◆ Human Brain Mapping 36:3213–3226 (2015) ◆

Monitoring the Growth of the Neural Representations of New Animal Concepts

Andrew James Bauer and Marcel Adam Just*

Department of Psychology, Center for Cognitive Brain Imaging, Baker Hall, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213

Abstract: Although enormous progress has recently been made in identifying the neural representations of individual object concepts, relatively little is known about the growth of a neural knowledge representation as a novel object concept is being learned. In this fMRI study, the growth of the neural representations of eight individual extinct animal concepts was monitored as participants learned two features of each animal, namely its habitat (i.e., a natural dwelling or scene) and its diet or eating habits. Dwelling/scene information and diet/eating-related information have each been shown to activate their own characteristic brain regions. Several converging methods were used here to capture the emergence of the neural representation of a new animal feature within these characteristic, a priori-specified brain regions. These methods include statistically reliable identification (classification) of the eight newly acquired multivoxel patterns, analysis of the neural representational similarity among the newly learned animal concepts, and conventional GLM assessments of the activation in the critical regions. Moreover, the representation of a recently learned feature showed some durability, remaining intact after another feature had been learned. This study provides a foundation for brain research to trace how a new concept makes its way from the words and graphics used to teach it, to a neural representation of that concept in a learner's brain. *Hum Brain Mapp 36:3213–3226, 2015.*

Key words: neural change; neural representation; concept representation; fMRI; MVPA; representational similarity analysis

INTRODUCTION

In August, 2013, the Smithsonian Institution announced the discovery of the olinguito, the first new species of carnivore to be identified in the Western hemisphere in 35 years. The olinguito is a mammal that eats mainly fruit instead of meat, and it lives by itself in the treetops of foggy rainforests. Millions of people encountered this semantic information and thereby permanently changed their own brains to encode the new animal concept. Our research happened to be examining this process at that time in a laboratory setting, using functional magnetic resonance imaging (fMRI) and multivoxel pattern analyses (MVPA) to discover how the neural representation of a novel animal concept arises in the brain as a person learns about the features of that animal.

A key goal of cognitive neuroscience is to delineate the nature, content, and anatomical distribution of the neural representation of knowledge in long-term semantic memory. The importance of concept knowledge is that it underlies human thought, communication, and daily activities, from small talk about well-worn topics to the learning of quantum physics or the pioneering of new scientific discoveries. Accordingly, research that uncovers the neural

Additional Supporting Information may be found in the online version of this article.

Contract grant sponsor: Office of Naval Research; Contract grant number: N00014-13-1-0250

*Correspondence to: Marcel A. Just; Center for Cognitive Brain Imaging, Baker Hall, Dept. of Psychology, Carnegie Mellon University, 5000 Forbes Ave., Pittsburgh, PA 15213.
E-mail: just@cmu.edu

Received for publication 6 April 2015; Accepted 8 May 2015.
DOI: 10.1002/hbm.22842
Published online 2 June 2015 in Wiley Online Library (wileyonlinelibrary.com).

representations of different concepts (such as that of a tool, emotion, or number) has made considerable headway, particularly research on object concepts [e.g. Kassam et al., 2013; Damarla and Just, 2013; for a review on object concepts see Martin, 2007].

Research to date has revealed that object concepts (such as the concept of a hammer) are neurally represented in multiple brain regions, corresponding to the various brain systems that are involved in the physical and mental interaction with the concept. The concept of a hammer entails what it looks like, what it is used for, how one holds and wields it, and soforth, resulting in a neural representation distributed over sensory, motor, and association areas. There is a large literature that documents the responsiveness (activation) of sets of brain regions to the perception or contemplation of different object concepts, including animals (animate natural objects), tools, and fruits and vegetables (for a comprehensive fMRI study see Huth et al., 2012]. For example, fMRI research has shown that nouns that refer to physically manipulable objects such as tools elicit activity in left premotor cortex in right-handers, and activity has also been observed in a variety of other regions to a lesser extent [Chao and Martin, 2000]. Clinical studies of object category-specific knowledge deficits have uncovered results compatible with those of fMRI studies. For example, damage to the inferior parietal lobule can result in a relatively selective knowledge deficit about the purpose and the manner of use of a tool [for a review of the clinical literature see Capitani et al., 2003].

The significance of such findings is enhanced by the commonality of neural representations of object concepts across individuals [Shinkareva et al., 2012]. For example, pattern classifiers of multivoxel brain activity trained on the data from a set of participants can reliably predict which object noun a new test participant is contemplating [Just et al., 2010]. Similarity in neural representation across individuals may indicate that there exist domain-specific brain networks that process information that is important to survival, such as information about food and eating or about enclosures that provide shelter [Mahon and Caramazza, 2003].

Although research on the neural representations of familiar object concepts has progressed considerably, relatively little is known about the changes that occur in a neural knowledge representation as a novel object concept is being learned. A small number of fMRI studies have explored changes in sites of activation after the learning of novel object concepts. For example, after a hands-on session of learning how to use novel tool-like objects, activation to pictures of the objects was found to shift predominantly to motor cortex (e.g., left premotor cortex) compared to prelearning [Weisberg et al., 2007]. In another study, after participants were verbally instructed about the kind of motion or sound that was associated with novel living objects, the brain activation elicited by the object pictures was localized to motion-specific or auditory cortex [James and Gauthier, 2003].

In the current fMRI study, the growth of the neural representations of individual novel natural concepts was tracked as they were enriched regarding two object features. By sequentially teaching two features, it was possible to both (i) study the emergence of the neural knowledge representation of a feature directly after instruction and (ii) assess the retention of the neural representation of a previously learned feature beyond its learning period.

The novel concepts that were taught in this study were animal concepts (derived from actual extinct animals). In the scanner, participants viewed pictures of the animals while they received written information about two types of feature of the animals, namely their habitats (i.e., natural dwellings or scenes) and diets or eating habits. After instruction about a feature, the neural representation of an emerging animal concept was assessed with fMRI as participants thought about that animal, including each feature that had been taught so far. The experimental paradigm is depicted in Figure 1. Activation was examined in a priori-specified regions-of-interest (ROIs), which are shown in Figure 2. The regions where *habitat*-related activation was expected included the parahippocampal gyrus, which is well known to activate to information about dwellings and scenes [Epstein and Kanwisher, 1998]. Other *habitat*-related areas included the precuneus, which also activates to information about dwellings [Just et al., 2010]. These areas were anatomically close to the retrosplenial cortex, which is thought to be involved in spatial updating between egocentric and allocentric points of view, and in localization of a scene within a larger, extended environment [Epstein and Higgins, 2007; for a review on the retrosplenial cortex see Vann et al., 2009]. These roles are consistent with the possible cognitive processes required to learn about an animal's habitat, such as imagining how an animal might spatially fit into its surrounding habitat. The regions where *diet*-related activation was expected included areas within the left inferior frontal gyrus because this region activates to words about foods and eating-related objects [Just et al., 2010] and face- and jaw-related actions [Hauk et al., 2004].

