ELSEVIER

Neuropsychologia 44 (2006) 576–585

NEUROPSYCHOLOGIA

www.elsevier.com/locate/neuropsychologia

The Cambridge Face Memory Test: Results for neurologically intact individuals and an investigation of its validity using inverted face stimuli and prosopagnosic participants

Brad Duchaine[a,*], Ken Nakayama[b]

[a] *Institute of Cognitive Neuroscience, University College London, 17 Queen Square, London WC1N 3AR, UK*
[b] *Department of Psychology, Harvard University, Vision Science Laboratory, 33 Kirkland Street, 7th Floor, Cambridge, MA 02138, USA*

Received 21 October 2004; accepted 1 July 2005
Available online 19 September 2005

Abstract

The two standardized tests of face recognition that are widely used suffer from serious shortcomings [Duchaine, B. & Weidenfeld, A. (2003). An evaluation of two commonly used tests of unfamiliar face recognition. *Neuropsychologia*, *41*, 713–720; Duchaine, B. & Nakayama, K. (2004). Developmental prosopagnosia and the Benton Facial Recognition Test. *Neurology*, *62*, 1219–1220]. Images in the Warrington Recognition Memory for Faces test include substantial non-facial information, and the simultaneous presentation of faces in the Benton Facial Recognition Test allows feature matching. Here, we present results from a new test, the Cambridge Face Memory Test, which builds on the strengths of the previous tests. In the test, participants are introduced to six target faces, and then they are tested with forced choice items consisting of three faces, one of which is a target. For each target face, three test items contain views identical to those studied in the introduction, five present novel views, and four present novel views with noise. There are a total of 72 items, and 50 controls averaged 58. To determine whether the test requires the special mechanisms used to recognize upright faces, we conducted two experiments. We predicted that controls would perform much more poorly when the face images are inverted, and as predicted, inverted performance was much worse with a mean of 42. Next we assessed whether eight prosopagnosics would perform poorly on the upright version. The prosopagnosic mean was 37, and six prosopagnosics scored outside the normal range. In contrast, the Warrington test and the Benton test failed to classify a majority of the prosopagnosics as impaired. These results indicate that the new test effectively assesses face recognition across a wide range of abilities.

Keywords: Face recognition; Prosopagnosia; Neuropsychology

Face recognition is one of the most intensively studied aspects of human cognition involving scientists from a wide range of related fields. Because of this, it is important that researchers have access to well-designed standardized tests of face recognition. Such tests would provide a means to compare the performance of participants in different laboratories. In addition, they would provide researchers with a ready-made tool so they would not need to create a test and develop norms. Lastly, neuropsychologists and neurologists require additional tests that can contribute to classifying individuals who have face recognition impairments.

Currently, there are two commonly used standardized tests of face recognition, the Benton Facial Recognition Test (BFRT) (Benton et al., 1983) and the Recognition Memory Test for Faces (RMF) (Warrington, 1984). They are widely used with normal participants and neuropsychological participants, but both suffer from serious shortcomings that make them potentially misleading tests of face recognition (Duchaine & Weidenfeld, 2003; Duchaine & Nakayama, 2004). In the BFRT, participants are simultaneously presented with a target face and six test faces. Participants must choose the three test faces that match the target face. Because the target face and the test faces are presented simultaneously, participants can use a feature matching strategy. An experiment with normal participants showed that a substan-

* Corresponding author. Tel.: +44 20 7679 1005; fax: +44 20 7916 8517.
E-mail address: brad.duchaine@gmail.com (B. Duchaine).

0028-3932/$ – see front matter
doi:10.1016/j.neuropsychologia.2005.07.001

tial proportion were able to score in the normal range on a modified version of the BFRT in which the face was occluded so that only the eyebrows and the hairline were presented (Duchaine & Weidenfeld, 2003). Furthermore, a number of prosopagnosics have been shown to score normally on the BFRT (Duchaine & Nakayama, 2004; Newcombe, 1979; Nunn, Postma, & Pearson, 2001) including some who in addition to their deficits with face memory tests also have deficits with face perception tests (Duchaine, Yovel, Butterworth, & Nakayama, in press). These converging results make it clear that the possibility of feature matching makes the BFRT a poor measure of face recognition ability. The RMF is limited by the nature of the images used in the test. During the inspection phase of the RMF, participants are presented with 50 target images for 3 s each. Following this, participants are presented with 50 forced-choice items consisting of a target face and a distracter face. Images contain substantial non-face information that can be used to discriminate between target and distracter images. These include hair, clothing, posture, emotional expressions, and image imperfections. When normal participants were presented with a modified version of the RMF that occluded only the facial information, many participants were able to score in the normal range (Duchaine & Weidenfeld, 2003). In addition, some prosopagnosics have scored normally on the test (Duchaine, 2000; Nunn et al., 2001), including one whose performance fell nearly to chance when the non-facial information was occluded (Nunn et al., 2001).

Because of the problems with these tests, scientists and practitioners are left without an effective standardized test of face recognition. To address this deficiency, we have created a test of face memory that maximizes the strengths of the BFRT and the RMF. Like the BFRT, our test will have sections with different levels of difficulty and test items with novel views of target faces. Like the RMF, our test will involve a memory paradigm with multiple faces, which will make simultaneous feature matching impossible. However, unlike in the BFRT and the RMF, the face stimuli will be limited strictly to facial information (e.g. no clothing, no hair line). Our test is also akin to everyday face recognition in that participants will have an opportunity to gradually acquire knowledge of target faces from a wide range of views. They will see each target face 17 times throughout the entire test. Although they do not receive feedback after seeing each test item, repetitive viewing should provide the opportunity to develop better representations after viewing images in test items.

Because the test will measure face memory, performance on the test will depend on both perceptual mechanisms and memory. As a result, the test will not provide a means to measure the perceptual processes uncontaminated by memory processes, and our laboratory is currently developing a test of face perception. However, face memory, not face perception, is the ability that determines our success in identity recognition in everyday life, and so it is especially important to measure it. When tests of face perception are developed, the combination of tests of face perception and face memory should provide a means to assess the contributions of perceptual processes and memory to variability in face recognition ability.

We call our test the Cambridge Face Memory Test (CFMT). The test will be available free of charge when used for research purposes. In the following sections, we will describe the results of testing with neurologically normal participants. Following this, we discuss experiments aimed at evaluating the validity of the test by testing neurologically normal participants with inverted face stimuli and examining whether individuals with face recognition impairments score poorly with the CFMT.

1. Method

1.1. Stimuli

The faces are those of men in their 20s and early 30s, and each individual was photographed in the same range of poses and lighting conditions. Men's faces were used, because men and women perform equivalently with men's faces whereas women show an advantage with women's faces (Lewin & Herlitz, 2002; McKelvie, Standing, St. Jean, & Law, 1993). All faces were cropped so that no hair was visible and facial blemishes were removed. The men posed with neutral expressions.

Six individuals were chosen as target faces. We used six targets, because it is a challenging yet manageable number of faces for normal subjects to encode after brief exposures. Twelve images of each target face were selected, and the same poses and lighting conditions were used for each target face. Test items consisted of a target face along with two distracter faces with the same pose and lighting. Forty-six individuals were used as distracter faces. Many of the distracter individuals were presented repeatedly, and this repetition meant that participants could not simply make a familiar/unfamiliar discrimination on test items with repeated distracters.

1.2. Procedure

The test consists of four stages (practice, introduction/same images, novel images, and novel images with noise). Completing the test takes between 10–15 min.

1.2.1. Practice

The practice stage familiarizes participants with the procedure used in the introduction/same images stage by presenting cartoon faces in the same fashion that the target faces will be presented. After instructing the participant to memorize the following faces, three study images of Bart Simpson are presented for three seconds each: a left 1/3 profile, a frontal view, and a right 1/3 profile. Then a test item consisting of one of the study views of Bart along with two other cartoon faces is presented. Participants are instructed to press the key corresponding to the number below the target face (1, 2, or

3). Two more test items follow, and each consists of one of the study faces along with two distracter faces.

1.2.2. Introduction/same images

Participants are instructed that they will now begin the test, and they are introduced to the first target face in the same way that they were introduced to Bart Simpson during the practice stage. Three study images are presented for three seconds each. The images are a left 1/3 profile, a frontal view, and a right 1/3 profile. Fig. 1 Panel A shows an individual in the three views. Three test items are then presented and participants are instructed to pick out the individual whom they were just shown (see Fig. 1, panel B). Each test item includes an item identical to a study item. Because the study and test images are the same, the participants could respond correctly by recognizing the image rather than face (Hay & Young, 1982). There are six target faces, and this procedure is repeated for the five remaining target faces. Target faces are never used as distracter faces. Feedback is not provided during the test.

1.2.3. Novel images

Immediately before this stage, participants are presented with a single review image that has a frontal shot of each target face. They are given 20 s to review this image. Following the review image, participants are presented with 30 forced-choice test items (6 target faces × 5 presentations) in a fixed, random order. Each test item contains three faces, one of which is a target face. Participants are instructed that each test item will contain one of the six target faces and told to respond with the key corresponding to the number under the target face. All are novel images in which the lighting, pose, or both vary (see Fig. 1, panel C). Appendix A displays examples of the poses and lighting used for target items in the novel images and novel images with noise sections, and the lighting and poses used were the same for all six target

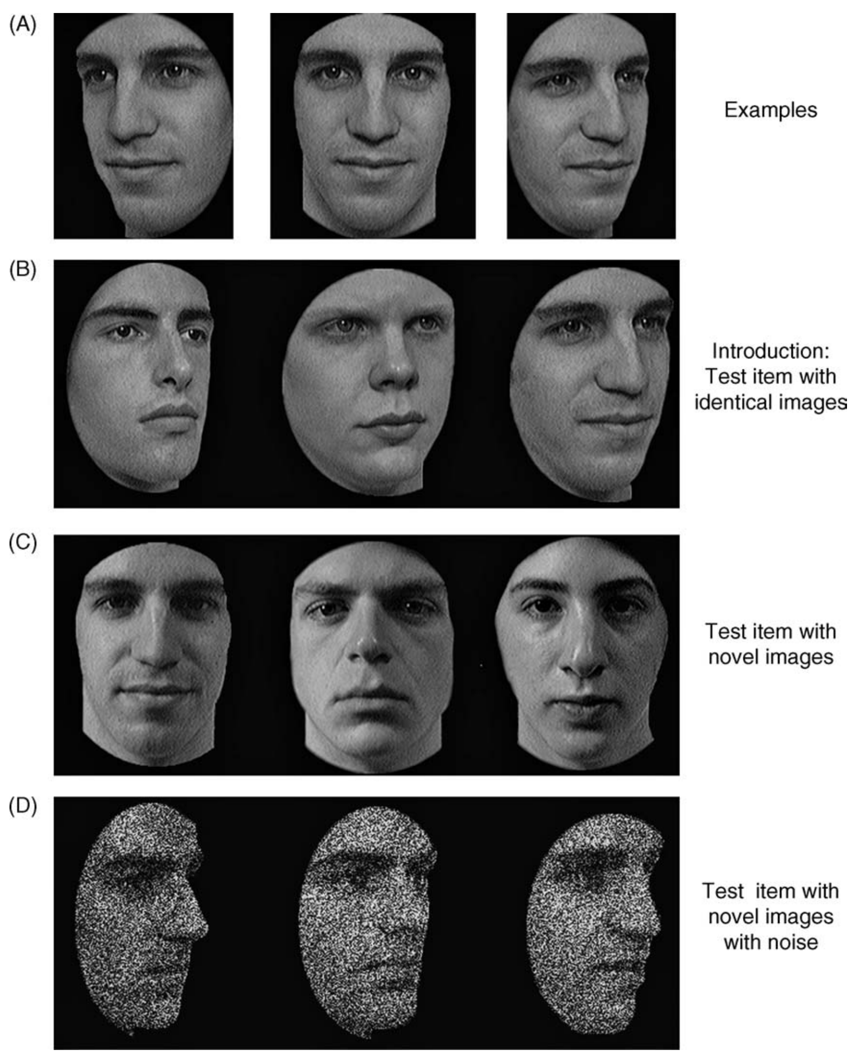


Fig. 1. Examples of stimuli similar to test stimuli. None of these items was used in the test. In the test, test faces are numbered 1, 2, and 3 from left to right, but we omitted this to save space. Panel A shows study views of a target face. Study views are presented for three seconds each. Panel B displays a test item from the introduction. Face 3 is the same image as the rightmost study view in Panel A. Panel C shows an item from the novel image section (face 1 is the target). Panel D displays a test item from the novel images with noise section (face 3 is target).

faces. When participants are presented with test items in the introduction, they know which target face will be the correct answer. However, during this stage and the final stage, the correct answer for an item can be any of the target faces, and so the items are much more difficult.

1.2.4. Novel images with noise

Participants are presented with the review image again for twenty seconds. Following this, 24 test items (6 target faces × 4 presentations) are presented in a fixed, random order. These items consist of novel images, and different levels of Gaussian noise were added to the face images (see Fig. 1, panel D). Levels of noise for the faces in a test item are identical. Noise was added to the faces for two reasons. At the beginning of the final section, participants will have seen each target image 13 times so the noise was added to keep performance away from ceiling. Second, studies with normal participants indicate that noise forces increased reliance on the special mechanisms that face recognition normally depends on (McKone, Martini, & Nakayama, 2001).

2. Results

In this section, we discuss the results from our normal participants. Following this, we discuss two conditions aimed at determining whether our test effectively measures face recognition. We do this by first administering the test when all faces are presented inverted, and by giving the test in its upright version to prosopagnosic individuals.

2.1. Performance of neurologically intact participants with upright faces

Our participants were 50 college age individuals ranging in age from 18–26 with a mean age of 20.2 (S.D. = 1.8). Twenty-nine of these participants were female while 21 were male. They were paid for their participation.

Fig. 2 displays the average cumulative performance along with the standard deviation (see Appendix C for means and standard deviations). The figure is divided into the three sections of the test, and the intersections of the dashed lines indicate perfect performance. Because participants knew ahead of time which target face would be present in each item in the introduction, we expected them to perform very well and in fact, they made few mistakes. However, they made many more errors in the second section as is evidenced by the decreasing slope in Fig. 2. For these items, participants did not know which target face would be presented and all of the images were novel views. Fig. 3 plots individual scores at the end of each section, and this figure makes it clear that there were a wide range of scores in the second section. In the final section, participants were presented with novel images degraded by noise. The slope of the line in Fig. 2 was even flatter for this section, so it appears that the noise made these items even more difficult.

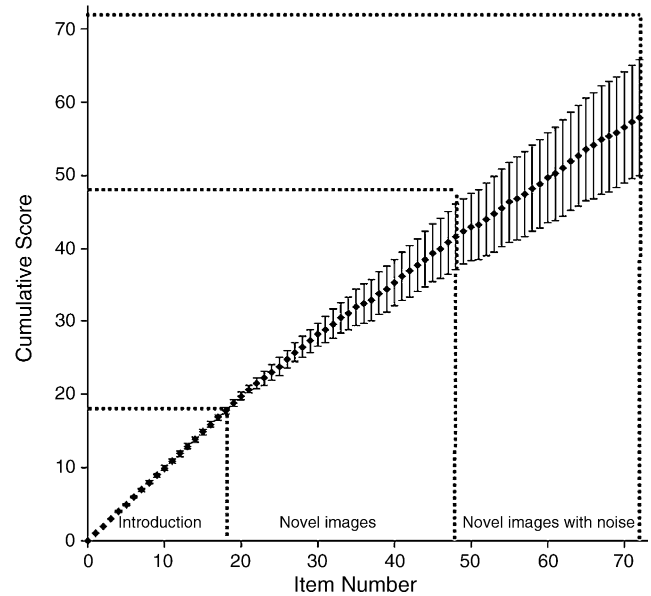


Fig. 2. Average cumulative performance for the 50 controls on the upright version. Points display the average cumulative score for controls at each test item. Error bars display the standard deviation for the cumulative scores. Dashed lines divide the figure into the three different sections. Deviation from perfect responding at the end of each section can be gauged by viewing the distance between the intersection of the dashed lines and the cumulative score for the final item in the section.

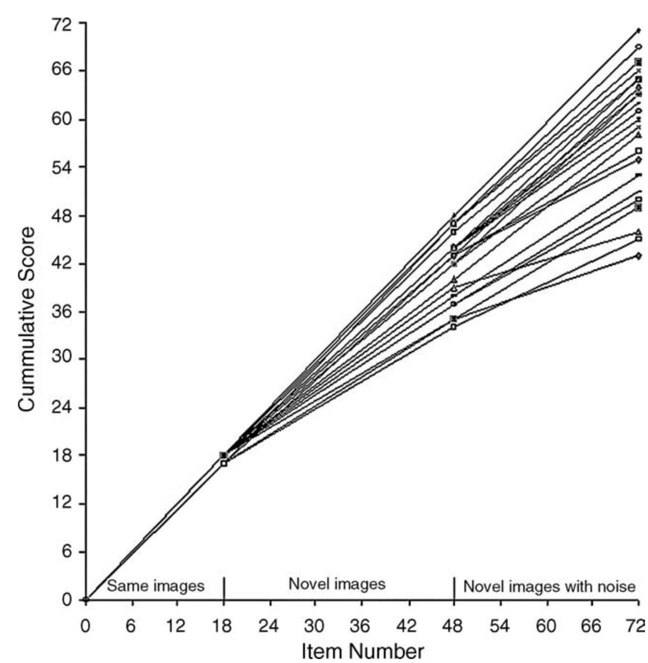


Fig. 3. Individual cumulative scores for controls. We included the scores of every other control so the figure is not overly cluttered. Scores at the end of each section were computed. The largely similar relative position of individual controls from section to section indicates that performance in different sections depended on the same abilities.

The average total score out of 72 for the controls was 57.9 (S.D. = 7.9), which converts to 80.4% (S.D. = 11.0). Total scores ranged from 43 to 72. The male participants averaged 56.5 (S.D. = 7.3) and the female participants averaged 58.9 (S.D. = 8.3). This difference was not significant.

2.1.1. *Consistency of scores from section to section*

To check if the different sections of the test relied on the same abilities and representations, we looked at the participants' consistency from section to section. We computed correlation coefficients using each participant's score for each test section. Because participants performed so well in the introduction/same images section, there was little variability, and so the correlation coefficients were relatively low for the same images-novel images comparison ($r = 0.27$, $p = 0.06$) and for the same images-novel images with noise comparison ($r = 0.35$, $p = 0.01$).

In contrast, the scores for the novel images section and the novel images with noise section were quite consistent, and the correlation coefficient for this comparison was 0.74 ($p = 0.001$). In Fig. 3, the consistency of the participants is apparent in that their rank at the end of the novel images with noise is quite similar to their rank at the end of the novel images section.

2.1.2. *Item analysis*

Next we conducted an item analysis to determine whether the test contained items that did not effectively discriminate between good performers and poor performers. To do this, we computed a correlation coefficient involving each participant's total score and their performance on each item (correct or incorrect). Because performance was nearly perfect or perfect for the same image items in the introduction, the correlations for these items were either not interesting or we were unable to compute them. However, there was variability for all of the items in the other two sections except for one so we were able to compute coefficients for 53 items. The average correlation for the novel image items was 0.35 (S.D. = 0.13) and was 0.35 (S.D. = 0.13) for the novel image with noise items. (Note that these equivalent values are not a typo.) All but one of the correlations was positive (it was −0.004) so 52 of 54 items contributed to the test's sensitivity.

2.1.3. *Analysis by face*

To analyze the difficulty of the six target faces, we computed the percent correct for each target face. These percentages were 77, 69, 80, 81, 88, and 88. An inspection of the two faces producing the two highest percentages leads us to believe that it was because these faces were the most distinctive, but the order may have contributed as well.

2.2. *Performance of neurologically intact participants when the faces are inverted*

The results in the previous section show that the test produces a nice range of scores that are consistent from section to section and that do not suffer from ceiling or floor effects. However, these scores do not demonstrate that the test actually assesses face recognition abilities. It could simply activate general-purpose visual recognition mechanisms.

To address this issue, we will first assess the effect of inverting all of the faces in the test. Typically, inversion decreases percent correct in face recognition experiments by 15–25% (Diamond & Carey, 1986; Scapinello & Yarmey, 1970; Yin, 1969) whereas inversion of many other objects classes affects percent correct far less (Diamond & Carey, 1986; Scapinello & Yarmey, 1970; Yin, 1969). This disproportionate effect has been used to argue that upright faces are processed in a manner that is qualitatively distinct from the processing applied to other objects (Yin, 1969). Further work has shown that the specialized processing which upright faces receive involves holistic or configural representation (Freire & Lee, 2000; McKone et al., 2001; Tanaka & Farah, 1993; Tanaka & Sengco, 1997; Young, Hellawell, & Hay, 1987) whereas most other types of objects, including inverted faces, are represented more as a collection of parts (Biederman, 1987). This distinction has also been supported by a double dissociation between upright face processing and inverted face processing (Farah, Wilson, Drain, & Tanaka, 1995; Moscovitch, Winocur, & Behrmann, 1997).

If the test relies on the mechanisms normally used for upright face recognition, then we should find a large decrement in performance when the faces are inverted. However, if we find the effect is not comparable to past face inversion effects, this will indicate that performance did not depend on the special mechanisms. We examined this prediction by testing 20 new participants drawn from the same population as the participants used for the upright version.

Fig. 4 plots the cumulative scores for participants in the upright and inverted conditions. Even by the end of the introduction/same images section, inverted scores are significantly worse than upright scores ($t(68) = 5.0$, $p < 0.0001$). This difference suggests that normal face recognition mechanisms were contributing to upright performance even in this very easy section. As Fig. 4 shows, the difference between upright and inverted scores became much more pronounced when novel images were presented, and this difference is highly significant ($t(68) = 7.9$, $p < 0.0001$). For the section with novel images with noise, the inverted average was only 10% above chance (43%), and the difference between upright and inverted was highly significant ($t(68) = 6.0$, $p < 0.0001$).

The average inverted score for the entire test was 42.1 (S.D. = 4.7) or 58.4% correct (S.D. = 6.5). The upright mean was 80.4% so inversion lowered performance by 22%, an effect comparable to previous inversion effects. This difference was highly significant ($t(68) = 8.4$, $p < 0.0001$). The inverted mean is two standard deviations below the upright mean. There was little overlap between the scores in the two conditions with inverted scores ranging from 33–50 whereas the upright scores ranged from 43–71.

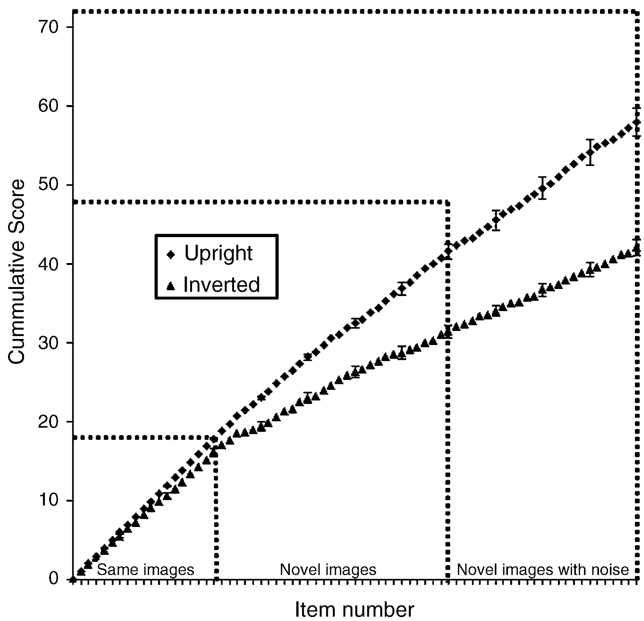


Fig. 4. Comparison of average upright cumulative scores to average inverted cumulative scores. Error bars display one standard error above and below the mean (upright $n = 50$; inverted $n = 20$). The figure is divided into the three sections, and deviation from perfect responding at the end of each section can be gauged by viewing the distance between the intersection of the dashed lines and the cumulative score for the final item in the section.

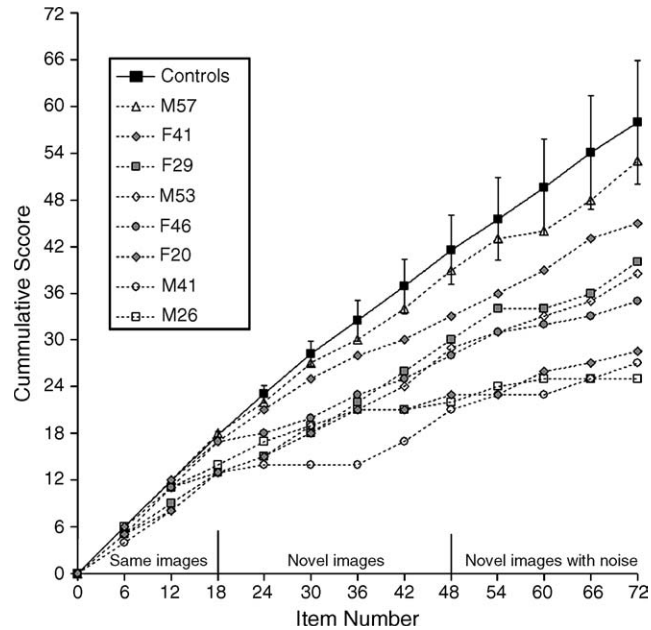


Fig. 5. Comparison of average upright cumulative scores for eight prosopagnosic participants and the control mean. The error bars for the controls display one standard deviation above and below the control mean. Cumulative score after every six items is displayed.

2.3. *Performance by individuals with face recognition impairments with upright faces*

The difference in performance for the upright and inverted versions of the test indicates that the test activates the special processes used to recognize upright faces. Next we address this same issue by assessing the performance of individuals with face recognition impairments on the normal, upright version of the test. Because the test appears to rely on the special processes used with upright faces, we expect that the individuals with impairments to these mechanisms will perform poorly on the test.

In addition, their scores will demonstrate whether the test can contribute to assessments of individuals who may have face recognition impairments. We will compare the prosopagnosics' scores on the test to their scores on the BFRT (Benton et al., 1983) and the RMF (Warrington, 1984). If scores on the CFMT better classify the prosopagnosics than the BFRT or the RMF, it will suggest that the CFMT could be a useful measure for neurologists and neuropsychologists.

The eight participants in this section contacted our laboratory, because they complained of significant problems in everyday face recognition. We will refer to these individuals with labels indicating their sex (F or M) and their age at the time of testing. Two out of this group suffered brain damage as young children (M26 and M41). The rest report no head trauma and so appear to be congenital prosopagnosics. Four of these individuals have been studied in other papers on prosopagnosia (M26—Kosslyn et al., 1995; Hadjikhani & de Gelder, 2002; M53—Duchaine, Dingle, Butterworth, & Nakayama, 2004; Harris, Duchaine, & Nakayama, 2005; M57—Duchaine, 2000; F46—Harris et al., 2005). To assess whether these individuals did, in fact, have face recognition impairments, we tested them with two memory tests with unfamiliar faces (Duchaine et al., 2003; Duchaine & Nakayama, 2005) and a famous face test (Duchaine & Nakayama, 2005). The *z* scores for the participants are presented in Appendix B along with their *z* scores on the CFMT, and this table shows that their performance was clearly impaired.

Fig. 5 shows the upright control average and the scores for each prosopagnosic participant. Whereas the controls scored nearly perfectly in the introduction/same images section, many of the prosopagnosics made errors and the prosopagnosic average was significantly lower than the control average ($t(56) = 9.2$, $p < 0.0001$). Like the inverted average, the prosopagnosic average in the section involving novel images plummeted relative to the upright control average ($t(56) = 6.2$, $p < 0.0001$). By the end of the novel images section, all but two of the prosopagnosics were more than two standard deviations below the mean. For the novel images with noise added, the prosopagnosic average was just above chance (34.9%), and this difference was quite significant ($t(56) = 5.3$, $p < 0.0001$). The overall prosopagnosic mean was 36.5 (S.D. = 9.7) or 50.7% (S.D. = 13.4), which is 2.7 standard deviations below the control mean ($t(56) = 6.9$, $p < 0.0001$). Scores for the prosopagnosic participant ranged from 25 to 53.

2.3.1. *Performance with different views*

Fig. 6 displays percent correct on test items with front views and those with side views in the novel views and novel

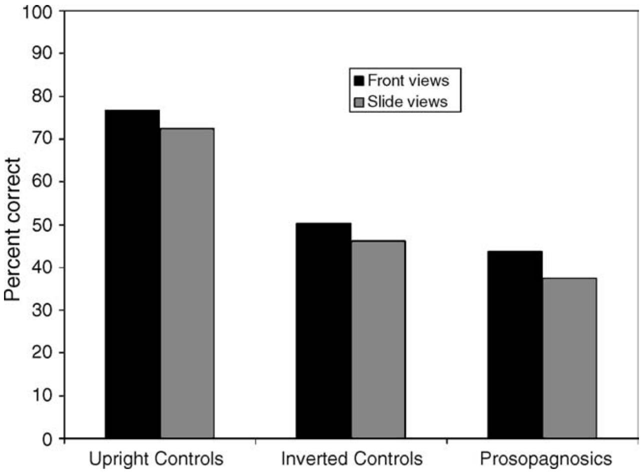


Fig. 6. Performance on items involving front views and side views from the novel items and novel items with noise for upright controls, inverted controls, and prosopagnosics.

views with noise sections. Among these items, there were 24 front views and 30 side views (18 right and 12 left). As is apparent in the figure, percent correct was slightly higher for the front views. The difference between the different views was nearly identical in our three participant groups.

