The Role of Experience in the Face-Selective Response in Right FFA

Rankin W. McGugin¹, Katie F. Ryan¹, Benjamin J. Tamber-Rosenau¹,² and Isabel Gauthier¹

¹Department of Psychology, Vanderbilt University, Wilson Hall, PMB 407817, 2301 Vanderbilt Place, Nashville, TN 37240-7817, USA and ²Department of Psychology, University of Houston, 126 Heyne Building, 3695 Cullen Blvd, Houston, TX 77204, USA

Address correspondence to Rankin W. McGugin, Department of Psychology, Vanderbilt University, Wilson Hall, PMB 407817, 2301 Vanderbilt Place, Nashville, TN 37240-7817, USA. Email: Rankin.McGugin@vanderbilt.edu

Abstract

The expertise hypothesis suggests the fusiform face area (FFA) is more responsive to faces than to other categories because of experience individuating faces. Accordingly, individual differences in FFA’s selectivity for faces should relate to differences in behavioral face- recognition ability. However, previous studies have not demonstrated this, while the comparable association is often observed with nonface objects. We created a training paradigm with conditions sufficient to observe the same effect with faces. First, we selected subjects with a wide range of behavioral face-recognition abilities, then we manipulated experience with an artificial race of faces based on subjects’ pretraining ability, maximizing variability in face individuation. Neural selectivity was measured for Caucasian faces and artificial-race faces relative to control objects. Selecting subjects for greater variability in face-recognition ability revealed an association between behavior and FFA selectivity for Caucasian faces, with an effect exclusive to the middle right FFA (FFA2). Manipulating experience with artificial-race faces led to stronger brain-behavior correlation for artificial-race faces, also in right FFA2. Group analyses showed an overlap of these effects for Caucasian and artificial-race faces in right FFA2. The right FFA2 appears particularly sensitive to experience with faces just as it is for nonface objects.

Key words: experience, faces, fMRI, fusiform, training

Introduction

One goal of cognitive neuroscience is to link brain activity and behavior to test and constrain theories. In the present work, we investigated the conditions under which face-recognition ability is related to selectivity for faces in the fusiform face area (FFA). In particular, we sought to test for the role of experience in this brain-behavior correlation.

While performance cannot be expected to always correlate in a direct fashion with local differences in brain activity, such relations are sometimes observed. One example is behavioral performance in object recognition tasks predicting the selective activity these objects elicit in the functionally defined FFA. In this case, an “expertise account” for why faces elicit more activity than objects in this cortical area has led to the prediction that experience individuating nonface objects from a visually homogenous category increases the number of neurons that respond to objects of this category in the FFA. This neural selectivity would, therefore, predict behavioral performance.

Functional magnetic resonance imaging (fMRI) studies have related behavioral performance to FFA’s selective response for various categories of expertise, including birds, cars, chess configurations, and radiological images (Gauthier et al. 2000; Xu 2005; Harley et al. 2009; Harel et al. 2010; Bilalić et al. 2011; 2016; McGugin et al. 2012). Typically, the FFA (including its posterior...
and middle portions, FFA1 and FFA2, respectively) is localized in each subject using a functional contrast of the response to faces versus nonface objects, and then faces and other object categories are used to probe responses in this functionally localized region. Ideally, behavioral performance in the domain of interest (e.g., birds, cars, X-rays, etc.) is measured in more than 1 task, the tasks are then averaged in an effort to overcome task-specific variance, and then performance on the same task(s) with “other” object categories is regressed out to target domain-specific ability. Finally, neural selectivity for objects in the domain of interest is often measured in a simple task to ensure visual attention (such as a 1-back identity judgment) and compared with the response for a control category (often faces, because variability in experience with faces is assumed to be minimal). For example, in McGugin et al. (2015), expertise with cars was indexed using 2 different tasks with cars, regressing out performance on several control categories. This car recognition index, controlling for variability in non-car recognition, was related to the FFA’s response to cars relative to faces in a 1-back task (e.g., rFFA1, r_{25} = 0.40; rFFA2, r_{25} = 0.51).

Paradoxically, while this relation between object recognition and neural selectivity in the FFA has been observed many times, it has been more difficult to relate “face-specific” responses in FFA to behavioral face-recognition ability—yet this relation is a central prediction of the expertise account. Few studies describe a positive correlation between face-recognition performance and FFA response. One such study did so, but with such a modest effect size (r_{504} = 0.13) that it would require 362 subjects to expect replicating this effect with 80% power at an alpha of 0.05, even with a 1-tail test (Huang et al. 2014). Another study included both prosopagnosic patients and controls to obtain a brain-behavior correlation of r_{50} = 0.57 but the correlation was much reduced (r_{52} = 0.28, n.s.) when the patients were not included (Furl et al. 2011). Studies in children have shown evidence for a relation between behavioral face perception and the brain’s response to faces, using the neural inversion effect (Aylward et al. 2005), or the size of face-selective activity in FFA (Golarai et al. 2007). A recent study using adults reported a correlation between the right FFA’s response to faces with phase noise relative to the response to normal faces, and the behavioral performance difference between noisy and normal faces (Hermann et al. 2015). However, these authors found no significant correlation between face discrimination performance and blood oxygen level-dependent (BOLD) selectivity relative to objects, with either normal or noisy faces (P. Hermann, personal communication). One group reported that performance on behavioral face tasks predicted the number of FFA voxels engaged during face-recognition, but not the magnitude of FFA activation (Elbich and Scherf 2017). To sum up, much effort has been devoted to relating functional activity in FFA to face-recognition ability and the results suggest that the relation between behavior and selectivity for faces in this area is not as strong as it is for objects of expertise. A few studies find that neural structure (cortical thickness or anisotropy of fiber tracts in FFA (Gomez et al. 2015; McGugin et al. 2016)) can predict face-recognition ability. Event-related potential work has found a relation between the N170 component (thought to correspond to the FFA) and face cognition abilities across several tasks: e.g., face performance predicted the latency and amplitude of the N170 component (Herzmann et al. 2010; Kaltwasser et al. 2014). While acknowledging that the FFA’s response to faces has been related to behavioral performance in other paradigms (for instance correlating performance in the scanner with brain activity on a trial-by-trial basis, see Grill-Spector et al. 2004), here, we ask why individual differences in face-selective neural responses in the FFA are not consistently related to face-recognition ability in the way that has been observed for other object categories.

We can rule out some reasons why past studies may have failed to find a strong relation between face-recognition performance and FFA response in adults. First, recent work quantifying the reliability of individual differences in selectivity for faces within face-selective patches of the fusiform gyrus revealed high internal consistency, generally above 0.7 in bilateral FFAs, suggesting a sufficient range of reliable individual differences to correlate with behavior (McGugin and Gauthier 2016). Second, performance on sensitive face-recognition tasks, such as the Cambridge Face Memory Test (CFMT; Duchaine and Nakayama 2006) shows great variability in the normal population and as a result, measurements on this task in samples of normal adults tend to be highly reliable (~0.9 in internal consistency) and stable (~0.70–0.76 in test-retest reliability; Wilmer et al. 2010), evidence of lasting and robust individual differences.

Given these reliable neural and behavioral individual differences, why are these effects for faces not as clearly related as they are for nonface objects of expertise? Consider that variability in performance on tests of visual recognition stems in part from stable, or trait-like abilities: for instance, both face and car recognition have been found to be highly heritable (Wilmer et al. 2010; Shakeshaft and Plomin 2015). In contrast, neuroimaging work so far found no effect of genetic or shared environment on FFA response in adults, suggesting that FFA selectivity is more dependent on experience even though face-recognition ability shows a strong heritable component (Pinel et al. 2015). In the case of nonface recognition studies, experience is more likely to carry individual differences in both behavioral and neural selectivities. In studies of car expertise, subjects are typically recruited to vary as much as possible in their performance with cars, which is typically done by first screening subjects with self-report of “experience” with cars and verifying these claims with behavioral tests. This could explain why in studies where subjects are selected to vary in their performance with cars but not faces, the brain-behavior correlation would be most robust for cars. Our working hypothesis in the present work is that this correlation between performance and selectivity in the FFA can be obtained for faces, and in particular that it should be facilitated by increasing variability in experience with faces. Failing to support this hypothesis would represent an important challenge to the expertise account of specialization for faces in the FFA.

We adopted a 2-step design in which we first attempted to select individuals to maximize the range of performance in face-recognition ability (CFMT+, Russell et al. 2009) to test whether a large range of CFMT+ scores is sufficient to observe reliable correlations. In a second step, we increased the range in performance by training subjects with a new race of faces, “as a function” of their face ability. Because all our subjects would have considerable experience with standard faces, we used an artificial race of faces that would be equally unfamiliar to all our subjects (Chua et al. 2014). Critically, we then provided subjects who demonstrated low face-recognition ability on the CFMT+ with very little experience with the artificial race (low ability, low experience), and those with high face-recognition ability much more training with the artificial race (high ability, high experience). In this second step, our design thereby intentionally confounds ability and experience so as to maximize the range in “performance” with artificial-race faces. In this way, we recognize that prior work on car expertise
selected subjects to maximize the range in performance with cars, without any way of knowing how much of this variability was due to intrinsic differences in ability or differences in skills acquired through experience with cars. In fact, it is not unlikely that these factors interact such that those who have better potential ability are those seeking the most experience in a domain.

Our fMRI experiment, conducted after completion of the behavioral training and tests, includes both Caucasian faces and untrained exemplars of the artificial race for which subjects have been trained on a distinct set of faces. A first hypothesis is that CFMT+ scores and FFA neural selectivity for Caucasian faces will be correlated. Comparing this effect to those we have obtained in prior work in which we did not select subjects to vary in face-recognition should give us a sense for the effect of subject selection. A second hypothesis is that behavioral performance for artificial-race faces and neural selectivity for artificial-race faces will be related, with an effect size comparable to prior studies with nonface objects of expertise. This should reveal the effect of maximizing the range of performance with faces (via a combination of selection and training). A third hypothesis is that this relation is primarily driven by experience, which leads to the prediction that regressing out CFMT+ scores and their interaction with training will not abolish the effects.

Methods and Materials

Subjects

We chose a sample size that would yield sufficient power for the correlation typically observed between car expertise and behavioral performance ($r = 0.54$; see McGugin et al. (2014), for a meta-analysis). With $n = 25$ and alpha of 0.05, power is 0.87 to detect such an effect. 25 subjects (ages 18–35, mean 23 years; 9 males; 21 Caucasian, 4 Asian-American; all right-handed) completed all portions of this study for monetary compensation. All subjects reported normal or corrected-to-normal vision, and no history of psychiatric or neurological illness. We obtained informed consent for each participant in accordance with the institutional review board guidelines for Vanderbilt University and Vanderbilt University Medical Center. One subject was removed from all fMRI analyses due to excessive head motion.

Lunari Face Stimuli

A set of faces (obtained from Jim Tanaka at the University of Victoria; made with the FaceGen software) were modified with Adobe Photoshop to create faces with unique top and bottom halves. For each of these faces, the mouth was stretched to be as wide as the eyes, the ears were pulled downwards so the bottoms were below the mouth, the distance between the eyes was stretched to form the impression of a large bone above the nose (the “sign of the moon” giving its name to the race), the distance between the eyes and the eyebrows was stretched, and a second eyebrow was added above the original eyebrow (Chua et al. 2014; Fig. 1). Images of between 3 and 33 Lunari individuals were used during training depending on the training group, with an additional 20 Lunari faces used in each of the pre- and post-test.

Behavioral Screening and Training Groups

We screened subjects using the CFMT long form (“CFMT+”; Russell et al. 2009). In the CFMT+, subjects first studied frontal views of unfamiliar male faces, followed by introductory learning trials. Subjects were then given forced-choice test displays each containing 1 target face and 2 distractor faces, where they were instructed to select the face that matched one of the original target faces. The matching faces varied from their original presentation in their lighting, pose, or both. In the third section, test displays were presented in Gaussian noise. The fourth and final section specifically manipulated test displays to be very difficult: target and distractor images varied significantly from the training images, distractor images were repeated much more frequently, and images contained increased visual noise relative to the third section (Fig. 2). For a complete description of the CFMT+ see Russell et al. (2009).

We created 5 bins of subjects, from exceptionally low to exceptionally high performance on the CFMT+, to exacerbate differences along the range of face-recognition performance (see Table 1). A total of 48 subjects were tested in the CFMT+, and the first 5 subjects to fill each of 5 bins were selected for further behavioral testing and MRI. Each bin received a separate training schedule for a artificial race of faces, Lunari (see above; Chua et al. 2014). To increase variability in performance between groups, individuals with low face-recognition ability as measured by the CFMT+ scores received very little experience with
Lunari faces (low ability and low experience). Conversely, individuals with the greatest face-recognition ability received much more training with Lunari faces (high ability and high experience). One subject was mistakenly placed in the wrong training bin (bin 5 instead of 2). However, the correlation between training and bin number for the entire sample remains 0.9.

While the amount of training varied by bin (see Table 1 for details), the basic training task was the same for all subjects, with only the number of faces learned and duration of training varying across bins. This operationalization of face experience combines 2 influences: the number of unique faces that are individuated and the number of trials over which they are experienced. These 2 influences are confounded, similar to variability in experience for real world categories (i.e., relative to a novice, an experienced bird watcher likely has seen more exemplars from different bird species and has also individuated these species with greater frequency).

Pretest/Post-Test (2-Alternative Forced-Choice Recognition Task)

On the first day, all subjects first completed a 2-alternative forced-choice (AFC) recognition memory pretest (McGugin et al. 2011; see Fig. 2) prior to any training. On each of 20 trials, we presented 1 of 20 Lunari faces for 125 ms followed by a dynamic mask for 500 ms. Then, subjects viewed 2 Lunari faces (the Lunari face presented before the mask and a novel Lunari face) and responded as to whether the previously viewed face was located on the left or the right. Subjects had a 2000 ms window to make a response; if they did not respond in time, the next trial began.

Immediately following their final training session, subjects completed a 2-AFC recognition memory post-test with the same format as the pretest, but with previously un-seen Lunari faces. The fMRI part of the experiment followed: in the scanner, all subjects performed a 1-back matching test on novel Lunari faces (along with Caucasian faces, butterflies, and mushrooms), while whole-brain images were acquired.

Lunari Training (Individuation Task)

Immediately following the pretest, subjects began the first training session based on their initial bin placement (Fig. 2). Each daily session lasted between 30–60 min, depending on the
training group and subjects’ face. The task was the same across each session, but the amount of time and number of Lunari faces varied based on the subject’s bin placement (see Table 1). Each training block began with a learning phase in which we presented a sequence of Lunari faces centrally on the screen for 500 ms per face, paired with nonsense-syllable names (e.g., “Vabo,” “Bifa,” etc.). In the response phase, a single Lunari face appeared with a question mark in the place of the name label, prompting subjects to identify the face. Subjects responded by pressing the first letter of the correct name, or the space bar if they had not previously viewed the face. Subjects completed 1 training session per day, with no more than 2 days between each session.

**MRI Acquisition**

We performed all scans after the completion of behavioral training and tests, at the Vanderbilt University Institute for Imaging Science on a Philips 3-T Intera Achieva MRI scanner with a 32-channel head coil. We acquired high-resolution (HR) T1-weighted anatomical volumes (time repetition [TR]: 9.037 ms; time echo [TE]: 35 s; field of view [FOV]: 256 × 256; voxel resolution: 0.8 × 0.8 mm in plane; slice thickness: 0.9 mm; flip angle: 9°). For all functional scanning (2 localizer scans and 6 experimental scans) we used standard gradient-echo echoplanar T2*-weighted imaging to obtain functional images (TR: 2000 ms; TE: 4.61 ms; scans) we used standard gradient-echo echoplanar T2*-weighted imaging to obtain functional images (TR: 2000 ms; TE: 4.61 ms; slices: 33; FOV: 240 × 240; voxel resolution: 3 mm in plane; slice thickness: 4 mm; flip angle: 79°).

**fMRI Stimuli and Design**

We presented all images with an Apple Macintosh computer running Matlab 2014 (MathWorks) using Psychophysics Toolbox (Brainard 1997). Stimuli were displayed on a rear-projection screen using an Eiki LC-X60 DLP projector with a Navitar zoom lens.

Immediately following the HR structural scan, each subject completed 2 runs of a functional localizer scan (224 dynamics/run). We used grayscale images (36 faces, 36 objects, 36 body parts, and 36 scrambled arrays) in a 1-back detection task across 28 alternating blocks of faces, objects, body parts, or scrambled images. Each image was presented for 1 s, with a 2 s fixation at the beginning and end of each category block.

Following the localizer, subjects completed 6 runs of the experimental task. Categories were blocked and subjects responded to an immediate repeat in identity during a 1-back memory task of grayscale images of 109 Caucasian male faces, 109 Lunari faces (a novel set that was not used during the behavioral training sessions), 109 butterflies, and 109 mushrooms. We superimposed the face images on phase-scrambled backgrounds (made from images of butterflies) so that the complexity of the backgrounds was equivalent across categories. All images were equated for low-level visual properties using the SHINE toolbox for Matlab (Willenbockel et al. 2010) (Fig. 1). All images were presented such that the inner edge was at 3 degrees of eccentricity, to either side of fixation. Each image subtended 6 × 6 degrees of visual angle. Subjects searched for an immediate within-category image repeat in a sequence of target images shown in isolation (runs 1 and 4) (Fig. 2). (See Supplementary text for description of the Attention conditions.) Images appeared alternately to the left and right of fixation (Fig. 2), and subjects were instructed to maintain central fixation, then to press their right index or middle finger following the trial if they detected a repeat on the left or right side of the screen, respectively.

Each run started and ended with 30 s of fixation, followed by 12 stimulus blocks. At the beginning of each block (28.5 s) an instruction cue appeared on the screen (1.33 s) noting the category to which the subject should attend during the following block (e.g., “search for Faces”, “search for Lunaris”, etc). Each block contained 3 rapid serial visual presentation (RSVP) search trials of 18 images, each image presented for ~333 ms, for a total of 6 s per RSVP trial. A fixation period (3.1 s) followed each trial. 1-back repeats occurred once per 18-image trial, never in the first or last image occurrence of a trial. Blocks were separated by an additional 1.5 s of fixation, for a total of 4.6 s of fixation between blocks. Blocks in the isolated runs consisted of Caucasian faces, Lunari faces, mushrooms, and butterflies, and each category block occurred 3 times per run.

**Data Analysis**

**Magnetic Resonance Imaging**

The HR T1-weighted structural scans were normalized to Talairach space. Functional data were analyzed using Brain Voyager (www.brainvoyager.com) and in-house Matlab scripts (MathWorks). Preprocessing in Brain Voyager included registration to the original (non-transformed) structural scan, slice scan time correction (cubic spline), 3D motion correction (sinc interpolation) and temporal filtering (high-pass criterion of 2 cycles per run) with linear trend removal. One subject’s data were excluded from further analyses due to excessive head motion. No spatial smoothing was applied.

Regions of interest (ROIs) were defined using the Face > Object contrast from the face-localizer scan. We localized ROIs that responded more to faces than objects in bilateral posterior and middle fusiform gyrus (FFA1/2, respectively) and occipital face area (OFA). We localized ROIs medial to FFA and responding more to objects than faces in posterior and middle parahippocampal gyrus (PHG1/2, respectively) (Table 2).

Our functionally defined regions were initially localized on the 1 mm (interpolated) statistical maps from the localizer scans as the contiguous 1-mm voxels passing statistical threshold within a fixed distance of the activation peak for each contrast. This approach ensured consistency with previous reports of these ROIs and ensured uniformity across subjects (Table 2). However, we then down-sampled all ROIs to the functional resolution of 3 mm to guarantee that the signal was appropriately weighted per functional voxel. Any functional voxel containing 1 or more 1 mm voxels from the initial ROI definition was considered part of the final ROI. We removed any functional voxels that were members of multiple ROIs, thereby avoiding partial-volume effects associated with functional region membership. In defining the ROIs, we started with the peak, that is, most face-(or object-) selective 3-mm (i.e., subtending a volume of 27 mm³) voxel. We then grew the region to include a total of four 3-mm voxels by successively adding the most selective voxel that was contiguous with the already-selected voxel(s). We also used the peak voxel in FFA ROIs, because prior work found that reliability of individual differences in face selectivity was at its highest in the center of FFA, with no benefit to using larger ROIs (McGugin and Gauthier 2016). In this way, we could investigate the effect of ROI size (1 vs 4 functional voxels, corresponding to 27 mm³ vs 108 mm³ regions) on the effect of face expertise.

We modeled the fMRI data using a general linear model (GLM) that included regressors for each stimulus category (Caucasian face, Lunari face, Mushroom, and Butterfly), as well as response and instruction/preparation periods. Fixation
blocks were left un-modeled. For each voxel of each ROI, beta weights were computed as the regression parameter weights corresponding to each relevant stimulus category of a given run. For each subject, the beta weights from all voxels within a pre-defined ROI were then averaged for a given condition. In a separate step, the isolated conditions of interest were defined as Caucasian face beta weights regressing out Object beta weights and Lunari face beta weights regressing out Object beta weights. We regressed Object (mushrooms and butter) responses from the conditions of interest (Caucasian faces and Lunari faces) to control for general object selectivity that was typically above 0.6, the rFFA1 is the only region that the scores of interest are uncorrelated with the baseline scores by statistically removing the baseline variance, ensuring that the scores of interest are uncorrelated with the baseline variance. While all regions except the more anterior object-selective PHG2 showed a reliability for Caucasian faces that was typically above 0.6, the rFFA1 is the only region with iterative Gaussian filtering to compute the cluster size that yields a corrected threshold of \( P < 0.05 \) separately for each contrast of interest: Lunari face—Object (151 mm\(^3\)) and Caucasian face—Object (144 mm\(^3\)).

Results

Reliability

First, we considered reliability of our brain and behavior measures separately. A prerequisite of individual differences analyses is the reliability of measurements. Importantly, reliability should not be attributed to a test, a region or a task, as the same test, task or region can be more or less reliable depending on the sample. Instead, reliability must be calculated in the study sample (Wilkinson, The American Psychological Association APA Task Force on Statistical Inference 1999). The maximum correlation that can be observed between 2 variables is the square root of the product of their reliabilities. As such, measurement error is an important limitation on observed effect sizes and power.

Reliability of Behavioral Measures

The reliability of behavioral measures used in correlations with measures of neural response was estimated by calculating the internal consistency of the 2-AFC Lunari post-test using Cronbach’s Alpha, which was revealed to be good (\( \alpha = 0.85 \)) and similar to the CFMT’s reliability in the present sample (\( \alpha = 0.89 \)). Internal reliability of the 2-AFC Lunari pretest showed a much lower Cronbach’s Alpha (\( \alpha = 0.55 \)). This was expected given subjects had no experience with Lunari faces at the time of the pretest, while after the training, the range of their ability with Lunari faces should have increased.

Reliability of Brain Selectivity

We computed the reliability of beta weights in all regions of interest as the split-half reliability across runs. Ideally, we would want measures to achieve at least 0.7 reliability, a common minimum standard in new domains in psychometrics. The rFFA2 achieved this level of reliability for both Caucasian and Lunari faces (Spearman–Brown split-half reliability; Table 3; see Supplementary Table S1 for reliability in Attention conditions). Given our sample size (which varies by ROI but averages 20), we can observe a correlation of 0.42 with 80% power. Given the reliability of the Lunari post-test (0.85), with a 0.5 threshold for the reliability of the neural data, we reach 80% power to detect true correlations of at least 0.65 (of course we do not know what the true correlation is but we do not expect that our short training can account for more than 50% of the variance in selectivity). While all regions except the more anterior object-selective PHG2 showed a reliability for Caucasian faces that was typically above 0.6, the rFFA1 is the only region besides the rFFA2 to surpass 0.5 reliability for Lunari faces.

Behavioral Results from the Lab

Pretest/Post-Test

Average accuracy on Lunari pretest and post-test was high regardless of CFMT+ performance or training bin: pretest

<table>
<thead>
<tr>
<th>Region</th>
<th>Mean Talairach coordinates for peak voxel ± SD</th>
<th>One-sample t-test (P-value) on mean face beta weights (relative to objects)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Right FFA1 (N = 22)</td>
<td>36.9, −60.4, −16.6 (3.7, 6.8, 6.1)</td>
<td>4.8 (0.0001)</td>
</tr>
<tr>
<td>Right FFA2 (N = 24)</td>
<td>36.7, −42.6, −18.7 (4.6, 6.1, 5.2)</td>
<td>5.0 (0.0001)</td>
</tr>
<tr>
<td>Right OFA (N = 19)</td>
<td>34.3, −77.5, −12 (5.4, 6.5, 6.7)</td>
<td>4.4 (0.0002)</td>
</tr>
<tr>
<td>Right PHG1 (N = 21)</td>
<td>20.9, −65.2, −11.6 (6.7, 6.8, 5.9)</td>
<td>−4.6 (0.0001)</td>
</tr>
<tr>
<td>Right PHG2 (N = 22)</td>
<td>25.8, −47.7, −13.7 (5.7, 7.7, 5.3)</td>
<td>−4.7 (0.0001)</td>
</tr>
<tr>
<td>Left FFA1 (N = 20)</td>
<td>−41.4, −60.7, −18 (4.5, 7.9, 6.6)</td>
<td>4.5 (0.0001)</td>
</tr>
<tr>
<td>Left FFA2 (N = 21)</td>
<td>−40.9, −42.9, −20.1 (3.6, 5.1, 4.9)</td>
<td>4.6 (0.0001)</td>
</tr>
<tr>
<td>Left OFA (N = 16)</td>
<td>−36.8, −80.2, −14 (7.9, 7.4, 6.3)</td>
<td>4.0 (0.0006)</td>
</tr>
<tr>
<td>Left PHG1 (N = 20)</td>
<td>−25.2, −61.5, −10.3 (6.9, 7.2, 4.3)</td>
<td>−4.5 (0.0001)</td>
</tr>
<tr>
<td>Left PHG2 (N = 24)</td>
<td>−33.1, −40.8, −15.5 (6.5, 6.8, 5.4)</td>
<td>−5.0 (0.0001)</td>
</tr>
</tbody>
</table>
(acc = 0.85, standard deviation = 0.12), post-test (acc = 0.90, standard deviation = 0.13). Table 4 shows a summary of accuracy and response time data for pretest and post-test separated by bin (response times computed for correct trials only). In general, subjects improved in accuracy and speed from pretest to posttest. Although post-test accuracy appears to be close to ceiling in several bins, the high internal consistency of this measure assures us that it was sufficiently sensitive for individual differences to correlate with other measures.

Performance on the CFMT+ was not correlated with performance on the Lunari pretest ($r_{25} = 0.22$, n.s.), though it was significantly correlated with Lunari post-test ($r_{25} = 0.65$, $P = 0.0007$); this was expected because training was administered as a function of CMFT+ scores.

Training

Tables 5a,b show average accuracy rates and response times for the individuation task in each training day, for all 25 subjects together and for bins of 5 subjects. In general, over the course of training, accuracy rates for individuating Lunari faces increased while response times decreased after session 1 and then remained relatively constant.

Behavioral Results from the Scanner

Localizer Runs

In the localizer runs, sensitivity to detect 1-back repetitions was high for all conditions, especially intact-object conditions: (hit rate, false alarm rate) Face (0.89, 0.07), Object (0.86, 0.07), Body Part (0.82, 0.03), and Scrambled (0.76, 0.07). There was no difference in hit rates between Face and Object blocks ($t_{25} = 1.670$, n.s.) but performance was greater for Faces relative to Scrambled ($t_{25} = 3.44$, $P = 0.002$) and for Faces relative to Body Parts ($t_{25} = 3.67$, $P = 0.001$). Hit rate was greater for Object blocks relative to Scrambled blocks ($t_{25} = 2.11$, $P = 0.04$), with no difference between Objects and Body Parts ($t_{25} = 1.45$, n.s.) or Body Parts and Scrambled ($t_{25} = 1.43$, n.s.).

Experimental Runs

Average response rate across all trials of a block was 0.89 (after excluding 2 subjects for whom responses were not recorded due to technical error). In about 10% of trials, subjects responded either at the time of the repeat or after the response window; responses were only considered when they occurred within the response window which immediately followed the trial. Considering the trials when responses were made at the appropriate time, subjects were able to perform the task although it was difficult: (hit rate and false alarm rate) Caucasian face (0.65 and 0.34), Lunari face (0.63 and 0.41), Butterfly (0.79 and 0.19), Mushroom (0.73 and 0.25). Paired t-test on hit rates revealed significantly better detection for Butterflies relative to Caucasian faces ($t_{25} = 3.55$, $P = 0.002$) and to Lunari faces ($t_{25} = 3.79$, $P = 0.001$) with no difference between other conditions.

Comparing Brain Response to Caucasian and Lunari Faces

First, we compared the mean response amplitude for Caucasian and Lunari faces within the face-selective ROIs that showed reliable neural response to both categories: right FFA1 and right FFA2. Consistent with standard analysis of GLM contrasts, for this group analysis, we used a baseline “subtraction” (rather than regression) so difference values can be more readily interpreted relative to zero. In right hemisphere FFAs, subjects responded significantly more to Lunari faces than to Objects ($r_{FFA1}: t_{25} = 2.57, P = 0.018$; $r_{FFA2}: t_{25} = 3.79, P < 0.001$), and significantly more to Caucasian faces than to Lunari faces ($r_{FFA1}: t_{25} = 4.77, P < 0.0001$; $r_{FFA224}: t = 3.67, P = 0.001$) (Fig. 3a).

We would expect faces to engage the FFA more than objects, and Caucasian faces are far more familiar than artificial Lunari faces even following training.

Next, we considered the variability in individual differences for Caucasian and Lunari faces by correlating the BOLD Caucasian face response with the BOLD Lunari face response, controlling for the BOLD Object response. Here and for all correlational analyses, we regressed the baseline (Objects) instead of subtracting it as we did for the contrasts above. Within right hemisphere FFAs, the responses to Lunari faces were significantly more to Lunari faces than to other classes ($r_{FFA1}: r_{xx} = 0.79$, $r_{FFA2}: r_{xx} = 0.74$, Fig. 3b), as expected given they are both types of faces. To compare this relation to a nonface category, we separated the Objects category into Butterflies and Mushrooms.

In contrast to what we observed between categories of faces, in the same regions, correlations between faces (either Caucasian or Lunari) and Butterflies, regressing out the response to Mushrooms from all categories, were much lower. Specifically, correlations were smaller in the case of $r_{FFA1}$ ($F(m) \times B(m): r = 0.40$; $L(m) \times B(m): r = 0.37$) and virtually zero for $r_{FFA2}$ ($F(m) \times B(m): r_{25} = -0.09$; $L(m) \times B(m): r = -0.09$).

Correlation of CFMT+ and Neural Selectivity for Caucasian Faces

Here and in all correlational analyses, we do not apply a correction for multiple comparisons. Our prediction of an effect in FFA was based on prior results with cars (McGugin et al. 2015).
Table 5a Accuracy rates (standard deviation) during Lunari face training

<table>
<thead>
<tr>
<th>Bin</th>
<th>n</th>
<th>Day 1</th>
<th>Day 2</th>
<th>Day 3</th>
<th>Day 4</th>
<th>Day 5</th>
<th>Day 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td>0.86 (0.12)</td>
<td>0.86 (0.08)</td>
<td>0.82 (0.13)</td>
<td>0.87 (0.15)</td>
<td>0.9 (0.13)</td>
<td>0.94 (0.12)</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>0.75 (0.07)</td>
<td>0.76 (0.11)</td>
<td>0.85 (0.07)</td>
<td>0.86 (0.11)</td>
<td>0.82 (0.15)</td>
<td>0.87 (0.15)</td>
</tr>
</tbody>
</table>

Table 5b Response times (standard deviation) during Lunari face training

<table>
<thead>
<tr>
<th>Bin</th>
<th>n</th>
<th>Day 1</th>
<th>Day 2</th>
<th>Day 3</th>
<th>Day 4</th>
<th>Day 5</th>
<th>Day 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td>1132 (175)</td>
<td>957 (179)</td>
<td>1048 (301)</td>
<td>1058 (355)</td>
<td>1085 (248)</td>
<td>1144 (240)</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>1239 (111)</td>
<td>1274 (224)</td>
<td>1093 (57)</td>
<td>997 (203)</td>
<td>970 (296)</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>1048 (136)</td>
<td>830 (67)</td>
<td>970 (296)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>1007 (172)</td>
<td>1045 (185)</td>
<td>1127 (318)</td>
<td>1058 (355)</td>
<td>1085 (248)</td>
<td>1144 (240)</td>
</tr>
</tbody>
</table>

and we aimed to have sufficient power to replicate the same effect with faces. In other ROIs, where we did not expect a correlation, reporting uncorrected correlations ensures that we are less susceptible to false negatives.

Previous studies in which subject populations are not screened for their face-recognition performance have—perhaps surprisingly—“not” found a correlation between behavioral face performance and the BOLD response to faces in face-selective brain regions (e.g., second pair of bars in Fig. 4, data from McGugin et al. 2015; also see Jiang et al. 2013; Hermann et al. 2015). Here, we can re-examine this relation in a sample selected to increase the range of individual differences in face-recognition ability scores, and in brain regions where reliability of the face response was sufficiently high (Table 3).

In bilateral face-selective regions (FFA1, FFA2, and OFA) and bilateral posterior object-selective parahippocampal regions (PHG1) (excluding bilateral middle object-selective parahippocampal regions where reliability was \( \alpha < 0.50 \)), we computed the correlation between CFMT+ and the subject-specific beta weights for Caucasian faces, regressing out those for Objects (see Table 6 and Fig. 4).

Given the reliability of the CFMT+ (\( \alpha = 0.89 \)) combined with the reliability of the BOLD response to Caucasian faces, we also computed the correlations disattenuated for measurement error (the true correlation expected if measurements were perfect, Muchinsky 1996). Table 6 shows the zero-order correlations and the disattenuated correlations, \( r_{\text{dis}} \). Disattenuated correlations are not suitable for statistical hypothesis testing but they indicate when observed correlations may be low primarily due to measurement error and provide a better estimate of true effect sizes. The only face-selective region to demonstrate a significant correlation between CFMT+ scores and BOLD face selectivity was rFFA2 (\( r_{24} = 0.35, P = 0.041; r_{\text{dis}} = 0.42 \)). (See Fig. 5 and Table 6 values in parentheses for face effects at the 27 mm\(^3\) peak of face-selectivity.)

Correlation of Lunari Face Recognition and Neural Selectivity for Lunari

Next, we examined the relation between post-test behavioral performance with Lunari faces and the BOLD response to Lunari faces (regressing out the BOLD response to Objects). Using the Lunari post-test is the closest equivalent to correlating performance on a test for a category like cars in a sample of subjects selected to vary in their ability with this category (i.e., it is a measure of performance, not of learning). We tested this relation for regions where the reliability of the BOLD response to Lunari faces was \( \alpha > 0.50 \); right FFA1, right FFA2, left FFA1, left OFA, and left PHG1. Results are given in Table 7 (also see Fig. 4). Right FFA2 showed a significant correlation between post-test performance and selectivity for Lunari faces (\( r_{24} = 0.40, P = 0.02, r_{\text{dis}} = 0.51 \)), and the same effect was marginally significant in right FFA1 (\( r_{22} = 0.33, P = 0.06, r_{\text{dis}} = 0.47 \); Table 7).

Not surprisingly, behavioral performance with Caucasian faces (CFMT+) and Lunari faces (post-test) were related (\( r_{24} = 0.40, P = 0.03 \)), but this correlation accounts for only 16% of the variance and so the 2 tests may explain different parts of the variance in neural response to Lunari faces.

Accordingly, we tested the extent to which the Lunari post-test predicted the neural response to Lunari faces after controlling for other factors in 2 different models. Model one partialed out the linear contributions of CFMT+ performance and Lunari pretest, whereas in Model 2, we assumed that ability could have an influence as a function of amount of training, and controlled for the interaction between CFMT+ performance with bin number, as well as the linear contribution of Lunari pretest. With regard to the assumptions tested in Model 2, note that we found no evidence using improvement in performance during training session 1 (the only one that all subjects performed) or improvements in accuracy or response times on the Lunari task that Bin number (as determined by ability) in any way predicted differential amounts of learning (correlations with Bin number were all lower then 0.05, n.s.). Therefore, while it would be reasonable to expect people with better face-recognition ability would learn more from experience, this was not observed here. Nonetheless, these learning estimates were based on only a small amount of data and should be interpreted with caution.

By regressing out pretraining tests, we control for the influence of pre-existing ability to estimate the influence of the Lunari training by itself. By regressing out the interaction...
between ability and training in Model 2, we control for the possibility that pre-existing ability predicts neural responses to Lunari faces as a function of training. In both cases, we find that the correlation between Lunari post-test in the center of rFFA2 remains as strong and significant (it is no longer significant in the larger ROI).

Whole-Brain Analyses

We conducted whole-brain analyses, though they are inherently less powerful than analyses in individually localized ROIs as a result of multiple comparisons correction and greater variance created by aligning subject data based on gross anatomical landmarks rather than functionally defined regions of interest. However, the results support individual-level analyses. We employed the same contrasts for whole-FOV analyses as for ROI-based analyses. In whole-FOV analyses only, an initial contrast with an uncorrected threshold of \( P < 0.05 \) was corrected for multiple comparisons using Brain Voyager’s Cluster Threshold Estimator (Forman et al. 1995; Goebel et al. 2006), which uses a Monte Carlo simulation with iterative Gaussian filtering to compute the cluster size yielding a corrected threshold of \( P < 0.05 \) separately for each contrast of interest: BOLD response to Lunari faces and Caucasian faces (regressing out BOLD response to Objects) as a function of Lunari post-test accuracy and CFMT+ Accuracy, respectively.

As can be appreciated from Figure 6, the BOLD response to Lunari faces relative to Objects correlated with Lunari post-test accuracy reveals a significant focus near the Middle Fusiform Sulcus and overlapping with the group-averaged coordinates for right FFA2. The Lunari face effect (red) in rFFA2 also overlaps with the Caucasian face effect (green), represented by the black outline. Correlations between Caucasian faces and CFMT+ as well as between Lunari faces and Lunari post-test accuracy reveal regions between right FFA1 and OFA, though non-overlapping. In contrast to these results from the right hemisphere, in the left hemisphere, we did not observe significant overlap between group-averaged face- or object-selective regions with group statistical maps of Caucasian or Lunari face expertise effects (see Supplementary Fig. S1).

Discussion

Prior studies that reported correlation between object recognition performance and the response to those objects in the FFA were generally motivated to explain why there is face-selectivity...
in FFA in the first place. Differences in experience are thought to account for some of the observed variability in performance with objects such as cars or birds, as well as intentionally manipulated variability in training studies. Thus, individual differences in experience with objects are thought to modulate both behavioral performance and neural selectivity in the FFA for these objects, supporting the role of experience in driving FFA’s responses (Gauthier et al. 2000; Xu 2009; Harley et al. 2009; Harel et al. 2010; Bilalic et al. 2011; McGugin et al. 2012, 2015). However, a correlation between face-recognition ability or face experience and selectivity in the FFA for faces has been harder to find, despite highly reliable neural estimates for face-selectivity (McGugin and Gauthier 2016) and face-recognition ability (Wilmer et al. 2010). This is illustrated in Figure 4 in the shaded area using the results of a prior study (McGugin et al. 2015) in which car performance predicted selectivity for cars in the FFA, but the same effect was not obtained for faces. Here, we asked whether we could observe a brain-behavior correlation between face-evoked fMRI signal in FFA and behavioral face ability by 1) selecting subjects to maximize the range of face-recognition “abilities”, 2) maximizing the range of variability in “performance” with a novel face category using the multiplicative effect of ability selection and experience, and 3) providing subjects with different amounts of “experience” with a novel face category (an effect estimated by controlling for initial differences in ability). We address each question in turn below.

When we measured selectivity for Caucasian faces after selecting subjects to maximize the range of face-recognition ability scores, we then observed a significant correlation in rFFA2 (third set of bars in Fig. 4). The effect size is greater than that reported in prior work in which the subjects were not selected for performance with faces. To detect a correlation of $r = 0.35$ with a 1-tail test (alpha = 0.05, power = 0.80) would require a sample of 46 subjects (compared with 362 for the effect obtained by Huang et al. (2014)). Importantly, while this correlation does not depend on a manipulation of experience, variability in experience with faces could contribute to the effect. On the behavioral side, gregariousness (Li et al. 2010) and hometown size (Balas and Saville 2015) have been related to face-recognition ability. On the neural side, a developmental fMRI study found that the rFFA2 was the locus of age-related increases in selectivity specifically for own-age faces with which children would have more experience (Golarai et al. 2017). These results suggest that selecting subjects for their ability in face processing might increase the variability from experience, but it is impossible to know the causes of the variability. Another study using the other-race effect in adults to study the role of experience found that the only brain-behavior correlation was in left FFA, and poorer performance with own-race faces was associated with weaker responses to own-race faces (relative to other-race faces; Feng et al. 2011). These results are consistent with other work (e.g., Golby et al. 2001), all failing to find a brain-behavior correlation in right FFA, which is the most consistent locus of expertise effects with nonface objects. This suggests that studying the role of experience in adults with natural categories, such as faces—for which exposure is generally very high—may be particularly difficult.

---

**Table 6** Correlation between CFMT+ and the BOLD response to Caucasian faces, regressing out the BOLD response to Objects.

<table>
<thead>
<tr>
<th></th>
<th>rFFA1</th>
<th>rFFA2</th>
<th>rOFA</th>
<th>rRHG1</th>
<th>lFFA1</th>
<th>lFFA2</th>
<th>IFPA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zero-order correlation*</td>
<td>0.15 (0.04)</td>
<td>0.35* (0.27)</td>
<td>−0.20</td>
<td>−0.26</td>
<td>−0.14 (−0.13)</td>
<td>0.13 (0.28)</td>
<td>−0.07</td>
</tr>
<tr>
<td>Disattenuated correlation</td>
<td>0.18 (0.05)</td>
<td>0.42 (0.32)</td>
<td>−0.26</td>
<td>−0.31</td>
<td>−0.16 (−0.15)</td>
<td>0.16 (0.34)</td>
<td>−0.09</td>
</tr>
</tbody>
</table>

*Values in parentheses are for the 1-voxel, 27 mm³ peak of face-selectivity.

*P < 0.05.
In an attempt to further understand this effect, we combined 1) a selection of subjects based on ability in face-recognition and 2) offering these subjects training as a function of their ability, to maximize the range of performance scores with faces from an artificial race. Using artificial faces should reduce the effect of previous experience to some extent, and prior behavioral work has found that training with these faces influenced the extent to which they were processed holistically, a hallmark of face expertise (Chua et al. 2014). Multiplying the effect of ability and experience could produce a situation not

| Table 7 Correlation between Lunari post-test and the BOLD response to Lunari faces, regressing out various control variables |
|-------------------------------------------------|-----------------|-----------------|-----------------|-----------------|
| Zero-order correlation*                        | rFFA1           | rFFA2           | lFFA2           | lOFA            | lPHG1           |
| Disattenuated correlation                      | 0.33 (0.25)     | 0.40* (0.47)    | −0.01 (0.03)    | 0.26            | −0.06           |
| Partial correlation (CFMT and Lunari pretest regressed out) | 0.30 (0.25)     | 0.24 (0.37)    | 0.01 (−0.04)    | 0.25            | −0.06           |
| Partial correlation (Lunari pretest and CFMT× Bin regressed out) | 0.30 (0.24)     | 0.35 (0.45)    | 0.00 (0.00)    | 0.24            | −0.10           |
| Disattenuated partial correlation              | 0.43 (0.33)     | 0.30 (0.46)    | 0.01 (−0.06)    | 0.36            | −0.08           |

Values in parentheses are for the 1-voxel, 27 mm³ peak of face-selectivity.
*P < 0.05.
Figure 6. Group-average correlation maps depicting the BOLD response to Lunari faces or Caucasian faces relative to Objects, with Lunari post-test accuracy overlaid on an individual's inflated right hemisphere. The maps are shown at a threshold of 0.05, corrected for multiple comparisons using a Monte Carlo simulation. Black lines indicate the overlap regions showing expertise effects for both Lunari faces and Caucasian faces. Group-average coordinates for right FFA1, FFA2, and OFA are marked with circles. Suli labels: Calcarine Sulcus, CaS; Collateral Sulcus, CoS; Middle Fusiform Sulcus, MFS; Occipitotemporal Sulcus, OTS; Middle Temporal Sulcus, MTS; Superior Temporal Sulcus, STS; Precentral Gyrus, PrG; Inferior Frontal Gyrus, IFG.

Unlike what is obtained when selecting subjects on the basis of performance for cars, it is plausible that those individuals who spent the most time learning about cars may also be individuals with a more general high potential for visual learning. In the present work, this resulted in brain-behavior correlations for faces that were significant in rFFA2 and nearly significant in rFFA1. This effect was greatest at the peak of the right FFA2 (where prior work also found the maximal effects for car expertise), approaching the effect size observed for comparable effects with cars in a meta-analysis by McGugin et al. (2014, r = 0.54).

This effect is the maximum correlation we would expect to observe in our study, as it combines ability and experience to create larger differences in performance. We sought to estimate the effect of experience alone by controlling for 2 measures of pretraining ability, the CFMT+ scores with Caucasian faces and the pretest scores with Lunari faces. This is the closest we can achieve here to estimating the effect of training on its own. At the peak of rFFA2, this correlation was 0.41 (average of the 2 models' partial correlations); to detect such a correlation in future work with a 1-tail test (alpha = 0.05, power = 0.80) would require a sample of 33 subjects. Although it is of a similar magnitude, this training effect is hard to compare to that observed in prior training studies with objects in which all subjects received the same amount of training (e.g., Gauthier et al. 1999; Wong et al. 2009). We note that it is possible our behavioral measures underestimate the interaction between ability and training, such that the effects of training on selectivity could still in fact be due to an interaction between training and ability. A design in which ability and amount of training are independent would help answer this question with more certainty.

One additional aspect of our design requires explanation. We measured neural selectivity for Caucasian and artificial-race faces in a task in which we could manipulate attention and clutter (i.e., another image aside from the one that is task relevant was sometimes presented on the screen, see Reddy and Kanwisher (2007) and McGugin et al. (2014) for similar designs). In prior work, we found robust effects of car expertise in the FFAs that were not dampened by clutter (extraneous visual information in the form of to-be-ignored distractors) but that were diminished by reduced attention (selectivity for to-be-ignored cars) and abolished when cars were shown together with faces (McGugin et al. 2014). It was our intention to measure the effects of clutter and attention manipulations with faces in the present work, but due to the difficulty in observing correlations for faces in prior work, we decided to first estimate the reliability of the category-specific BOLD response in each condition. Unfortunately, the reliability of face-selective responses was low across FFAs during the cluttered conditions, especially for the artificial-race faces (see Supplementary Text). When neural effects are not reliable, they cannot be expected to correlate with behavior. For this reason, we do not address the effects of clutter and attention and only report and analyze selectivity in the condition where only one object is shown on each trial. In addition, even in this condition, face selectivity showed very poor reliability in some areas outside of FFA, especially for the artificial Lunari faces. fMRI studies that report correlations with behavior rarely report the reliability of the BOLD effects. The present results illustrate that it is important to verify the reliability of specific measurements in each ROI and condition.

There are limitations to the current work, including the fact that estimating the role of experience by regressing out ability scores that were by design confounded with the amount of training may lead to an underestimate of training effects. Experience effects may also be limited by our use of artificial faces, since familiarity with faces in general may create a ceiling effect for what can be learned. The limited reliability of the responses to Lunari faces in areas other than the right FFAs limits what we can say about experience outside this region.

In summary, our work points to an interpretation of the selectivity in FFA, especially in right FFA2, as being in large part driven by experience. Some authors attempt to reconcile the body of work on face selectivity and expertise effects for non-face objects in FFA by suggesting that we may be born with the former mechanisms that are later co-opted for other stimulus classes (Adams 2011; Forsythe et al. 2014). Recent work quantifying the heritability of face and object recognition ability found them to be equally high (Shakeshaft and Plomin 2015). Here, we found evidence that this part of the ventral visual stream is also as sensitive to experience with faces as it appears to be for objects. Comparing effect sizes between different areas can be difficult, but consistent with other work pointing to the middle FFA as more sensitive to experience than its posterior counterpart (Golarai et al. 2017; McGugin et al. 2014, 2015), it is interesting that here we found the most robust brain-behavior correlations in the right FFA2.

Supplementary Material
Supplementary material is available at Cerebral Cortex online.

Notes
Conflict of Interest: None declared.