There were two goals of the current study: one goal was to monitor the growth of a neural representation after the integration of knowledge of a feature into each animal concept. It was hypothesized that activation levels would increase in brain regions a priori predicted to encode newly acquired knowledge of a feature (Hypothesis 1A). This expectation is consistent with the results of previous research that compared univariate measures of activation before and after learning [for example, James and Gauthier, 2003]. It was also hypothesized that a classifiable multivoxel representation of the new feature knowledge of each animal concept would emerge within the same brain regions (Hypothesis 1B). Detection of a unique neural encoding of each concept would go beyond a mean activation-based finding that the brain areas are somehow *involved* in the processing of a class of concepts [Mur et al., 2009]. Furthermore, a posteriori, it was examined whether

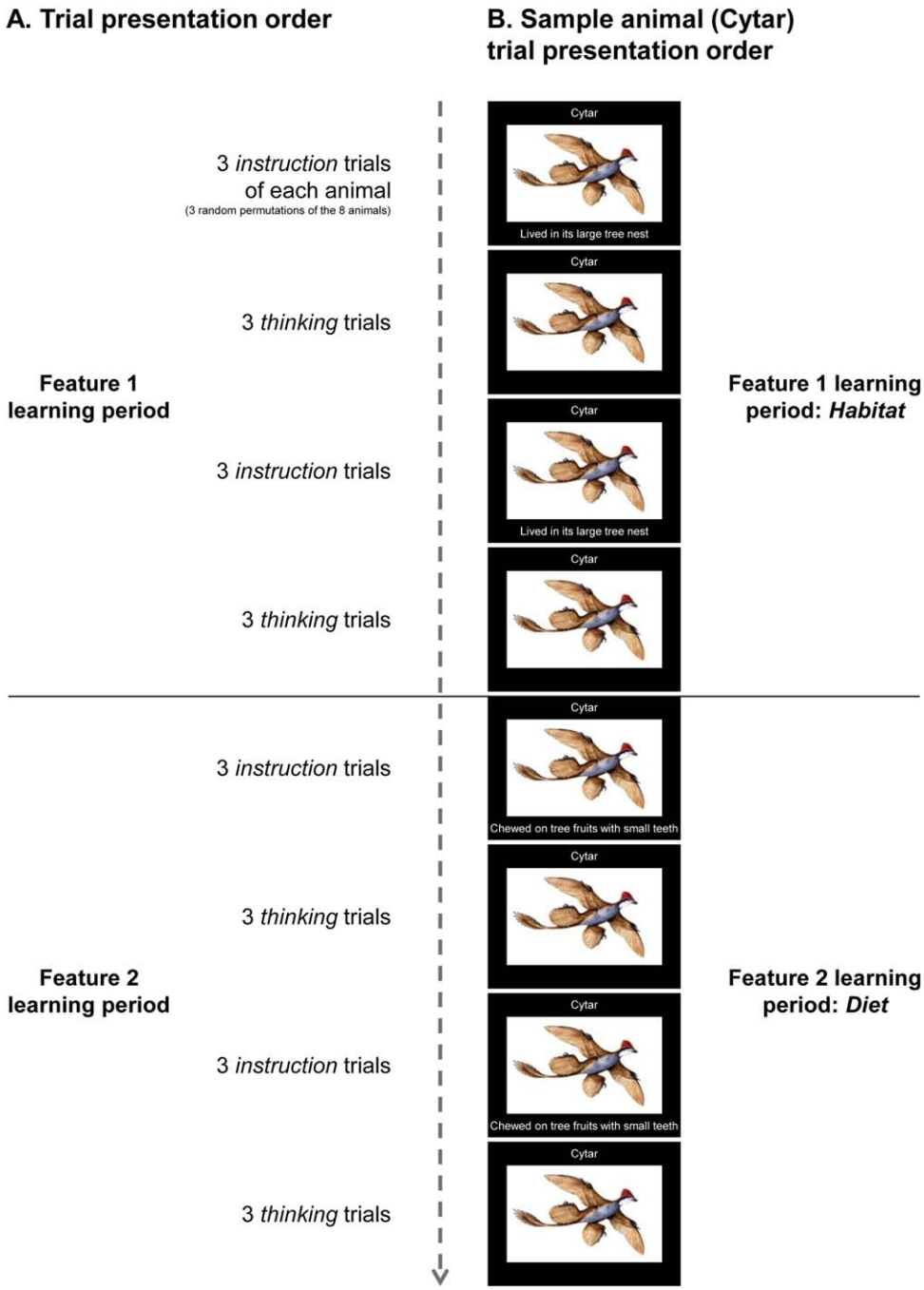
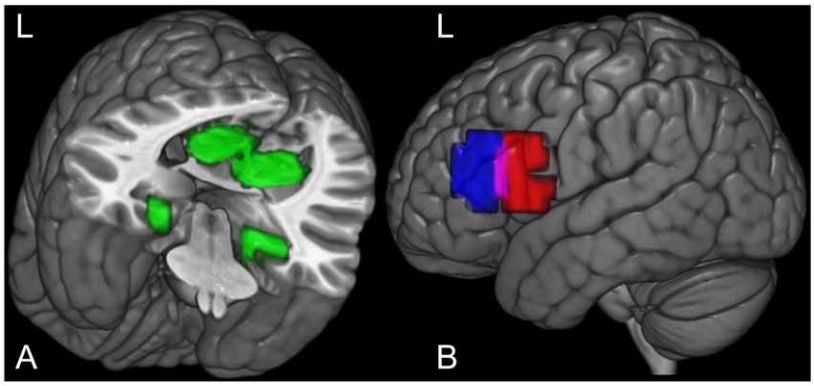


Figure 1.

Experimental paradigm and trial presentation order. **A**: Schematic of the scanning experimental paradigm and trial presentation order. One feature knowledge type (*habitat or diet*) was taught about all the animals before the other knowledge type. Two participant groups were taught the knowledge types in different orders. During *instruction* trials, participants were asked to silently read and remember the animal name, and imagine and think about the animal embodying only the new feature description. During *thinking* trials, participants were asked to imagine and think about the animal embodying every feature taught about thus far for that animal. The main focus of the data analysis was on activation from the thinking trials, which was hypothesized to constitute the neural representations of the animal concepts. **B**: Schematic of the full *instruction* trial and *thinking* trial presentation order for Cytar, a sample animal (in this example, *habitat* is learned before *diet*). Figure S1 contains detailed information for all eight of the animal concept stimuli.



Feature knowledge type	Cluster location	Cluster centroid (MNI)			Total no. voxels in cluster		Mean no. voxels in cluster analyzed (*standard deviation*)	
		x	*y*	*z*				
Habitat	L parahippocampal gyrus	-28	-37	-10	23		2 (*1*)	
	R parahippocampal gyrus	26	-33	-13	37		2 (*1*)	
	L precuneus	-11	-61	16	212		19 (*4*)	
	R precuneus	15	-55	13	203		17 (*3*)	
Diet	L inferior frontal gyrus	-53	11	18	163	41 in shared	17 (*4*)	4 (*1*) in
	L mid/inferior frontal gyrus	-47	27	18	172	area	18 (*4*)	shared area

Figure 2.

A priori-specified ROIs predicted to encode the new feature knowledge. The set of (**A**) *habitat* and (**B**) *diet* clusters used as ROIs (clusters adapted from Just et al., 2010). The *diet* clusters overlapped partially (purple area in **B**). The clusters were rendered on an MNI template brain using the 3D medical imaging software MRIcroGL (Rorden and Brett, 2000). For data analysis, 40 voxels were selected from each set of ROIs for a given feature knowledge type. The right-most column of the table above shows the mean (over participants) distribution of the 40 voxels over the ROIs of a given feature knowledge type, for the analyses of the activation levels and classification accuracies of the *thinking* trial data. (Information about the voxel selection criterion is found in the section of the text on the classification procedures.). *L*: left; *R*: right; *ROI*: region of interest; *MNI*: Montreal Neurological Institute template.

animals that were described as having similar features became neurally more similar to each other, namely, have similar multivoxel activation patterns in brain areas that encode that feature (Hypothesis 1C). Verification of this hypothesis would further indicate that the emergence of distinguishable activation patterns of the animal concepts was driven by the specific feature information that was taught about each of the various animals.

Another research goal was to track the fate of the neural knowledge representation of a previously learned feature into the learning period of a different feature. In brain regions a priori postulated to encode the first feature, it was expected that the recently heightened activation levels of the first feature learning period would persist after instruction about the second feature (Hypothesis 2A). Similarly, it was hypothesized that the multivoxel representations of the first taught feature of each animal concept would remain classifiable during the learning period of the second feature (Hypothesis 2B). Finally, there was no hypothesis about whether there would be a change in the similarity relations among the neural representations of the concepts with respect to the first feature after the second feature was taught.