2.3.2. Performance on different face tests

Fig. 7 displays the scores for each prosopagnosic participant on the CFMT, BFRT, and RMF as standard deviations from the normal control mean. Control means and standard deviations for the BFRT and the RMF were obtained from the manuals (Benton et al., 1983; Warrington, 1984). Neuropsychologists often classify scores two standard deviations below the mean as impaired, and Fig. 7 shows that for the CFMT scores for six of the eight prosopagnosic participants

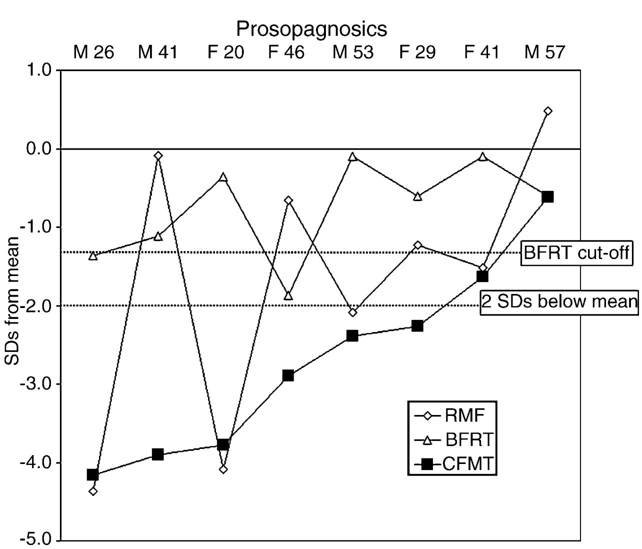


Fig. 7. Comparison of performance on the CFMT, BFRT, and RMF for the prosopagnosic participants. Scores are displayed as standard deviations below the control mean.

were below this cut-off. F41's score on the CFMT was 1.6 standard deviations below the control mean.

If we next consider the scores on the RMF, Fig. 7 shows that scores on the CFMT and RMF for a number of prosopagnosics (M26[1], F20, M53, and F41) were very similar. However, only three of the eight prosopagnosics scored more than 2 standard deviations below the mean. Especially problematic are the scores of participants such as M41 and F46. They did very poorly on the CFMT and other tests of face recognition, yet scored normally on the RMF. Their normal performance appeared to rely on non-facial information. M41 commented that he was doing photograph recognition rather than face recognition, and F46 remarked that she recognized the clothing and haircuts on many of the items. Other prosopagnosics were also able to score well on the RMF. M53's RMF score and F29's RMF score were near the control mean yet they were clearly impaired on other face memory tests.

The BFRT suggests classifying scores of 40 and below as impaired, so we have placed a dashed line in Fig. 7 at the standard deviation corresponding to a score of 40.5. The mean score for the prosopagnosics was 42.4 (S.D. = 2.5), and six of the eight prosopagnosics had scores classified as normal. None had scores more than 2 standard deviations below the control mean. The BFRT has three different types of items: matching of identical front-views, matching of front-views with three-quarter views, and matching front views under different lighting. All of the prosopagnosics scored perfectly on items requiring matching of identical front-views. For matching different views, they average 19.4 out of 24 while their mean was 17 out of 24 for matching under different lighting.

To compare how well each test discriminates between individuals with normal face recognition and those with impaired face recognition, we have computed d' for each test. d' is a bias-free measure of discrimination (Green and Swets, 1966). The CFMT classified all 50 normal participants correctly (hits) so its specificity is 100%. It correctly classified six of the eight prosopagnosics (correct rejections) so its sensitivity is 75%. Because d' cannot be computed when there are zero hits or false alarms, we changed the false alarm rate to one out of 50. This produces an d' for the CFMT of 2.7. Because we did not run controls on the RMF and the BFRT, we will charitably assume that, like the CFMT, no controls would have been classified as impaired. On this assumption, the d' score for the BFRT is 1.4 and the d' score for the RMF is 1.7.

3. Discussion

We created a new test of face memory in hopes that it can supplement standardized tests of face recognition. The results discussed in the previous section are very encouraging. Fig. 8 displays performance on the three sections of the test for the three conditions. First consider the upright percent correct. Because these scores are far off of the floor and the ceiling, the

[1] M26's score for the RMF was taken from Hadjikhani & de Gelder (2002).

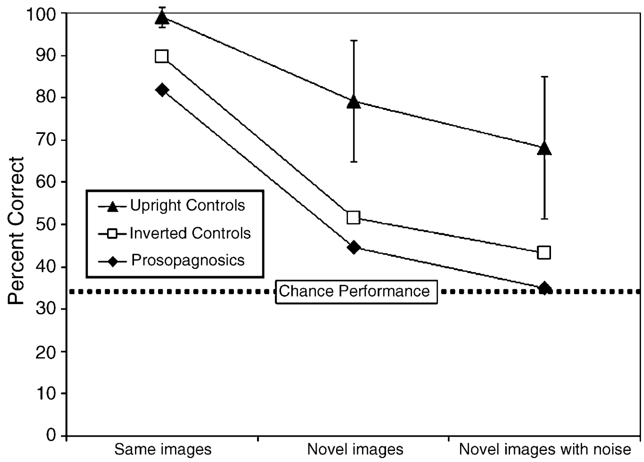


Fig. 8. Percent correct for each test section for the three conditions. The error bars show the standard deviation for each section for the upright condition. The dashed line indicates chance performance.

test can assess a wide range of abilities. Each of the top five possible scores (68–72) was achieved by only one participant so the test challenges even individuals with very good face recognition. Similarly, only five participants scored below 49 so it appears to discriminate in the low range of normal face recognition abilities as well.

The test produced similar scores for men and women. While the scores for the women were approximately two points higher than the scores for the men, this difference was not significant. A small sample of middle-aged participants also suggests that the test can be used with older participants as well. Nine middle-aged, college educated participants with an average age of 46.6 (S.D. = 7.7) produced a mean slightly higher (61.8) than our college age mean.

We predicted that inversion of the images would lead to a drop in performance if the upright version activates the special processes that contribute to upright face recognition. Fig. 8 makes it clear that inversion affected performance in every section. The difference between total scores for upright and inverted was 22%, and this drop is comparable to or greater than that seen in other recognition memory experiments comparing upright and inverted performance (Yin, 1969; Scapinello & Yarmey, 1970; Diamond & Carey, 1986).

We also investigated the validity of the test by testing eight prosopagnosic individuals with the upright version. Fig. 8 shows that the prosopagnosic mean on each section was slightly lower than the inverted means, and their overall mean was 2.7 standard deviations below the control mean. Six of eight participants were also outside the range of control scores. All of the normal participants scored better than 2 standard deviations below the mean, and all but two of the prosopagnosic participants were more than two standard deviations below the mean.

It is particularly interesting to examine the scores of prosopagnosic individuals who have shown normal performance with object recognition tests, because their normal or at least relatively normal object recognition mechanisms may provide an alternative means to achieve a good score on the test. One of the participants (F46) has performed normally on a number of tests of object discrimination (Duchaine & Nakayama, 2005)[2] while another (M53) has performed normally on every non-face test on which we have tested him (Duchaine et al., in press). Despite their good abilities with many categories of objects, F46 scored 2.9 standard deviations below the normal mean while M53 was 2.4 below. This suggests that the test forced reliance on the special processes which are impaired in these individuals.

Two of the prosopagnosic participants (F41 and M57), however, had scores within 2 standard deviations of the mean and within the normal range. F41's score of 45 is 1.6 standard deviations below the mean while M57's score of 53 was only slightly below the control mean. However, they had the best scores among the prosopagnosics on the two tests of face memory that we used to classify prosopagnosics (see Appendix B for the scores and Duchaine & Nakayama, 2005 for details on the tests). Given that we tested a number of prosopagnosics, a score like M54's, which places her in the bottom 5% of normal participants, is not particularly surprising. However, M57's score gives no hint of his impairment. The experimenter checked M57's score immediately after the test, and after seeing how good it was, asked M57 how he had done so well. M57 responded that he intentionally attempted to "lust" after the faces rather than simply memorize them. He has been in many faces tests and he was interested in how this would affect his performance, because he believes that he processes faces differently when he is attracted to them. His score suggests his encoding strategy may have worked, and other experiments indicate that more attractive faces are better remembered than unattractive faces (Cross, Cross, & Daly, 1971; Shepherd & Ellis, 1973). Given all of the different types of information that can be extracted from faces (emotionality, masculinity–femininity, attractiveness, etc.), our test, like all current tests with faces, presents information that mechanisms other than those used for face recognition can operate on. Because of these alternative routes to recognition, potential prosopagnosics should always be tested with a range of tests.

Our comparison of the CFMT, BFRT, and RMF showed that the CFMT classified 75% of the prosopagnosics correctly while only 25% were correctly classified by the BFRT and only 38% were correctly classified by the RMF. Because the CFMT and the RMF both test face memory, the disparity between these classifications is very problematic for the RMF. The BFRT is sometimes used as a test of face recognition, and the normal performance by the prosopagnosics demonstrates that it does not effectively classify individuals with face recognition impairments. However, the BFRT despite its name was designed as a test of face perception, and so normal performance on it by prosopagnosics along with deficits for face memory performance could be explained as a dissociation between intact face perception and impaired

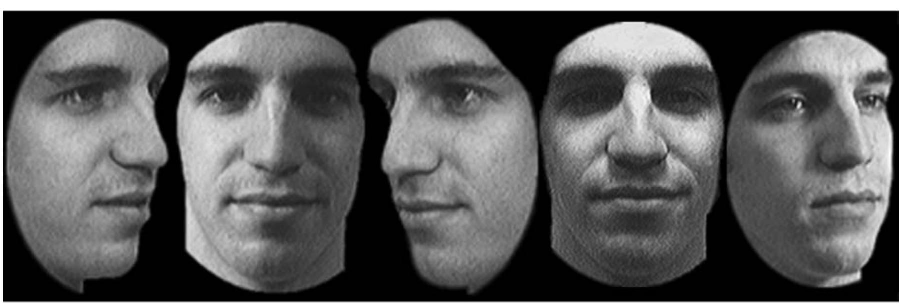
[2] Participant F46 was called F2 in Duchaine & Nakayama (2005).

face memory. This may account for some of the scores in the normal range, but some of the prosopagnosics tested appear to have impaired face perception. For example, M53 scored 45 on the BFRT, yet he shows no face inversion effect and is impaired on a range of face processing tasks (emotion, gender, attractiveness)(Duchaine et al., in press). Past results showing that normal participants can score normally when the majority of the face is occluded also indicate that normal scores do not demonstrate normal face perception (Duchaine & Weidenfeld, 2003). Thus, our results suggest that normal scores on the BFRT and the RMF should be interpreted cautiously.

In summary, these results indicate that the Cambridge Face Memory Test is a valid measure of face recognition ability that is sensitive to a wide range of abilities. The test is available free of charge for research purposes. Because it will be freely available, we hope to rapidly generate norms for different demographic groups.

