

## The Philadelphia Face Perception Battery

Amy L. Thomas<sup>a</sup>, Kathy Lawler<sup>a</sup>, Ingrid R. Olson<sup>b</sup>, Geoffrey K. Aguirre<sup>a,b,\*</sup>

<sup>a</sup> *University of Pennsylvania, Neurology Department, United States*

<sup>b</sup> *University of Pennsylvania, Center for Cognitive Neuroscience, United States*

Accepted 23 October 2007

### Abstract

The Philadelphia Face Perception Battery (PFPB) tests four aspects of face perception: discrimination of facial similarity, attractiveness, gender, and age. Calibration with 116 neurologically intact subjects yielded average performance of ~90%. Across subjects, there was a low correlation (<0.22) in performance between the tests (with the exception of the attractiveness and age discrimination tests) suggesting that the tests measure independent aspects of face perception. There were modest effects of subject demographic factors upon performance, and test–retest reliability scores (between 0.37 and 0.75) were comparable to other neuropsychological batteries. Modification of the stimuli to obscure internal facial features lowered performance on the age, gender, and attractiveness discrimination tests between 2 and 4 standard deviations. The clinical sensitivity of the battery was demonstrated by testing a patient with acquired prosopagnosia. She showed performance impairments of between 2 and 4 standard deviations on all sub-tests. The PFPB is freely available for non-commercial use.

© 2007 National Academy of Neuropsychology. Published by Elsevier Ltd. All rights reserved.

*Keywords:* Face perception; Prosopagnosia; Facial beauty; Neuropsychological testing

### 1. Introduction

Neuropsychological tests provide standardized measures of cognitive processes. Face perception ability, and its neural correlates, has been the subject of intense study and frequently the target of neuropsychological investigation. Well described is the clinical deficit of prosopagnosia: an inability to perceive facial information despite ostensibly intact rudimentary visual perception. Initially understood as an acquired deficit following focal, ventral occipito-temporal lesions, prosopagnosia is now also recognized as a developmental condition in both general (Duchaine & Nakayama, 2006a) and special (e.g. autistic) populations (Blair, Frith, Smith, Abell, & Cipolotti, 2002; Cipolotti, Robinson, Blair, & Frith, 1999). Despite the great research interest in face processing deficits, formal neuropsychological tests to detect impairments in these core abilities of human social behavior have been fairly limited. The two most commonly used neuropsychological tests that employ face stimuli are Warrington's Recognition Memory Test (RMT) and the Benton Facial Recognition Test (BFRT). Despite their widespread application these tests have specific limitations for the measurement of face perception ability per se.

The RMT uses face stimuli and was well standardized with 300 intact and impaired individuals. The test involves presenting faces and words to a subject and asking them to judge if each stimulus is pleasant or unpleasant. Following a delay, the subjects are then asked to determine which of two faces or words they had been shown previously. This

\* Corresponding author at: 3400 Spruce St., 3W. Gates, Philadelphia, PA 19104, United States. Tel.: +1 215 662 3390; fax: +1 215 349 8260.  
E-mail address: [aguirreg@mail.med.upenn.edu](mailto:aguirreg@mail.med.upenn.edu) (G.K. Aguirre).

test was originally designed to evaluate deficits of verbal vs. non-verbal memory (Sweet, Demakis, Ricker, & Millis, 2000; Warrington, 1984), although recent evidence has suggested that the RMT may not be reliably sensitive to these deficits (Sweet et al., 2000). As the test employs face stimuli, it has often been used to measure impairments in face perception (e.g., Nunn, Postma, & Pearson, 2001). By design, however, the test is clearly not a process-pure measure of face perception, as impairments in memory alone could contribute to impaired performance. The 3rd edition of the Wechsler Memory Scale includes a subtest of face memory and similarly does not provide a pure measure of face perception.

The BFRT was designed to evaluate unfamiliar face perception. It is easily administered to a clinical population and is also well standardized (Benton, Sivan, Hamsher, Varney, & Spreen, 1983). Both short (27 items) and long form (54 items) versions of the test are used. The patient is first shown gray scale photographs of several faces under severe lighting conditions designed to eliminate the perception of hairstyle and other features not intrinsic to the face. Initially, the task is to identify a face that matches a target face from a set of six. Additional trials in the long form involve identifying three of the six faces that match a target face, but are displayed in different orientations. The BFRT is commonly used to measure face perception deficits (Nunn et al., 2001; Weniger, Boucsein, & Irle, 2004). While ostensibly a more direct measure of face perception per se, there is evidence that the test may be solved without reference to the internal facial features. Prosopagnosia is generally conceived as a specific impairment of perception of internal facial features (eyes, nose, mouth) and it has long been recognized that such patients can demonstrate some intact performance with faces by using other perceptual features of the head (e.g., hairline, shape of the jaw).

Duchaine and Weidenfeld (2003) administered a modified form of the BFRT (and the RMT) that occluded internal facial features to normal undergraduate volunteers. These subjects were able to maintain performance within the normal range using only non-internal face features, such as hairline and eyebrows. Duchaine and Nakayama (2004) then tested developmental prosopagnosics (DP), individuals with a congenital deficit of facial perception, on the BFRT and ascertained that they are able to achieve scores within the normal range.

In response to these findings, and to assist in the identification of neurologically intact individuals who have developmental impairments in face perception, Duchaine and Nakayama (2006a) created a test of face memory focused upon internal facial features. The Cambridge Face Memory Test (CFMT) first presents six target faces to the participant and then 72 target detection, three-alternative forced choice trials, each involving one target face with views identical to the target stimuli, novel views, or novel views with noise. In tests of eight subjects who self-describe as having developmental prosopagnosia, six were found to be impaired on the CFMT.

While a substantial advance over extant methods, the CMFT is not suitable for all applications. First, the CMFT relies upon a single aspect of face perception to identify impairments. LeGrand et al. (2006) have argued that sensitivity to developmental impairments in face perception are best identified using multiple measures of face information. They tested DP patients on a series of tasks to evaluate their sensitivity to global form and motion, face detection, holistic face processing, perception of facial identity, the ability to determine gender from face stimuli, and whether attractiveness ratings were similar to controls. Their results demonstrated that DP patients are sensitive to global motion and form and perform normally in detection of faces, holistic processing, and gender discriminations. But they were impaired in processing of facial identity and judgments of attractiveness. LeGrand et al. (2006) concluded that facial deficits, specifically DP, cannot be reliably determined by poor performance on any single task.

Second, the CMFT is explicitly a memory test, and has fairly complicated instructions and motor response requirements. While ideal for screening an otherwise normal population for isolated face perception deficits, it is less well suited to neurologically impaired populations. For example, it could not be used to examine perceptual impairments in patients with organic dementing illnesses such as Alzheimer's disease.

We describe here the creation of a set of tests to examine face perception ability. Ideally, these tests would be sensitive to and specific for impairments in the perception of internal facial features. To allow broad application to neurologically impaired populations, the tests would have minimal memory, instruction, and motor response requirements, with a roughly 30 min administration time. Following the lead of Le Grand and colleagues, the measurement of several different aspects of face perception might provide for improved sensitivity. Human observers can rapidly derive from facial appearance information about facial identity, gender, age, and relative attractiveness. Using photo-realistic, synthetic face stimuli, we created a set of two-alternative forced-choice discrimination tasks that draw upon these aspects of facial perception. As many as 116 subjects from a broad range of demographic backgrounds completed the tasks, allowing the identification of a subset of trials that provide good test performance. We then demonstrated the relative independence of performance on each subcomponent of the test, the good test-retest reliability of the

measures, and the specificity of the tests for loss of internal facial feature processing. Finally, the sensitivity of the measures for clinical impairments in face perception was demonstrated in a patient with acquired prosopagnosia from a right temporal-occipital stroke. These tests are freely available for non-commercial use, and we have named the collection the Philadelphia Face Perception Battery (PFPB).

## 2. Methods

### 2.1. General experimental features

All artificial face stimuli were created using commercial software (GenHead by Genemation, <http://www.genemation.com/>) that has been modified for use in our lab. The software creates human faces in which the appearance of facial identity is determined by settings on each of 114 parameters, each an eigenvector derived from a principal components analysis of a large database of face photographs. Additional parameters allow control over ethnicity, age and gender. The ethnicity parameter was set to be Caucasian and held constant across tasks. Pilot behavioral studies were used to normalize the perceptual salience of changes in each of the 114 parameters, and to remove those parameters that had an obvious effect upon the direction of gaze. All face stimuli were created in the full frontal view. The program creates faces that are cropped to remove hair, although the ears and jaw line remain. Tasks were designed to investigate a subjects' ability to discriminate facial similarity, beauty, gender, and age. Each subject completed these tasks in a random order and was paid for their participation. Approval was obtained from the University of Pennsylvania's Institutional Review Board for human participation and all subjects provided a signed informed consent prior to participating.

All experiments were designed using Dell Intel Pentium (R) computers and E-Prime software (Schneider, Eschman, & Zuccolotto, 2002a, 2002b) for the presentation of stimuli. All experiments were run on a laptop 14.1 in. LCD screen using 640 × 480 resolution and full color (16 bits/pixel). The laptop was placed on a desk, and subjects were approximately 20 in. from the screen.

### 2.2. Four tasks of face perception

A total of 124 controls completed one or more of the following tasks (Table 1). While the population was drawn from a wide demographic pool, there was an over-representation of younger, Caucasian women with higher levels of education.

#### 2.2.1. Similarity task

**2.2.1.1. Subjects.** Seventeen subjects, eight male and nine female, completed an initial rating task (age range 20–29, mean 23.1 years). Data from these initial subjects were used to determine the relative similarity of a set of face stimuli, and construct the full test. A different group of subjects participated in the second stage of the study. This sample consisted of 116 subjects, 46 male and 70 female (age range 18–81 years, mean 37.1 years).

**2.2.1.2. Stimuli.** A set of 27 faces were created for this task. All appeared male with an age of 20–30 years. The faces varied in appearance across three parameters; skin tone, facial thickness, and a third ineffable parameter appearing to modify the overall appearance. There were three levels for each parameter, yielding a set of 27 faces. The stimuli were 313 × 313 pixels and in 16-bit color. The program displayed each face to take up 30% of the screen width and 40% of height.

**2.2.1.3. Procedure. Initial similarity ratings:** An initial group of 17 subjects provided similarity ratings for all possible pairs of the 27 stimuli not including identical pairs. On each of 351 trials, two faces from the set were presented side-by-side on the screen. Subjects were instructed to provide a rating of the similarity of the two faces on a scale of 1 (most similar) to 10 (zero key) (most different). Responses were made using the numeric keys on the keyboard. No time limit was imposed. The order of stimuli was randomized across subjects, as was the assignment of any particular face to the right or left side of the screen. Ratings from each subject were *z*-transformed and then averaged across subjects to yield an average similarity rating for each pair of faces. A set of “triads” of faces was then considered, and ranked by the difficulty of the judgment required to match one face to each of two other choices, based upon the relative

Table 1  
Subject demographics

Gender	
Male	46
Female	78
Handedness	
Right	109
Left	11
Ambidex	1
Age (years)	
<35	59
35–49	33
50–65	29
>65	3
Ethnicity	
White	72
AfricanAmerican	25
Asian	17
Other/Unknown	10
Education (years)	
≤12	22
13–16	60
16–19	12
>19	10

similarity ratings. A subset of 197 of the triads was selected evenly spaced across the range of predicted difficulty of discrimination.

*Similarity triad task:* On each of 197 trials, a given triad of faces was presented. The subject's task was to determine which of the two faces at the bottom of the screen was most similar to the simultaneously presented "target" face at the top center of the screen. Subjects responded by clicking with a mouse over their selection. Subjects were encouraged to respond accurately but quickly. Each triad remained on the screen until the subject provided a valid response.

### 2.2.2. Beauty task

*2.2.2.1. Subjects.* Initial beauty ratings were collected from 13 subjects, eight male and five female (age range 18–32, mean 23.4). These ratings were used to construct the complete test. Eighty-nine subjects, 32 male and 57 female, consisting of subjects who participated in the similarity task also participated in this task (age range 18–67, mean 38.6 years).

*2.2.2.2. Stimuli.* Faces used for the beauty task were taken from a set of 400 face pairs of roughly equivalent distinctiveness. The stimuli were 288 × 288 pixels and in 16-bit color. The program displayed each face to take up 40% of the screen width and 50% height.

*2.2.2.3. Procedure. Initial beauty ratings:* An initial group of 13 subjects provided judgments of whether each of 333 single stimuli were "more attractive than average" on 702 trials. Subjects responded by pressing a button on a response pad while each stimulus was displayed (1000 ms). These data were acquired as part of a separate study (Chatterjee, Thomas, Smith, & Aguirre, 2005). Beauty ratings were calculated as the proportion of subjects that judged a face to be more attractive than average. Ratings provided by the initial task allowed for the computation of subjects percentage agreement for all trials.

*Beauty task:* The highest 83 rated female stimuli were sorted in ascending order and paired with the 83 female stimuli rated least attractive sorted in descending order. For example, the highest rated face was paired with the least attractive, the second highest with the second least attractive, and so on. Another 83 stimuli were created by performing the same manipulation with the male stimuli, yielding a total of 166 stimuli. Subjects were presented with the stimulus pairs and directed to identify the more attractive face. Subjects responded by clicking on a mouse over their selection.

Subjects were encouraged to respond quickly without sacrificing accuracy. Each pair remained on the screen until the subject provided a valid response. The trials were ordered in ascending difficulty (difference in beauty ratings being greatest to face pairs with the most similar beauty ratings). Correct responses were determined by majority subject agreement.

### 2.2.3. Gender task

2.2.3.1. *Subjects.* Eighty-seven subjects (31 male, 56 female) completed the gender task. The subjects ranged from 18 to 67 years old (mean 38.3 years).

2.2.3.2. *Stimuli.* Faces used for the gender task were taken from the same set of stimuli used for the beauty task. Initial testing allowed for the sets to be narrowed to 166 single stimuli. The stimuli were 288 × 288 pixels and in 16-bit color. The program displayed each face to take up 40% of the screen width and 50% of height.

2.2.3.3. *Procedure.* Each of the 166 single stimuli was presented on the screen above the words “male” and “female”. Subjects were instructed to respond by clicking a mouse over the word “male” if the face appeared to be male and “female” if the face appeared to be female. The stimulus remained on the screen until a valid response was provided. Trials were presented in a random order and correct responses were determined to be those in agreement with the majority of subjects.

### 2.2.4. Age task

2.2.4.1. *Subjects.* Sixty-two subjects (23 male, 39 female) completed the age task. Subjects ranged from 18 to 64 years (mean 37.5 years).

2.2.4.2. *Stimuli.* Another set of 16 Caucasian males and females were created for each of five age groups (20, 30, 40, 50, 60 years of age) resulting in a total stimuli set of 160 faces. All possible 195 face pairs of the gender matched faces were included. The face stimuli were full color (16 bits/pixel), and set to be a uniform 300 × 300 pixel size. All faces were generated in the full-frontal orientation and displayed on 40% of the screen width and 50% of height.

2.2.4.3. *Procedure.* Subjects were then presented with the stimulus pairs and directed to identify the face which was older. Subjects responded by clicking a mouse over their selection. Subjects were also encouraged to respond quickly, but accurately. Each pair remained on the screen until the subject provided a valid response. Correct responses were determined to be those in agreement with the majority of subjects. The trials were in ascending order of difficulty (i.e., starting with trials with the largest apparent age difference between stimuli).

## 2.3. Task retesting

### 2.3.1. Subjects

Nineteen subjects (8 male, 11 female) who had completed the four face perception tasks were invited to participate in a retest session. These subjects ranged in age from 18 to 45 years (mean 21.6). The average time between initial and retest was 29.3 days (range 22–46 days).

### 2.3.2. Stimuli

Faces used for these retest tasks were identical to those used in the original tasks.

### 2.3.3. Procedure

For each task, a subset of 75 trials was identified. Trials were selected so that, on average, subjects in the initial testing demonstrated at least 80% agreement on each trial, and had overall average agreement across all 75 trials of approximately 90%. Further details are provided below in the results section. The 19 retest subjects completed these abridged versions of the tasks.

## 2.4. Occluded internal features tasks

### 2.4.1. Subjects

Twenty new subjects (10 male, 10 female) completed the set of four tasks using a modified version of the stimuli to eliminate internal face features. These tasks were created to determine the specificity of the intact tasks for loss of internal facial feature perception. The subjects' ages ranged from 19 to 32 years (mean 23.2).

### 2.4.2. Stimuli

The initial four tasks were administered to the subjects in the intact (original) and this modified form. The stimuli for the intact versions were identical to those described above. The stimuli for the modified versions were modified to occlude the internal facial features (eyes, nose, mouth) with an oval of the average skin tone of the intact stimuli.

### 2.4.3. Procedure

All 75 trials of the intact and occluded tasks were administered under the same testing conditions and the intact and modified test order was counter-balanced across subjects. The task order continued to be randomized within each version type.

## 2.5. Case study DK

To examine the sensitivity of the tests to clinical face perception deficits, we administered the final versions of all tasks to a patient with new-onset prosopagnosia. DK is a 63 year-old woman with a previous left occipital lobe hemorrhage producing a partial right inferior quadrantanopsia. She presented with a new right temporal-occipital hemorrhage producing left hemianopsia and a stated inability to recognize friends and family despite reportedly intact general object perception. The final set of four tasks, each composed of 75 trials, was administered at the bedside to DK 2 days after symptom onset. A subsequent brain biopsy diagnosed amyloid angiopathy as the cause of hemorrhage.

## 3. Results and discussion

### 3.1. Trial reduction

The initial versions of the four tests contained between 166 and 351 trials (Fig. 1). As there is no objective "correct" or "incorrect" answer regarding, e.g., the male or female appearance of a face, the percent agreement across subjects was calculated for each trial. For the four tasks, across-subject agreement ranged from 51% on some trials to 100% on others. To serve as a sensitive tool for the identification of subjects with perceptual abilities different from the normal population, trials with very high and relatively low subject agreement should be eliminated. Trials with low (approaching chance at 50%) agreement are non-diagnostic of face perception ability, as the trial cannot distinguish between an impaired individual and a normal population. Further, a test composed solely of very easy trials, in which 100% agreement across subjects was achieved, would be susceptible to ceiling effects and provide relatively poor sensitivity to subtle deficits. Therefore, we first identified the set trials for each task in which across-subject agreement was at least 80%. Next, 75 trials from this subset for each task were selected, such that the average subject agreement across all trials for a given task was approximately 90%. Table 2 provides the mean and standard deviation of across-subject agreement for the subset of 75 trials for each task. We will hereafter refer to a subject's "performance" as the extent to which his or her responses match those selected by the majority (specifically, >80%) of control subjects.

Beyond theoretical improvement in sensitivity, the reduction of the trials to 75 for each task allows for completion of the tasks, by a majority of subjects, within 30 min. Finally, the 75 trials were placed in order of ascending difficulty. This standard practice in neuropsychological instruments (e.g., BFRT, BLO, BNT, WAIS, WIAT; Benton et al., 1983; Spreen & Strauss, 1991) allows clinical populations to more quickly grasp the design of the task, increase patient confidence, and provides the most complete data if the testing must be discontinued (Giannakou & Kosmidis, 2006).

### 3.2. Test-retest

Within the normal population, the standard deviation of performance on each test was roughly 9%. Does this variability represent true differences in performance between subjects within the normal population, or is it the result

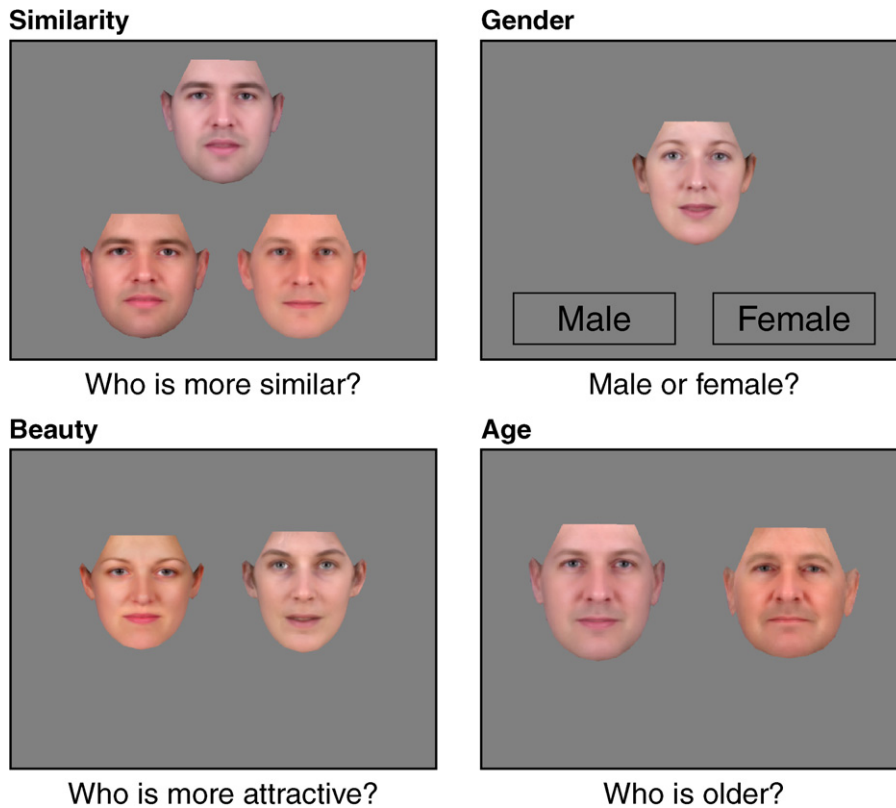


Fig. 1. Example trial for each task type. Instructions regarding the discrimination to be performed for each test were provided to the subject prior to testing. A description of the judgment to be made is shown below the stimuli, although this was not present during testing. Subjects provided their response by clicking on the appropriate face except for the Gender task during which the “Male” or “Female” box was selected on each trial.

of other random factors (e.g., time of day of administration, alertness) that do not replicate within subject? The assignment of variance in performance to within or between subjects is accomplished by measuring the test–retest reliability of a measure.

A group of 19 subjects completed the subset of 75 trials of each test at least 3 weeks after their initial participation. We examined the correlation across subjects of performance upon the second test to performance on the subset of 75 trials contained within the first testing session. Fig. 2 presents the scatter-plots for each subtest, and the non-parametric (Spearman) correlation between the two testing sessions. As can be seen, test–retest reliability ranged from a low of 0.37 on the gender test to 0.75 on the similarity test. This indicates that differences between subjects within the normal population in performance may be attributed both to true individual differences as well as between-subject random factors. It should be noted that low test–retest reliability does not impact the capacity of the measure to identify a test subject as not belonging to the normal population, which is instead determined by the standard deviation of performance within the normal population. Levine, Miller, Becker, Selnes, and Cohen (2004) examined eight widely used neuropsychological measures and determined that the Spearman test–retest correlation ranged from 0.47 and 0.88. Salinsky, Storzbach, Dodrill, and Binder (2001) found similar correlations for two neuropsychological tests examined

Table 2  
Across-subject agreement on the 75 trial subset for each test

	Mean $\pm$ S.D.
Similarity	89% $\pm$ 8.9%
Beauty	88% $\pm$ 8.3%
Gender	93% $\pm$ 8.5%
Age	90% $\pm$ 9.8

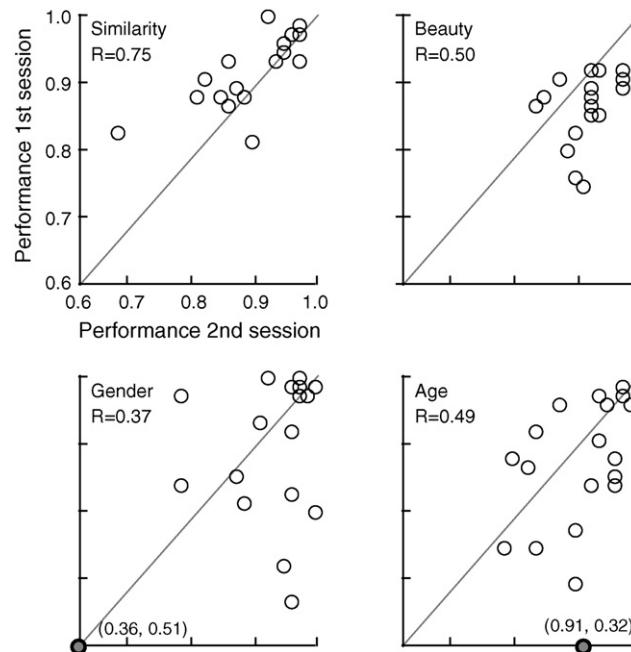


Fig. 2. Test–retest scatter-plots. Each data point corresponds to one of the 19 subjects who completed each of the four tasks during two separate testing sessions. Performance during the first session is plotted against second session performance. The Spearman (non-parametric) correlation is given for each plot. Two data-points fell outside of the plotted range (indicated by filled, gray circles). The actual scores for these points are given in parentheses.

by Levine et al. (2004), but also examined six other measures and found those correlations to range between 0.39 and 0.80. Therefore, the test–retest correlations of these tasks are similar to other neuropsychological instruments commonly used.

### 3.3. Between tests performance correlations

Given some degree of test–retest reliability within each test, we may examine the degree to which performance is correlated between tests across subjects. The presence of a correlation in performance between two neuropsychological measures can be taken as evidence that those two tests draw upon common mental processes for their accurate completion (Kahana, Rizzuto, & Schneider, 2005). The PFPB is composed of four tests of facial perception, although each requires the subject to derive different information from facial appearance. We examined the correlation between task performances to determine the extent to which the four test components assess independent aspects of facial perception. Table 3 presents the correlation across subjects in performance for the different tests. As can be seen, there was a generally low correlation in performance between tests, with the exception of the Beauty and Age discrimination tasks which were correlated at  $R = 0.6$ . Given previous work that shows that age modulates the perceived attractiveness of a face (Henss, 1991; Mckelvie, 1993; Tatarunaite, Playle, Hood, Shaw, & Richmond, 2005; Wernick & Manaster, 1984) the Age and Beauty judgment tasks may draw upon shared perceptual processes. Regardless, given the relatively low between-test correlations, there is little risk of redundancy in administering all four-test components to subjects.

Table 3  
Task performance cross-correlations

	Similarity	Beauty	Gender	Age
Similarity	–			
Beauty	0.21	–		
Gender	0.09	0.20	–	
Age	0.01	0.60	0.10	–

### 3.4. Influence of demographic factors

As “correct” responses to task trials were determined by consensus across the population, we wished to determine if demographic factors would lead to systematically different performance for some subjects. We examined the effects of age, education, gender, ethnicity, and handedness upon subject performance on the 75 trial sub-sets of the PFPB. Two demographic variables were significant in the ANOVA. First, there was a decrease in performance (i.e., agreement with the population) on the beauty task by approximately 0.2% for each year of subject age ( $t = 2.18, p = 0.03$ ). Equivalently, an 84-year-old subject would be expected to perform only one standard deviation below average. Previous studies have also observed systematic changes in assessment of facial attractiveness with subject age (Henss, 1991; Saykin et al., 1995). Subject age did not modulate performance on the other face tasks. Second, the ANOVA found a significant improvement in performance on the age task by approximately 1.4% for each additional year of education ( $t = 2.56, p = 0.01$ ). Effectively, seven additional years of education (e.g., completion of post-doctoral work) to the average would be expected to increase performance on the age task by one standard deviation. Subject education levels did not modulate performance on the other face tasks.

The other demographic variables – gender, ethnicity, and handedness – did not significantly modulate performance on any task. Given previous studies of own-race performance bias in face perception (Brigham & Malpass, 1985), the absence of a relationship between ethnicity of the subject and task performance is noteworthy. The equivalent performance of non-Caucasian subjects with the Caucasian face stimuli may be the result of two factors. First, test answers were defined by population consensus, which included subjects of different ethnicities. Second, the own-race performance bias has been demonstrated primarily in discrimination of individuals, while the components of our battery require categorical and similarity judgments. We do note, however, that despite the equivalent overall performance across ethnicities, it remains possible that differences in performance dependent upon subject ethnicity may exist for individual test items (Reynolds, 2000).

Table 4 provides the expected average performance on the beauty and age discrimination tasks as a function of subject age and years of education.

### 3.5. Occluded internal features tasks

Duchaine and Weidenfeld (2003) have previously shown that some tests that ostensibly measure internal face perception may be performed entirely using external facial features (e.g., hair, eyebrows, etc.). The stimuli used in the PFPB omit the hair, but present the outline of the face including the ears and jaw-line. Additionally, the faces are presented in color with differences in skin tone between stimuli. We elected to use these nearly whole faces (as opposed to severely cropped versions) to encourage naturalistic face behavior. Doing so, however, runs the risk of producing tests that are not strictly specific for deficits in the perception of only internal facial features. To evaluate the extent to which non-internal facial feature information can contribute to test performance, we created modified versions of the face stimuli in which the internal facial features were obscured by a uniform oval that reflected average skin tone (Fig. 3). Twenty subjects completed the four tasks of 75 trials each using the unaltered and feature-obscured versions of the stimuli. We then compared performance across subjects on the two versions of the test.

Average performance on the intact versions of the tests was not significantly different from that seen in the original control population. With internal features obscured, performance was significantly impaired on the similarity

Table 4  
Expected performance by subject age and education for the beauty discrimination and age discrimination tasks

Test	Subject age (years)			
	<35	35–49	50–65	>65
Beauty discrimination	90.3%	88.8%	84.1%	81%
Test	Subject education (years)			
	≤12	13–16	16–19	>19
Age discrimination	83.7%	89.3%	93.5%	99.8%

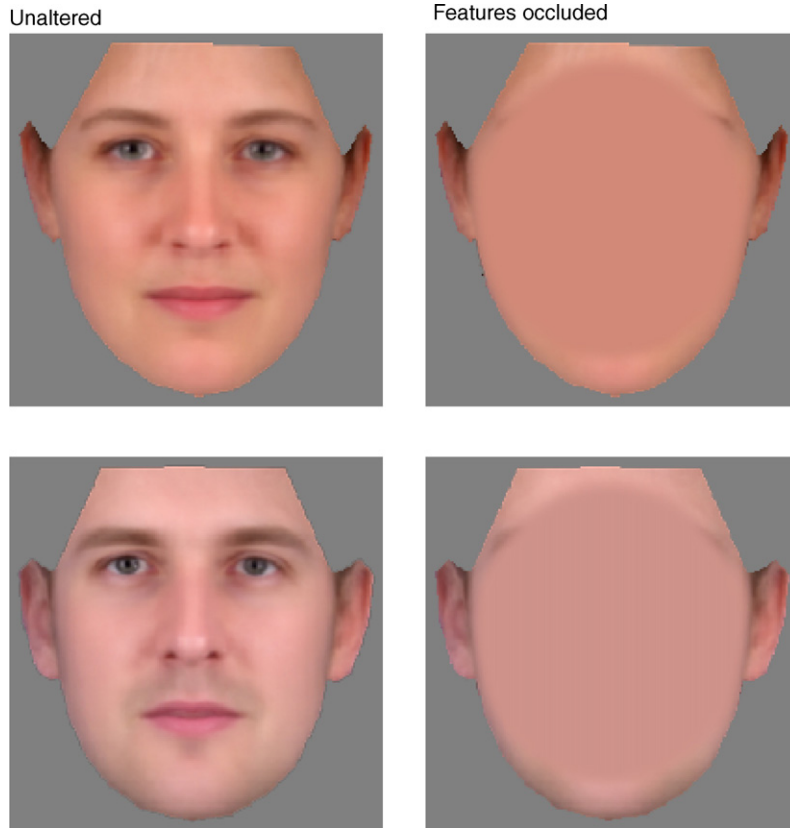


Fig. 3. To examine the potential contribution of external facial features to task performance, a version of the stimuli were created in which the internal facial features were obscured by an oval of uniform color matching average skin tone.

( $t = -2.864$ ,  $p < .01$ ), beauty ( $t = -8.073$ ,  $p < .001$ ), gender ( $t = -6.463$ ,  $p < .001$ ), and age ( $t = -14.154$ ,  $p < .001$ ) discrimination tasks. Fig. 4 shows the average performance on each of the subtests using the intact and obscured stimuli. For the gender, age, and beauty discrimination tasks, average performance was reduced between 2 and 4 standard deviations. This suggests that these tests would be both sensitive to and specific for impairment of internal facial feature perception. Performance on the similarity discrimination test, however, was less affected by obscuring the facial features, with an average performance reduction of 0.5 standard deviations. For these stimuli, skin tone and the outline of the face are sufficient to make accurate judgments of facial similarity. As we will see below, the test may nonetheless be sensitive to clinical impairments in face perception, as damage may also impair judgments of facial similarity using these more coarse components of facial appearance. An impairment in the similarity test using these stimuli may not,

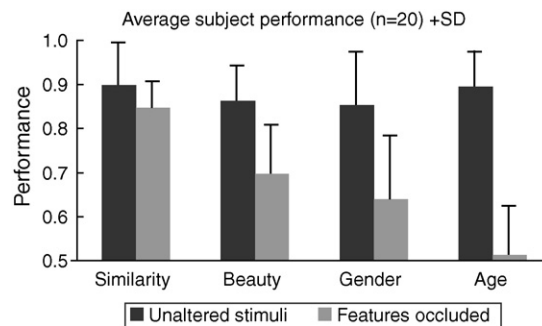


Fig. 4. Average subject performance on the intact and occluded feature versions of each task. Bars indicate one standard deviation of performance in this set of 20 subjects.

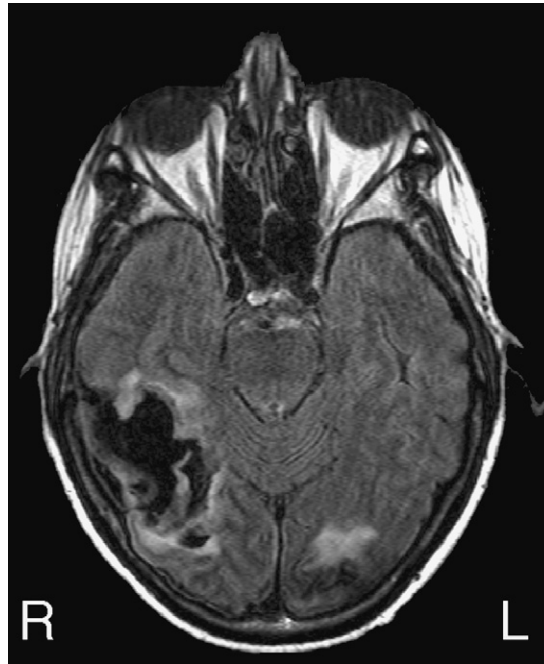


Fig. 5. T2-weighted (FLAIR), axial MRI scan of patient DK. Seen is a recent right temporal-occipital hemorrhage (dark on this sequence) and an area of signal abnormality in the left occipital lobe from a prior hemorrhage.

Table 5  
Performance of prosopagnosic subject DK

	Similarity	Beauty	Gender	Age
Accuracy	57	61	75	68
S.D.	−3.6	−3.2	−2.2	−2.2

however, be taken as specific evidence of impairment in internal facial feature perception. A modified version of the similarity task, which eliminates variation in skin tone and facial outline, could be created for this purpose.

### 3.6. Subject DK

Finally, to demonstrate that sensitivity of the tests to clinical impairments in facial perception, we administered the PFPB to a patient with acquired prosopagnosia. Patient DK developed an inability to recognize otherwise familiar friends and family following a right temporal-occipital hemorrhage (Fig. 5). Bedside testing indicated an intact ability to name visually presented objects, but she was unable to identify family members in the room. The PFPB was administered at the bedside, and she was found to have a substantial (>2 standard deviation) impairment in all subtests (Table 5). Further testing with neurological populations is necessary to demonstrate that impairments on the PFPB can be seen in the presence of objectively intact performance on other measures of visual perception, but the results with patient DK are sufficient to demonstrate the sensitivity of the test.

## 4. Conclusions

The Philadelphia Face Perception Battery can be administered within approximately 30 min and is sensitive to clinical impairments in facial perception while minimizing the memory requirements present in other measures that use face stimuli (RMF; Warrington, 1984, CFMT; Duchaine and Nakayama, 2006b). In the presence of moderate

test–retest reliability, the four subtests appear to measure generally dissociable aspects of face perception. There is only a small effect of demographic factors upon performance on the PFPB.

As seen in other measures of facial perception, the naturalistic stimuli presented in the PFPB may permit the use of external facial features to complete the similarity component of the battery, although this does not apply to the other components. Other potential limitations of the PFPB are the relatively large standard deviations in performance in the normal population and the limited demographic range of the control population. Finally, while we have demonstrated the sensitivity of the test in one patient with clinical face perception impairments, we have not yet objectively demonstrated that this impairment may be observed in the presence of intact elementary visual perception. In future studies we will address these limitations and apply the PFPB to patient populations with focal cortical lesions as well as diffuse degenerative pathology.

The PFPB is freely available for non-commercial use and the stimuli and testing program are available for download at <http://cfn.upenn.edu/aguirre>.

## Acknowledgements

This research was supported by the National Institutes of Health (K08 MH72926) and the Burroughs Wellcome Fund.

## References

- Benton, A. L., Sivan, A. B., Hamsher, K., Varney, N. R., & Spreen, O. (1983). *Contributions to Neuropsychological Assessment*. New York: Oxford University Press.
- Blair, R. J. R., Frith, U., Smith, N., Abell, F., & Cipolotti, L. (2002). Fractionation of visual memory: Agency detection and its impairment in autism. *Neuropsychologia*, *40*, 108–118.
- Brigham, J. C., & Malpass, R. S. (1985). The role of experience and contact in the recognition of faces of own- and other-race persons. *Journal of Social Issues*, *41*, 139–155.
- Chatterjee, A., Thomas, A. L., Smith, S. E., & Aguirre, G. K. (2005). *Beauty in the Brain of the Beholder*. Poster session presented at the annual meeting of the Cognitive Neuroscience Society, New York, NY.
- Cipolotti, L., Robinson, G., Blair, J., & Frith, U. (1999). Fractionation of visual memory: Evidence from a case with multiple neurodevelopmental impairments. *Neuropsychologia*, *37*, 455–465.
- Duchaine, B. C., & Nakayama, K. (2004). Developmental prosopagnosia and the Benton Facial Recognition Test. *Neurology*, *62*, 1219–1220.
- Duchaine, B. C., & Nakayama, K. (2006a). Developmental prosopagnosia: A window to content-specific face processing. *Current Opinion in Neurobiology*, *16*, 166–173.
- Duchaine, B. C., & Nakayama, K. (2006b). 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*, 576–585.
- Duchaine, B. C., & Weidenfeld, A. (2003). An evaluation of two commonly used tests of unfamiliar face recognition. *Neuropsychologia*, *41*, 713–720.
- Giannakou, M., & Kosmidis, M. H. (2006). Cultural Appropriateness of the Hooper Visual Organization Test? Greek Normative Data. *Journal of Clinical and Experimental Neuropsychology*, *28*, 1023–1029.
- Hess, R. (1991). Perceiving age and attractiveness in facial photographs. *Journal of Applied Social Psychology*, *21*(11), 933–946.
- Kahana, M. J., Rizzuto, D. S., & Schneider, A. R. (2005). Theoretical correlations and measured correlations: Relating recognition and recall in four distributed memory models. *Journal of Experimental Psychology, Learning, Memory, and Cognition*, *31*(5), 933–953.
- LeGrand, R., Cooper, P. A., Mondloch, C. J., Lewis, T. L., Sagiv, N., de Gelder, B., et al. (2006). What aspects of face processing are impaired in developmental prosopagnosia? *Brain and Cognition*, *61*(2), 139–158.
- Levine, A. J., Miller, E. N., Becker, J. T., Selnes, O. A., & Cohen, B. A. (2004). Normative data for determining significance of test-retest differences on eight common neuropsychological instruments. *The Clinical Neuropsychologist*, *18*, 373–384.
- Mckelvie, S. J. (1993). Stereotyping in perception of attractiveness, age, and gender in schematic faces. *Social Behavior and Personality*, *21*(2), 121–128.
- Nunn, J. A., Postma, P., & Pearson, R. (2001). Developmental prosopagnosia: Should it be taken at face value? *Neurocase*, *7*, 15–27.
- Reynolds, C. R. (2000). Methods for detecting and evaluating cultural bias in neuropsychological tests. In E. Fletcher-Janzen, T. L. Strickland, & C. R. Reynolds (Eds.), *Handbook of cross-cultural neuropsychology*. New York: Kluwer Academic/Plenum Publishers.
- Salinsky, M. C., Storzach, D., Dodrill, C. B., & Binder, L. M. (2001). Test-retest bias, reliability, and regression equations for neuropsychological measures repeated over a 12–16-week period. *Journal of the International Neuropsychological Society*, *7*, 597–605.
- Saykin, A. J., Gur, R. C., Gur, R. E., Shtasel, D. L., Flannery, K. A., Mozley, L. H., et al. (1995). Normative neuropsychological test performance: Effects of age, education, gender and ethnicity. *Applied Neuropsychology*, *2*, 79–88.
- Schneider, W., Eschman, A., & Zuccolotto, A. (2002a). *E-Prime User's Guide*. Pittsburgh: Psychology Software Tools Inc.
- Schneider, W., Eschman, A., & Zuccolotto, A. (2002b). *E-Prime Reference Guide*. Pittsburgh: Psychology Software Tools Inc.
- Spreen, O., & Strauss, E. (1991). *A compendium of neuropsychological tests: Administration, norms and commentary*. New York, NY: Oxford University Press.

- Sweet, J. J., Demakis, G. J., Ricker, J. H., & Millis, S. R. (2000). Diagnostic efficiency and material specificity of the Warrington Recognition Memory Test: A collaborative multisite investigation. *Archives of Clinical Neuropsychology*, *15*(4), 301–309.
- Tatarunaite, E., Playle, R., Hood, K., Shaw, W., & Richmond, S. (2005). Facial attractiveness: A longitudinal study. *American Journal of Orthodontics and Dentofacial Orthopedics*, *127*, 676–682.
- Warrington, E. K. (1984). *Recognition Memory Test*. Windsor, UK: NFER-Nelson.
- Weniger, G., Boucsein, K., & Irle, E. (2004). Impaired associative memory in temporal lobe epilepsy subjects after lesions of Hippocampus, Parahippocampal Gyrus, and Amygdala. *Hippocampus*, *14*, 785–796.
- Wernick, M., & Manaster, G. J. (1984). Age and the perception of age and attractiveness. *The Gerontologist*, *24*(4), 408–414.