A secondary hypothesis was that activation levels during *instruction* periods would be elevated in the same brain regions predicted to represent that feature information *after* instruction and learning (Hypothesis 3). Participants were also expected to behaviorally demonstrate learning of the animal concepts during a recall task at the end of the experiment (Hypothesis 4).

MATERIALS AND METHODS

Participants

Sixteen right-handed adults (ten males, six females; mean age of 22.4 years, ranging from 18 to 34) from Carnegie Mellon University and the Pittsburgh community participated and gave written informed consent approved by the Carnegie Mellon Institutional Review Board. Three additional participants' data were excluded because of excessive head motion (greater than 4mm total

displacement in any dimension). Three other participants' data were discarded due to chance-level multivoxel pattern classification accuracy of the animal concepts (classification features were the 120 most "stable" voxels selected from anywhere in the brain excluding the occipital lobe; more detail concerning classification is provided below). This classification, which differed from the classification analyses that tested the hypotheses, was used to check for systematicity in a participant's activation patterns regardless of its correspondence to the hypotheses.

Experimental Paradigm

Functional images were acquired for the entire duration of the learning of the animal concepts. There were a total of eight novel animals (four mammals and four birds), all of which were based on real extinct animals. The two feature knowledge types that were taught were *habitat* and *diet*. In a feature learning period, the same feature type was taught for all the animals, one animal at a time, before proceeding to the second feature. The order in which the two features were taught was balanced across participants (two groups of eight participants). The fMRI data associated with the two features were acquired in two separate 21.8-min scans.

Interleaved sets of *instruction* and *thinking* trials were presented within each feature learning period. The *instruction* trials conveyed the new information about the feature currently being taught. During the *thinking* trials, which followed each round of *instruction* trials, participants were prompted to imagine and think about an animal and the feature(s) that they had been taught about thus far. The *thinking* trials constitute the type of paradigm that has previously been used to evoke activation that is amenable to classification [e.g., Just et al., 2010; Kassam et al., 2013]. Thus the analysis of newly acquired feature information focused on the activation patterns obtained during the *thinking* trials.

The scanning experimental paradigm and trial presentation order is shown in Figure 1A. Within a feature learning period, each set of *instruction* and *thinking* trials cycled through each animal three times; the trial presentation order randomly intermixed the repetitions of a given animal with the repetitions of other animals. Thus, in a learning period, participants (*i*) were taught about one feature for each animal six times to allow for sufficient learning; and (*ii*) imagined and thought about each animal and its associated feature(s) six times so that there was enough data for the multivoxel pattern classification analysis.

Each *instruction* trial displayed information about one animal, including (i) a name (significantly shortened and changed from the animal's scientific taxonomic name), (ii) a picture, and (iii) a short phrase describing a feature. For example, the four-winged dinosaur bird species *Microraptor zhaoianus* was called Cytar in the experiment, and the descriptions of its habitat and diet were as follows: "Lived in its large tree nest" and "Chewed on tree fruits with small teeth." Figure 1B depicts the *instruction* trial (left-hand column) and *thinking* trial presentation order for Cytar (in this example, *habitat* is learned before *diet*). During the *instruction* trials, participants were asked to silently read and remember the animal name, and imagine and think about the animal including the feature. A trial consisted of 5s display-time and 7s off-time (a fixation "X" was shown during off-time).

Each *thinking* trial consisted of only a picture and name for one animal, with no feature description. The same timing parameters were used for the *thinking* trial presentations as for the *instruction* trials. Participants were prompted by the picture and name to imagine and think about the animal including the feature(s) that had been taught thus far for that animal. This consisted of just one feature after information about the first feature was taught and both features after the second feature was taught. Participants were asked to use mental imagery so as to maximize the amount of semantic information retrieved. It was emphasized to participants that they think the same thoughts for each repetition of an animal, to ensure comparability of the data across repetitions. Participants were free to choose which specific details to think about, in which sensory modalities, and whether and how to include motor imagery.

Visual depictions of the animals were included in the experiment to facilitate a detailed instantiation of each novel animal. Participants in a pilot study reported difficulty in thinking about a novel animal's eating habits, for example, when they did know what the animal looked like. Instantiation of the animals in terms of a picture improved classification accuracies in the pilot study and hence pictures were included in the study.

There were ten total presentations of an "X" alone in the center of the screen, 24s each, distributed evenly throughout the two scans to provide a baseline measure for calculating percent signal change (PSC) in the fMRI signal and for statistical parametric mapping. During these fixation periods and the off-time portion of each *instruction* and *thinking* trial, participants were instructed to fixate on the "X" and clear their minds.

Supporting Information Figure S1 contains detailed information for all eight of the animal concept stimuli. Participants were informed at the end of the experiment that the animals taught to them were derived from real extinct animals, but that the animal names used were not the actual scientific taxonomic names.

fMRI Scanning Parameters and Data Preprocessing

Functional blood oxygen level-dependent (BOLD) images were acquired on a 3T Siemens Verio Scanner and 32-channel phased-array head coil (Siemens Medical Solutions, Erlangen, Germany) at the Scientific Imaging and Brain Research (SIBR) Center of Carnegie Mellon University using a gradient echo EPI sequence with TR = 1000 ms, TE = 25ms, and a 60° flip angle. Seventeen 5-mm thick

oblique-axial slices were imaged with a gap of 1 mm between slices, starting at the bottom in an interleaved spatial order. The acquisition matrix was 64 × 64 with 3.125 × 3.125 × 6 mm voxels.

Data preprocessing was performed with the Statistical Parametric Mapping software (SPM2, Wellcome Department of Cognitive Neurology, London, UK). Images were corrected for slice acquisition timing, motion, and linear trend; temporally smoothed with a high-pass filter using a 190s cutoff; and normalized to the Montreal Neurological Institute (MNI). The *thinking* trial data were analyzed with a multivoxel pattern classification analysis to study the neural representations of the animal concepts; these images were not spatially smoothed.

The *instruction* trial data were analyzed using SPM2 using the general linear model (GLM) and Gaussian random field theory. The voxels were smoothed with an 8-mm full-width half-maximum (FWHM) Gaussian kernel to decrease spatial noise.

Regions of Interest

A set of a priori ROIs for each feature knowledge type was generated and used for analysis; these included voxels predicted to encode *habitat*, and a separate group of voxels predicted to encode *diet*. The ROIs were adapted from a previous study that localized each of these factors to a set of six brain regions [Just et al., 2010]. The ROIs that were adapted included the bilateral parahippocampal gyrus and bilateral precuneus clusters (labeled "shelter" in the original study) to create the *habitat* set of ROIs; and the left inferior frontal gyrus clusters (labeled "eating") were adapted to create the *diet* set.

Adaptation of the clusters from the original study proceeded as follows: first, gray matter spheres were created using the centroids of the original clusters. The radius of each sphere was set to 15 mm, resulting in increased volumes to allow the ROIs to be more generalizable to the participants in the current study. Enlargement of the parahippocampal gyrus ROIs in this way initially resulted in inclusion of voxels from the fusiform gyrus and cerebellum [identified as such using Automated Anatomic Labeling (AAL), Tzourio-Mazoyer et al., 2002]; these latter voxels were then excluded from the ROI. Figure 2 depicts and provides more information about the ROIs used in the current study. The ROIs served as a pool of voxels from which a subset was selected for data analysis. (The section below on the classification procedures contains a description of how this subset of voxels was defined.)

Data Analysis

Overview: monitoring the growth of the neural representations of the animal concepts

A combination of analyses of activation levels (percent signal change), accuracies of multivoxel pattern classification, and representational similarity relations among the activation patterns were used to test the hypotheses about the growth of the neural representations of the animal concepts. After the teaching of feature information about a knowledge type (*habitat* or *diet*), activation levels and classification accuracies of the concepts were expected to increase in brain regions a priori predicted to encode that knowledge type (Hypotheses 1A-B). In addition, it was examined a posteriori whether animals that were described as having similar features became neurally more similar to each other with respect to that feature, after instruction about that feature (Hypothesis 1C). Finally, activation levels and classification accuracies in the regions that encode the first-instructed knowledge type were hypothesized to retain their recently heightened levels into the learning period of the second knowledge type (Hypotheses 2A-B). Data from the *thinking* trials (when feature knowledge was being recalled) were analyzed. To correct for possible drift in the baseline signal levels throughout the scans, each acquired image was normalized (mean = 0, SD = 1) across gray matter voxels from the whole brain. Analyses of activation levels and classification accuracies used data from the same set of voxels that was selected for each partitioning of the data into classification cross-validation training and test sets. The section below on the classification procedures contains more information about the voxel selection criterion. The representational similarity analysis was performed using voxels selected according to the same voxel selection criterion that was applied once across all the data.

Measuring the emergence of the neural knowledge representation of a learned concept feature

Activation levels in brain regions a priori predicted to encode the second feature were compared before versus after the time when the information about the second feature had been taught. This comparison was made for each of the eight animals. (Note that there was no assessment of the acquisition of the first feature because there was no baseline for comparison obtained before the first feature learning period.) The activation levels were analyzed with a mixed-model ANOVA defined by participant group (two groups) and preinstruction and postinstruction about the second feature.

Additionally, a 16-way multivoxel pattern classification analysis was performed, based on the activation for the eight animal concepts both before and after instruction about the second feature within brain regions specific to the second feature. A classification accuracy was obtained for each of the 16 items. The same voxels were used to assess the neural representations before and after instruction. The mean accuracies across participants over the eight animals before instruction about the second feature were compared to the mean accuracies over these eight animals after instruction. The classification accuracies were

TABLE I. Animal feature descriptions including feature pairs rated highly similar

Habitat feature descriptions		*Diet* feature descriptions
Lived in swamps encircled by trees		Swallowed mud containing vegetation
Lived within the rainforest canopy		Swallowed small seeds and nuts
Lived around and between steep hills		Sucked large berries into its mouth
Lived around water in ice caverns		Ate eel meat with serrated teeth
Lived near beaches in caves		Ate crab meat with sharp teeth
Lived on grasslands around mountains		Ate rabbit meat with its sharp beak
Lived in its large and shallow burrow		Chewed on grass with flat teeth
Lived in its very large tree nest		Chewed on tree fruits with small teeth
The four pairs of *habitat* features rated highly similar		
Lived around and between steep hills	←→	Lived on grasslands around mountains
Lived within the rainforest canopy	←→	Lived in its very large tree nest
Lived around water in ice caverns	←→	Lived near beaches in caves
Lived near beaches in caves	←→	Lived in its large and shallow burrow
The four pairs of *diet* features rated highly similar		
Swallowed small seeds and nuts	←→	Sucked large berries into its mouth
Ate eel meat with serrated teeth	←→	Ate crab meat with sharp teeth
Swallowed mud containing vegetation	←→	Swallowed small seeds and nuts
Chewed on grass with flat teeth	←→	Chewed on tree fruits with small teeth

analyzed with a mixed-model ANOVA similar to that used to analyze the activation levels.

Representational similarity analysis was used to determine whether the animals that were described as having similar second-taught features became neurally more similar after the feature instruction, based on activation patterns from the same a priori-specified brain regions as in the analyses above. An independent group of eight raters was instructed to select only four to six pairs of animal feature descriptions that were highly similar, out of the set of 28 possible pairs between the eight animals (for *habitat* and *diet* separately). (Prior examination of the features identified between four to six pairs that were highly similar.) The modal number of pairs selected by the raters was four, and none of the raters selected more than six pairs. The four pairs that were submitted for analysis were the ones most frequently selected across the raters, as shown in Table I.

A permutation test was used to determine whether the change in pairwise correlation distance from preinstruction to postinstruction differed between these four pairs of animals and the remaining 24 pairs. The difference in the change in pairwise correlation distance between the two sets of pairs of animals (averaged over *habitat* and *diet*) was compared to an empirical distribution of difference scores, where each difference score was calculated after the pairs were randomly assigned to either the "most similar" set of four pairs or the "least similar" set of 24 pairs. The preinstruction and postinstruction neural representations of the animal concepts were calculated by averaging over all the preinstruction and postinstruction repetitions, and over the participants. The correlation distance (i.e. 1 – the Pearson correlation between two vectors) was calculated between all pairs of animals separately for the four most similar pairs and the remaining 24 pairs. [Correlation distance has been shown to provide better accounts than other dissimilarity metrics, such as Euclidean distance; see Kriegeskorte et al., 2008.]

Assessing retention of the neural knowledge representation of a previously taught feature into the learning period about a different feature

Activation levels in brain regions postulated a priori to encode the first-instructed feature were compared before versus after the second feature had been taught. In addition, classification accuracies for the 16 items, using voxels from only the brain regions specific to the first feature, were compared before versus after the second feature had been taught, using ANOVA.

Neural evidence of the processing of knowledge about a feature during its instruction period

During the period of *instruction* about a feature, elevated activation levels were expected in the same brain regions a priori predicted to represent that feature knowledge *after* instruction (Hypothesis 3). To test this secondary hypothesis, a random-effects analysis was run on individual participants' GLM contrast images between instruction about *habitat* and instruction about *diet* (voxel-wise paired-sample *t*-tests). A regressor was specified for a given feature knowledge type, which consisted of the entire set of *instruction* trials about that feature. Each trial was convolved with the canonical hemodynamic response function. Because the *instruction* trials about each feature were equivalent across the two participant groups except with respect to presentation order, the data from the two groups were combined.

Behavioral evidence of learning the animal concepts

After the scanning, participants were asked to write down the two features for each of eight animals, prompted by only the animal names and not the pictures. Responses that were verbatim or that contained the salient aspects of the features were counted as correct (out of a possible total of 16).

Multivoxel pattern classification procedures

The classifier used was support vector machines with a multiway classification decision and linear kernel. The implementation was a modified version of SVM-light [Joachims, 1999] using MATLAB 6.5 (Mathworks, MA). Classification proceeded through three stages: (1) algorithmic selection of a set of voxels to be used for classification; (2) training of a classifier on a subset of the data; and (3) testing of the classifier on the remaining subset of the data. The training and testing used cross-validation procedures that iterated through all possible partitionings of the data into training and test sets, always keeping the training and test sets separate.

The 40 most stable voxels were selected from the *diet* ROIs and 40 from the *habitat* ROIs. (The set size of 40 voxels was convenient, but other sizes, such as 30–80, resulted in similar outcomes.) The stability of a voxel was computed as the average pairwise correlation between its activation profiles (vector of its activation levels across the 16 items) across the repetitions in a training data subset [Just et al., 2010]. A voxel's activation level was its mean across the seven brain images acquired within a 7s window, offset 4s from the stimulus onset (to account for the delay in hemodynamic response). Figure 2 details the distribution of the selected voxels over the ROIs.

For each partitioning into training and test data, the voxel selection criterion was applied to the training set and the classifier was trained to associate an activation pattern to each of the 16 item labels. Four (out of the six) repetitions of each item were used for training and the mean of the remaining two repetitions was used for testing, thus making 15 total partitionings into training and test data. The activation values of the 40 voxels were normalized (mean = 0, SD = 1) across all the items, separately for the training and test sets, to correct for possible drift in the signal across the six repetitions. Rank accuracy (referred to as *accuracy*) was the percentile rank of the correct item in the classifier's ranked output [Mitchell et al., 2004].

RESULTS

After Instruction about a Feature, There Were Changes in Brain Regions That Encode That Feature Knowledge, Which Were Manifested as (I) Increases in Activation Levels and (II) Increases in the Accuracies of multi-Voxel Pattern Classification of the Animal Concepts

One main finding was that during the *thinking* trials that followed instruction about the second feature, there were increases in the activation levels [$F(1, 14) = 5.01$, $P < 0.05$] in the a priori-specified brain regions that encode the instructed feature knowledge (as predicted by Hypothesis 1A). Activation increased bilaterally in parahippocampal gyrus and precuneus after instruction about *habitat*, and in the left inferior frontal gyrus after instruction about *diet*. Figure 3A shows that the activation level in the specified regions increased after information about the feature relevant to the region (*habitat* or *diet*) had been taught.

A second main finding was that the animal concepts enriched by knowledge about the second-taught feature became reliably more classifiable, using voxels from within these brain regions as classifier features [$F(1, 14) = 7.16$, $P < 0.05$; increase in mean accuracy from 0.52 to 0.56], as predicted by Hypothesis 1B. (The $P < 0.05$ probability threshold for a rank accuracy being greater than chance level is 0.546.) Note that an increase in the mean activation level for the animal concepts would not in itself necessarily make the concepts more discriminable. The finding of greater classification accuracy after instruction indicates that the patterns of activation in the regions used by the classifier became more systematically individuated after instruction. As shown in Figure 3B, the classification accuracy based on the specified regions reliably increased after information about the feature relevant to the region (*habitat* or *diet*) had been taught.

In summary, the integration of new feature knowledge into each animal concept was reflected in increases in mean activation levels and classification accuracies within the a priori-specified brain regions. The findings occurred for two types of feature knowledge (*habitat* and *diet*) and two animal categories (mammals and birds).

After Instruction about a Feature, Animals that were described as Having Similar Features Became Neurally More Similar to Each Other with Respect to That Feature

Although the animal features that were taught were not intended to have any particular semantic relation to each other, there were several cases in which different animals had similar features. It seemed plausible that animals that were similar to each other based on a given feature would become neurally more similar following instruction about that feature, in terms of their voxel activation patterns in the a priori-specified *habitat*- or *diet*-related brain regions. For example, two of the *habitat* features that were similar are "Lived around water in ice caverns" and "Lived near beaches in caves"; two of the *diet* features that were similar are "Ate eel meat with serrated teeth" and "Ate crab meat with sharp teeth." An independent group of raters selected four pairs of animal feature descriptions that were highly similar, out of the 28 possible pairs between the eight animals (for *habitat* and *diet* each). The pairs that were selected the most frequently across the raters were designated the most similar animals; Table I contains the

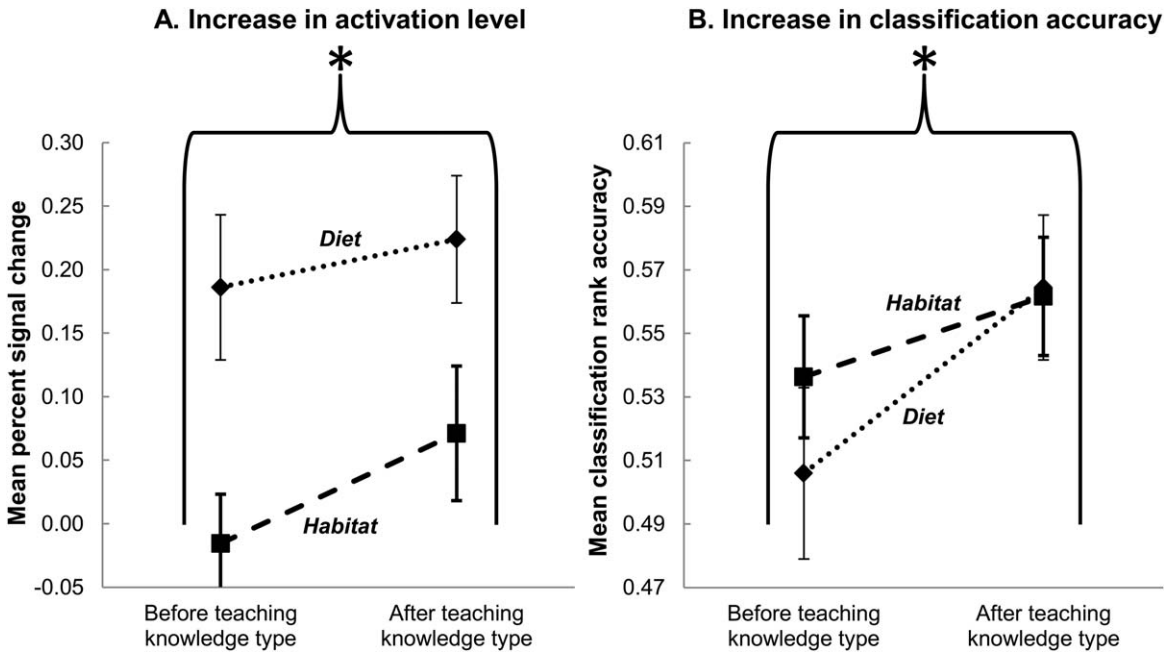


Figure 3.

Emergence of the neural knowledge representation of a concept feature. After the teaching of feature information about a knowledge type (*habitat* or *diet*), there were increases in (**A**) activation levels and (**B**) multivoxel pattern classification accuracies of the animal concepts in brain regions a priori predicted to encode that knowledge type. Rank accuracy was the percentile rank of the correct item in the classifier's ranked output. The $P < 0.05$ probability threshold for an accuracy being greater than chance level is 0.546. Error bars are standard error of the mean. *$P < 0.05$, main effect of teaching of feature knowledge.

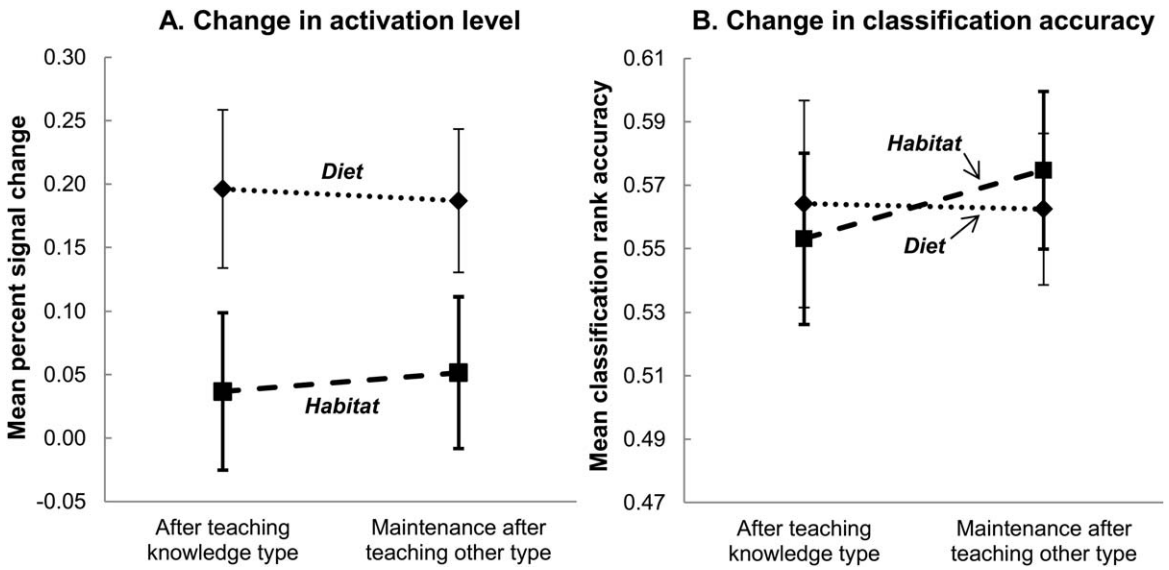


Figure 4.

Retention of the neural knowledge representation of a recently taught feature. The (**A**) activation levels and (**B**) classification accuracies in brain regions a priori predicted to encode the first-instructed feature knowledge type retained their recently heightened levels into the learning period of the second knowledge type (no main effect of instruction about the second knowledge type, as expected). Error bars are standard error of the mean.

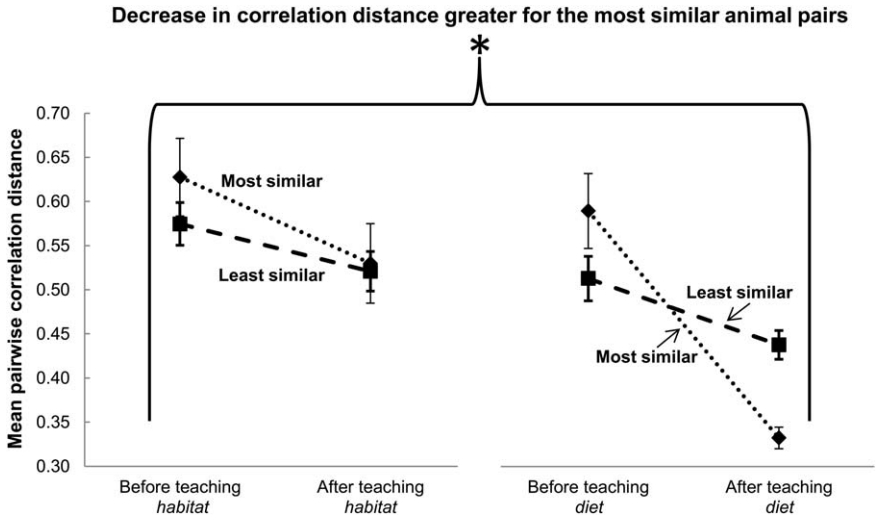


Figure 5.

Decrease in neural dissimilarity between animals with similar features. After instruction about a feature (*habitat* or *diet*), the pairwise correlation distance (neural dissimilarity) decreased for the pairs of animals judged to be most similar with respect to that feature, relative to the pairs of animals that were judged to be least similar. The decrease occurred in the brain regions a priori predicted to encode the new feature knowledge. Error bars are standard error of the mean. $*P < 0.05$ using permutation testing.

feature descriptions for all the animals and indicates these four similar pairs for each feature type. The results showed that after instruction about the second feature, the pairs of animals with the most similar features became more neurally similar than did the other pairs ($P < 0.05$ using permutation testing). Note that the features were presented only in association with an animal, and never as pairs of features. As shown in Figure 5, there was a greater decrease in the pairwise correlation distance (i.e., greater similarity) between the activation patterns of the pairs of animals with similar features than for the other pairs (Hypothesis 1C). Six of the eight total pairs of animals with similar features consisted of a mammal and a bird, which indicates that the increase in neural similarity with learning was based on feature commonality irrespective of the superordinate category of an animal.

Note that the *decrease* in dissimilarity (increased similarity) among the animals' activation patterns is compatible with the finding of an *increase* in the discriminability of the animals (increased classification accuracy) following instruction on the second feature. Classification accuracy is determined not only by the dissimilarities among the animals being classified, but also by the level of noise in the neural representation (measured by 1—the mean correlation among all the repetitions of a given animal's activation patterns). Thus, the increase in classification accuracy following instruction is likely accounted for by a decrease in the noise level. The noise level was indeed generally lower (but not reliably so) after instruction versus before instruction.

The Neural Representation of a Previously Taught Feature Remained in Place after Instruction about a Different Feature Had Occurred

During the *thinking* trials following instruction about the second feature, the neural representation of the first-taught feature remained at its recently heightened activation level, indicating that the previously acquired feature knowledge had been retained (as predicted by Hypothesis 2A). Activation levels were maintained in bilateral precuneus and bilateral parahippocampal gyrus (which encode *habitat*), and in the left inferior frontal gyrus (*diet*). (The mean activation levels did not change and there was no main effect of instruction about the second feature: $F(1, 14) = 0.03$, $P = 0.89$.) Figure 4A depicts the maintenance of the recently heightened activation level in the voxels that encode the first-instructed feature knowledge (*habitat* or *diet*).

Another main finding was that even after instruction about the second feature, the animal concepts remained reliably classifiable using features (voxels) from within the brain regions specified for the first feature, as predicted by Hypothesis 2B [$F(1, 14) = 0.31$, $P = 0.59$, with a mean accuracy of 0.56 before and 0.57 after instruction about the second feature]. Figure 4B shows that the accuracy of the classification based on the voxels specific to the first feature was maintained. Thus, retention of the neural representation of the first-instructed feature was indicated by the persistence of both the activation levels and

classification accuracies within the brain regions specific to the first feature. These findings occurred for two types of feature knowledge and two animal categories.

An additional analysis tested the combined predictions that there would be evidence of an increment in learning after instruction on a feature (Hypotheses 1A-B—emergence of new feature knowledge) and evidence of the maintenance of learning on a previously instructed feature (Hypothesis 2—retention of formerly learned feature information). The standardized classification accuracies and brain activation levels were submitted to a single ANOVA as repeated-measures data. The ANOVA revealed an interaction effect that was statistically significant ($P < 0.05$, one-tailed) in the a priori-hypothesized direction. In the brain regions corresponding to the second-taught feature there was an increase in classification accuracy and activation level after instruction about the second feature, whereas in the regions corresponding to the first-taught feature there was a maintenance of classification accuracy and activation level.

During Instruction about a Feature, Activation Was Elevated in the Brain Regions Where the Neural Representation of That Feature Knowledge Eventually Emerged

Activation during instruction about a feature was elevated in the same brain regions that were a priori predicted and confirmed (as described above) to encode that learned feature knowledge (Hypothesis 3). The GLM contrast between the two types of *instruction* trial data (*diet* minus *habitat*) showed that activation was elevated in the left inferior frontal gyrus during instruction about *diet*, and not in the *habitat*-specific brain regions. Correspondingly, the opposite contrast (*habitat* minus *diet*) revealed that during instruction about *habitat*, activation was elevated in bilateral parahippocampal gyrus and precuneus (*habitat* regions), and not in *diet* brain regions. Figure 6B shows the activation clusters that survived an uncorrected height threshold of $t(15) = 2.13$ ($P < 0.05$, two-tailed) and an extent threshold of 5 voxels. The clusters were situated near the a priori ROIs that were used in the analysis of the *thinking* trial data, which are shown again in Figure 6A. The mean Euclidean distance between the centroids of the a priori regions and the observed *instruction*-trial corresponding clusters was 11.5 mm, indicating an activation commonality between learning about a type of feature and thinking about an animal that possesses that feature. (Interestingly, despite the activation in these regions being elevated during *instruction* trials, the activation patterns during those trials were not sufficiently distinguishable from each other to permit reliably accurate classification of which animal was the target of the instruction.)

The *diet* minus *habitat* contrast also revealed increased activation in an additional brain region that was not in the a priori set, namely the left postcentral gyrus, which has been implicated in thinking about physical manipulation of objects [Just et al., 2010]. It is possible that a description of what an animal ate entailed activation of a region that encodes the use of hands, feet, and limbs (for manipulation of food).

