Cerebral Cortex Advance Access published June 10, 2011

Cerebral Cortex
doi:10.1093/cercor/bhr036

Learning the Exception to the Rule: Model-Based fMRI Reveals Specialized Representations for Surprising Category Members

Tyler Davis[1], Bradley C. Love[1] and Alison R. Preston[1,2,3]

[1]Department of Psychology, [2]Center for Learning and Memory and [3]Institute for Neuroscience, The University of Texas at Austin, Austin, TX 78712, USA

Address correspondence to Tyler Davis, Department of Psychology, The University of Texas at Austin, 1 University Station, A8000, Austin, TX 78712, USA. Email: Thdavis@mail.utexas.edu.

Category knowledge can be explicit, yet not conform to a perfect rule. For example, a child may acquire the rule "If it has wings, then it is a bird," but then must account for exceptions to this rule, such as bats. The current study explored the neurobiological basis of rule-plus-exception learning by using quantitative predictions from a category learning model, SUSTAIN, to analyze behavioral and functional magnetic resonance imaging (fMRI) data. SUSTAIN predicts that exceptions require formation of specialized representations to distinguish exceptions from rule-following items in memory. By incorporating quantitative trial-by-trial predictions from SUSTAIN directly into fMRI analyses, we observed medial temporal lobe (MTL) activation consistent with 2 predicted psychological processes that enable exception learning: item recognition and error correction. SUSTAIN explains how these processes vary in the MTL across learning trials as category knowledge is acquired. Importantly, MTL engagement during exception learning was not captured by an alternate exemplar-based model of category learning or by standard contrasts comparing exception and rule-following items. The current findings thus provide a well-specified theory for the role of the MTL in category learning, where the MTL plays an important role in forming specialized category representations appropriate for the learning context.

Keywords: category learning, category representation, exception learning, hippocampus, medial temporal lobe, SUSTAIN

Introduction

Is this person a friend or a foe? Is this plant edible or poisonous? Organisms use categories to make decisions that range from critical to mundane. Categories facilitate generalization based on previous experiences, thus enabling inferences about future events. A central challenge in category learning research is determining how category knowledge is represented. Many category learning models propose a single fixed form of representation, such that all category representations share a common form, irrespective of the objective structure of the category information in the environment. For example, exemplar models always represent categories as collections of previously encountered category examples (Medin and Schaeffer 1978; Nosofsky 1986; Kruschke 1992), prototype models always represent categories by a single average (i.e., abstraction) of category members (Posner and Keele 1968; Rosch 1973; Smith and Minda 1998), and rule-based (Bruner et al. 1956; Trabasso and Bower 1968) and decision bound models (Ashby and Gott 1988; Ashby and Lee 1991) always represent categories by decision criteria that determine category membership.

Using these fixed representational approaches as building blocks, multiple systems' models (We use the term system to refer to systems that are theorized to contain functionally separate representations [e.g., rules vs. prototypes] as opposed to "systems" that may instantiate different component processes of categorization [e.g., object representation vs. decision processing] but operate on common representations.) combine 2 or more single system models to address behavioral findings outside the scope of most single system models (e.g., Nosofsky et al. 1994; Erickson and Kruschke 1998; Ashby et al. 1998). For example, Erickson and Kruschke (1998) combine exemplar and rule approaches to capture patterns of data that neither single system model can account for alone. In contrast to most single and multiple systems' models that represent all forms of categories in the same way, cluster-based models build appropriate representations for a given category structure (Anderson 1991; Love et al. 2004). This flexibility allows cluster-based models to manifest characteristics of various single system models; the model's behavior is determined by the demands of the learning context. While it is well established that people do build category representations that are appropriate for a category's structure, there is little work describing the neural processes that accomplish this feat.

Many of the fixed representational forms proposed by category learning models have been ascribed to systems in the brain that are associated with learning categories (for reviews, see Ashby and Maddox 2005; Ashby and O'Brien 2005; Poldrack and Foerde 2008; Smith and Grossman 2008). The prefrontal cortex and head of the caudate nucleus are theorized to engage a rule-based category learning system that depends on working memory to support maintenance of rules and new hypothesis testing (Ashby et al. 1998; Seger et al. 2000; Monchi et al. 2001; Patalano et al. 2001; Maddox and Ashby 2004; Seger and Cincotta 2006). The tail and body of the caudate nucleus are theorized to support a category learning system that involves the strengthening of associations between individual stimuli and category responses, often described as procedural learning (Knowlton et al. 1994, 1996; Ashby et al. 1998; Poldrack et al. 2001; Maddox and Ashby 2004; Shohamy et al. 2004; Foerde et al. 2006, 2007). Exemplar representations have been associated with regions in the temporal and occipital lobes that are associated with processing objects (e.g., Sigala and Logothetis 2002; Palmeri and Gauthier 2004; Palmeri and Tarr 2008).

One neurobiological system that has proven difficult to characterize in terms of its role in category learning is the medial temporal lobe (MTL). The essential role of the MTL for encoding and retrieval of declarative memories—long-term memory for facts and events—is well established (Scoville and

Milner 1957; for reviews, see Squire 1992; Preston and Wagner 2007). However, the role of the MTL in category learning remains controversial; each of the major fixed representational forms have been ascribed to the function of the MTL by different groups of researchers. For example, many theories suggest that the MTL uses exemplar-based representations (Pickering 1997; Ashby and Maddox 2005; Ashby and O'Brien 2005). However, empirical work has suggested that the MTL may be essential for the storage of category rules (Seger and Cincotta 2006; Nomura et al. 2007) or representations of category prototypes (Aizenstein et al. 2000; Reber et al. 2003; Zaki et al. 2003; Zeithamova et al. 2008). In contrast, other theories question whether the MTL is involved in category learning at all (Ashby et al. 1998; Maddox and Ashby 2004). Given these difficulties in ascribing a single fixed representational type to the function of the MTL, one plausible alternative that may integrate these disparate theories is that the MTL builds representations that are appropriate for a specific learning context, like those proposed by clustering models (e.g., Anderson 1991; Love et al. 2004).

In the current study, we present a theory of MTL function in category learning that relies on a category learning model, SUSTAIN (Love et al. 2004). This model-based approach provides a method for formalizing our hypothesis that the MTL builds representations that are appropriate for a category's structure (Love and Gureckis 2007). SUSTAIN is a network-based category learning model that works by comparing the similarity between incoming stimuli and category representations that are stored in memory. SUSTAIN represents categories as clusters that code the features and categories of items that it has experienced before. SUSTAIN initially represents categories as simply as possible but is able to dynamically recruit new category representations (clusters), if needed, to solve a task. For example, for a task in which subjects learn categories that are defined by rules or a simple prototype structure, SUSTAIN will encode a single cluster to represent each category. For tasks in which accurate performance cannot be achieved by storing a single cluster per category, SUSTAIN will encode additional clusters as needed to represent the category. If a category structure is complex enough, SUSTAIN will act like an exemplar model storing every category member in its own cluster.

Our cluster-based theory of MTL function in category learning is informed by current theories of the representational and functional roles of the MTL and its subregions in declarative memory. In terms of declarative memory, the hippocampus is theorized to play a critical role in rapidly forming conjunctive representations that bind together different sources of information into a single flexible memory (Brown and Aggleton 2001; Norman and O'Reilly 2003; Eichenbaum et al. 2007). Conjunctive representations are thought to be encoded by the hippocampus in response to novelty (Stern et al. 1996; Tulving et al. 1996; Yamaguchi et al. 2004), in as little as a single trial (Morris et al. 1982; Rutishauser et al. 2006), as well as code information about the spatiotemporal context in which an item occurred (Wallenstein et al. 1998; Staresina and Davachi 2009). SUSTAIN's clusters resemble hippocampal conjunctive representations in that they can be dynamically recruited in response to novelty on a single trial and that they bind together multiple item features and category information into a single flexible representation that can promote generalization to novel contexts (e.g., Yamauchi et al. 2002; Love et al. 2004).

In addition to the hippocampus, components of SUSTAIN's architecture may relate to functionality in other regions of the MTL. For example, the perirhinal cortex is associated with representing object-level information (Wan et al. 1999; Davachi et al. 2003; Staresina and Davachi 2008), which contrasts with hippocampal representations that bind object-level and category/contextual information. Stimulus matching mechanisms in SUSTAIN that compute an item's similarity to stored representations may relate to object-based representations in the perirhinal cortex (Love and Gureckis 2007).

Our theory relating SUSTAIN's representational properties to the MTL may help to explain how factors like expectations, goals, and category structure combine to influence how category representations are formed. To foreshadow our study design, many real world categories often appear to be describable by simple representations, such as logical rules, but upon closer inspection are found to be more complex (Wittgenstein 1953/2001). For example, natural categories such as birds and mammals are often associated with verbalizable rules such as, "If it has wings, it is a bird" but also contain violations of these rules, such as bats. People can verbally report descriptions of bats and explicitly relate bats to other mammals, but these descriptions are not rules, per se. In order for people to learn that examples as diverse as bats and ponies are all members of the category mammals, people need to build representations of the category mammals that are appropriate for this goal. SUSTAIN would predict that people achieve this goal by forming a separate cluster for birds and mammals and then creating additional specialized clusters for exceptions, like bats, as they are encountered. Other models that are able to learn the task, such as exemplar models, may accomplish the goal of categorizing birds and mammals but would do so by storing each member separately, making all items equally differentiated in memory. Given the need to store exceptions as specialized representations, we predict exceptions and rule-following category members will be differentiated to the extent that encoding and retrieval processes in the hippocampus and surrounding MTL cortex are engaged.

We use a model-based functional magnetic resonance imaging (fMRI) approach to test the proposed mapping between MTL function and SUSTAIN's representational properties by collecting fMRI data during a rule-plus-exception task that preserves critical aspects of the mammals and birds example. During scanning, subjects learn to sort schematic beetles (Fig. 1*A*) into categories based on trial and error. Each trial of the category learning task contains a stimulus presentation period in which subjects try to predict the correct category assignment for a single beetle followed by a feedback period during which subjects are told whether they are right or wrong and the correct category assignment (Fig. 1*B*). Subjects are instructed that most of the beetles (i.e., rule-following items) in each category can be classified accurately by using a rule based on a single stimulus dimension that they were cued to attend to but also that each category contains an exception item that appears as if it should belong to the opposing category based on the rule. After scanning, a surprise memory test queried subjects' memory for both the exception and the rule-following items.

SUSTAIN learns rule-plus-exception tasks in a manner analogous to the birds and mammals example. Thus, it can provide a mechanistic account of how psychological processes associated with category learning unfold in a rule-plus-exception task. SUSTAIN predicts that subjects will build representations

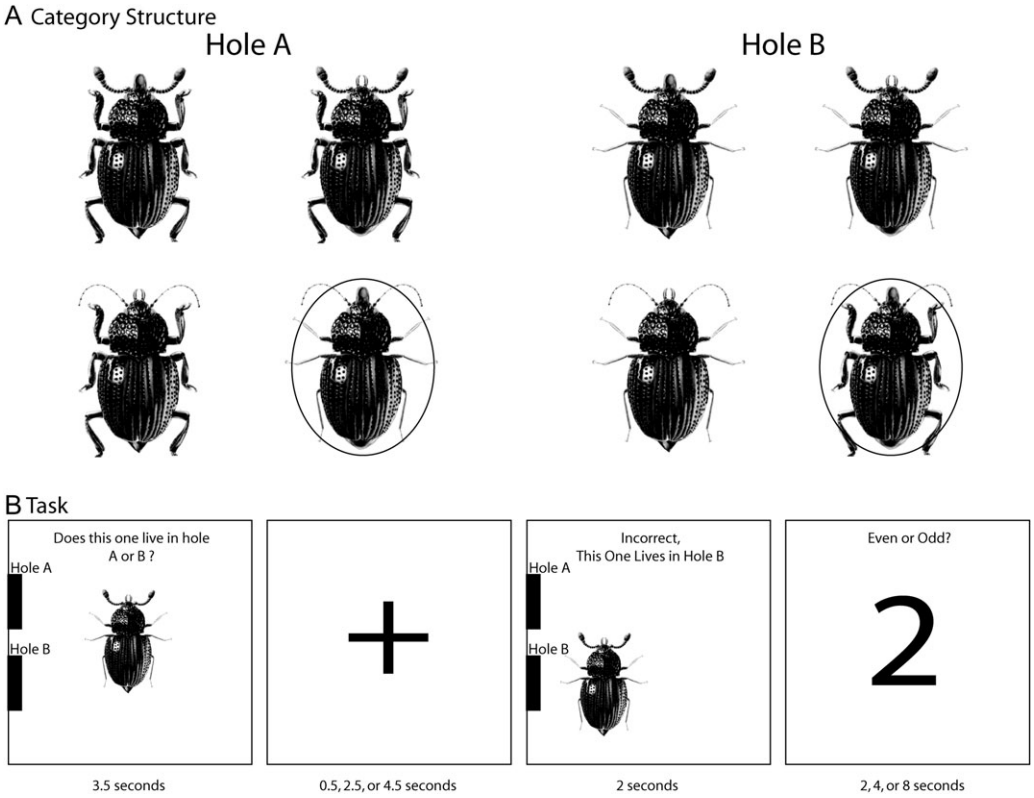

Figure 1. (*A*) An example category structure. The beetles vary on 4 of the following 5 perceptual dimensions, where the fifth dimension is held fixed: eyes (green or red), tail (oval or triangular), legs (thin or thick), antennae (spindly or fuzzy), and fangs (pointy or round). The rule-relevant dimension in this example is legs. Most (3/4) of Hole A beetles have thick legs, whereas most (3/4) of Hole B beetles have thin legs. The 2 stimuli circled are the exceptions because they have legs consistent with the opposing category. The rest of the features are evenly distributed across the exemplars, with the exception of eyes, which is held constant in this example (for abstract structure, see Supplementary Table S6). (*B*) Trial structure. During stimulus presentation, a beetle was presented, and subjects were asked to classify an either Hole A or Hole B beetle. Following a variable fixation period, subjects received feedback about their response. Feedback was followed by a variable number of even–odd digit trials that served as baseline.

that are appropriate for the task by encoding rule-following items in common clusters and recruiting additional clusters to store exception items. Based on the theoretical relationship between SUSTAIN and the brain, encoding and retrieval of both item types are hypothesized to be dependent on the MTL. However, because exception items are predicted to be more differentiated than rule-following items in memory, SUSTAIN predicts different levels of MTL involvement in encoding and retrieval processes for the 2 item types.

We test our theory relating SUSTAIN's representational properties to MTL function through model-based analyses of fMRI data. We focus on SUSTAIN's characterization of recognition and error correction processes, which we hypothesize to correspond to the function of MTL in this task. For each of these 2 processes, we specify a quantitative measure that characterizes SUSTAIN's moment-to-moment operation and relate these measures to brain activation during stimulus presentation and feedback. Figure 2*A*,*B* illustrates these 2 measures, which capture SUSTAIN's psychological account of how encoding and retrieval processes dynamically change during learning.

According to our quantitative measures of SUSTAIN, exception and rule-following items should elicit different levels of activation across learning trials. We test whether subjects exhibit patterns of brain activation that are consistent with the recognition strength and error correction measures by including the measures directly in the fMRI analysis as parametric regressors. Parametric regression is used in fMRI data analysis to assess whether activation within a condition varies according to a specified measure. While this analysis technique has been used extensively in model-based fMRI paradigms in reinforcement learning (for review, see Daw forthcoming; O'Doherty et al. 2007; Gläscher and O'Doherty 2010), it has not yet been used to relate predictions from category learning models to fMRI data.

Model-based fMRI techniques offer a powerful method for assessing whether similar computational processes underlie SUSTAIN's operations and MTL involvement in rule-plus-exception learning. Because we fit the model to behavioral data and not to fMRI data directly, in contrast to other methods, there is no risk of over fitting the brain data. Indeed, predictions from SUSTAIN and subjects' pattern of brain activation during the task are measured independently, and thus finding a statistical relationship between them strongly suggests that subjects are using similar mechanisms to guide performance in the task.

In our analysis, the recognition strength measure derived from SUSTAIN (Fig. 2*A*) is used to predict activation during stimulus presentation, when we believe subjects are attempting to retrieve stored category representations to predict category membership. The recognition strength measure indicates the extent to which a stimulus matches SUSTAIN's stored cluster representations and is proposed to reflect stimulus matching and cluster retrieval processes that may occur in the hippocampus and MTL cortex. Exception and rule-

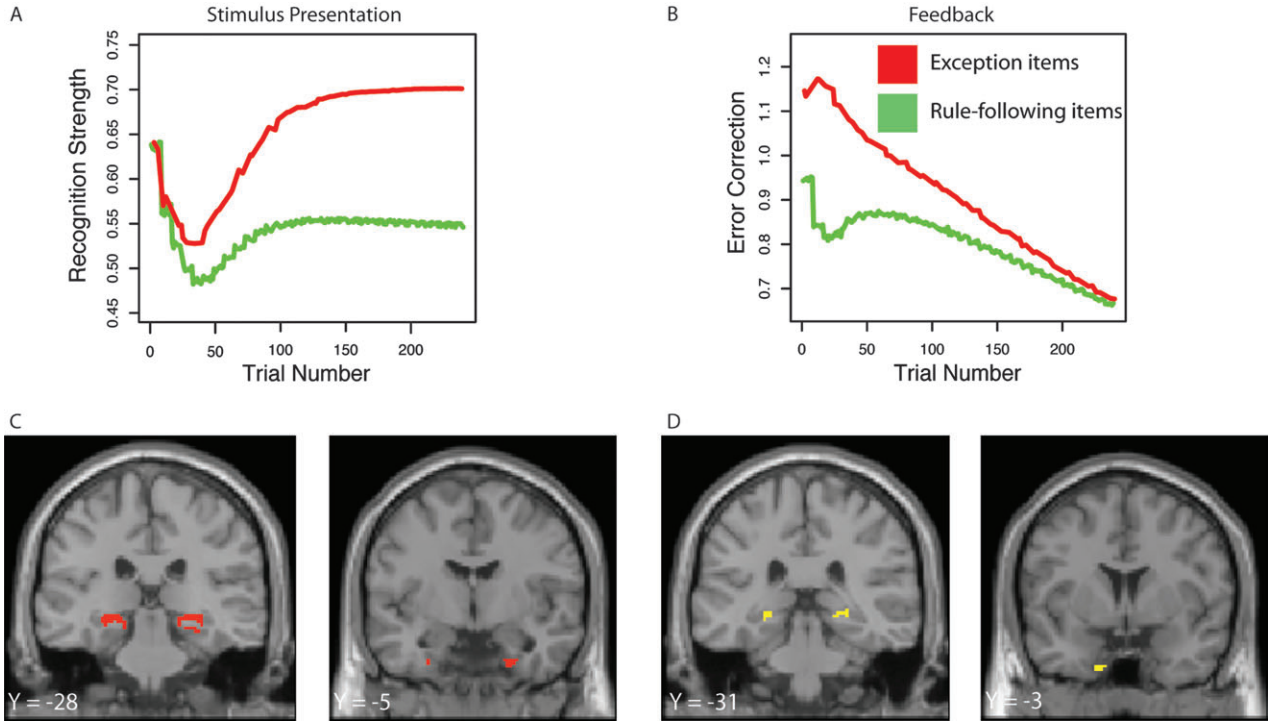

Figure 2. Illustrations of recognition strength (*A*) and error correction (*B*) measures derived from SUSTAIN that were used as predictors of brain activation during stimulus presentation and feedback. Recognition strength is the sum of the total cluster activation for a given stimulus/trial, and error is the absolute value of the model's output error on a given trial. Below the measures are the corresponding statistical maps associated with each regressor. Activation during stimulus presentation is presented in red and activation during the feedback period in yellow. (*C*) MTL regions exhibiting a significant ($P < 0.05$, false discovery rate [FDR] corrected) correlation between activity during the categorization period and the predicted recognition strength measure. (*D*) MTL regions exhibiting a significant ($P < 0.05$, FDR corrected) correlation between during feedback and the predicted error correction measure.

following items yield different recognition strength measures across trials such that exception items are predicted to have greater overall recognition strength. SUSTAIN predicts that exception items more closely match clusters stored in memory because these items match their own cluster perfectly and have moderate matches to clusters representing rule-following items from the opposing category. In contrast, rule-following items tend to only moderately match stored clusters because the rule-following clusters represent abstractions, or averages, of the rule-following items within a category rather than representations of the items themselves. The recognition strength measure predicts that this difference between rule and exception recognition will increase over trials as subjects learn.

The error correction measure derived from SUSTAIN (Fig. 2*B*) is used to predict brain activation during the feedback portion of a trial when subjects are told whether their response was correct or incorrect. The error correction measure indicates a mismatch between SUSTAIN's prediction for category membership and the actual category assignment. Thus, the error correction measure is predicted to relate to prediction error or novelty signals (Ranganath and Rainer 2003; Strange et al. 2005; Kohler et al. 2005; Kumaran and Maguire 2006, 2007a, 2007b) in the MTL cortex and hippocampus that lead to the encoding of new memories. SUSTAIN predicts that exceptions should lead to greater prediction error than rule-following items because exception items are somewhat confusable with rule-following items from the opposing category (see above discussion of the recognition measure). In contrast, rule-following items significantly activate only those clusters associated with correct rule application. For both items types, SUSTAIN's error correction measure decreases over trials as subjects learn.

SUSTAIN's predictions for exception learning have been validated in a number of behavioral studies and are consistent with the prediction that exceptions recruit MTL-based processes more than rule-following items. One robust finding from this literature is that exceptions items are better remembered than rule-following items at the end of learning (Sakamoto and Love 2004, 2006), even when subjects are not explicitly encouraged to use a rule (Palmeri and Nosofsky 1995). Early attempts to model subjects' behavior in rule-plus-exception tasks relied on a multiple system model that combined rule and exemplar-based representations (RULEX; Palmeri and Nosofsky 1995). Later research found that SUSTAIN was able to account for these findings and additional patterns of behavior in rule-plus-exception tasks not predicted by other models (Sakamoto and Love 2004, 2006). For example, a critical question in early research on exceptions was whether the high number of errors subjects make in learning exceptions drove the increase in exception recognition or whether it was because exceptions violate a salient knowledge structure (i.e., the rule). SUSTAIN successfully predicts that items that violate a knowledge structure result in stronger memories than those that are simply associated with high errors (Sakamoto and Love 2004). Finally, SUSTAIN is able to predict differences in exception learning performance between older adults, who have putatively impaired MTL function, and young adults by varying

a parameter corresponding to the degree of cluster recruitment, which is hypothesized to relate to MTL function (Love and Gureckis 2007).

In addition to evaluating our predictions derived from SUSTAIN, we also conduct model-based analyses based on an exemplar model, ALCOVE (Kruschke 1992), and standard condition-based fMRI analyses. In contrast to our view of MTL function in category learning, the prevailing view proposes that MTL function is best described by exemplar models (Pickering 1997; Ashby and Maddox 2005; Ashby and O'Brien 2005). Standard exemplar models provide a strong theoretical contrast to SUSTAIN. Whereas SUSTAIN builds representations appropriate for the structure of a category, exemplar models store all items separately in memory for all categories, regardless of their structure. Thus, where SUSTAIN is able to differentiate between rule-following and exception items, standard exemplar models hold that both types of items have the same status in memory and thus incorrectly predict that exception and rule-following items will be recognized at the same rates (One caveat comes from exemplar model variants that allow for exemplar-specific attention weights (Sakamoto and Love 2004; Rodrigues and Murre 2007) or update item representations on the basis of how much error they elicit (Sakamoto and Love 2004). These nonstandard exemplar models make predictions that are virtually indistinguishable from SUSTAIN because such models embody similar principles at the computational level (Sakamoto and Love 2004). ALCOVE was chosen to provide a strong theoretical contrast to SUSTAIN, not to serve as the standard-bearer for all possible variants of exemplar model.).

Model-based fMRI analyses using SUSTAIN should offer several advantages over standard general linear model (GLM)-based contrasts that sort items into experimentally defined conditions. Model-based analysis should better characterize within-condition variance due to learning and avoid some of the risks associated with standard fMRI analyses. Standard condition-based comparisons can mistakenly lead to a conclusion that a region exhibits functional specificity for a particular condition when the underlying processing difference between 2 conditions is more accurately described as a matter of degree (see Fig. 2). Our model-based analyses using recognition strength and error correction measures derived from SUSTAIN, along with their comparisons to ALCOVE and standard fMRI analysis techniques, should provide a window into the underlying computations and representations used by the MTL during category learning.

Materials and Methods

Subjects

Twenty-two healthy right-handed volunteers (13 females) ages 19–28 participated in the current experiment after giving informed consent in accordance with a protocol approved by the University of Texas at Austin Institutional Review Board. Each subject received a $50 payment for his or her participation. Seven additional subjects were excluded for failing to achieve greater than 50% performance on exception items in the final (sixth) run.

Materials

Subjects completed a rule-plus-exception category learning task (Love and Gureckis 2007) during fMRI scanning. The stimuli used in the task were schematic beetles that varied along 4 perceptual dimensions (see Fig. 1*A*) and were assigned to categories (Hole A or Hole B) based on their combinations of feature values. For each of the stimuli, 4 of 5 possible dimensions (eyes, tail, legs, antennae, and fangs) were randomly selected to vary and the unselected dimension was held fixed at a constant value. Six of the stimuli were rule-following items and could be categorized correctly based on the value of a single rule-relevant dimension. In the example in Figure 1*A*, the rule-relevant dimension was the legs; all but one of the beetles in Hole A had thick legs and all but one of the beetles in Hole B had thin legs. The other 2 beetles (red circles in Fig. 1*A*) served as exceptions to the rule and appeared to belong to the opposing category based on their value on the rule-relevant dimension (legs). An abstract representation of the category structure is given in Supplementary Table S6. In order to minimize the effects of feature salience, the mapping of each abstract dimension to a physical dimension was randomized for each subject.

Procedures

On each trial of the category learning task, a single beetle was presented in the center of the screen, and subjects were asked to decide whether it was a Hole A or Hole B beetle (Fig. 1*B*). Each stimulus was presented for 3.5 s during which time subjects had to indicate category membership via button boxes held in their left and right hands. After a brief fixation (0.5, 2.5, or 4.5 s; mean = 2.5 s), feedback was presented for 2.0 s during which time the beetle would appear next to the correct category (i.e., the correct Hole), and subjects were informed whether their response on that trial was correct or incorrect. In between categorization trials, subjects completed between 1 and 4 trials (2 s each) of an even/odd digit task that served as baseline (mean baseline time per trial = 4 s). Such active baselines are often used in memory research because the MTL has high resting state activity (Stark and Squire 2001). No feedback was given during the even/odd digit task.

Subjects were trained using the rule-plus-exception procedure for 6 functional runs, each lasting 8 min and 27 s. During each run, the 8 stimuli (beetles) were presented 5 times sequentially in a pseudorandom order. Trial order and duration were optimized for each of the 6 functional runs to allow for efficient deconvolution of the hemodynamic response using standard optimization techniques. A Latin square design was used to balance the order of the 6 functional runs across subjects. The first 12 s of each run, consisting of fixation, were discarded. Prior to beginning the task, subjects were given explicit instructions indicating the rule-relevant dimension for category membership and were encouraged to memorize the exceptions to the rule (Love and Gureckis 2007).

Following the category learning task, subjects completed a self-paced, 2-alternative forced choice recognition memory task outside of the scanner. On each trial of the recognition task, subjects were presented with 2 beetles: one that was presented during the category learning phase and a foil (see Supplementary Material) that was not presented during the category learning phase. Subjects were asked to identify the old item presented during the scanned rule-plus-exception task.

fMRI Data Acquisition

Whole-brain imaging data were acquired on a 3.0 T GE Signa MRI system (GE Medical Systems). Structural images were acquired during a T_2-weighted flow-compensated spin-echo pulse sequence (time repetition [TR] = 3 s; time echo [TE] = 68 ms; 256 × 256 matrix, 1 × 1-mm in-plane resolution) using thirty-one 3-mm thick oblique axial slices (0.6 mm gap), approximately 20 degrees off AC–PC line, oriented for optimal whole-brain coverage. Functional images were acquired using a multiecho GRAPPA parallel imaging echo planar imaging (EPI) sequence using the same slice prescription as the structural images (TR = 2 s; TE = 30 ms; 2 shot; flip angle = 90°; 64 × 64 matrix; 3.75 × 3.75-mm in-plane resolution). For each functional scan, the first 6 EPI volumes corresponding to the initial 12-s fixation period were discarded to allow for T_1 stabilization. Head movement was minimized using foam padding.

fMRI Data Analysis

Data were preprocessed and analyzed using SPM5 (Wellcome Department of Cognitive Neurology) and custom Matlab routines.

Functional images were corrected to account for the differences in slice acquisition times by interpolating the voxel time series using sinc interpolation and resampling the time series using the first slice as a reference point. Functional images were then realigned to the first volume in the time series to correct for motion. The T_2-weighted structural image was coregistered to the mean T_2^*-weighted functional volume computed during realignment. The structural image was then spatially normalized into common stereotactic space using the Montreal Neurological Institute template brain. The spatial transformation calculated during the normalization of the structural image was then applied to the functional time series and resampled to 2-mm isotropic voxels. After normalization, the functional images were spatially smoothed using an 8-mm full-width at half-maximum Gaussian kernel. Voxel-wise analysis was performed under the assumptions of the GLM. Regressor functions were constructed by modeling condition related activation as an impulse function convolved with a canonical hemodynamic response function.

For the model-based analysis, model-based measures of recognition strength and error correction were fit as parametric modulators of the stimulus presentation and feedback periods separately. In all cases, the model-based measures were obtained by fitting the model to subjects' aggregate behavioral performance (see Supplementary Material) and thus were not biased when interrogating fMRI data. In the primary analysis, the recognition strength and error correction measures of SUSTAIN (see Supplementary Methods) were included as parametric modulators of the stimulus presentation and feedback period of the trial, respectively. SUSTAIN's recognition strength measure indexes the sum of the cluster outputs across trials, providing a measure of the degree of match between a presented item and a stored memory representation. SUSTAIN's error correction measure indexes the difference between SUSTAIN's expectations for a queried dimension on a given trial and the actual outcome (i.e., the expected and given category label). Measures are averaged over 1000 runs of the model.

One key question is whether SUSTAIN's measures of recognition and error correction are specific to particular phases of the trial as predicted by our model and hypotheses regarding MTL function. To evaluate the explanatory power of each hypothesized process and its relationship to MTL activation during stimulus presentation and feedback, the role of the 2 measures in the parametric analyses was swapped such that the recognition strength measure was implemented as a parametric modulator of feedback and the error correction measure was implemented as a parametric modulator during stimulus presentation. If our predictions regarding the proposed mapping between psychological processes and MTL function are correct, this swapped analysis should be less effective in identifying MTL involvement in the task.

Another goal of the current study was to test the predictions for MTL function derived from SUSTAIN with alternate accounts of MTL function in category learning. To this end, an additional set of parametric analyses assessed whether predictions derived from ALCOVE, an exemplar model of category learning, isolated MTL involvement in exception learning. The parametric analysis using ALCOVE was performed in the same manner as the analysis using SUSTAIN. Paralleling the SUSTAIN analysis, ALCOVE's recognition strength and error correction measures (see Supplementary Material) were fit to the categorization and feedback periods, respectively.

Finally, a set of standard linear contrast analyses was conducted comparing exception and rule-following items. These analyses provide an important comparison to the model-based analyses. Unlike model-based analyses, simple linear contrasts do not model changes in the time course of activation as subjects learn the task. For these linear contrasts, we examined a GLM model that included regressors for each stimulus (exception or rule following) and subjects' behavior (correct or incorrect) at each trial period (stimulus presentation or feedback).

For each subject, fixed effects analysis tested the effects of interest (e.g., parametric modulation with recognition strength and error correction measures derived from SUSTAIN or contrasts between exception and rule-following items). The resulting contrast images generated in the individual subject analysis were analyzed across subjects using a mixed effects GLM, treating subjects as a random effect to allow for population inference. Activation in the MTL was identified in all group analyses by masking the whole-brain results at a threshold of 20 or more contiguous voxels exceeding a false-discovery-rate corrected threshold of $P < 0.05$ with an anatomical MTL gray matter mask derived from the statistical parametric mapping template. A whole-brain threshold of 20 or more contiguous voxels exceeding a false-discovery-rate corrected threshold of $P < 0.01$ was used for all other regions.

Results

Behavioral Results

Subjects remembered exception items (mean = 0.75, standard deviation [SD] = 0.20) more accurately than rule-following items (mean = 0.45, SD = 0.13) during the postscan recognition memory test, $t_{14} = 5.09$, $P < 0.001$ (Fig. 3). In contrast, during the learning phase, subjects were less accurate at categorizing exception items (mean = 0.66, SD = 0.19) in comparison to rule-following items (mean = 0.87, SD = 0.07), $t_{14} = 4.80$, $P < 0.001$. The enhanced memory for exception items is consistent with previous work on exception learning (Palmeri and Nosofsky 1995; Sakamoto and Love 2004, 2006) and predictions from SUSTAIN that exception representations are more differentiated in memory than representations of rule-following items. In addition, the behavioral data served as the basis for estimating SUSTAIN's and ALCOVE's parameters to create model-based trial-by-trial predictions for brain activation (see Supplemental Methods for modeling procedures).

MTL Activation Tracks Model-Based Estimates of Recognition Strength and Error

First, we identified brain regions in which activation during the stimulus presentation period correlated with the measure of recognition strength derived from SUSTAIN (Fig. 2*A*). Psychologically, the recognition strength measure relates to the amount of information that subjects retrieve or remember on a given trial and should therefore correlate with MTL regions such as the hippocampus that are thought to support retrieval of stored memories. SUSTAIN also predicts how recognition should change over the course of the experiment. The recognition strength measure predicts that subjects should not recognize the exceptions early in learning because they

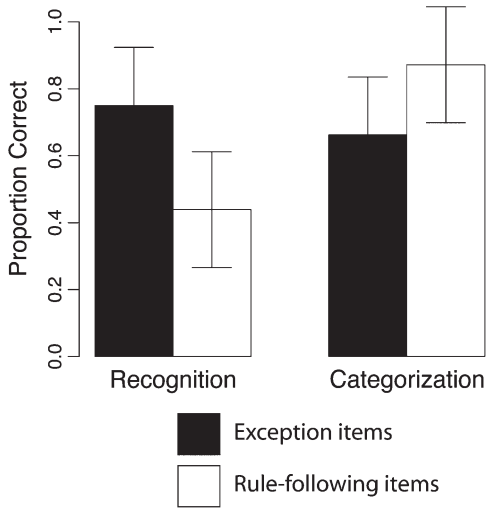

Figure 3. Behavioral results for the postscanning recognition phase and the scanned category learning phase. Error bars represent 95% confidence intervals.

have not yet formed representations for them. Likewise, subjects should recognize exceptions more so than rule-following items later in learning because, unlike exception items, rule-following items tend to share common clusters. These common clusters match on the rule dimension but tend to mismatch on the other stimulus dimensions, which obscures information that individuates items. In contrast, exception items must be fully differentiated from rule-following items in memory for subjects to categorize them accurately. As a result, exception items will be recognized to a greater extent later in the experiment when exception representations have been established. Consistent with our predictions, activation in several MTL regions was correlated with the recognition strength measure derived from SUSTAIN, including bilateral hippocampus, parahippocampal cortex, and perirhinal cortex (Fig. 2*C*; Supplementary Tables S1 and S5, which control for reaction time).

Next, we identified regions in which activation correlated with SUSTAIN's error correction measure (Fig. 2*B*). The error correction measure gives the difference between SUSTAIN's expectations for category membership and the correct category membership on a given trial. This error signal is used by the model to determine the extent to which it updates category representations based on feedback. Psychologically, the error correction measure is similar to mismatch or associative novelty signals that are thought to engage MTL-based encoding processes leading to the formation of new memory representations (Kohler et al. 2005; Kumaran and Maguire 2006). The similarity between these 2 constructs derived from the category learning and declarative memory literatures suggests a unifying principle that describes MTL activation during feedback based learning tasks such as the one in the current study. To this end, SUSTAIN makes intuitive predictions for how psychological processes related to error-driven learning change over time. Early in learning, subjects should make large numbers of errors for both item types because the task is novel and category memberships are unknown. As learning progresses, the rule-following items should produce less error than exception items because, unlike exceptions, they tend to only match clusters for their own category. Consistent with these predictions, activation in bilateral hippocampus and perirhinal cortex was correlated with the error correction measure (Fig. 2*D*; Supplementary Table S2).

Early in learning, exceptions are unexpected in the context of their category, and recruitment of hippocampus during feedback may reflect an associative novelty response that a particular exception item is a novel example of a specific categorical context (Ranganath and Rainer 2003; Kohler et al. 2005; Kumaran and Maguire 2007a, 2007b). This mismatch signal may lead to the formation of a new representation of the specific exception item in perirhinal cortex. While associative mismatch signals are often considered in the context of declarative memory, they nonetheless bear striking resemblance to SUSTAIN's error-driven learning mechanisms and may highlight important commonalities in MTL function that are shared across category learning and declarative memory tasks.

Model-based analyses using SUSTAIN revealed distinct psychological processes related to recognition and feedback that occurred at distinct points within a trial and changed in different ways across trials. These results suggest that MTL activation may shift during learning such that regions active in response to prediction error early in learning are recruited in later trials to recognize exception items after the necessary representations have been formed. While both of SUSTAIN's measures predict that exception items will elicit greater activation than rule-following items across trials, the 2 measures strongly differ in how they vary over the course of learning. Overall recognition strength increases in later trials, whereas error correction decreases. Important differences are also manifest between rule and exception items across trials. The recognition strength measure predicts that the difference between exceptions and rule-following items during stimulus presentation will increase throughout the experiment as subjects learn, whereas the error correction measure predicts that the difference between exceptions and rule-following items during the feedback decreases as subjects learn. Interestingly, when we swapped the role of the 2 measures in our parametric analyses by fitting the recognition measure to the feedback trial component and the error correction measure to the stimulus presentation trial phase, we failed to reveal significant effects in the MTL except at an extremely liberal threshold of $P < 0.05$, uncorrected. This finding provides additional support for the explanatory power of SUSTAIN in describing the contributions of MTL structures to forming and retrieving category representations.

While our model-based analyses using SUSTAIN focused on the MTL, activation in a number of additional brain regions was correlated with SUSTAIN's recognition strength and error correction measures (Fig. 4*A*,*B*; Supplementary Tables S1 and S2), including prefrontal cortex, posterior parietal cortex, and the lateral occipital complex. Each of these regions has been implicated in recognition memory and controlled retrieval processes (Malach et al. 1995; Grill-Spector et al. 2001; Badre and Wagner 2002; Dobbins et al. 2002; Moscovitch and Winocur 2002; Sommer et al. 2005; Wagner et al. 2005; Fleck et al. 2006; Cabeza et al. 2008; Hutchinson et al. 2009). In addition, correlations with SUSTAIN's recognition strength and error correction measures were found in the midbrain, insula, and regions of the ventral striatum, which have been associated with reward processing and reinforcement learning (Schultz et al. 1997; Knutson et al. 2001; Bayer and Glimcher 2005; Aston-Jones and Cohen 2005; Bechara and Damasio 2005; O'Doherty et al. 2006) as well as the anterior cingulate which has been differentially described as being involved in reward learning and conflict resolution (Botvinick et al. 2001; Holroyd and Coles 2002; Kerns et al. 2004; for relations between these perspectives, see Botvinick 2007).

Comparison between SUSTAIN and ALCOVE's Ability to Model MTL Contributions to Exception Learning

Recognition strength and error correction measures for ALCOVE, an exemplar model, were generated in the same manner as those used in the SUSTAIN model-based analyses. The key difference between SUSTAIN and ALCOVE is that the representational forms used by ALCOVE are fixed such that, regardless of the category structure, ALCOVE assumes that all items are stored individually in memory. ALCOVE does not predict a recognition advantage for exceptions because exceptions and rule-following items are equally differentiated in ALCOVE's representational space. Accordingly, the recognition strength measure of ALCOVE (Fig. 5*A*) failed to reveal significant MTL activation during the categorization period, even at a liberal threshold of $P < 0.05$, uncorrected. However,

like SUSTAIN, ALCOVE was able to predict that the exception items result in more prediction error than rule-following items (Fig. 5*B*). ALCOVE's error correction measure identified bilateral clusters of activity in the MTL (Fig. 5*D*; Supplementary Table S3) similar to those found in the SUSTAIN error correction analysis.

Comparisons between Model-Based and Standard Condition-Based Regressors

An important point highlighted by the model-based results is that activation associated with experimentally defined conditions is not constant throughout an experiment as is commonly assumed by simple linear contrasts comparing 2 (or more) conditions. Thus, standard contrast-based analyses are less able to predict changes in activation over time when compared with measures derived from models like ALCOVE and SUSTAIN that incorporate empirically derived learning functions that change across learning trials. For example, accuracy increases throughout the experiment for exception items, whereas accuracy peaks early and asymptotes for rule-following items. Directly comparing exceptions and rule-following items thus confounds differences between items with differences in the engagement of psychological processes across learning trials. Accordingly, a linear contrast of exception and rule-following items during the stimulus presentation and feedback periods failed to detect MTL activation during learning except at a liberal threshold of $P < 0.05$, uncorrected. The failure to detect MTL activation using standard linear contrasts suggests an important advantage for our model-based approach in isolating the contributions of MTL structures to category learning.

A number of post hoc comparisons are possible that can improve our ability to identify regions associated with

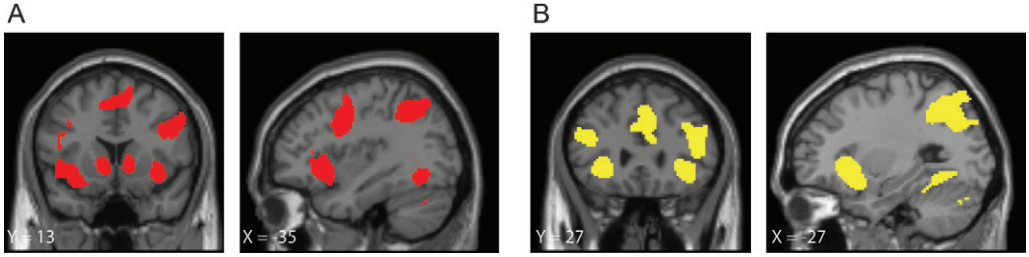

Figure 4. Whole-brain statistical maps for regions exhibiting a significant ($P < 0.01$, FDR corrected) correlation between (*A*) the recognition strength measure of SUSTAIN during the categorization period and (*B*) the error correction measure of SUSTAIN during feedback. Red indicates activation present during stimulus presentation and yellow indicates activation present during feedback.

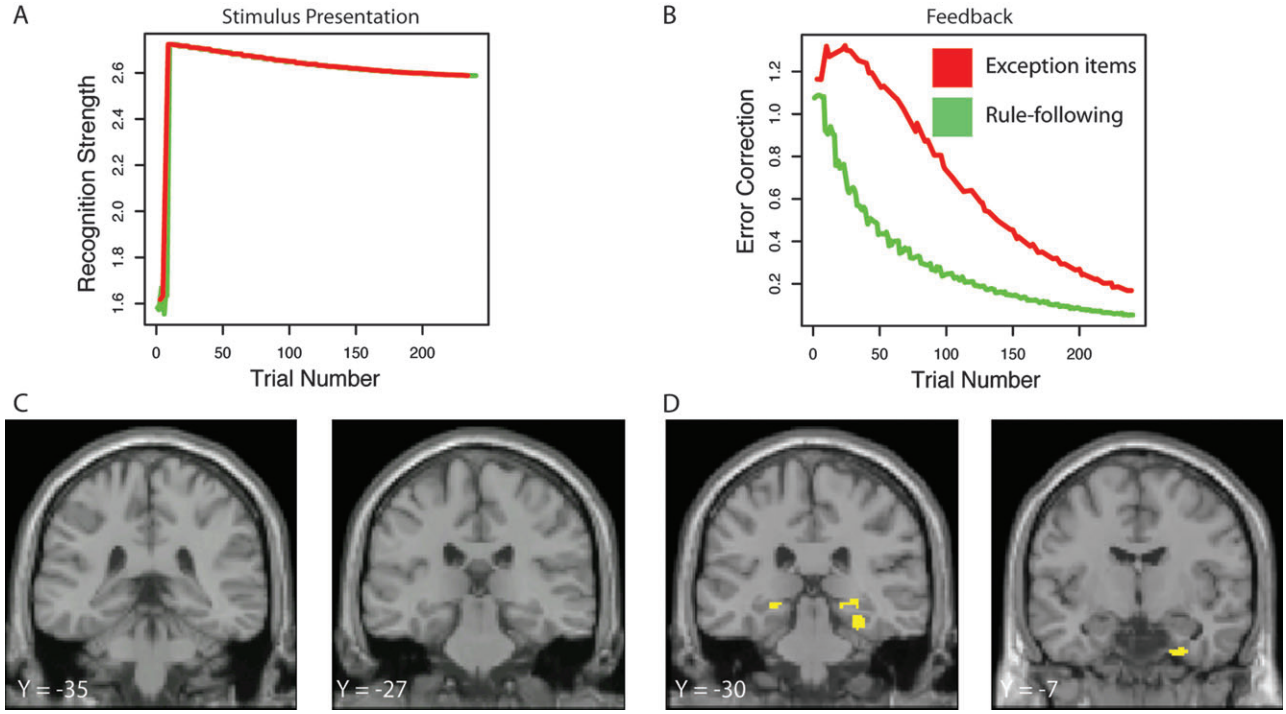

Figure 5. Illustrations of the recognition strength (*A*) and error correction (*B*) measures derived from ALCOVE that were used to predict activation during stimulus presentation and feedback. Recognition strength is the sum of the similarities between a given stimulus/trial and all items stored in memory, and error correction is the absolute value of the model's output error on a given trial. Below the measures are the corresponding statistical maps associated with each regressor. Activation during stimulus presentation is presented in red and activation during the feedback period in yellow. (*C*) No MTL regions exhibited a significant correlation between activation during stimulus presentation and the predicted recognition strength measure. (*D*) MTL regions exhibiting a significant ($P < 0.05$, FDR corrected) correlation between during feedback and the predicted error correction measure.

exception processing when using linear contrasts. For example, it is both intuitive and predicted by SUSTAIN that recognition differences between exception and rule-following items should be highest late in the experiment when subjects have mastered the task. Accordingly, contrasting correct exception trials with correct rule-following trials reveals similar statistical maps to those found for the recognition strength measure of SUSTAIN (Fig. 6*A*,*C*; Supplementary Table S4). However, no intuitive post hoc contrasts (correct exception > correct rule following; incorrect exception > incorrect rule following) recovered MTL activation associated with error correction at feedback except at a liberal threshold of $P < 0.05$, uncorrected. One general weakness of such post hoc linear contrasts is that, rather than utilizing all of the data, biased samples of data are used. For example, because subjects learn rule-following items faster than exceptions, the correct > incorrect analysis includes rule-following item trials from throughout the experiment, but the majority of exception trials included are only from later stages of the experiment. In contrast, the model-based analysis uses all of the data when predicting patterns of activation.

Another weakness of standard linear contrasts is that they can lead to potentially questionable and incorrect conclusions regarding the functional specificity of a particular brain region. For example, SUSTAIN predicts that MTL structures will contribute to successful categorization of both exception and rule-following items but predicts a quantitative difference in the degree of MTL involvement with exception items recruiting MTL processing to a greater degree. In contrast, the post hoc comparison between correct exceptions and correct rule-following items could lead to the incorrect conclusion that there is a qualitative difference between exceptions and rule-following item such that the MTL is not engaged during categorization of rule-following items. As evidenced by the success of our measures from SUSTAIN in fitting not only the exception items but also the rule-following items, rule-following items simply have a different time course and, on average, recruit the MTL less strongly than exception items, as predicted by our theory.

Discussion

The role of the MTL in category learning has long been debated, and central questions remain regarding the functional contributions of MTL structures to the acquisition, representation, and use of novel category information. A number of recent observations have implicated MTL structures in a variety of category learning paradigms from prototype learning (Reber et al. 2003; Zeithamova et al. 2008) to rule storage (Nomura et al. 2007) and probabilistic categorization (Poldrack et al. 2001; Hopkins et al. 2004). Rather than characterizing these discrepant findings as contradictory, our view is that the MTL builds category representations that are tailored to the requirements of the learning context. In other words, as a function of the category learning task, MTL representations can mimic that of an exemplar-, prototype-, or rule-based model. On this view, one key challenge is to specify a model that is able to flexibly adapt category representations to the nature of the learning context and thus capture the essential function of the MTL in category learning.

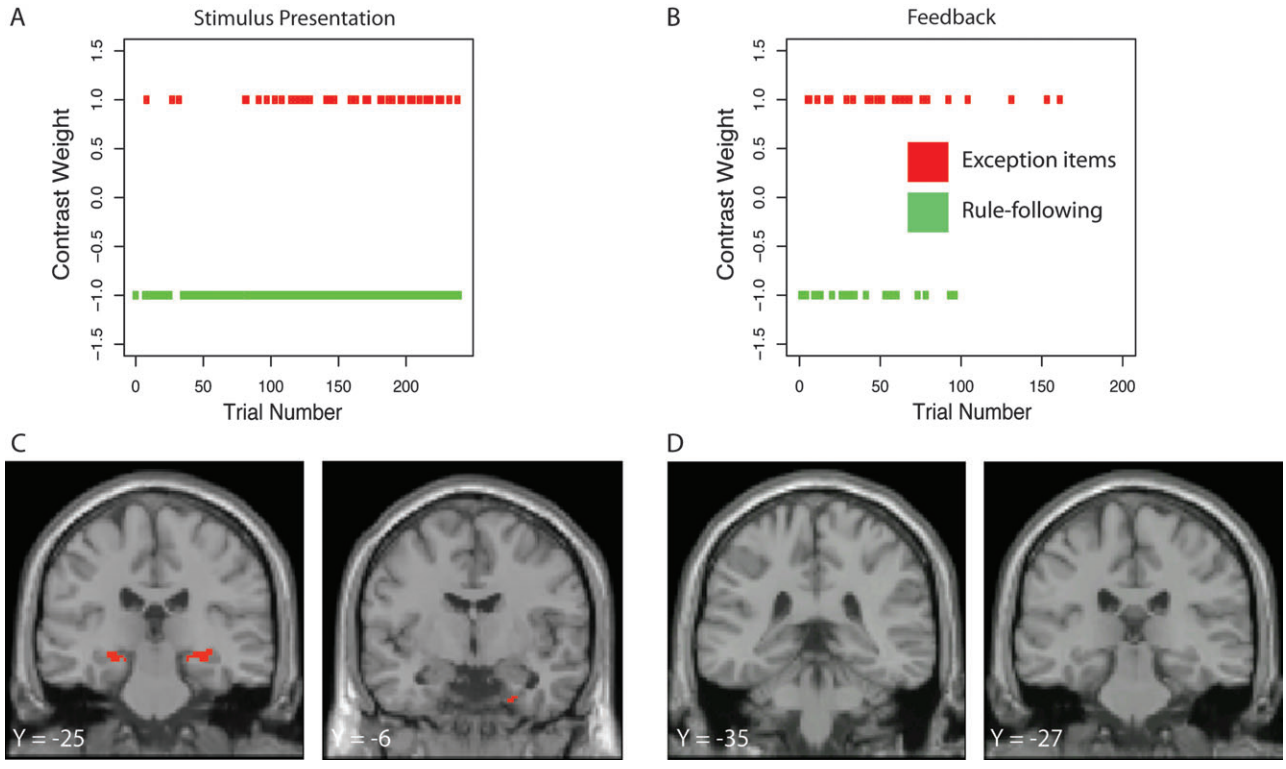

Figure 6. Illustrations of standard linear contrasts as applied in the present paradigm. (*A*) Contrast weights for correct exceptions > correct rule-following items and (*B*) contrast weights for incorrect exceptions > incorrect rule-following items. Below the contrast weights are the corresponding statistical maps associated with each comparison. Activation during stimulus presentation is presented in red and activation during the feedback period in yellow. (*C*) MTL regions exhibiting significant ($P < 0.05$, FDR corrected) activation during stimulus presentation for correct exceptions compared with correct rule-following items. (*D*) No MTL activation was found for comparisons during feedback.

We suggest that SUSTAIN, a mathematical model of category learning, addresses this challenge by building category representations that are tailored to the learning context. Consistent with our view, model-based fMRI analyses using SUSTAIN highlight the role of the MTL in the formation and retrieval of specialized representations for exception items that violate a category rule. By combining predictions from SUSTAIN with fMRI data from a rule-plus-exception learning task, we isolated trial-by-trial fluctuations in MTL activation that were associated with encoding (via error-driven learning) and retrieval (via recognition) of novel category information. In this task, exception and rule-following items fundamentally differ in their representational requirements. Exceptions are both more difficult to learn and easier to recognize (once learned) than rule-following items. SUSTAIN captures the representational differences between exception and rule-following items and accurately predicts greater MTL engagement for exception items during category decisions and feedback. These findings suggest that, like SUSTAIN, the MTL contributes to category learning by forming specialized category representations appropriate for the learning context.

One interesting result of our model-based approach is that in many of the analyses, activation was observed not only in the hippocampus, the region thought to be primarily involved in the encoding and retrieval of cluster representations (Brown and Aggleton 2001; Norman and O'Reilly 2003; Eichenbaum et al. 2007; Love and Gureckis 2007) but also in regions of the MTL cortex, including perirhinal and parahippocampal cortices. While historically the MTL as whole was considered a system for declarative memory (Squire 1992), current research suggests that the hippocampus and MTL cortical regions may differ in their functional roles. For example, the perirhinal cortex is often implicated in stimulus-based familiarity processes that rely on the global similarity between a stimulus and stored representations, while hippocampus has been associated with encoding and retrieval processes that underlie associative memory and recollection of event details (for review, see Brown and Aggleton 2001; Davachi 2006; Eichenbaum et al. 2007; Diana et al. 2007).

One possibility for the role of the MTL cortex in the present task is that it provides representations of stimulus features to the hippocampus, which then retrieves or encodes additional associated information such as an item's category label. Indeed, the recognition strength and error correction measures are aggregate measures and contain information about global matching processes as well as local associative processes that are likely implemented in different MTL regions. Future model-based fMRI analyses could develop measures to separately interrogate these theoretical processes and dissociate the function of hippocampus from MTL cortex in category learning.

While model-based analysis is a powerful tool for localizing mental function, care must be taken in interpreting results. Like other fMRI analysis techniques, model-based analysis is correlational. Therefore, it is possible that areas can correlate with a model measure yet not instantiate the corresponding mental processes in the brain. One strength of model-based analysis is that it can help identify such situations. Whenever 2 measures (from the same or different models) correlate with one another, one can expect that overlapping brain areas will be recovered by the 2 analyses. Furthermore, models can be compared with help design future studies that tease apart (i.e., decorrelate) measures of interest.

In the present study, it is likely that SUSTAIN's regressors correlated with brain areas not directly related to the targeted processes. For example, areas related to reward processing (e.g., ventral striatum, midbrain, insula) were recovered by SUSTAIN's recognition strength measure. One possible interpretation is that SUSTAIN's recognition measure in a rule-plus-exception task tracks processes related to category uncertainty (for supportive results, see Grinband et al. 2006). As learning progresses, exceptions items should be high on measures tapping recognition, uncertainty, and response conflict at stimulus presentation (Davis et al. 2009). Given the multitude of processes involved in any category learning task, any one measure in any one study is likely to correlate with activity in regions that are not of direct interest. This observation underscores the importance of integrating findings across multiple studies and methods to formulate predictions. Along these lines, applying the same set of models across studies offers an effective means for triangulating mental function.

In other cases, SUSTAIN's regressors may correlate with activation in particular regions because they support psychological processes that do have a direct relationship with the measure of interest. To this end, a number of other brain regions, including the prefrontal cortex, posterior parietal cortex, and lateral occipital complex, were found to correlate with measures of recognition strength and error correction. Prefrontal and posterior parietal cortices are known to support attentional processes (Posner and Petersen 1990; Miller and Cohen 2001) that extend to the domain of memory (Wagner et al. 2005; Fleck et al. 2006; Cabeza et al. 2008). Likewise, lateral occipital complex has been associated with object representation (Malach et al. 1995; Grill-Spector et al. 2001) and may provide object-level information to MTL structures in memory tasks (Sommer et al. 2005). Accordingly, prefrontal and parietal regions may correlate with the recognition strength and error correction measures in the present task because they support attentional processes needed for encoding and retrieval of clusters. Lateral occipital complex may correlate with these measures because it encodes information about specific stimuli (Sigala and Logothetis 2002; Palmeri and Gauthier 2004) that becomes bound in the higher-level mnemonic representations in the MTL (i.e., clusters). Again, deciding whether a region that correlates with a model measure is directly involved with the computational processes of interest requires careful integration across research studies. While the same issues arise when interpreting standard imaging analyses, model-based analyses can provide a more powerful and theoretically motivated means for integrating past and present findings, and charting a way forward.

The preceding discussion makes clear that multiple brain areas are likely engaged to support category learning. While the dominant view in the neuroscience community is that different brain systems support performance in different types of categorization problems (for reviews, see Ashby and Maddox 2005; Ashby and O'Brien 2005; Poldrack and Foerde 2008; Smith and Grossman 2008), many behavioral findings thought to indicate the need for multiple systems of representation have subsequently been shown to be consistent with a single system interpretation (Nosofsky and Zaki 1998; Nosofsky and Johansen 2000; Johansen and Palmeri 2002). Our approach is to specify model-based mechanisms and relate these mechanisms to brain function as opposed to arguing for or against

a particular number of learning systems, as we believe that, in practice, the criteria for delineating separate systems is often underspecified and can lead to needless controversy. Indeed, SUSTAIN is a single system model, which forms representations (i.e., clusters) that can behave like exemplar-, prototype-, or rule-based representations depending on the nature of the category learning task. We relate the properties of SUSTAIN's learning mechanisms primarily to MTL function and theorize that this mapping will hold in many category learning tasks. It is possible, however, that other learning systems are better characterized by alternate mechanisms and forms of representation (cf. Love and Gureckis 2007). SUSTAIN may not be the preferred model for experimental manipulations and category structures that preferentially tap learning systems outside the MTL.

Tasks that could potentially be better characterized by mechanisms other than those proposed by SUSTAIN include "information integration" or procedural learning tasks (Nomura et al. 2007). Models like the covering map version of ALCOVE (Kruschke 1992) and the Striatal Pattern Classifier (SPC; Ashby and Waldron 1999) may provide a better characterization of procedural learning mechanisms and tasks. Unlike SUSTAIN, these models do not build specialized representations for a category learning problem. Instead, these models learn to associate various visual inputs with a behavioral (category) response in the same manner for all learning tasks, much like how exemplar and prototype models always represent categories in the same format. This mode of incremental associative learning is sufficiently powerful that these models can eventually learn any possible category structure (Ashby and Waldron 1999), consistent with theories of procedural category learning that focus on the striatum (Ashby et al. 1998).

However, this form of learning differs from the way in which SUSTAIN builds category representations. For example, whereas SUSTAIN builds specialized representations for the exception items in our study, covering map models of procedural learning, such as SPC, do not; therefore, covering map models do not predict the recognition advantage observed for exception items. In models of procedural learning, regions activated by stimuli, regardless of whether these items are exceptions or rule following, are associated with the reinforced response. More generally, MTL representations are hypothesized to be more readily adapted to novel uses beyond the original learning circumstances than the representations formed by other learning systems (Eichenbaum and Cohen 2001; Preston et al. 2004). Consistent with this notion, SUSTAIN is able to use previously acquired clusters in novel contexts (Yamauchi et al. 2002; Love et al. 2004), whereas procedural learning models would have to begin anew in associating stimuli with novel responses. Thus, while more than one learning system may be able to master a particular category learning task, the nature of the representations across systems and the adaptive functions they can serve may dramatically differ.

We favor a model-based approach that makes strong theoretical connections to the broader literature. The theory we forward relating SUSTAIN to the MTL, much like the theory relating the SPC to the striatum, goes beyond the model's equations by tying model operations to brain regions (see Love and Gureckis 2007). For instance, although not reflected in any equation, one would expect SUSTAIN's parameters to change given task manipulations that are known to reduce MTL involvement. Conversely, manipulations that affect striatal mediated learning should alter the SPC model's operation. The encompassing theory linking brain and computation can guide the application and interpretation of model fits.

A final issue in category learning that model-based approaches are well suited to address is whether the brain systems that support category learning change as subjects learn a task. Representational shifts in category learning are hotly debated in the behavioral and modeling literature where ultimately the behavioral data alone underdetermine whether such shifts occur (Johansen and Palmeri 2002). Similarly, in the literature on MTL involvement in category learning, some studies suggest that the MTL is involved only early in learning (Poldrack et al. 2001; Poldrack and Rodriquez 2004; Little et al. 2006) or late in learning (Knowlton et al. 1994; Knowlton et al. 1996), whereas other results suggest that the MTL is involved throughout a categorization task (Zeithamova et al. 2008; Degutis and D'Esposito 2009). Our results highlight that regions recruited for category learning, such as the MTL, may be differentially recruited during different stages of learning, during different trial components, and for different items within categories (e.g., exceptions). Early MTL involvement may reflect formation of specialized representations (consistent with the error correction measure of SUSTAIN) and later MTL involvement may reflect retrieval of those representations (consistent with the recognition strength measure of SUSTAIN). Commonly employed fMRI methods, such as block designs or modeling stimulus presentation and feedback portions of a trial with a single regressor, average over trial-level information as well as across trial components and may thus obscure important information about different mechanisms that support performance at the level of individual trials. Model-based fMRI approaches thus have additional power to resolve existing debates about the engagement of different learning systems during distinct stages of learning.

In conclusion, we present a model-based approach that proposes a role for the MTL in learning categories that require subjects to build cluster-based category representations that fit the needs of the learning context. Using quantitative predictions from SUSTAIN, a mechanistic category learning model, we observed activation in hippocampus and MTL cortex consistent with the notion that the MTL builds specialized representations appropriate for the categories to be learned. In the present study, such specialized representations were required to master exceptions to a category rule. Our results extend current neurobiological approaches of category learning by providing a well-specified theory for the role of the MTL in the formation and use of novel category information, and in doing so, have the potential to unite the results from a number of disparate category learning studies in which MTL activation has been observed. Importantly, our model-based analysis suggests that MTL involvement in category learning can vary across items within a task and thus likely across different category learning tasks themselves. More broadly, model-based methods will prove critical for integrating results across different category learning paradigms and for reaching a general consensus regarding the mechanisms and brain systems that underlie category learning.

Supplementary Material

Supplementary material can be found at: http://www.cercor.oxfordjournals.org/

Funding

Air Force Office of Scientific Research (FA9550-07-1-0178 to B.C.L.); Army Research Laboratory (W911NF-09-2-0038 to B.C.L); and National Alliance for Research on Schizophrenia and Depression (to A.R.P).

Notes

Thanks to David Schnyer, Ross Otto, and Molly Ireland for comments on a previous draft of this manuscript and Frances Fawcett for use of beetle stimuli. *Conflict of Interest*: None declared.

References

Aizenstein HJ, MacDonald AW, Stenger VA, Nebes RD, Larson JK, Ursu S, Carter CS. 2000. Complementary category learning systems identified using event-related functional MRI. J Cogn Neurosci. 12:977–987.

Anderson JR. 1991. The adaptive nature of human categorization. Psychol Rev. 98:409–429.

Ashby FG, Alfonso-Reese LA, Turken AU, Waldron EM. 1998. A neuropsychological theory of multiple systems in category learning. Psychol Rev. 105:442–481.

Ashby FG, Gott RE. 1988. Decision rules in the perception and categorization of multidimensional stimuli. J Exp Psychol Learn Mem Cogn. 14:33–53.

Ashby FG, Lee WW. 1991. Predicting similarity and categorization from identification. J Exp Psychol Gen. 120:150–172.

Ashby FG, Maddox WT. 2005. Human category learning. Annu Rev Psychol. 56:149–178.

Ashby FG, O'Brien JB. 2005. Category learning and multiple memory systems. Trends Cogn Sci. 9:83–89.

Ashby FG, Waldron EM. 1999. On the nature of implicit categorization. Psychon Bull Rev. 6:363–378.

Aston-Jones G, Cohen JD. 2005. An integrative theory of locus coeruleus-norepinephrine function: adaptive gain and optimal performance. Annu Rev Neurosci. 28:403–450.

Badre D, Wagner AD. 2002. Semantic retrieval, mnemonic control, and prefrontal cortex. Behav Cogn Neurosci Rev. 1:206–218.

Bayer HM, Glimcher PW. 2005. Midbrain dopamine neurons encode a quantitative reward prediction error signal. Neuron. 47:129–141.

Bechara A, Damasio AR. 2005. The somatic marker hypothesis: a neural theory of economic decision. Games Econ Behav. 52:336–372.

Botvinick MM. 2007. Conflict monitoring and decision making: reconciling two perspectives on anterior cingulate function. Cogn Affect Behav Neurosci. 7:356–366.

Botvinick MM, Braver TS, Barch DM, Carter CS, Cohen JD. 2001. Conflict monitoring and cognitive control. Psychol Rev. 108: 624–652.

Brown MW, Aggleton JP. 2001. Recognition memory: what are the roles of the perirhinal cortex and hippocampus? Nat Rev Neurosci. 2:51–61.

Bruner JS, Goodnow JJ, Austin GA. 1956. A study of thinking. New York: Wiley.

Cabeza R, Ciaramelli E, Olson I, Moscovitch M. 2008. The parietal cortex and episodic memory: an attentional account. Nat Rev Neurosci. 9:613–625.

Davachi L. 2006. Item, context, and relational episodic encoding in humans. Curr Opin Neurobiol. 16:693–700.

Davachi L, Mitchell JP, Wagner AD. 2003. Multiple routes to memory: distinct medial temporal lobe processes build item and source memories. Proc Natl Acad Sci U S A. 100:2157–2162.

Davis T, Love BC, Maddox TM. 2009. Anticipatory emotions in decision tasks: covert markers of value or attentional processes? Cognition. 112:195–200.

Daw ND. 2011. Trial-by-trial data analysis using computational models. In: Delgado MR, Phelps EA, Robbins TW, editors. Decision making, affect, and learning: Attention and Performance XXIII. Oxford: Oxford University Press.

DeGutis J, D'Esposito M. 2009. Network changes in the transition from initial to well-practiced visual categorization. Front Neurosci. 3:1–14.

Diana RA, Yonelinas AP, Ranganath C. 2007. Imaging recollection and familiarity in the medial temporal lobe: a three-component model. Trends Cogn Sci. 11:379–386.

Dobbins IG, Foley H, Schacter DL, Wagner AD. 2002. Executive control during episodic retrieval: multiple prefrontal processes subserve source memory. Neuron. 35:989–996.

Eichenbaum H, Cohen NJ. 2001. From conditioning to conscious recollection: memory systems of the brain. Oxford: Oxford University Press.

Eichenbaum H, Yonelinas AP, Ranganath C. 2007. The medial temporal lobe and recognition memory. Annu Rev Neurosci. 30:123–152.

Erickson MA, Kruschke JK. 1998. Rules and exemplars in category learning. J Exp Psychol Gen. 127:107–140.

Fleck MS, Daselaar SM, Dobbins IG, Cabeza R. 2006. Role of prefrontal and anterior cingulate regions in decision-making processes shared by memory and nonmemory tasks. Cereb Cortex. 16:1623–1630.

Foerde K, Knowlton BJ, Poldrack RA. 2006. Modulation of competing memory systems by distraction. Proc Natl Acad Sci U S A. 103: 11778–11783.

Foerde K, Poldrack RA, Knowlton BJ. 2007. Secondary-task effects on classification learning. Mem Cogn. 35:864–874.

Gläscher JP, O'Doherty JP. 2010. Model-based approaches to neuroimaging: combining reinforcement learning theory with fMRI data. WIREs Cogni Sci. 1:501–510.

Grill-Spector K, Kourtzi Z, Kanwisher N. 2001. The lateral occipital complex and its role in object recognition. Vision Res. 41:1409–1422.

Grinband J, Hirsch J, Ferrera VP. 2006. A neural representation of categorization uncertainty in the human brain. Neuron. 49: 757–763.

Holroyd CB, Coles MGH. 2002. The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity. Psychol Rev. 109:67–709.

Hopkins RO, Myers CE, Shohamy D, Grossman S, Gluck M. 2004. Impaired probabilistic category learning in hypoxic subjects with hippocampal damage. Neuropsychologia. 42:524–535.

Hutchinson JB, Uncapher MR, Wagner AD. 2009. Posterior parietal cortex and episodic retrieval: convergent and divergent effects of attention and memory. Learn Mem. 16:343–356.

Johansen MK, Palmeri TJ. 2002. Are there representational shifts during category learning? Cogn Psychol. 45:482–553.

Kerns JG, Cohen JD, MacDonald AW, Cho RY, Stenger VA, Carter CS. 2004. Anterior cingulate conflict monitoring and adjustments in control. Science. 303:1023–1026.

Knowlton BJ, Mangels JA, Squire LR. 1996. A neostriatal habit learning system in humans. Science. 273:1399–1402.

Knowlton BJ, Squire LR, Gluck MA. 1994. Probabilistic classification in amnesia. Learn Mem. 1:106–120.

Knutson B, Adams CM, Fong GW, Hommer D. 2001. Anticipation of increasing monetary reward selectively recruits nucleus accumbens. J Neurosci. 21:RC159.

Kohler S, Danckert S, Gati JS, Menon RS. 2005. Novelty responses to relational and nonrelational information in the hippocampus and the parahippocampal region: a comparison based on event-related fMRI. Hippocampus. 15:763–774.

Kumaran D, Maguire EA. 2006. An unexpected sequence of events: mismatch detection in the human hippocampus. PLoS Biol. 4:e424.

Kumaran D, Maguire EA. 2007a. Match-mismatch processes underlie human hippocampal responses to associative novelty. J Neurosci. 27:8517–8524.

Kumaran D, Maguire EA. 2007b. Which computational mechanisms operate in the hippocampus during novelty detection? Hippocampus. 17:735–748.

Kruschke JK. 1992. ALCOVE: an exemplar-based connectionist model of category learning. Psychol Rev. 99:22–44.

Little DM, Shin SS, Sisco SM, Thulborn KR. 2006. Event-related fMRI of category learning: differences in classification and feedback networks. Brain Cogn. 60:244–252.

Love BC, Gureckis TM. 2007. Models in search of a brain. Cogn Affect Behav Neurosci. 7:90–108.

Love BC, Medin DL, Gureckis TM. 2004. SUSTAIN: a network model of category learning. Psychol Rev. 111:309–332.

Maddox WT, Ashby FG. 2004. Dissociating explicit and procedural-learning based systems of perceptual category learning. Behav Process. 66:309–332.
Malach R, Reppas JB, Benson RR, Kwong KK, Jiang H, Kennedy WA, Ledden T, Brady TJ, Rosen BR, Tootell RB. 1995. Object-related activity revealed by functional magnetic resonance imaging in human occipital cortex. Proc Natl Acad Sci U S A. 92:8135–8139.
Medin DL, Schaffer MM. 1978. Context theory of classification learning. Psychol Rev. 85:207–238.
Miller EK, Cohen JD. 2001. An integrative theory of prefrontal cortex function. Annu Rev Neurosci. 24:167–202.
Monchi O, Petrides M, Petre V, Worsley K, Dagher A. 2001. Wisconsin Card Sorting revisited: distinct neural circuits participating in different stages of the task identified by event-related functional magnetic resonance imaging. J Neurosci. 21:7733–7741.
Morris RGM, Garrud P, Rawlins JNP, O'Keefe J. 1982. Place navigation impaired in rats with hippocampal lesions. Nature. 297: 681–683.
Moscovitch M, Winocur G. 2002. The frontal cortex and executive control processes. In: Stuss DT, Knight RT, editors. Principles of frontal lobe function. Oxford: Oxford University Press.
Nomura EM, Maddox WT, Filoteo JV, Ing AD, Gitelman DR, Parrish TB, Mesulam MM, Reber PJ. 2007. Neural correlates of rule-based and information-Integration visual category learning. Cereb Cortex. 17:37–43.
Norman KA, O'Reilly RC. 2003. Modeling hippocampal and neocortical contributions to recognition memory: a complementary-learning-systems approach. Psychol Rev. 110:611–646.
Nosofsky RM. 1986. Attention similarity and the identification—categorization relationship. J Exp Psychol Gen. 115:39–57.
Nosofsky RM, Johansen MK. 2000. Exemplar-based accounts of multiple-system phenomena in perceptual categorization. Psychon Bull Rev. 7:375–402.
Nosofsky RM, Palmeri TJ, Mckinley SC. 1994. Rule-plus-exception model of classification learning. Psychol Rev. 104:266–300.
Nosofsky RM, Zaki SR. 1998. Dissociations between categorization and recognition in amnesic and normal individuals: an exemplar-based interpretation. Psychol Sci. 9:247–255.
O'Doherty JP, Buchanan TW, Seymour B, Dolan RJ. 2006. Predictive neural coding of reward preference involves dissociable responses in human ventral midbrain and ventral striatum. Neuron. 49: 157–166.
O'Doherty JP, Hampton A, Kim H. 2007. Model-based fMRI and its application to reward learning and decision making. Ann N Y Acad Sci. 1104:35–53.
Palmeri TJ, Gauthier I. 2004. Visual object understanding. Nat Rev Neurosci. 5:291–303.
Palmeri TJ, Nosofsky RM. 1995. Recognition memory for exceptions to the category rule. J Exp Psychol Learn Mem Cogn. 21: 548–569.
Palmeri TJ, Tarr MJ. 2008. Visual object perception and long-term memory. In: Luck S, Hollingsworth A, editors. Visual memory. New york: Oxford University Press.
Patalano AL, Smith EE, Jonides J, Koeppe RA. 2001. PET evidence for multiple strategies of categorization. Cogn Affect Behav Neurosci. 1:360–370.
Pickering AD. 1997. New approaches to the study of amnesic patients: what can a neurofunctional philosophy and neural network methods offer? Memory. 5:255–300.
Poldrack RA, Clark J, Pare-Blagoev EJ, Shohamy D, Creso Moyano J, Myers C, Gluck MA. 2001. Interactive memory systems in the human brain. Nature. 414:546–550.
Poldrack RA, Foerde K. 2008. Category learning and the memory systems debate. Neurosci Biobehav Rev. 32:197–205.
Poldrack RA, Rodriguez P. 2004. How do memory systems interact? Evidence from human classification learning. Neurobiol Learn Mem. 8:324–332.
Posner MI, Keele SW. 1968. On the genesis of abstract ideas. J Exp Psychol. 77:353–363.
Posner MI, Petersen SE. 1990. The attention system of the human brain. Annu Rev Neurosci. 13:25–42.
Preston AR, Shrager Y, Dudukovic NM, Gabrieli JDE. 2004. Hippocampal contribution to the novel use of relational information in declarative memory. Hippocampus. 14:148–152.
Preston AR, Wagner AD. 2007. The medial temporal lobe and memory. In: Kesner RP, Martinez JL, editors. Neurobiology of learning and memory. 2nd ed. New York: Elsevier. p. 305–337.
Ranganath C, Rainer G. 2003. Neural mechanisms for detecting and remembering novel events. Nat Rev Neurosci. 4:193–202.
Reber PJ, Gitelman DR, Parish TB, Mesulam MM. 2003. Dissociating explicit and implicit category knowledge with fMRI. J Cogn Neurosci. 15:574–583.
Rosch EH. 1973. Natural categories. Cogn Psychol. 4:328–350.
Rodrigues PM, Murre JMJ. 2007. Rules-plus-exception tasks: a problem for exemplar models? Psychon Bull Rev. 14:640–646.
Rutishauser U, Mamelak AN, Schuman EM. 2006. Single-trial learning of novel stimuli by individual neurons of the human hippocampus-amygdala complex. Neuron. 49:805–813.
Sakamoto Y, Love BC. 2004. Schematic influences on category learning and recognition memory. J Exp Psychol Gen. 133:534–553.
Sakamoto Y, Love BC. 2006. Vancouver Toronto Montreal Austin: enhanced oddball memory through differentiation not isolation. Psychon Bull Rev. 13:474–479.
Schultz W, Dayan P, Montague PR. 1997. A neural substrate of prediction and reward. Science. 275:1593–1599.
Scoville WB, Milner B. 1957. Loss of recent memory after bilateral hippocampal lesions. J Neurol Neurosurg Psychiatry. 20:11–21.
Seger CA, Cincotta CM. 2006. Dynamics of frontal, striatal, and hippocampal systems during rule learning. Cereb Cortex. 16: 1546–1555.
Seger CA, Poldrack RA, Prabhakaran V, Zhao M, Glover GH, Gabrieli JD. 2000. Hemispheric asymmetries and individual differences in visual concept learning as measured by functional MRI. Neuropsychologia. 38:1316–1324.
Sigala N, Logothetis NK. 2002. Visual categorization shapes feature selectivity in the primate temporal cortex. Nature. 415:318–320.
Shohamy D, Myers CE, Grossman S, Sage J, Gluck MA, Poldrack RA. 2004. Cortico-striatal contributions to feedback-based learning: converging data from neuroimaging and neuropsychology. Brain. 127:851–859.
Smith EE, Grossman M. 2008. Multiple systems of category learning. Neurosci Biobehav Rev. 32:249–264.
Smith JD, Minda JP. 1998. Prototypes in the mist: the early epochs of category learning. J Exp Psychol Learn Mem Cogn. 24: 1411–1436.
Sommer T, Rose M, Weiller C, Büchel C. 2005. Contributions of occipital, parietal and parahippocampal cortex to encoding of object-location associations. Neuropsychologia. 43:732–743.
Squire LR. 1992. Memory and the hippocampus: a synthesis from findings with rats, monkeys, and humans. Psychol Rev. 99:195–231.
Staresina BP, Davachi L. 2008. Selective and shared contributions of the hippocampus perirhinal cortex to episodic item and associative encoding. J Cogn Neurosci. 20:1478–1489.
Staresina BP, Davachi L. 2009. Mind the gap: binding experience across space and time in the human hippocampus. Neuron. 63:267–276.
Stark CEL, Squire LR. 2001. When zero is not zero: the problem of ambiguous baseline conditions in fMRI. Proc Natl Acad Sci U S A. 98:12760–12766.
Stern CE, Corkin S, González RG, Baker JR, Carr CA, Sugiura RM, Vedantham V, Rosen BR. 1996. The hippocampal formation participates in novel picture encoding: evidence from functional magnetic resonance imaging. Proc Natl Acad Sci U S A. 93:8660–8665.
Strange B, Duggins A, Penny W, Dolan R, Friston K. 2005. Information theory, novelty and hippocampal responses: unpredicted or unpredictable? Neural Netw. 18:225–230.
Trabasso T, Bower GH. 1968. Attention in learning: theory and research. New York: Wiley.
Tulving E, Markowitsch HJ, Craik FE, Habib R, Houle S. 1996. Novelty and familiarity activations in PET studies of memory encoding and retrieval. Cereb Cortex. 6:71–79.

Wagner AD, Shannon BJ, Kahn I, Buckner RL. 2005. Parietal lobe contributions to episodic memory retrieval. Trends Cogn Sci. 9:445–453.
Wallenstein GV, Hasselmo ME, Eichenbaum H. 1998. The hippocampus as an associator of discontiguous events. Trends Neurosci. 21:317–323.
Wan H, Aggleton JP, Brown MW. 1999. Different contributions of the hippocampus and perirhinal cortex to recognition memory. J Neurosci. 19:1142–1148.
Wittgenstein L. 1953/2001. Philosophical investigations. Oxford: Blackwell Publishing.
Yamaguchi S, Hale LA, D'Esposito M, Knight RT. 2004. Rapid prefrontal-hippocampal habituation to novel events. J Neurosci. 24:5356–5363.
Yamauchi T, Love BC, Markman AB. 2002. Learning nonlinearly separable categories by inference and classification. J Exp Psychol Learn Mem Cogn. 28:585–593.
Zaki SR, Nosofsky RM, Stanton RD, Cohen AL. 2003. Prototype and exemplar accounts of category learning and attentional allocation: a reassessment. J Exp Psychol Learn Mem Cogn. 29:1160–1173.
Zeithamova D, Maddox WT, Schnyer DM. 2008. Dissociable prototype learning systems: evidence from brain imaging and behavior. J Neurosci. 28:13194–13201.