Appendix A

Novel images



Novel images with noise

Appendix B. *z* scores for prosopagnosics on four tests of face memory

	CFMT	Famous faces	Old–new 1	Old–new 2
M26	−4.2	−11.6	<Chance	<Chance
M44	−3.9	−6.2	−11.8	na
F20	−3.8	−5.7	−4.1	−4.0
F46	−2.9	−5.3	−1.5	−6.7
M53	−2.4	−7.3	−3.4	−9.4
F29	−2.3	−4.7	−4.7	−3.9
F41	−1.6	−2.2	−1.4	−3.4
M57	−0.6	−8.5	−2.8	−1.6

Appendix C. Summary of scores for each condition and section

	Mean	S.D.	Range
Upright			
Introduction	17.82	0.44	16–18
Novel images	23.74	4.31	17–30
Novel images with noise	16.36	4.02	7–24
Total	57.92	7.91	43–71
Inverted			
Introduction	16.15	2.28	8–18
Novel images	15.50	2.76	11–19
Novel images with noise	10.40	2.96	4–15
Total	42.05	4.71	33–50
Prosopagnosics			
Introduction	15.31	2.21	12–18
Novel images	14.15	3.93	8–21
Novel images with noise	8.77	3.11	3–14
Total	38.23	7.52	25–53

References

Benton, A. L., Sivan, A. B., Hamsher, K. De S., Varney, N. R., & Spreen, O. (1983). *Contribution to Neuropsychological Assessment*. New York: Oxford University Press.

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review*, *94*, 115–147.

Cross, J. F., Cross, J., & Daly, J. (1971). Sex, age and race and beauty as factors in recognition of faces. *Perception and Psychophysics*, *10*, 393–396.

Diamond, R., & Carey, S. (1986). Why faces are and are not special: An effect of expertise. *Journal of Experimental Psychology*, *115*, 107–117.

Duchaine, B. (2000). Developmental prosopagnosia with normal configural processing. *Neuroreport*, *11*, 79–83.

Duchaine, B., Dingle, K., Butterworth, E., & Nakayama, K. (2004). Normal greeble learning in a severe case of developmental prosopagnosia. *Neuron*, *43*, 469–473.

Duchaine, B., & Nakayama, K. (2004). Developmental prosopagnosia and the Benton Facial Recognition Test. *Neurology*, *62*, 1219–1220.

Duchaine, B., & Nakayama, K. (2005). Dissociations of face and object recognition in developmental prosopagnosia. *Journal of Cognitive Neuroscience*, *17*, 249–261.

Duchaine, B., & Weidenfeld, A. (2003). An evaluation of two commonly used tests of unfamiliar face recognition. *Neuropsychologia*, *41*, 713–720.

Duchaine, B. C., Yovel, G., Butterworth, E. J., & Nakayama, K. (in press). Prosopagnosia as an impairment to face-specific mechanisms: Elimination of the alternative hypotheses in a developmental case. *Cognitive Neuropsychology*.

Farah, M. J., Wilson, K. D., Drain, H. M., & Tanaka, J. R. (1995). The inverted face inversion effect in prosopagnosia: Evidence for mandatory, face-specific perceptual mechanisms. *Vision Research*, *35*, 2089–2093.

Freire, A., Lee, K., & Symons, L. A. (2000). The face-inversion effect as a deficit in encoding of configural information: Direct evidence. *Perception*, *29*, 159–170.

Green, D. M., & Swets, J. A. (1966). *Signal Detection Theory*. New York: Wiley.

Hadjikhani, N., & de Gelder, B. (2002). Neural basis of prosopagnosia: an FMRI study. *Human Brain Mapping*, *16*, 176–182.

Harris, A., Duchaine, B., & Nakayama, K. (2005). Normal and abnormal face selectivity of the M170 response in developmental prosopagnosics. *Neuropsychologia*, *43*, 2125–2136.

Hay, D. C., & Young, A. W. (1982). The human face. In A. W. Ellis (Ed.), *Normality and Pathology in Cognitive Functions*. London: Academic Press.

Kosslyn, S. M., Hamilton, S. E., & Bernstein, J. H. (1995). The perception of curvature can be selectively disrupted in prosopagnosia. *Brain and Cognition*, *27*, 36–58.

Lewin, C., & Hertlitz, A. (2002). Sex differences in face recognition—women's faces make the difference. *Brain and Cognition*, *50*, 121–128.

McKelvie, S. J. (1987). Recognition memory for faces with and without spectacles. *Perceptual and Motor Skills*, *65*, 705–706.

McKelvie, S. J., Standing, L., St. Jean, D., & Law, J. (1993). Gender differences in recognition memory for faces and cars: Evidence for the interest hypothesis. *Bulletin of the Psychonomic Society*, *31*, 447–448.

McKone, E., Martini, P., & Nakayama, K. (2001). Categorical perception of face identity in noise isolates configural processing. *Journal of Experimental Psychology: Human Perception and Performance*, *27*, 373–399.

Moscovitch, M., Winocur, G., & Behrmann, M. (1997). What is special about face recognition? Nineteen experiments on a person with visual object agnosia and dyslexia but normal face recognition. *Journal of Cognitive Neuroscience*, *9*, 555–604.

Newcombe, F. (1979). The processing of visual information in prosopagnosia and acquired dyslexia: Functional versus physiological interpretation. In D. J. Oborne, M. M. Gruneberg, & J. R. Eiser (Eds.), *Research in Psychology and Medicine*. London: Academic Press.

Nunn, J. A., Postma, P., & Pearson, R. (2001). Developmental prosopagnosia: should it be taken at face value? *Neurocase*, *7*, 15–27.

Scapinello, K. F., & Yarmey, A. D. (1970). The role of familiarity and orientation in immediate and delayed recognition of pictorial stimuli. *Psychonomic Science*, *21*, 329–331.

Shepherd, J. W., & Ellis, H. D. (1973). The affect of attractiveness on recognition memory for faces. *American Journal of Psychology*, *86*, 627–633.

Tanaka, J. W., & Farah, M. J. (1993). Parts and wholes in face recognition. *Quarterly Journal of Experimental Psychology*, *46A*, 225–245.

Tanaka, J. W., & Sengco, J. (1997). Features and their configuration in face recognition. *Memory and Cognition*, *25*, 583–592.

Warrington, E. K. (1984). *Recognition Memory Test*. Windsor, England: NFER-Nelson.

Yin, R. K. (1969). Looking at upside-down faces. *Journal of Experimental Psychology*, *81*, 141–145.

Young, A. W., Hellawell, D., & Hay, D. (1987). Configurational information in face perception. *Perception*, *16*, 747–759.