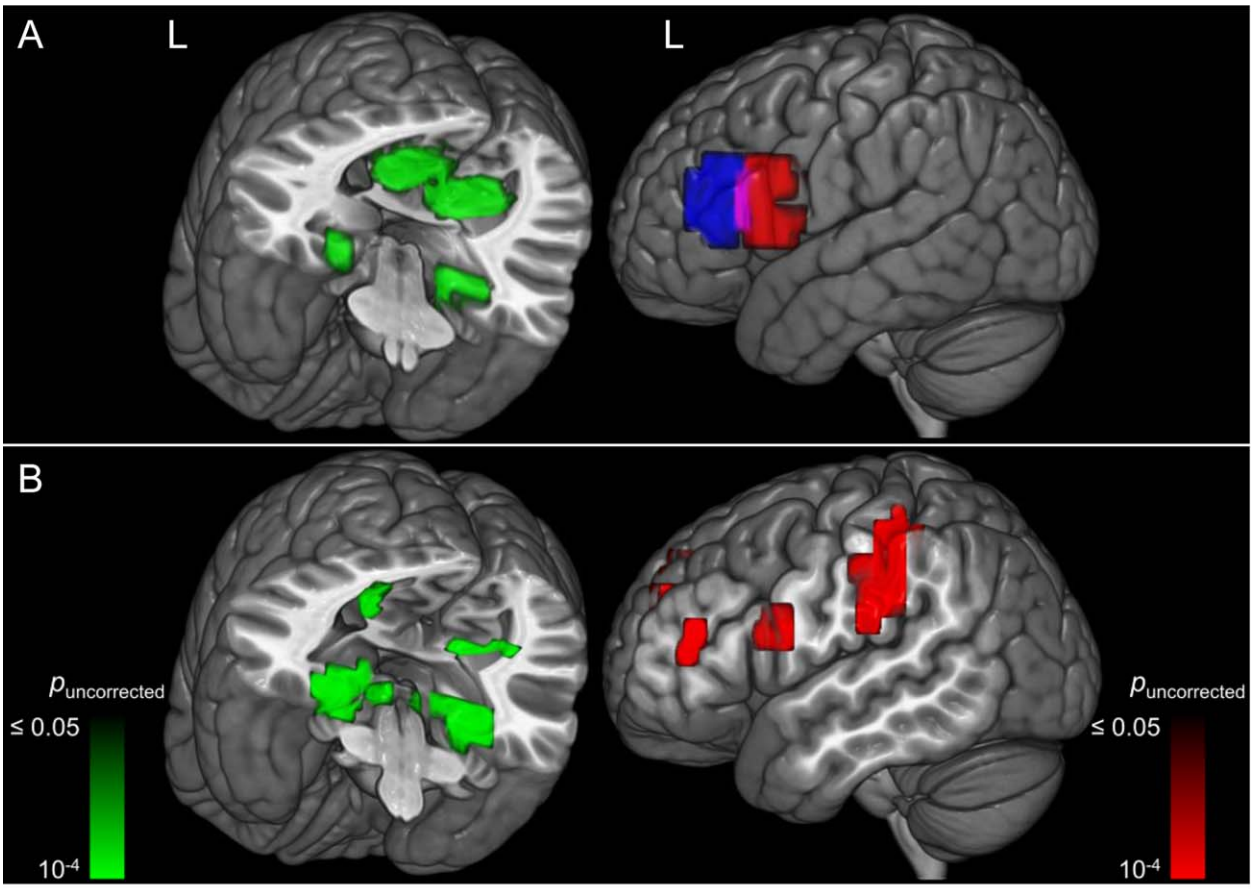
The Recall of the Features of Each Animal Provided Behavioral Evidence of Learning the Animal Concepts

After the scanning session, participants were asked to write down their recall of the two features for each of the eight animals, cued by only the name of each animal. The modal accuracy of the participants was 100% (recall of all 16 features). The mean accuracy was 75% (SD = 35%) because four participants had very poor recall, possibly because they thought of the animals more in terms of their pictures than their names, thereby making the name cue ineffective (one of these four participants volunteered this rationale during debriefing). There was no relation between a participant's recall accuracy either with his or her activation levels in the critical brain regions or with the classification accuracies. In summary, these results provided behavioral evidence that the participants learned the knowledge whose neural representation was investigated (as predicted by Hypothesis 4).

DISCUSSION

We can now specify much more precisely than ever before what happened in the brains of the millions of people who learned that the olinguito eats mainly fruit instead of meat. A region of the left inferior frontal gyrus as well as several others encoded this semantic information, and the information encoding remained intact as they continued to learn other facts about the olinguito. The new knowledge gained from the Smithsonian Institution's announcement became encoded in the brain areas that were predicted to contain this type of information.

This study is among the first to document the *establishment* of a neural knowledge representation of a newly learned concept, and furthermore, the representation was present within brain regions that were predicted a priori to represent the new knowledge about a particular feature type. Although the absolute classification accuracy of the animal concepts after learning was low, it is possible that it would have been higher had the paradigm permitted more time for consolidation of the new knowledge, such as time for intervening sleep [Stickgold, 2005]. But the important finding was that the *increase* in accuracy after learning was statistically reliable, reflecting the emergence of the neural representation of the learned feature knowledge. Furthermore, animals that were described in similar terms with respect to a feature type became neurally more similar to each other within the brain regions that encode that feature, demonstrating a close correspondence



Feature knowledge type	Cluster location	Cluster centroid (MNI)			No. voxels in cluster
		x	*y*	*z*	
Habitat	L parahippocampal gyrus	-21	-41	-14	101
	R parahippocampal gyrus	23	-36	-16	54
	L precuneus	-26	-62	9	20
	R precuneus	30	-53	3	37
Diet	L inferior frontal gyrus	-59	10	20	9
	L mid/inferior frontal gyrus	-45	40	15	8
	L postcentral gyrus	-58	-29	40	69

Figure 6.

Activation commonality between learning and thinking about a feature. There was a commonality of locations between (**A**) the a priori ROIs (*habitat*, left) used in the analysis of the *thinking* trial data and (**B**) the activation clusters of the *instruction* trials. The table above contains information about the activation clusters in **B**. Rendering was performed on an MNI template brain using the 3D medical imaging software MRIcroGL (Rorden and Brett, 2000). *L*: left; *R*: right; *ROI*: region of interest; *MNI*: Montreal Neurological Institute template.

between the neural changes and the specific information that was taught.

These results thus constitute a first step in documenting how a simple addition to the knowledge about a concept can selectively change one part of the neural representation of that concept. Moreover, the change in one part of a neural representation that was brought about by instruction remained intact after an addition to the concept knowledge (pertaining to a different feature) had been made. In this way, the growth of the neural representation of an individual novel animal concept was monitored across successive stages of feature learning.

Although the reported results focused on the representation of new knowledge as it was manifested during the *thinking* trials (when participants thought about the animals concepts), there was also evidence that participants processed information about the feature that was being taught during the *instruction* periods. Specifically, activation levels during instruction about a feature were elevated precisely in those brain regions that were here confirmed to encode the acquired knowledge about that feature. These were the same regions that showed increased activation in the *thinking* trials after instruction. These results offer convergent evidence for the conclusions.

Collectively, the results show that *before* instruction about a feature, there were no stored representations of the new feature knowledge; and *after* instruction, the feature information had been acquired and stored in the critical brain regions. The activation patterns in the regions that encode the semantic information that was taught (*habitat* and *diet*) changed, reflecting the specific new concept knowledge. This study provides a novel form of evidence (i.e., the emergence of new multivoxel representations) that newly acquired concept knowledge comes to reside in brain regions previously shown to underlie a particular type of knowledge [for a review of theories of semantic representation in the brain see Meteyard et al., 2010 and Kiefer and Pulvermüller, 2012]. Furthermore, this study provides a foundation for brain research to trace how a new concept makes its way from the words and graphics used to teach it, to a neural representation of that concept in a learner's brain.

Monitoring the developmental trajectory of the neural representation of a learned concept may constitute a method for investigating this transduction process. For example, changes in the anatomical location and particular configuration of the activation pattern underlying a new concept could be tracked as learners deepen their understanding of the concept. A learner could be scanned after each exposure to an unchanging concept; or, as in the current study, a concept could be built up over time as different features of that concept are taught. In either case, a characterization of each stage of the learner's understanding of the concept would be related to its corresponding neural representation, which could reveal how and why the brain activation is reconfigured with learning.

Monitoring the growth of a new neural knowledge representation may also complement methods that investigate the structure of existing knowledge. It may eventually be possible to characterize the transition from newly learned to old knowledge in terms of its degree of usage, relatedness to other concepts, elapsed time, or other properties yet to be discovered. It may also be possible to use an analogous approach to characterize the decline of knowledge (i.e., forgetting) in terms of similar properties. Furthermore, characterizations of pathological declines of knowledge (e.g., frontotemporal dementia) may be amenable to this approach.

Several previous studies have demonstrated structural brain changes that accompanied learning. For example, a longitudinal study using diffusion tensor imaging showed that instruction-based improvements in reading ability changed the structural integrity of the cortical white matter in children with poor reading skills [Keller and Just, 2009]. In that case, the amount of change in white matter was correlated with the amount of change in reading skill. Several other studies have reported changes in grey matter morphology as a result of learning (see Fields, 2011, for a review of learning-induced structural changes in grey and white matter). For example, an MRI voxel-based morphometric comparison of gray matter before and after learning to juggle revealed structural changes in left posterior intraparietal sulcus, a brain area that underlies the ability to track moving objects [Draganski et al., 2004]. The new results here complement this body of research by documenting a change in brain function (activation patterns) accompanying the emergence of new conceptual knowledge, versus the enhancement of an intellectual skill (reading) or motor skill (juggling). It may soon become possible to measure activation-related changes in knowledge representation simultaneously with gray and white matter changes, in conjunction with various types of knowledge acquisition.

The findings of the current study may foreshadow a capability to apply brain imaging and multivoxel pattern analyses to assess the progress in learning a complicated concept—such as that of a high-school physics lesson—by monitoring the changes in the concept's neural representation as new features or aspects of the concept are learned. Furthermore, real-time feedback throughout learning could be administered, based on brain activity in earlier portions of learning a concept's features, to guide the ensuing portions of learning. A recent fMRI study in which real-time measurement of brain activation identified mental states that were either "prepared" or "unprepared" for encoding a new stimulus, lends credence to this possibility [Yoo et al., 2012]. Recognition memory for the stimulus was higher in the case of the prepared brain state condition, demonstrating that brain activation measures could be used to identify when instruction should be administered (versus when it should not be), thus increasing the efficiency of the learning. Multivoxel pattern analyses could be used to diagnose which aspects of a concept a student misunderstands (or lacks), in a manner that might be more fundamental and accurate than traditional test-based assessment. The ability to track the growth of a neural knowledge representation speaks to the foundation of cognitive neuroscience research, which seeks to understand the neural basis of knowledge acquisition.

Future study is needed to determine how the multivoxel representational pattern of a recently acquired concept changes with further increments in concept knowledge. In the current study, retention of the neural representation of the first-taught feature was confirmed by the persistence of the multivoxel pattern classification accuracy over time.

But in other situations, the accumulation of concept knowledge need not be additive: a newly added feature could potentially modify a previous feature representation or otherwise modify the entirety of the concept representation. It may be fruitful to apply the current techniques to the study of different types of concept growth in future research.

Research in these directions and others will add to an emerging understanding of the growth of neural knowledge representations. This understanding could help integrate two fundamental dimensions of human cognition—namely knowledge representation and concept learning—and also perhaps enable the application of neuroscientific results to improvements in instructional design.

ACKNOWLEDGMENTS

The authors thank Charles Kemp and Tom Mitchell for helpful suggestions regarding the experimental design, data analysis, and writing of the paper.

REFERENCES

Capitani E, Laiacona M, Mahon B, Caramazza A (2003): What are the facts of semantic category-specific deficits? A critical review of the clinical evidence. Cogn Neuropsychol 20:213–261.

Chao LL, Martin A (2000): Representation of manipulable man-made objects in the dorsal stream. NeuroImage 12:478–484.

Damarla SR, Just MA (2013): Decoding the representation of numerical values from brain activation patterns. Hum Brain Mapp 34:2624–2634.

Draganski B, Gaser C, Busch V, Schuierer G, Bogdahn U, May A (2004): Neuroplasticity: Changes in grey matter induced by training. Nature 427:311–312.

Epstein RA, Higgins JS (2007): Differential parahippocampal and retrosplenial involvement in three types of visual scene recognition. Cereb Cortex 17:1680–1693.

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

Fields RD (2011): Imaging learning: The search for a memory trace. Neuroscientist 17:185–196.

Hauk O, Johnsrude I, Pulvermuller F (2004): Somatotopic representation of action words in human motor and premotor cortex. Neuron 41:301–307.

Huth AG, Nishimoto S, Vu AT, Gallant JL (2012): A continuous semantic space describes the representation of thousands of object and action categories across the human brain. Neuron 76:1210–1224.

James TW, Gauthier I (2003): Auditory and action semantic features activate Sensory-specific perceptual brain regions. Curr Biol 13:1792–1796.

Joachims T (1999): Making large-scale SVM learning practical. In: Schölkopf B, Burges C, Smola A, editors. Advances in Kernel Methods: Support Vector Learning, Cambridge, MA: MIT Press. pp 169–184.

Just MA, Cherkassky VL, Aryal S, Mitchell TM (2010): A neurosemantic theory of concrete noun representation based on the underlying brain codes. PLos One 5:e8622

Kassam KS, Markey AR, Cherkassky VL, Loewenstein G, Just MA (2013): Identifying emotions on the basis of neural activation. PLos One 8:e66032

Keller TA, Just MA (2009): Altering cortical connectivity: Remediation-induced changes in the white matter of poor readers. Neuron 64:624–631.

Kiefer M, Pulvermüller F (2012): Conceptual representations in mind and brain: Theoretical developments, current evidence and future directions. Cortex 48:805–825.

Kriegeskorte N, Mur M, Bandettini P (2008): Representational similarity analysis – connecting the branches of systems neuroscience. Front Syst Neurosci 2:4 doi:103389/neuro06002008.

Mahon BZ, Caramazza A (2003): Constraining questions about the organisation and representation of conceptual knowledge. Cogn Neuropsychol 20:433–450.

Martin A (2007): The representation of object concepts in the brain. Annu Rev Psychol 58:25–45.

Meteyard L, Cuadrado SR, Bahrami B, Vigliocco G (2010): Coming of age: A review of embodiment and the neuroscience of semantics. Cortex 48:788–804.

Mitchell TM, Hutchinson R, Niculescu RS, Pereira F, Wang X, Just M, Newman S (2004): Learning to decode cognitive states from brain images. Mach Learn 57:145–175.

Mur M, Bandettini PA, Kriegeskorte N (2009): Revealing representational content with pattern-information fMRI–An introductory guide. Soc Cogn Affect Neurosci 4:101–109.

Rorden C, Brett M (2000): Stereotaxic display of brain lesions. Behav Neurol 12:191–200.

Shinkareva SV, Malave VL, Just MA, Mitchell TM (2012): Exploring commonalities across participants in the neural representation of objects. Hum Brain Mapp 33:1375–1383.

Stickgold R : (2005): Sleep-dependent memory consolidation. Nature 437:1272–1278.

Tzourio-Mazoyer N, Landeau B, Papathanassiou D, Crivello F, Etard O, Delcroix N, Mazoyer B, Joliot M (2002): Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. NeuroImage 15:273–289.

Vann SD, Aggleton JP, Maguire EA (2009): What does the retrosplenial cortex do? Nat Rev Neurosci 10:792–802.

Weisberg J, van Turennout M, Martin A (2007): A neural system for learning about object function. Cereb Cortex 17:513–521.

Yoo JJ, Hinds O, Ofen N, Thompson TW, Whitfield-Gabrieli S, Triantafyllou C, Gabrieli JDE (2012): When the brain is prepared to learn: Enhancing human learning using real-time fMRI. NeuroImage 59:846–852.