |
|---|-------------|----------------|-------------|-----------------|--------|-----------------|
| Cutoff score | Sensitivity | 95% CI | Specificity | 95% CI | PPV | 95% CI |
| 65 | 8.47% | 3.67% – 18.35% | 98.09% | 96.11% – 99.07% | 41.67% | 19.33% – 68.05% |

Note. PPV = Positive Predictive Value

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript