Multivoxel Object Representations in Adult Human Visual Cortex Are Flexible: An Associative Learning Study

Mehdi Senoussi[1,2], Isabelle Berry[3,4], Rufin VanRullen[1,2], and Leila Reddy[1,2]

Abstract

■ Learning associations between co-occurring events enables us to extract structure from our environment. Medial-temporal lobe structures are critical for associative learning. However, the role of the ventral visual pathway (VVP) in associative learning is not clear. Do multivoxel object representations in the VVP reflect newly formed associations? We show that VVP multivoxel representations become more similar to each other after human participants learn arbitrary new associations between pairs of unrelated objects (faces, houses, cars, chairs). Participants were scanned before and after 15 days of associative learning. To evaluate how object representations changed, a classifier was trained on discriminating two nonassociated categories (e.g., faces/houses) and tested on discriminating their paired associates (e.g., cars/chairs). Because the associations were arbitrary and counterbalanced across participants, there was initially no particular reason for this cross-classification decision to tend toward either alternative. Nonetheless, after learning, cross-classification performance increased in the VVP (but not hippocampus), on average by 3.3%, with some voxels showing increases of up to 10%. For example, a chair multivoxel representation that initially resembled neither face nor house representations was, after learning, classified as more similar to that of faces for participants who associated chairs with faces and to that of houses for participants who associated chairs with houses. Additionally, learning produced long-lasting perceptual consequences. In a behavioral priming experiment performed several months later, the change in cross-classification performance was correlated with the degree of priming. Thus, VVP multivoxel representations are not static but become more similar to each other after associative learning. ■

INTRODUCTION

We can rapidly and accurately detect and categorize objects even when they are flashed for just a fraction of a second. This astonishing ability relies on the ventral visual pathway (VVP), a neural system that extends from the occipital cortex to lateral and ventral regions of the temporal lobe (Grill-Spector, 2003). The VVP is not organized in a homogenous fashion (Grill-Spector & Malach, 2004). Instead, this expanse of cortex is dotted with several smaller regions that respond preferentially to specific classes of stimuli (e.g., faces, places, objects, or bodies; Downing, Jiang, Shuman, & Kanwisher, 2001; Epstein & Kanwisher, 1998; Kanwisher, McDermott, & Chun, 1997). This underlying architecture is remarkably consistent across normal, healthy participants (Haxby et al., 2011).

Object category representations in the VVP can be described at two different levels: in the activity of large-scale multivoxel patterns (MVPs) or at the level of the object selectivity of individual neurons (Reddy & Kanwisher, 2006). Although it is difficult to measure the selectivity of single neurons in the human brain, it is now well established that object category information is also reflected in the large-scale MVPs of activity that can be recorded with fMRI. Indeed, decoding studies have shown that category information is explicit in these response patterns (Op de Beeck, Brants, Baeck, & Wagemans, 2010; Reddy & Kanwisher, 2007; Spiridon & Kanwisher, 2002; Haxby et al., 2001). Here we ask if MVPs for well-learned categories still maintain flexibility related to visual experience in the adult brain.

Specifically, in this study, we directly test if large-scale representations for highly familiar categories in the VVP become more similar to each other when pairs of categories are behaviorally associated through extensive training. At the neuronal level, anterior ventral temporal cortex and medial-temporal lobe (MTL) structures have been implicated in associative learning in both monkeys (Wirth et al., 2003; Messinger, Squire, Zola, & Albright, 2001; Miyashita & Chang, 1988) and humans (Ison, Quian Quiroga, & Fried, 2015; Reddy et al., 2015). However, here we show that preexisting multivoxel representations for familiar objects (faces, houses, chairs, cars) in ventral visual cortex shift in a concerted way in a high-dimensional multivoxel space once two categories become perceptually related.

We scanned 20 observers before and after they learned arbitrary associations between different object categories (faces, houses, cars, chairs) and investigated changes in the large-scale category representations with MVP analysis

[1]Université de Toulouse, [2]CNRS, CerCo, Toulouse, France, [3]Inserm Imagerie cérébrale et handicaps neurologiques UMR 825, Toulouse, France, [4]Centre Hospitalier Universitaire de Toulouse Pôle Neurosciences CHU Purpan

Journal of Cognitive Neuroscience 28:6, pp. 852–868
doi:10.1162/jocn_a_00933

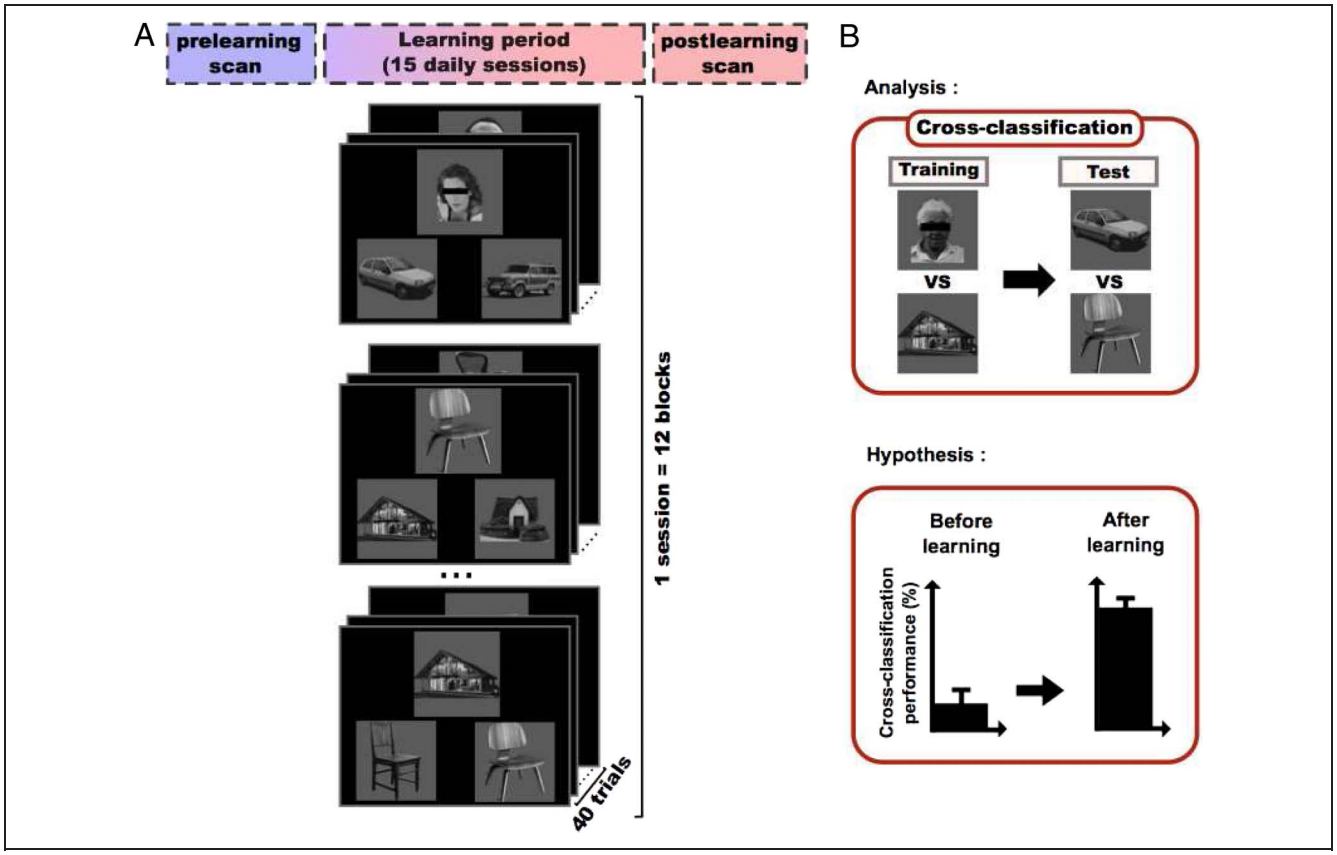

Figure 1. Experimental protocol and hypothesis. (A) Each participant followed a three-step procedure. In the first step, participants performed a prelearning scan in which they viewed blocks of faces, houses, chairs, cars, and scrambled images. Next, in 15 daily sessions, participants performed a learning task in which they learned arbitrary associations between members of the different categories. In this example, faces are paired with cars and houses with chairs. Category pairings were counterbalanced across subjects. Each learning session consisted of 12 blocks of 40 trials. On each trial, participants were presented with a main stimulus (e.g., a face) and two choice stimuli from the associated category (e.g., two cars) and had to decide which of the choice stimuli was paired with the main stimulus (by pressing the left or right arrow keys on the keyboard). After the learning sessions, participants performed a postlearning scan that was identical to the prelearning scan except that the block order was randomized. (B) To evaluate the similarity between category representations before and after learning we used a cross-classification procedure with the searchlight method. An SVM classifier was trained to distinguish between two categories and tested on their associated categories. We hypothesized that after learning, we would see an increase in cross-classification performance suggesting that the multivoxel representations of the paired categories had become more similar to each other.

methods. In particular, we trained a support vector machine (SVM) classifier to discriminate between two nonassociated object categories (e.g., houses vs. faces) and then tested it on discriminating between their paired associates (e.g., cars vs. chairs). We hypothesized that after learning we would see an increase in this cross-classification performance. Because a classification decision reflects the distance and the relative position of test patterns in a multidimensional space, an increase in cross-classification performance after learning would imply that the representations of the paired categories had moved in a high-dimensional multivoxel space or, equivalently, had become more similar to each other.

Using this cross-classification approach, we found an increase in decoding performance after learning, suggesting that large-scale fMRI response patterns in the VVP for associated object categories become more similar to each other. In other words, in an example participant who associated faces with chairs and houses with cars, face MVPs became more similar to chair MVPs and house MVPs became more similar to car MVPs after learning. This shift in category representations had perceptual consequences, as measured by a behavioral priming task performed several months after the associations had been learned. Specifically, we found that a given category facilitated the processing of its paired associate relative to the processing of a nonassociated category. In addition, this priming effect was correlated across participants with the overall amount by which the category representations shifted as a result of learning.

METHODS

Participants and Stimuli

Twenty-one participants were recruited for this study (10 women, mean age = 24 years, range = 19–35 years). One participant was excluded from the study because of excessive motion in the scanner. All participants had normal or corrected-to-normal vision and reported no

history of neurological problems. All participants provided written informed consent and received monetary compensation for their participation. The local ethics committee for human experimentation approved all procedures.

Ten stimuli from each of four categories (faces, houses, chairs, cars) were gathered from different sources on the Internet. These images were then transformed to gray-scale and pasted on a 500 × 500 pixels gray canvas. To avoid low-level category confounds, we normalized categories in luminance, contrast, and size. We then generated a scrambled version of each image for the functional ROI localizers.

Experimental Protocol

The experimental protocol consisted of three phases: a prelearning fMRI scan, an associative learning task outside the scanner over 15 days, and a postlearning fMRI scan.

During the fMRI scans, stimuli were presented with the VisionEgg toolbox (Straw, 2008). Each fMRI run consisted of four blocks each of the four categories (faces, houses, cars, and chairs) and scrambled images and five blocks of fixation. Each block was 16 sec long. The fixation blocks occurred after every five visual stimulation blocks. In each visual stimulation block, 16 stimuli were presented, each for 800 msec followed by an ISI of 200 msec. Participants were instructed to press a button when the same image was presented on two successive trials (1-back task). Each fMRI session consisted of eight runs that lasted approximately 6 min and 45 sec each. The pre- and postlearning fMRI sessions were identical, except for the block and stimulus order, which were randomized in each run.

In between the fMRI sessions, participants underwent 15 daily learning sessions during which they learned associations between exemplars of the object categories (e.g., each face was associated with a given car, and each house with a given chair). Each 20-min session consisted of 12 blocks of 40 trials. Each trial lasted up to 3 sec with an intertrial interval of 0.750 sec. On each trial, participants were presented with a main stimulus (e.g., a chair) and two choice stimuli (e.g., two houses) and had to decide (by pressing one of two keys on the keyboard) which of the choice stimuli was the correct associate of the main stimulus (Figure 1A). Exemplars of each category served as the main stimulus or choice stimuli on different blocks. Ten exemplars per category were used. Learning was achieved by trial and error, and negative auditory feedback was provided on incorrect trials. The category pairings were counterbalanced across participants: Half of the participants associated faces with cars and houses with chairs, and the other half associated faces with chairs and houses with cars.

Priming Experiment

The priming experiment was performed on average 14.1 months after the postlearning fMRI scan on 14 of the original 20 participants. Before participants performed the priming experiment, they underwent three training sessions on the main associative learning paradigm. They then performed two sessions of the priming experiment on two days.

To avoid low-level priming effects, we equalized all stimuli in the Fourier amplitude spectrum. On each trial of the priming experiment, participants were presented with a prime stimulus for 100 msec followed by a target stimulus for 2 sec and instructed to report the category of the target stimulus (Figure 2). The intertrial interval was 1000 msec, with a jitter of 500, 750, or 1000 msec. After each trial, the fixation cross turned to a dash for 1 sec and turned back to a cross to signal the beginning of the next trial. The prime and target stimuli were exemplars of the four object categories (faces, houses, chairs, cars). Within a block of trials, only two categories were targets (e.g., cars and chairs in blocks when participants were asked to

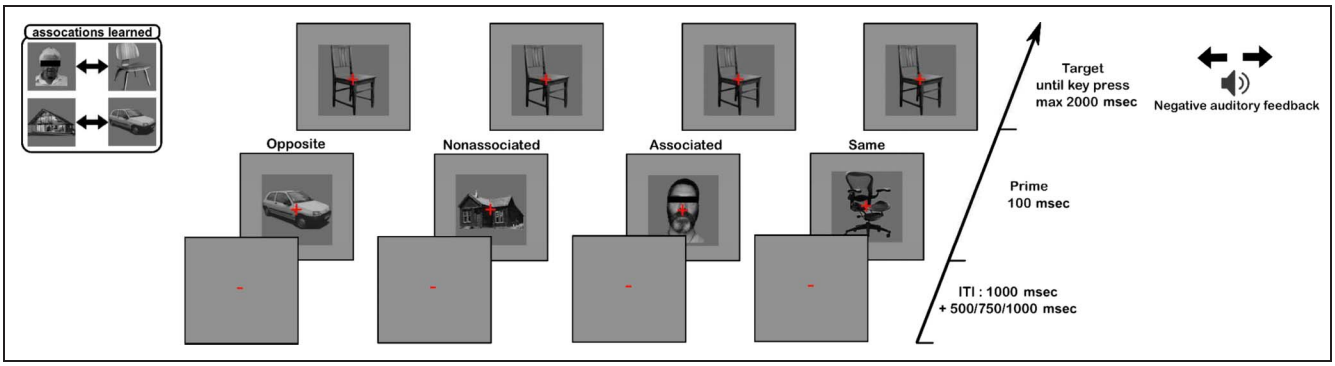

Figure 2. Behavioral priming task experiment design: On each trial of the priming experiment, participants were presented with a prime stimulus followed by a target stimulus and instructed to report the category of the target stimulus. The prime and target stimuli were exemplars of the four object categories (faces, houses, chairs, cars). There were four types of trials: When the primes and targets were different exemplars from the same category ("same" trials), when the prime and target were from opposite categories with respect to the category discrimination task ("opposite" trials), and when the prime and target were from associated/nonassociated categories.

discriminate cars from chairs), whereas all four categories could serve as primes. There were four types of trials: when the prime and target category matched ("same" trials), when the prime and target were from opposite categories with respect to the category discrimination task (e.g., the prime was a car and the target was a chair in a block when participants were instructed to discriminate cars from chairs, "opposite" trials), and when the prime and target were from associated/nonassociated categories (e.g., faces/houses, "associated"/"nonassociated" trials). For "associated" trials, the primes and targets were the pairs learned during the associative learning paradigm, for example, a participant who had learned to associate face1 with chair5 was presented with face1 as a prime when chair5 was the target on "associated" trials, in a block where participants were instructed to discriminate cars from chairs. Participants were instructed to respond as fast as possible on each trial. Each participant performed eight blocks of 250 trials. Trials were randomized within each block. Participants performed the priming experiment over 2 days. On the first day, the targets were cars and chairs, each with their own response button (left and right, respectively). On the second day, the targets were faces and houses, each with their own response button (up and down, respectively). We chose this design to avoid confusing participants by switching instructions within a single experiment session.

fMRI Data Acquisition and Analysis

fMRI data were collected on a 3T Philips (Amsterdam, The Netherlands) ACHIEVA scanner (gradient-echo pulse sequence, repetition time = 2 sec, echo time [TE] = 35 msec, 30 slices with a 32-channel head coil, slice thickness = 2 mm, in-plane voxel dimensions 1.88 × 1.88 mm). The slices were positioned to cover the entire temporal and occipital lobes. High-resolution anatomical images were also acquired per participant (1 × 1 × 1 mm voxels, repetition time = 8.13 msec, TE = 3.74 msec, 170 sagittal slices). Data analysis was performed with FreeSurfer and the FreeSurfer Functional Analysis Stream (FS-FAST; surfer.nmr.mgh.harvard.edu), custom Matlab scripts, and the PyMVPA toolbox (www.pymvpa.org/; Hanke et al., 2009). Similar results were also obtained with the SearchMight Toolbox (www.princeton.edu/~fpereira/searchmight/).

Preprocessing followed the FS-FAST processing stream. All images were motion-corrected (using AFNI with standard parameters), slice time-corrected, intensity-normalized, and smoothed with a 3-mm FWHM Gaussian kernel. We then estimated the beta weights using a general linear model (GLM) for the five stimulus conditions (faces, houses, cars, chairs, and scrambled) in each participant. The betas were computed on whole-run data. There were eight runs in each scan session and four blocks of 16 sec of each condition in each run. We obtained eight beta images per condition (i.e., one from each run) from each scanning session from the FS-FAST processing stream. The GLM fit the hemodynamic response with a gamma function (delta = 2.25, tau = 1.25) and modeled the drift with an order 1 polynomial. For all other parameters of the GLM, we used the default settings from FS-FAST. Finally, the beta-weight volumes were normalized on the MNI305 brain, and we used these volumes as inputs for the searchlight analysis. Similar results were obtained when the searchlight analysis was performed in the native space of each participant.

ROIs

ROIs were defined manually in each participant's native space using an independent analysis. Fusiform face area (FFA) was defined as the set of contiguous voxels in the fusiform gyrus that exhibited greater activation for faces than houses ($p < 10^{-5}$, uncorrected). Parahippocampal place area (PPA) was defined as the set of contiguous voxels in the parahippocampal gyrus that exhibited greater activation for houses than faces ($p < 10^{-5}$, uncorrected). lateral occipital complex (LOC) was defined as the set of voxels in the inferior occipital and temporal cortices that exhibited greater activation for cars and chairs than scrambled images ($p < 10^{-5}$, uncorrected). The anterior and posterior subdivisions of LOC (lateral occipital [LO] and posterior fusiform [pF]) were also identified for each participant. The hippocampus, V1, and V2 were defined using anatomical landmarks for each participant in FreeSurfer. The average ROIs displayed in Figure 7 were computed by selecting voxels that were common to at least 60% of the ROIs defined in individual participants. Note that the ROI analyses were performed in each participant's individual ROIs, and the average ROI is used for display purposes only.

Multivariate Analysis

The searchlight analysis was performed with the *CrossValidation*, *HalfPartitioner*, *LinearCSVMC*, and *sphere_searchlight* functions of the PyMVPA toolbox using default settings. A linear SVM with default settings from the PyMVPA toolbox was used to perform a cross-classification analysis within each searchlight. We used searchlights of different radii (1–14 voxels) that we moved along the MNI305 volumes. For each participant, within each searchlight, an SVM classifier was trained on the fMRI patterns for two nonassociated categories for that participant (e.g., faces vs. houses) and tested on the corresponding associated categories (e.g., cars vs. chairs). Additionally, the symmetric classification was also performed (i.e., in the example here, a car–chair classifier was tested on a face–house discrimination). The average of the two classification scores was reported as the cross-classification performance for the voxel at the center of the searchlight. The input to the classifiers were eight MVPs for each condition. For example, when training a classifier on a face versus house discrimination and testing it on a car versus chair discrimination, the

classifier was trained on eight MVPs of face betas and eight MVPs of house betas and tested on eight MVPs of car betas and eight MVPs of chair betas.

The same stimuli and data sets were used for the experimental sessions and for defining the functional ROIs. However, our analysis is free of the double-dipping problem (Kriegeskorte, Simmons, Bellgowan, & Baker, 2009) because orthogonal contrasts were used in defining the ROIs versus in the cross-classification analysis. For instance, when defining the FFA, we used a faces–houses contrast. On the other hand, the cross-classification analysis tested a face/house classifier on a car/chair discrimination. Defining our FFA with a face–house contrast guarantees that a face/house classifier in these voxels would perform superbly on a face/house discrimination of the same data (i.e., a circular analysis). However, there is no reason for performance of the face/house classifier on a car/chair discrimination task to benefit from this method of voxel selection.

Correct or incorrect classification depended on the association learned by the particular participant. For example, for participants who learned face–car and house–chair associations, cross-classification would be deemed as correct if the face–house classifier classified the MVP elicited by cars as faces and the MVP elicited by chairs as houses. This procedure produces cross-classification performances ranging from 0 to 1. A performance of 1 means that the classifier always considered patterns of associated categories as being more similar, a performance of 0 means that it always considered patterns of nonassociated categories as being more similar, and a score of 0.5 means that the classifier did not have any bias between the categories. This procedure could be done in two ways, because there were two pairs of associations: training the classifier on faces and houses and testing it on cars and chairs, or training it on cars and chairs and testing it on faces and houses. The results of these two analyses were equivalent so the final cross-classification performance values were averaged across the two analyses.

Note that the cross-classification approach might be a more sensitive test of learning-induced flexibility than a direct classification test on the associated category pairs because a priori, a chair pattern should fall roughly halfway between a face and a house pattern (i.e., 50% classification performance), so a small shift of the chair pattern toward the face pattern could result in a sizeable change in cross-classification performance. On the other hand, if face and chair patterns become more similar in a multidimensional space, they might still be far enough apart that a direct face/chair classifier would never confuse a chair with a face and thus learning would not seem to modify classification accuracy.

The searchlight analysis was performed across the entire scanned functional volume as well as in the specific

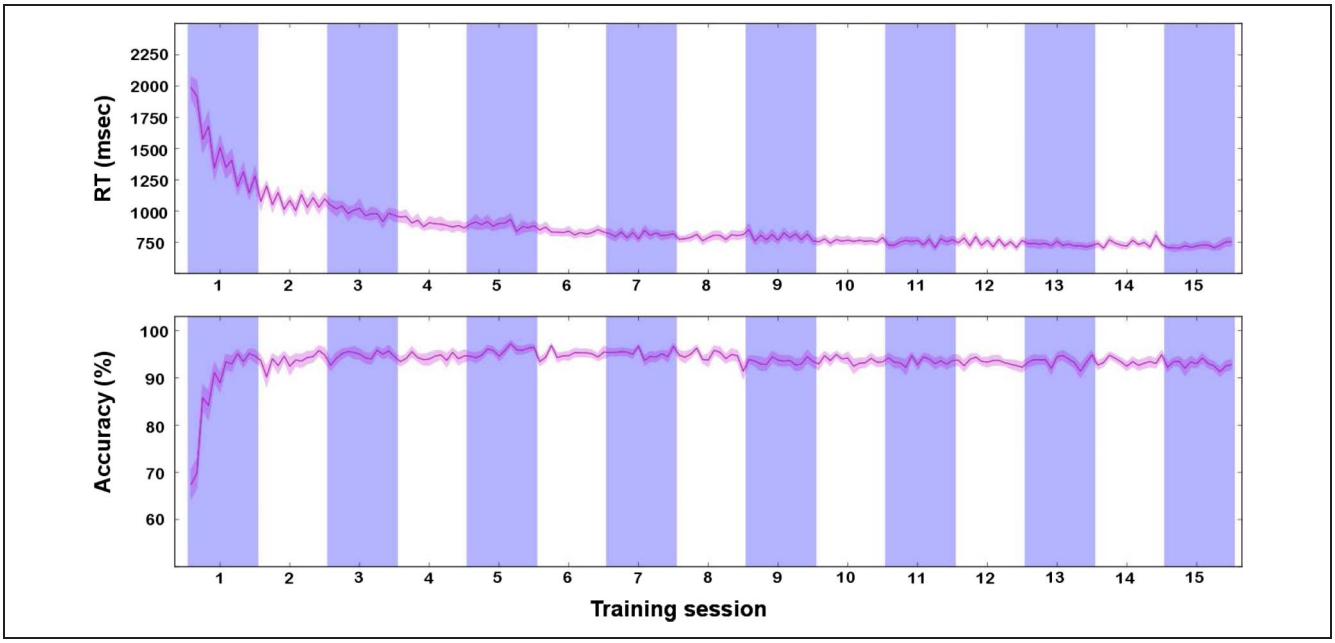

Figure 3. Behavioral results during learning. Each participant performed 15 learning sessions outside the scanner. The RTs and accuracies in each session are shown here (individual sessions indicated by the blue and white areas). RTs (top plot) decreased steadily (one-way, random-effects ANOVA on log(RT): $F(14, 266) = 128.61, p < 10^{-6}$), and stabilized after the tenth session (post hoc Tukey's HSD ($p < .05$)). For statistical tests only (but not for display purposes), the RTs were log-transformed to satisfy the constraints of normality. Accuracy (bottom plot) was computed for each session as the proportion of trials where the participant responded correctly. Accuracy on the learning task was at or above 90% by the end of the first learning session for most participants (19 of 20) and then stabilized by the second session (one-way, random-effects ANOVA: $F(14, 266) = 9.38, p < 10^{-6}$, post hoc Tukey's HSD ($p < .05$)). The pink lines correspond to the average across participants of all trials of each block in each session, and the shaded area is the *SEM*.

ROIs defined for each participant. For the ROI-specific analyses, we used a fixed-size searchlight of radius 3 voxels (i.e., a searchlight consisting of 123 voxels).

Statistical Analysis

The statistical significance of the difference between the pre- and postlearning distributions was evaluated using two-tailed one-sample t tests across participants. To assess the statistical significance of the voxels that showed the largest cross-classification shifts in Figure 4 (and in the corresponding surface maps in Figure 7), we used a nonparametric test in which we shuffled the labels of the pre- and postlearning sessions for each voxel and for each participant independently to simulate the null hypothesis that there was no difference between these sessions for each voxel. The surrogate distributions were computed 2000 times per participant. The p value of each voxel was assigned by comparing this voxel's cross-classification shift to the corresponding surrogate values (i.e., 2000 iterations × 78,842 voxels).

Correlation Analysis

In the pF, FFA, PPA, and LO, a searchlight of radius 3 voxels, centered on each voxel, was trained and tested on discriminating the four categories (faces, houses, chairs, cars) from each other prior to learning. This four-way classification analysis was performed individually for each participant in the MNI305 space and then averaged across participants to obtain an average four-way classification performance value for each voxel. This average four-way classification performance value was then correlated with the change in cross-classification performance of each voxel (also averaged across participants to obtain one performance value per voxel). The parameters of the four-way classifier were identical to the cross-classification classifier (see above). The four-way classifier was trained on blocks of data from all runs but one and tested on the remaining run (leave-one-run-out cross-validation). This correlation analysis was performed on the average ROIs computed by selecting voxels that were common to at least 60% of the ROIs defined in individual participants.

To make the correlations comparable across ROIs, we equalized the number of voxels in each ROI before computing the correlation value. Specifically, we resampled each ROI 100,000 times, each time randomly choosing 162 voxels (that corresponded to the size of the smallest ROI, pF) and computing the Pearson's r value in each resample. The reported r^2 values correspond to the square of the average r values of these resamples.

RESULTS

Twenty observers were scanned before and after they learned arbitrary associations between pairs of different object categories (Figure 1A). During these pre- and postlearning fMRI scans, the participants viewed 10 exemplars each of faces, houses, chairs, cars, and scrambled images in different blocks. Participants performed a 1-back task, in which they responded if the same image had been presented on two successive trials. Note that, in the scanner, the image presentation order and the 1-back behavioral task were independent of the associations learned by the participants outside the scanner. These scans simply allowed us to obtain pre- and postlearning MVPs for the four object categories.

In between these two scan sessions, participants learned arbitrary associations between different exemplars of the four object categories (e.g., each face was associated with a car/each house with a chair; Figure 1A). Most participants achieved greater than 90% accuracy by the end of the first session and continued to improve until behavioral measures of learning stabilized by the tenth session. Participants continued to train even after performance had stabilized (Figure 3).

As mentioned above, we trained an SVM classifier to discriminate between two nonassociated categories (e.g., faces and houses) and tested it on discriminating their paired associates (e.g., cars vs. chairs). We hypothesized that after learning we would see an increase in this cross-classification performance (Figure 1B), suggesting that the multivoxel representations of the paired categories had become more similar to each other. Because we had no strong a priori expectation about where these learning-related changes might occur, we used the searchlight method to explore different areas of the VVP (Kriegeskorte, Goebel, & Bandettini, 2006).

To perform the cross-classification procedure, we realigned each participant's functional volume to the MNI305 brain to make comparisons across participants. We moved a spherical searchlight along each participant's realigned functional volume and, at each voxel, calculated the cross-classification performance from the MVPs falling within the searchlight centered on that voxel. More specifically, for an example participant who had learned to associate faces with cars and houses with chairs, we tested the performance of a face–house classifier on car–chair discrimination and a car–chair classifier on face–house discrimination within the searchlight. Note that there is no "correct" answer for either of these classifiers, as the associations were arbitrarily determined—we simply assumed that, faced with a meaningless choice, the classifier would tend to choose the label of the associated category. The average of these two classification scores was the cross-classification score attributed to the voxel at the searchlight center. We performed this analysis separately on the MVPs from the pre- and postlearning scans and evaluated how cross-classification performance changed after learning.

The pre- and postlearning distributions of cross-classification performance across all the voxels in the scanned volume were averaged across the 20 participants and are shown in Figure 4A. To obtain optimal classification

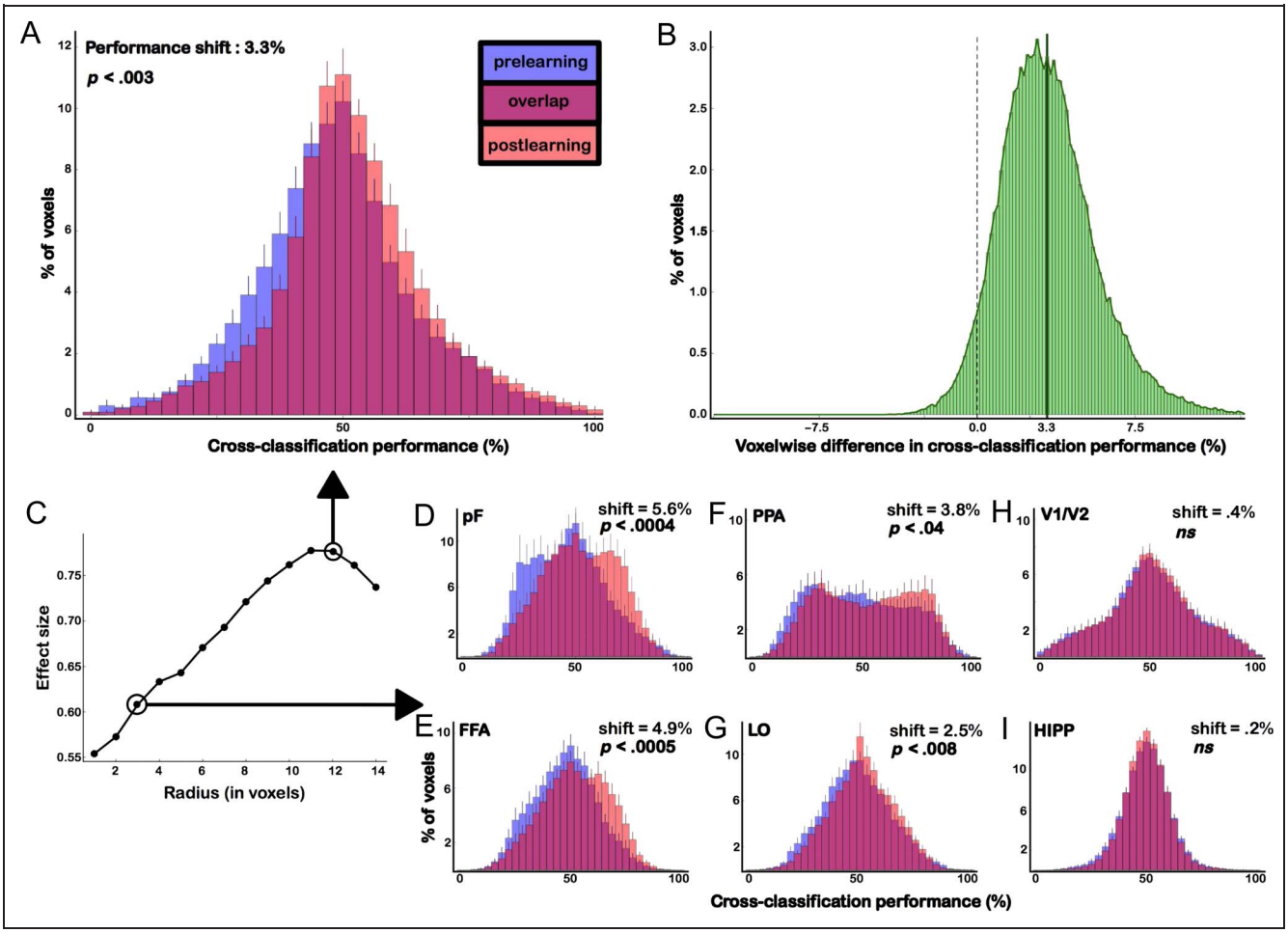

Figure 4. Cross-classification performance before and after learning. (A) Histograms showing the distribution of cross-classification performance across all voxels in the scanned volume, obtained with a searchlight of radius 12 voxels, averaged across 20 participants. Prelearning performance values are in blue, and postlearning performance values are in pink. The overlap between the pre- and postlearning distributions is shown in purple. Across all voxels, there was a significant increase in performance after learning (3.3 ± 0.95%; ($t(19) = 3.39$, $p < .005$). (B) Distribution of the voxelwise difference between the pre- and postlearning performance values. As expected from A, the average voxelwise difference was 3.3%. However, although the shift was absent or only moderate for some voxels, a number of voxels shifted by more than 10% on average. (C) The effect size of the difference between pre- and postlearning cross-classification distributions obtained with searchlights of different radii. (D–G) Same as in A for different ROIs, obtained with a searchlight of radius 3 voxels. The significance and shift of the difference between the pre- and postlearning distributions are indicated for each panel.

performance, the size of the searchlight must be commensurate with the size of the region where the effects occur (Kriegeskorte et al., 2006). Accordingly, we tested the effect of varying the searchlight radius on cross-classification performance. In the whole-brain analysis we obtained optimal cross-classification performance with a searchlight of radius 12 voxels (Figure 4A), but similarly significant effects were also obtained with searchlights of other radii from 1 to 14 voxels (Figure 4C). The pre- and postlearning average cross-classification performance values were 48.3% and 51.7%, respectively, and not significantly different from chance levels (50%, $t(19) = 1.55$, $p = .13$ for the prelearning distribution and $t(19) = 1.82$, $p = .08$ for the postlearning distribution). However, between the two learning sessions, we observed a significant increase of 3.3 ± 0.95% (mean ± *SEM*) in the average cross-classification performance over all voxels in the scanned volume. We statistically evaluated this difference in mean cross-classification performance between the two scan sessions using a two-tailed, paired t test of average pre- and postlearning performances (with each participant contributing one global cross-classification performance value to the statistical test, thus avoiding multiple comparisons across voxels or brain regions and warranting the assumption of independence between measurements ($t(19) = 3.39$, $p < .003$). Furthermore, the increase in cross-classification performance after learning was not driven by the type of association learned by participants (Figure 5): Similar increases were observed for the participants who had associated faces with cars and houses with chairs (increase of 3.4 ± 1.4%, two-tailed, paired t test; $t(9) = 2.26$, $p = .05$) as for the participants who had associated faces with chairs and houses with cars (increase of 3.2 ± 1.2%, $t(9) = 2.42$, $p < .04$).

As argued above, the shift in the distribution of cross-classification performance suggests that multivoxel object category patterns become more similar to each other after participants learn associations between the categories. How sparse are these learning-related changes? On the one hand, the representational changes could potentially be highly variable across voxels, with voxels in some areas showing a large increase in performance after learning and others showing no change at all. Alternatively, at the other extreme, every voxel in the scanned volume could shift by the same amount. To determine how specific the learning-induced changes were, we evaluated the distribution of voxelwise differences between the pre- and postlearning classification performances (Figure 4B). As expected from the results in Figure 4A, the average voxelwise increase in cross-classification performance after learning was 3.3%. However, the shift was variable: Some voxels showed an increase in cross-classification performance of more than 10%. Figure 6 shows the scatter of pre- and postlearning cross-classification performances across all voxels, and its relationship to the histograms shown in Figure 4A and B.

We next asked how the voxels that showed the largest shifts in performance were organized in cortex. In other words, did they occur together in localized regions or were they dispersed all across cortex? Some authors have suggested that expertise-related changes might occur in specific ROIs, for example, in the fusiform gyrus (Gauthier, Tarr, Anderson, Skudlarski, & Gore, 1999). In a first step, we thus evaluated how learning affected object representations in functionally defined regions of ventral temporal cortex that are known to be important for processing visual categories. In particular, we identified four functionally defined regions in each participant's native space: the FFA (Kanwisher et al., 1997), the PPA (Epstein & Kanwisher, 1998), and the pF and LO subdivisions of the LOC (Grill-Spector, Kourtzi, & Kanwisher, 2001). In addition, we anatomically identified the hippocampus (because of its implication in the acquisition of new associations), and the early visual cortex (V1 and V2) as a control region. In each of these areas, we performed the same analysis as in Figure 4A (Figure 4D–G). However, because we were considering smaller ROIs, we restricted this analysis to a smaller searchlight of radius 3 voxels. Note that we retained

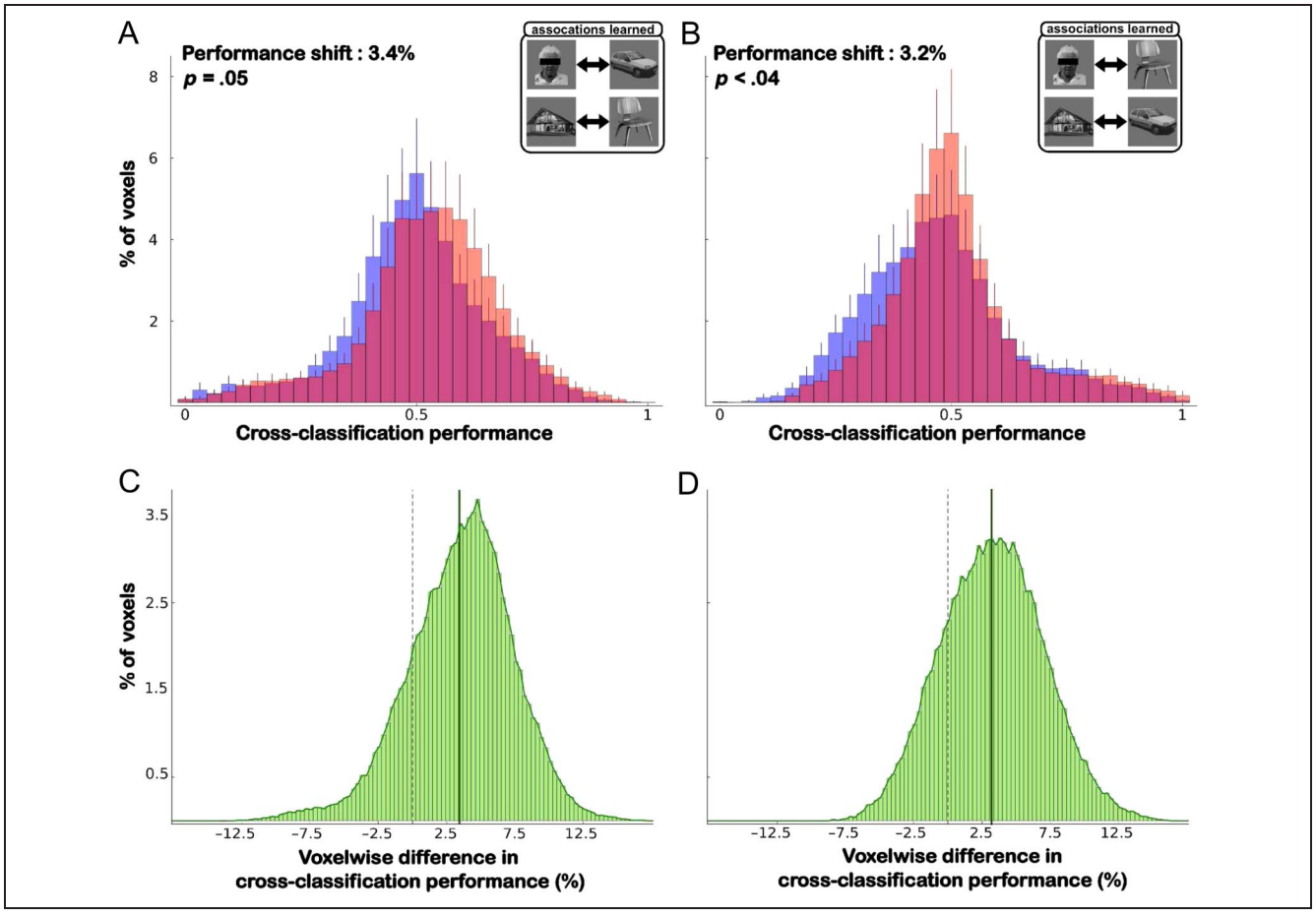

Figure 5. Cross-classification performance by association type. In our group of 20 participants, half the participants associated faces with cars and houses with chairs (Group 1; A), whereas the other half associated faces with chairs and houses with cars (Group 2; B). Both groups of participants showed similar effects of associative learning (independent samples t test, $t(9) = 0.11$, $p = .9$). A and B have the same format as in Figure 4A. C and D have the same format as in Figure 4B.

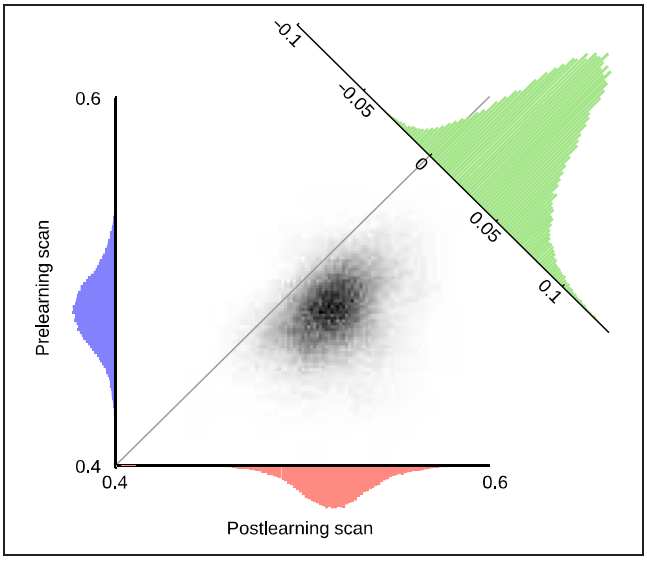

Figure 6. Scatter plot of voxelwise cross-classification performance in the prelearning versus postlearning scan sessions. The blue and pink histograms are the projections of the data on the *y*- and *x*-axes, respectively, and are similar to the data shown in the blue and pink histograms of Figure 4A (save for the fact that here the data points and corresponding histograms represent mean classification performance of each voxel across participants). The green histogram corresponds to the data in Figure 4B and is the projection of the data perpendicular to the diagonal.

a higher-resolution searchlight approach rather than testing a whole ROI classifier, because it is conceivable that over an entire ROI the most informative voxels (i.e., those that will dominate the classifier's decision) may not be those that show the strongest learning effect (and indeed, this possibility was confirmed in a subsequent analysis; see Figures 8 and 9). In that case, a whole ROI classifier may not show any learning-induced change in cross-classification (Figure 9), although individual voxels within the corresponding ROI could have significantly altered their response pattern; the searchlight method, on the other hand, would still reveal the changes in those voxels (Figure 4). We observed a statistically significant increase in cross-classification performance in all ROIs (pF: 5.6 ± 1.2%; $t(19) = 4.33$, $p < .0005$; FFA: 4.9 ± 1.1%; $t(19) = 4.24$, $p < .0005$, PPA: 3.8 ± 1.6%; $t(19) = 2.3$, $p < .05$, LO: 2.5 ± 0.8%; $t(19) = 2.99$, $p < .01$), but not in the hippocampus (0.2 ± 0.8%; $t(19) = 0.33$, $p > .7$) and V1/V2 (0.4 ± 1.2%; $t(19) = 0.72$, $p > .36$).

In a complementary analysis, we asked where the voxels that showed the largest increase in performance were localized. Figure 7 shows the average voxelwise performance differences (obtained with a searchlight of radius 3 voxels) projected on the inflated brain. To assess the statistical significance of the cross-classification shifts, we used a nonparametric test in which we shuffled the labels of the pre- and postlearning sessions for each voxel independently to simulate the null hypothesis that there was no difference between these sessions. Voxels that shifted significantly ($p < .001$, uncorrected) corresponded to a performance shift of at least 9.25% and were clustered principally in the left and right fusiform gyri. In particular, the largest group of these voxels overlapped with the functionally defined left and right FFA and an anterior subdivision of the LOC known as pF (Grill-Spector et al., 1999).

As can be seen in Figures 4 and 7, the learning-induced changes were not uniform within each ROI. Instead, some voxels exhibited greater shifts in cross-classification performance than others. We next investigated what characterized those voxels that showed higher flexibility. We reasoned that flexibility might be inversely related to initial selectivity, that is, that the voxels that originally provided the most information about object category (i.e., the most specialized voxels) would retain their selectivity, whereas the least informative voxels would be most sensitive to category associations during the learning phase. Thus, in each of the previously identified ROIs (pf, FFA, PPA, and LO), we compared the voxelwise increase in cross-classification performance after learning with the ability of that voxel to provide category-specific information before learning (i.e., the performance of a classifier trained on a set of patterns and tested on patterns from the same category). For each voxel, the performance of a four-way classifier (3-voxel radius spherical searchlight centered on that voxel), trained and tested on discriminating the four categories (faces, houses, chairs,

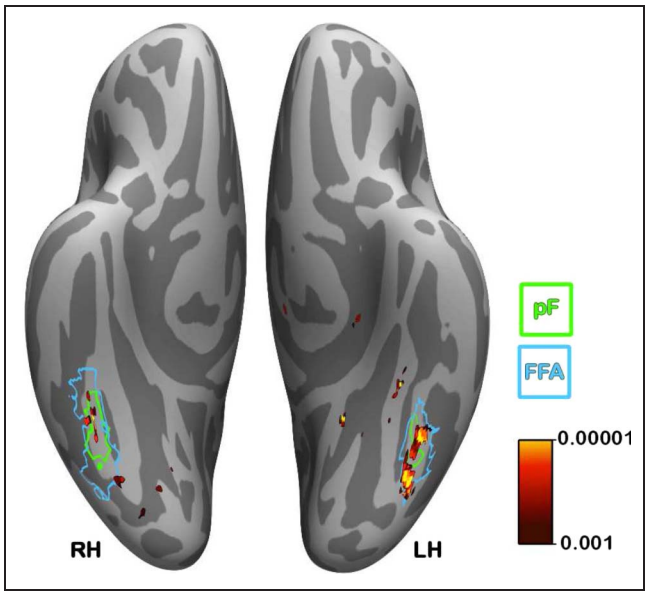

Figure 7. Localization of voxels that showed the largest increase in cross-classification accuracy. The voxels that showed the largest increase in cross-classification accuracy across all participants after learning were in relatively localized regions of the left and right fusiform gyri ($p < .001$ uncorrected; corresponding to an increase in cross-classification accuracy of 9.25% or more). The colorbar corresponds to p values (uncorrected) determined from a nonparametric test. The outline of the functionally defined FFA (averaged across participants) is shown in blue, and the average pF is shown in green.

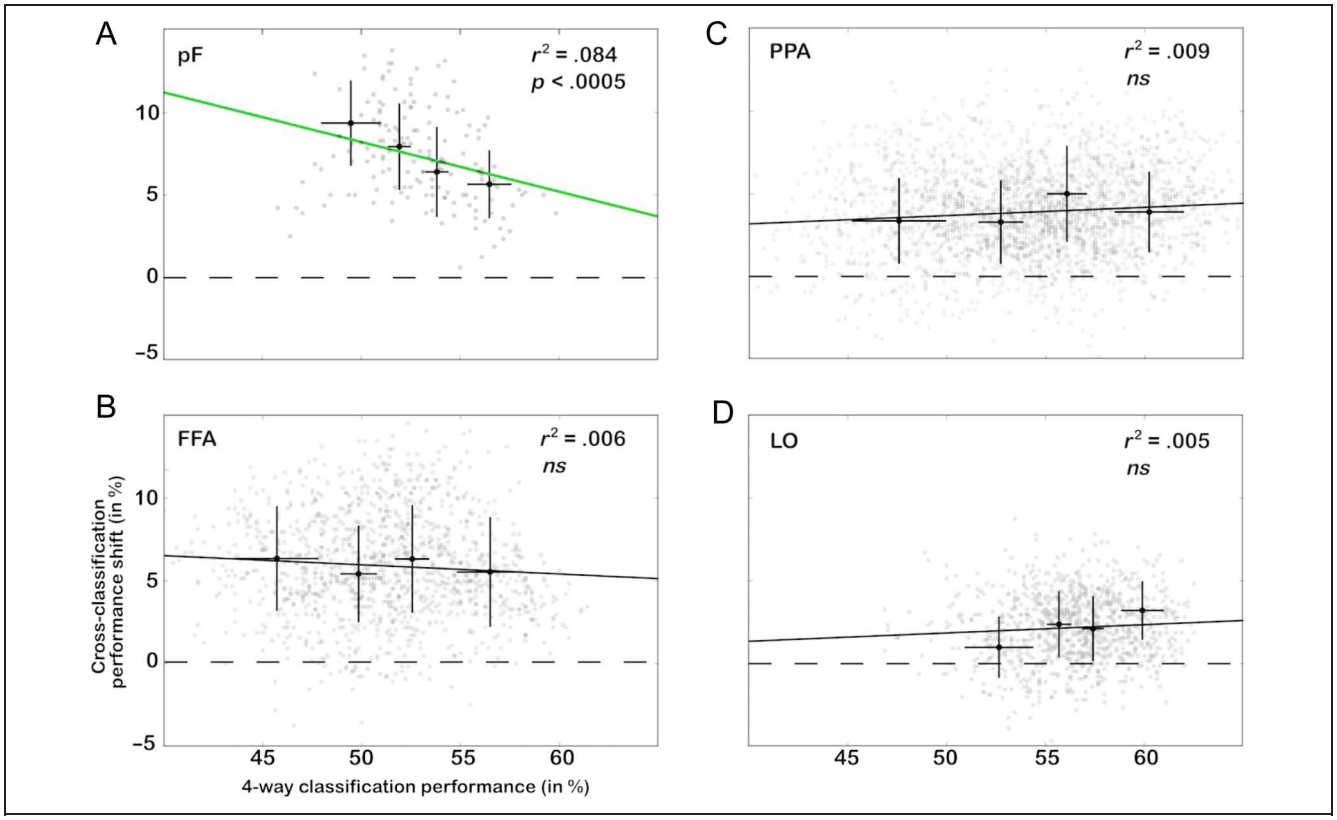

Figure 8. Voxelwise correlation of changes in cross-classification performance (Session 2–Session 1) with the initial performance of a four-way classifier. A four-way classifier (chance at 25%) was trained and tested on prelearning data to discriminate the four object categories from each other (i.e., trained on a set of patterns and tested on patterns of the same categories). Its performance was correlated with the learning-induced changes in cross-classification performance for each voxel in the (A) pF, (B) FFA, (C) PPA, and (D) LO. Increase in cross-classification performance as a result of learning was significantly negatively correlated with the category discrimination performance in area pF (p < .0005). The gray points correspond to the individual voxels in each ROI on which correlations were computed. For visibility only, the voxels were split into quartiles according to four-way classification performance. The mean performance for each quartile is shown by the black points (error bars correspond to *SD* across voxels).

cars) during the first fMRI recording session (prelearning), was correlated with the learning-induced change in cross-classification performance. Consistent with our hypothesis, increase in cross-classification performance was significantly negatively correlated with the initial performance of the four-way classifier in area pF (Figure 8). This finding indicates that, in this ROI, the voxels that exhibited the most flexibility during the learning procedure were the ones with the lowest category-specific information prelearning (albeit four-way classification performance in these voxels was much higher than the 25% chance level; Figure 8A). Note also that, although in pF

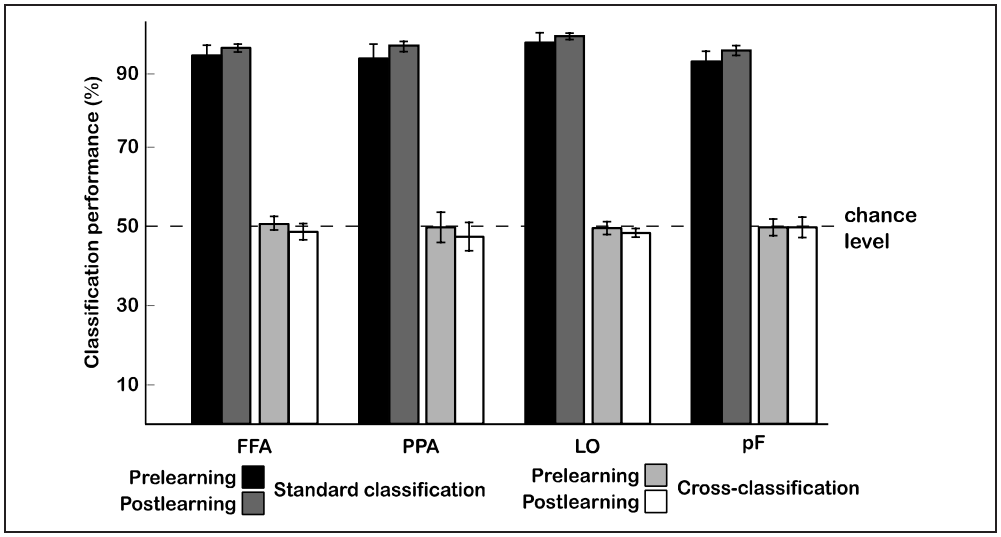

Figure 9. Prelearning and postlearning SVM classification performance in the FFA, PPA, LO, and pF, performed at the level of the entire ROI, that is, without a searchlight method. "Standard classification" refers to the average performance of a face–house (FH) classifier tested on FH discrimination and a car–chair (CC) classifier tested on CC discrimination (using a leave-one-run-out approach). "Cross-classification" refers to the average performance of the FH classifier on CC discrimination and the CC classifier on FH discrimination.

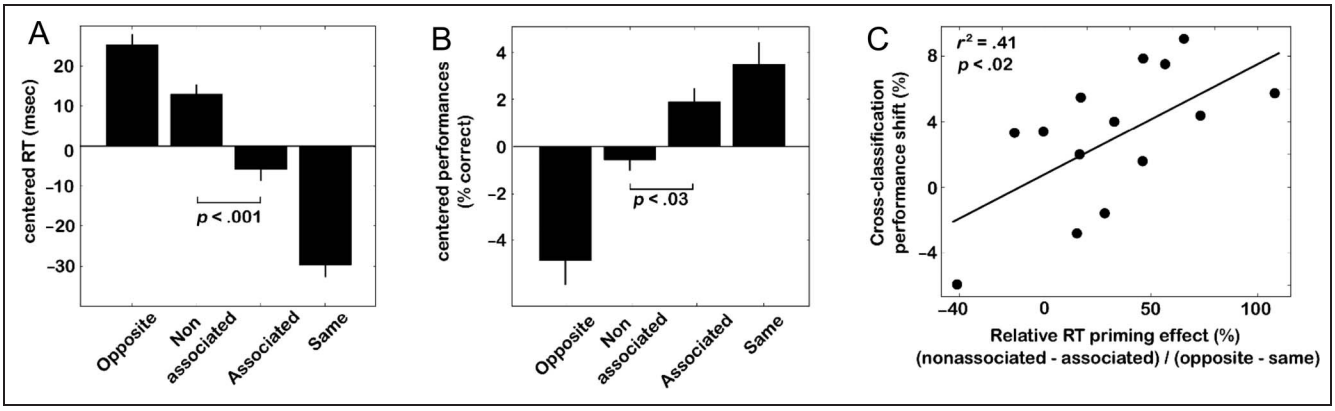

Figure 10. Behavioral performance on a priming task. Participants performed a priming task and were instructed to prioritize response times over accuracy. (A) RTs were significantly shorter (21.4 ± 5.4 msec, mean ± *SEM*, two-tailed, paired *t* test; $t(13) = 4.47$, $p < .001$) for the associated categories versus the nonassociated categories. The magnitude of the priming effect on RTs was approximately 32% of the maximal priming that could be observed between "same" and "opposite" trials (66.3 ± 5.0 msec). For each participant, the mean RT across all conditions was subtracted from each condition to obtain the centered RT displayed here. For statistical tests only (but not for display), the RT data were log-transformed to satisfy the constraints of normality. (B) Although RTs were our main dependent variable, a compatible difference was also observed for accuracies on "associated" versus "nonassociated" trials (2.5 ± 0.9%, two-tailed, paired *t* test; $t(13) = 2.53$, $p < .03$). (C) The priming effect on RTs (expressed relative to the maximal priming) was statistically correlated over the group of participants with the difference in cross-classification performance before and after learning ($r^2 = .41$; $p < .05$).

the voxels with the least category selectivity showed the largest learning effects, at the level of the entire ROI pF itself was highly category selective (Figure 9).

One may question the validity of using a local selectivity measure (the searchlight method) to draw global conclusions about the entire ensemble of recorded voxels across occipital and temporal cortex: global measurements (such as a classifier trained and tested on the entire set of voxels) may appear more appropriate for that purpose. In fact, however, the searchlight method allowed us to obtain a global measure of learning over the whole brain (by averaging across voxels) and to then subsequently hone in on more localized effects. Note that, as opposed to this approach of the searchlight method, a classification analysis performed over all the voxels in the entire scanned volume (or even in a specific ROI, as alluded to above; see Figure 9) could potentially fail to find the voxels that show the biggest changes in cross-classification. For example, a global classifier trained to discriminate faces versus houses across a large swath of cortex would identify the voxels that are the most informative (i.e., category selective) for this face/house discrimination task and disregard the voxels that are the least category selective. However, as we observed in Figure 8, the voxels that were the most prone to learning in pF (i.e., showing the most significant learning effects in a cross-classification task) were precisely the ones that were the least category selective. Thus, although a whole-brain classifier might assign negligible weights to these voxels and consequently fail to identify learning effects, the searchlight method would not because it is constrained to learn from local patterns.

Do the multivoxel representational shifts have perceptual consequences at the behavioral level? In a priming task, performed ~14 months after the associative learning had occurred, we investigated whether perception of one category facilitated the behavioral processing of its associated category, relative to its nonassociated category (Figure 2). An examination of participants' behavioral performance revealed that RTs were significantly shorter (two-tailed, paired *t* test; $t(13) = 4.47$, $p < .001$) on trials when the prime stimulus was a paired associate versus a nonassociate (Figure 10A). The average magnitude of this priming effect (21.4 ± 5.4 msec, mean ± *SEM*) was approximately 32% of the maximal priming (66.3 ± 5.0 msec, mean ± *SEM*) that could be observed between "same" and "opposite" trials. Although RTs were our main dependent variable (because participants were explicitly instructed to prioritize response speed over accuracy), a compatible difference was also present for accuracies on "associated" versus "nonassociated" trials (2.5 ± 0.9%, two-tailed, paired *t* test; $t(13) = 2.53$, $p < .03$), with a priming effect for associated categories that was 22.5% of the corresponding maximal priming (Figure 10B).

Could this priming effect represent a behavioral correlate of the cortical representational shifts observed in the fMRI? In support of this idea, we found that the priming effect on RTs (expressed relative to the maximal priming) was statistically correlated over the group of participants with the difference in cross-classification performance before and after learning ($r^2 = .41$ $p < .05$; 95% confidence interval: $.07 \leq r^2 \leq .75$; Figure 10C). In other words, the participants who had displayed the maximal shifts in multivoxel representations were also those who showed the largest priming effects. Thus, we found that even several months after the associative learning had occurred, changes in neural representations of the associated categories were accompanied by significant and commensurate response priming at the behavioral

level (although, as with all effects based on a correlation analysis, these results cannot provide evidence for a direct link between the changes observed in fMRI and the behavioral priming effects).

DISCUSSION

In this study, we asked how associative learning changes large-scale multivoxel representations in ventral temporal cortex. After learning, we observed an average increase of 3.3% in cross-classification performance of multivoxel category representations, with some voxels showing shifts of up to 10%. Because our experiment used a block design, it remains an open question whether these multivoxel category patterns arise spontaneously in the brain under different testing regimes (Kriegeskorte, Mur, & Bandettini, 2008; Kriegeskorte, Mur, Ruff, et al., 2008). Nevertheless, our results suggest that in conditions where category-specific MVPs can be recorded, the multivoxel representations for associated categories in object-selective cortex become more similar to each other after associations are learned. In a behavioral experiment, we verified the perceptual consequences of the shifts in multivoxel representations several months after the learning had occurred. Not only did paired associates produce significant cross-category priming, but also, the participants who had displayed the maximal shifts in multivoxel representations were those who showed the largest priming effects. Note however that we cannot exclude other factors that might also have contributed to the significant correlation between fMRI effects and behavior, for example, participants' motivation levels and their ability to follow task instructions.

Cross-classification performance after learning was significantly higher than before learning. However, when averaged over all participants and voxels (Figure 4), neither the pre- nor postlearning cross-classification performance values (48.3% and 51.7%, respectively) were significantly different from chance level (50%; $t(19) = 1.55$, $p = .13$, for the prelearning distribution and $t(19) = 1.82$, $p = .08$, for the postlearning distribution). We believe that the initially low performance value was caused by spontaneous biases in category associations occurring in many brain areas. In the PPA, for example, on average about 62% of the chair-category patterns tended to be spontaneously associated with house (rather than face) patterns and cars with faces (rather than houses). Of course, the counterbalanced set of participants was designed to minimize the effects of any such initial bias (because for one half of the participants, this bias would result in lower-than-chance prelearning cross-classification and higher than chance for the other half). However, in our limited participant population, it is not altogether surprising that the initial bias of a few participants could have been overly represented in the grand average (e.g., because of a higher signal-to-noise ratio during scanning or because the relative volume of specific ROIs was bigger in these participants), leading (in our case) to an average prelearning cross-classification below 50%. If we take this initially low value as the chance level (or baseline) for postlearning cross-classification, therefore, we observe a truly significant ($p < .003$) cross-classification improvement due to learning. It must also be emphasized that our findings are not contingent on below-chance prelearning cross-classification: Similar learning-induced improvements were registered for several brain regions and participants whose cross-classification accuracy started off above chance. This can be easily visualized in Figure 6: Even the voxels with the highest initial cross-classification performance demonstrated a learning-induced improvement (i.e., a shift to the right of the diagonal).

The shifts observed in the MVPs could reflect different mechanisms by which object representations change as a result of learning. For instance, the new patterns could reflect a link (or a coactivation) between the two (unchanged) initial representations of the associated categories or signal entirely new representations that combine information about the associated categories. Although it would be interesting to compare what category information is encoded in the initial versus changed representations, we must note that any comparison of MVPs across the two sessions (i.e., training on patterns from one session and testing on patterns from the other) would confound learning effects with pattern and classification differences that are simply due to the fact that the two scan sessions were obtained on different days. However, the finding that, in area pF, the voxels that showed the greatest flexibility during learning were the ones that originally provided the least (albeit still much greater than the 25% chance level) category-specific information (Figure 8) suggests that the voxels that are the most informative in encoding category information mostly preserve their response profiles whereas the least informative voxels are more readily modulated by associative learning.

Object representations in the VVP can be described both at the level of individual neuronal selectivities as well as in large-scale multivoxel activation patterns (Reddy & Kanwisher, 2006). Indeed, in the human brain, MVPs are often used as a proxy for understanding the neuronal codes underlying object representation (Stansbury, Naselaris, & Gallant, 2013; Kriegeskorte, Mur, Ruff, et al., 2008; Haynes & Rees, 2005; Kamitani & Tong, 2005; Carlson, Schrater, & He, 2003; Spiridon & Kanwisher, 2002; Haxby et al., 2001). As explained earlier, the observed increase in cross-classification performance after learning can be described in mathematical terms as a shift of the MVPs in a high-dimensional multivoxel space. However, this shift of MVPs could arise from different mechanisms at the neuronal level, and we can only speculate here about such neuronal properties. For instance, individual neurons within each voxel could change their tuning curve as a result of learning, such that initially face-selective neurons (for example) would also now respond to the associated chairs (Ison et al., 2015; Reddy et al., 2015). Such a

change in tuning is equivalent at the neuronal level to the coactivation account alluded to above. That is, when participants view one stimulus (e.g., a chair), neurons that are normally selective to the associated stimulus (e.g., a face) could also be partially and automatically activated, occasioning a change of their tuning curve. In turn, this would imply that the recorded MVP in response to viewing a chair is composed of a combination of chair and face representations. Alternatively, the newly learned associations could be encoded within each voxel by a new set of neurons that were previously nonselective for either stimulus of the associated pair. In other words, when viewing a face or a chair, in addition to the original populations of face- and chair-selective neurons (respectively), a new subpopulation of neurons encoding the face–chair relation would also be activated. Although our data do not allow us to tease apart these different (and nonexclusive) mechanisms at the neuronal level, they do provide evidence that object representations as measured by MVPs are not static. Recent studies have shown that multivoxel representations of objects in ventral temporal cortex are not fixed but can be modulated by top–down signals such as task goals (Harel, Kravitz, & Baker, 2014). Our findings add to this body of work and show that object representations of highly familiar categories can flexibly move in a high-dimensional multivoxel space as a result of associative learning.

During tasks of explicit memory recall, when participants learn to pair two stimuli together (e.g., a word and a scene), the presentation of a cue stimulus (e.g., the word) can reactivate the fMRI representation of the associated stimulus (Gordon, Rissman, Kiani, & Wagner, 2014; Kuhl & Chun, 2014; Kuhl, Rissman, Chun, & Wagner, 2011; Johnson, McDuff, Rugg, & Norman, 2009; Polyn, Natu, Cohen, & Norman, 2005; Nyberg, Habib, McIntosh, & Tulving, 2000; Wheeler, Petersen, & Buckner, 2000). This reactivation of associated stimuli during explicit recall appears to resemble the results reported here and is compatible with the coactivation account discussed above. Note, however, that this study differs in one crucial aspect from past studies of explicit recall. In the studies cited above, the reactivation of the associated MVP occurred as the participants were explicitly instructed to perform a recall task (and thus retrieve the corresponding stimulus in memory). In contrast, in our study, participants were not instructed to perform a recall task of associated stimuli. Instead they performed a 1-back task on the currently viewed images that was independent of any recall or associative learning. The changes in fMRI representations were observed while participants performed this 1-back task and in the presence of competing visual stimuli (e.g., information about chair stimuli could be decoded while participants were actually viewing and performing a task on faces). Thus, although we cannot discount the possibility that participants automatically recalled a chair while viewing the associated face, this recall must necessarily have occurred in the presence of competing visual input and simultaneously with the performance of a nontrivial, independent task performed on the currently perceived stimuli and that did not require explicit recall. In the end, as discussed above, although such automatic recall could be one of the possible mechanisms underlying the increase in cross-classification performance in our experiment, it is still consistent with the conclusion that multivoxel object representations can be flexibly modified through associative learning.

The finding that the largest learning-dependent changes (>9% increase in cross-classification performance) were observed in clusters of voxels in the left and right fusiform gyri is consistent with a previous study showing associative learning effects in the left fusiform cortex (Park, Shannon, Biggan, & Spann, 2012). The voxels showing the largest changes overlapped substantially with our functionally defined FFA, as well as with an anterior subdivision of the LOC located in the fusiform gyrus (pF; Grill-Spector et al., 1999). The object-selective pF itself partially overlapped with the FFA (Grill-Spector et al., 2001), but we were unable to further segregate these two ROIs in the native space of each participant. Other recent studies have also reported a mix of face- and object-selective voxels in the traditionally defined FFA (Cukur, Huth, Nishimoto, & Gallant, 2013; Hanson & Schmidt, 2011). It has been argued that increased expertise with a class of objects is correlated with the level of activation in the FFA (McGugin, Gatenby, Gore, & Gauthier, 2012; Gauthier et al., 1999), although this claim is still debated (McKone, Kanwisher, & Duchaine, 2007; Kanwisher, 2000). Although our data are unable to shed light on this debate because of the spatial overlap between the FFA and the pF, we find that face- and object-selective representations in the fusiform gyrus show the strongest changes in representational similarity as a result of associative learning.

Previous studies have investigated the effects of training on object representations in object-selective cortex. In general, these studies reveal that training-related changes occur in a distributed fashion in inferotemporal cortex and that these changes are often modest (Op de Beeck & Baker, 2010). In monkeys, training changes the selectivity and strength of neuronal responses in inferotemporal cortex (Li & DiCarlo, 2008; Freedman, Riesenhuber, Poggio, & Miller, 2006; Baker, Behrmann, & Olson, 2002; Sigala & Logothetis, 2002; Logothetis, Pauls, & Poggio, 1995). Human fMRI studies have shown that learning is associated with increases or decreases in the overall amplitude of the average BOLD response (Op de Beeck, Baker, DiCarlo, & Kanwisher, 2006; Kourtzi, Betts, Sarkheil, & Welchman, 2005; Gauthier et al., 1999), as well as with a sharpening of neural tuning (Zhang, Meeson, Welchman, & Kourtzi, 2010; Gillebert, Op de Beeck, Panis, & Wagemans, 2009; Jiang et al., 2007). The current study extends this previous work by investigating the effects of associative learning on preexisting, well-established response patterns for pairs of familiar categories (rather than extensive practice with a single category).

Previous monkey studies have investigated a class of neurons known as "pair-coding neurons" that respond similarly to pairs of stimuli that have been associated together (Sakai & Miyashita, 1991). In these studies, monkeys learned associations between novel, meaningless fractal patterns that they had been exposed to on a relatively short timescale (i.e., in recent experimental sessions). After learning, a neuron that was originally selective to a cue stimulus showed selective responses to its paired associate as well. However, neuronal selectivity for novel stimuli (e.g., the cue stimuli in the aforementioned studies) can flexibly develop as a result of recent exposure (Logothetis et al., 1995), suggesting that the pair-coding task principally modified neuronal responses in recently created representations. In contrast, our participants learned novel associations between already overlearned categories of stimuli, with which they had lifelong exposure. After learning, we found that category selectivity was modified in well-established (and hence presumably less flexible) multivoxel representations that are thought to contribute to visual categorization and object representation. Additionally, pair-coding neurons show significantly correlated responses to pairs of pictures (i.e., at the exemplar level) in a stimulus–stimulus association task. In contrast, we found that category level multivoxel representations change, although the associations were created between exemplars of the two categories. Finally, pair-coding neurons have typically been found in the anterior ventral portion of area TE and in the perirhinal cortex (although a larger proportion of these neurons and stronger pair-coding effects were found in the perirhinal cortex; Naya, Yoshida, & Miyashita, 2003). Other studies have also found evidence for associative learning in perirhinal cortex and anterior ventral IT neurons in monkeys (Eifuku, Nakata, Sugimori, Ono, & Tamura, 2010; Erickson & Desimone, 1999) and in single neurons in the human MTL (Ison et al., 2015; Reddy et al., 2015). However, information about associated stimuli has not been found in single neurons in more posterior portions of TE (Gochin, Colombo, Dorfman, Gerstein, & Gross, 1994). In this study, we observed the strongest effects of associative learning in voxels in the fusiform cortex, overlapping with the FFA and pF. Although it is difficult to establish exact homologies between the monkey and human brains, the human LOC and FFA are thought to correspond to the posterior and dorsal part of the monkey inferotemporal complex (Tsao, Moeller, & Freiwald, 2008; Denys et al., 2004). Our findings thus suggest that information about associated stimulus pairs is also observed in human visual regions more caudal to those previously reported in single neurons in monkey anterior ventral inferotemporal cortex.

Acquiring new associations depends critically on MTL structures, including the hippocampus (Squire, Stark, & Clark, 2004; Fortin, Agster, & Eichenbaum, 2002). As mentioned above, single-neuron recordings in monkeys (Wirth et al., 2003; Erickson & Desimone, 1999; Sakai & Miyashita, 1991; Miyashita, 1988) and humans (Ison et al., 2015; Reddy et al., 2015) show that MTL neurons change their selectivity as a result of learning associations between pairs of stimuli. Human fMRI studies have implicated different MTL structures in associative learning, sequence learning, and relational memory (Schapiro, Kustner, & Turk-Browne, 2012; Turk-Browne, Scholl, Chun, & Johnson, 2009; Haskins, Yonelinas, Quamme, & Ranganath, 2008; Aminoff, Gronau, & Bar, 2007; Diana, Yonelinas, & Ranganath, 2007; Davachi, 2006; Prince, Daselaar, & Cabeza, 2005). In particular, hippocampal fMRI activity patterns become more similar to each other as a result of incidental sequence learning (Schapiro et al., 2012). In our study however, object category multivoxel representations in the hippocampus were essentially unmodified during the postlearning scan. This difference between the two studies could be accounted for by differences in the learning protocols. For instance, in the previous study participants viewed sequences of items but were unaware of the relationships between them. In our study however, participants were explicitly instructed to make associations between the object categories. Additionally, in our study learning occurred over a much longer time frame, with the result that the associations were overlearned (Figure 3) when postlearning brain activity was measured. Thus, although the hippocampus undoubtedly plays an active role during the acquisition of new associations, for instance by differentially activating for successfully learned versus unlearned associations (Davachi, 2006), it is possible that the relevant information was processed and stored in other cortical areas once the associations were overlearned. Indeed, although it is not known how long memory traces need to remain active in MTL structures before being committed to long-term storage in anterior inferotemporal cortex, the representational changes we observe in the VVP could be consistent with such a reorganization of learned information.

Participants were explicitly asked to learn arbitrary associations between unrelated object categories, and we measured changes in neural response patterns in an fMRI scan session at the end of learning. Learned associations in this case could be direct and automatic or mediated by explicit strategies such as recall (as described earlier) and/or visual imagery. Visual perception and visual imagery of familiar categories of objects have been shown to elicit similar patterns of fMRI activity in ventral visual cortex (Reddy, Tsuchiya, & Serre, 2010). Recall of past visual stimuli also reactivates their representations in visual cortex (Wheeler et al., 2000). It is conceivable that during the postlearning scan of the current study, while viewing one category of images (e.g., chairs), participants brought the associated category (e.g., faces) to mind, although they performed a 1-back task on the images that was independent of any associative learning. However, note that even if participants could not avoid recall and mental imagery of the associated categories, the very experience of a stimulus "bringing another to mind" when the task (1-back)

does not require such recall is arguably a manifestation of a well-learned association.

To conclude, we show that associative learning is accompanied by large-scale neural changes in the VVP. Specifically, multivoxel activity patterns for associated object categories become more similar to each other with learning. An interesting open question that we have not addressed here is whether these representational changes are specific to the stimuli with which learning occurred, or whether they generalize to other exemplars in the category. Additionally, how long do these changes persist after the learned associations are no longer behaviorally relevant? Although these questions remain exciting topics for future research, here we show evidence for flexible and dynamic representations in ventral temporal cortex that could support the daily process of learning new relationships between different events.

Acknowledgments

This work was supported by funding from an ANR-JCJC (2012) to L. R. and funding from the Institute des Sciences du Cerveau de Toulouse to L. R. and R. V. We thank Francisco Pereira for sharing his SearchMight toolbox with us. L. R. and R. V. designed the research. The authors would like to thank the staff of the Imaging Center, INSERM/UPS UMR 825 MRI platform for their assistance in acquiring the data.

Reprint requests should be sent to Dr. Leila Reddy, CNRS-Centre de Recherche Cerveau et Cognition, Pavillon Baudot CHU Purpan, 31052 Toulouse Cedex, France, or via e-mail: leila.reddy@cerco.ups-tlse.fr.

REFERENCES

Aminoff, E., Gronau, N., & Bar, M. (2007). The parahippocampal cortex mediates spatial and nonspatial associations. *Cerebral Cortex, 17,* 1493–1503.

Baker, C. I., Behrmann, M., & Olson, C. R. (2002). Impact of learning on representation of parts and wholes in monkey inferotemporal cortex. *Nature Neuroscience, 5,* 1210–1216.

Carlson, T. A., Schrater, P., & He, S. (2003). Patterns of activity in the categorical representations of objects. *Journal of Cognitive Neuroscience, 15,* 704–717.

Cukur, T., Huth, A. G., Nishimoto, S., & Gallant, J. L. (2013). Functional subdomains within human FFA. *Journal of Neuroscience, 33,* 16748–16766.

Davachi, L. (2006). Item, context and relational episodic encoding in humans. *Current Opinion in Neurobiology, 16,* 693–700.

Denys, K., Vanduffel, W., Fize, D., Nelissen, K., Peuskens, H., Van Essen, D., et al. (2004). The processing of visual shape in the cerebral cortex of human and nonhuman primates: A functional magnetic resonance imaging study. *Journal of Neuroscience, 24,* 2551–2565.

Diana, R. A., Yonelinas, A. P., & Ranganath, C. (2007). Imaging recollection and familiarity in the medial temporal lobe: A three-component model. *Trends in Cognitive Sciences, 11,* 379–386.

Downing, P. E., Jiang, Y., Shuman, M., & Kanwisher, N. (2001). A cortical area selective for visual processing of the human body. *Science, 293,* 2470–2473.

Eifuku, S., Nakata, R., Sugimori, M., Ono, T., & Tamura, R. (2010). Neural correlates of associative face memory in the anterior inferior temporal cortex of monkeys. *Journal of Neuroscience, 30,* 15085–15096.

Epstein, R., & Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature, 392,* 598–601.

Erickson, C. A., & Desimone, R. (1999). Responses of macaque perirhinal neurons during and after visual stimulus association learning. *Journal of Neuroscience, 19,* 10404–10416.

Fortin, N. J., Agster, K. L., & Eichenbaum, H. B. (2002). Critical role of the hippocampus in memory for sequences of events. *Nature Neuroscience, 5,* 458–462.

Freedman, D. J., Riesenhuber, M., Poggio, T., & Miller, E. K. (2006). Experience-dependent sharpening of visual shape selectivity in inferior temporal cortex. *Cerebral Cortex, 16,* 1631–1644.

Gauthier, I., Tarr, M. J., Anderson, A. W., Skudlarski, P., & Gore, J. C. (1999). Activation of the middle fusiform 'face area' increases with expertise in recognizing novel objects. *Nature Neuroscience, 2,* 568–573.

Gillebert, C. R., Op de Beeck, H. P., Panis, S., & Wagemans, J. (2009). Subordinate categorization enhances the neural selectivity in human object-selective cortex for fine shape differences. *Journal of Cognitive Neuroscience, 21,* 1054–1064.

Gochin, P. M., Colombo, M., Dorfman, G. A., Gerstein, G. L., & Gross, C. G. (1994). Neural ensemble coding in inferior temporal cortex. *Journal of Neurophysiology, 71,* 2325–2337.

Gordon, A. M., Rissman, J., Kiani, R., & Wagner, A. D. (2014). Cortical reinstatement mediates the relationship between content-specific encoding activity and subsequent recollection decisions. *Cerebral Cortex, 24,* 3350–3364.

Grill-Spector, K. (2003). The neural basis of object perception. *Current Opinion in Neurobiology, 13,* 159–166.

Grill-Spector, K., Kourtzi, Z., & Kanwisher, N. (2001). The lateral occipital complex and its role in object recognition. *Vision Research, 41,* 1409–1422.

Grill-Spector, K., Kushnir, T., Edelman, S., Avidan, G., Itzchak, Y., & Malach, R. (1999). Differential processing of objects under various viewing conditions in the human lateral occipital complex. *Neuron, 24,* 187–203.

Grill-Spector, K., & Malach, R. (2004). The human visual cortex. *Annual Review of Neuroscience, 27,* 649–677.

Hanke, M., Halchenko, Y. O., Sederberg, P. B., Hanson, S. J., Haxby, J. V., & Pollmann, S. (2009). PyMVPA: A python toolbox for multivariate pattern analysis of fMRI data. *Neuroinformatics, 7,* 37–53.

Hanson, S. J., & Schmidt, A. (2011). High-resolution imaging of the fusiform face area (FFA) using multivariate non-linear classifiers shows diagnosticity for non-face categories. *Neuroimage, 54,* 1715–1734.

Harel, A., Kravitz, D. J., & Baker, C. I. (2014). Task context impacts visual object processing differentially across the cortex. *Proceedings of the National Academy of Sciences, U.S.A., 111,* E962–E971.

Haskins, A. L., Yonelinas, A. P., Quamme, J. R., & Ranganath, C. (2008). Perirhinal cortex supports encoding and familiarity-based recognition of novel associations. *Neuron, 59,* 554–560.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science, 293,* 2425–2430.

Haxby, J. V., Guntupalli, J. S., Connolly, A. C., Halchenko, Y. O., Conroy, B. R., Gobbini, M. I., et al. (2011). A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron, 72,* 404–416.

Haynes, J. D., & Rees, G. (2005). Predicting the orientation of invisible stimuli from activity in human primary visual cortex. *Nature Neuroscience, 8,* 686–691.
Ison, M. J., Quian Quiroga, R., & Fried, I. (2015). Rapid encoding of new memories by individual neurons in the human brain. *Neuron, 87,* 220–230.
Jiang, X., Bradley, E., Rini, R. A., Zeffiro, T., Vanmeter, J., & Riesenhuber, M. (2007). Categorization training results in shape- and category-selective human neural plasticity. *Neuron, 53,* 891–903.
Johnson, J. D., McDuff, S. G., Rugg, M. D., & Norman, K. A. (2009). Recollection, familiarity, and cortical reinstatement: A multivoxel pattern analysis. *Neuron, 63,* 697–708.
Kamitani, Y., & Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nature Neuroscience, 8,* 679–685.
Kanwisher, N. (2000). Domain specificity in face perception. *Nature Neuroscience, 3,* 759–763.
Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience, 17,* 4302–4311.
Kourtzi, Z., Betts, L. R., Sarkheil, P., & Welchman, A. E. (2005). Distributed neural plasticity for shape learning in the human visual cortex. *PLoS Biology, 3,* e204.
Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences, U.S.A., 103,* 3863–3868.
Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis—Connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience, 2,* 4.
Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., et al. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron, 60,* 1126–1141.
Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S., & Baker, C. I. (2009). Circular analysis in systems neuroscience: The dangers of double dipping. *Nature Neuroscience, 12,* 535–540.
Kuhl, B. A., & Chun, M. M. (2014). Successful remembering elicits event-specific activity patterns in lateral parietal cortex. *Journal of Neuroscience, 34,* 8051–8060.
Kuhl, B. A., Rissman, J., Chun, M. M., & Wagner, A. D. (2011). Fidelity of neural reactivation reveals competition between memories. *Proceedings of the National Academy of Sciences, U.S.A., 108,* 5903–5908.
Li, N., & DiCarlo, J. J. (2008). Unsupervised natural experience rapidly alters invariant object representation in visual cortex. *Science, 321,* 1502–1507.
Logothetis, N. K., Pauls, J., & Poggio, T. (1995). Shape representation in the inferior temporal cortex of monkeys. *Current Biology, 5,* 552–563.
McGugin, R. W., Gatenby, J. C., Gore, J. C., & Gauthier, I. (2012). High-resolution imaging of expertise reveals reliable object selectivity in the fusiform face area related to perceptual performance. *Proceedings of the National Academy of Sciences, U.S.A., 109,* 17063–17068.
McKone, E., Kanwisher, N., & Duchaine, B. C. (2007). Can generic expertise explain special processing for faces? *Trends in Cognitive Sciences, 11,* 8–15.
Messinger, A., Squire, L. R., Zola, S. M., & Albright, T. D. (2001). Neuronal representations of stimulus associations develop in the temporal lobe during learning. *Proceedings of the National Academy of Sciences, U.S.A., 98,* 12239–12244.
Miyashita, Y. (1988). Neuronal correlate of visual associative long-term memory in the primate temporal cortex. *Nature, 335,* 817–820.
Miyashita, Y., & Chang, H. S. (1988). Neuronal correlate of pictorial short-term memory in the primate temporal cortex. *Nature, 331,* 68–70.
Naya, Y., Yoshida, M., & Miyashita, Y. (2003). Forward processing of long-term associative memory in monkey inferotemporal cortex. *Journal of Neuroscience, 23,* 2861–2871.
Nyberg, L., Habib, R., McIntosh, A. R., & Tulving, E. (2000). Reactivation of encoding-related brain activity during memory retrieval. *Proceedings of the National Academy of Sciences, U.S.A., 97,* 11120–11124.
Op de Beeck, H. P., & Baker, C. I. (2010). The neural basis of visual object learning. *Trends in Cognitive Sciences, 14,* 22–30.
Op de Beeck, H. P., Baker, C. I., DiCarlo, J. J., & Kanwisher, N. G. (2006). Discrimination training alters object representations in human extrastriate cortex. *Journal of Neuroscience, 26,* 13025–13036.
Op de Beeck, H. P., Brants, M., Baeck, A., & Wagemans, J. (2010). Distributed subordinate specificity for bodies, faces, and buildings in human ventral visual cortex. *Neuroimage, 49,* 3414–3425.
Park, H., Shannon, V., Biggan, J., & Spann, C. (2012). Neural activity supporting the formation of associative memory versus source memory. *Brain Research, 1471,* 81–92.
Polyn, S. M., Natu, V. S., Cohen, J. D., & Norman, K. A. (2005). Category-specific cortical activity precedes retrieval during memory search. *Science, 310,* 1963–1966.
Prince, S. E., Daselaar, S. M., & Cabeza, R. (2005). Neural correlates of relational memory: Successful encoding and retrieval of semantic and perceptual associations. *Journal of Neuroscience, 25,* 1203–1210.
Reddy, L., & Kanwisher, N. (2006). Coding of visual objects in the ventral stream. *Current Opinion in Neurobiology, 16,* 408–414.
Reddy, L., & Kanwisher, N. (2007). Category selectivity in the ventral visual pathway confers robustness to clutter and diverted attention. *Current Biology, 17,* 2067–2072.
Reddy, L., Poncet, M., Self, M. W., Peters, J. C., Douw, L., van Dellen, E., et al. (2015). Learning of anticipatory responses in single neurons of the human medial temporal lobe. *Nature Communications, 6,* 8556.
Reddy, L., Tsuchiya, N., & Serre, T. (2010). Reading the mind's eye: Decoding category information during mental imagery. *Neuroimage, 50,* 818–825.
Sakai, K., & Miyashita, Y. (1991). Neural organization for the long-term memory of paired associates. *Nature, 354,* 152–155.
Schapiro, A. C., Kustner, L. V., & Turk-Browne, N. B. (2012). Shaping of object representations in the human medial temporal lobe based on temporal regularities. *Current Biology, 22,* 1622–1627.
Sigala, N., & Logothetis, N. K. (2002). Visual categorization shapes feature selectivity in the primate temporal cortex. *Nature, 415,* 318–320.
Spiridon, M., & Kanwisher, N. (2002). How distributed is visual category information in human occipito-temporal cortex? An fMRI study. *Neuron, 35,* 1157–1165.
Squire, L. R., Stark, C. E., & Clark, R. E. (2004). The medial temporal lobe. *Annual Review of Neuroscience, 27,* 279–306.
Stansbury, D. E., Naselaris, T., & Gallant, J. L. (2013). Natural scene statistics account for the representation of scene categories in human visual cortex. *Neuron, 79,* 1025–1034.
Straw, A. D. (2008). Vision egg: An open-source library for realtime visual stimulus generation. *Frontiers in Neuroinformatics, 2,* 4.

Tsao, D. Y., Moeller, S., & Freiwald, W. A. (2008). Comparing face patch systems in macaques and humans. *Proceedings of the National Academy of Sciences, U.S.A., 105,* 19514–19519.

Turk-Browne, N. B., Scholl, B. J., Chun, M. M., & Johnson, M. K. (2009). Neural evidence of statistical learning: Efficient detection of visual regularities without awareness. *Journal of Cognitive Neuroscience, 21,* 1934–1945.

Wheeler, M. E., Petersen, S. E., & Buckner, R. L. (2000). Memory's echo: Vivid remembering reactivates sensory-specific cortex. *Proceedings of the National Academy of Sciences, U.S.A., 97,* 11125–11129.

Wirth, S., Yanike, M., Frank, L. M., Smith, A. C., Brown, E. N., & Suzuki, W. A. (2003). Single neurons in the monkey hippocampus and learning of new associations. *Science, 300,* 1578–1581.

Zhang, J., Meeson, A., Welchman, A. E., & Kourtzi, Z. (2010). Learning alters the tuning of functional magnetic resonance imaging patterns for visual forms. *Journal of Neuroscience, 30,* 14127–14133.