Journal of Experimental Psychology:
Learning, Memory, and Cognition
2012, Vol. 38, No. 4, 821–839

0278-7393/12/$12.00 DOI: 10.1037/a0027865

Striatal and Hippocampal Entropy and Recognition Signals in Category Learning: Simultaneous Processes Revealed by Model-Based fMRI

Tyler Davis, Bradley C. Love, and Alison R. Preston
The University of Texas at Austin

Category learning is a complex phenomenon that engages multiple cognitive processes, many of which occur simultaneously and unfold dynamically over time. For example, as people encounter objects in the world, they simultaneously engage processes to determine their fit with current knowledge structures, gather new information about the objects, and adjust their representations to support behavior in future encounters. Many techniques that are available to understand the neural basis of category learning assume that the multiple processes that subserve it can be neatly separated between different trials of an experiment. Model-based functional magnetic resonance imaging offers a promising tool to separate multiple, simultaneously occurring processes and bring the analysis of neuroimaging data more in line with category learning's dynamic and multifaceted nature. We use model-based imaging to explore the neural basis of recognition and entropy signals in the medial temporal lobe and striatum that are engaged while participants learn to categorize novel stimuli. Consistent with theories suggesting a role for the anterior hippocampus and ventral striatum in motivated learning in response to uncertainty, we find that activation in both regions correlates with a model-based measure of entropy. Simultaneously, separate subregions of the hippocampus and striatum exhibit activation correlated with a model-based recognition strength measure. Our results suggest that model-based analyses are exceptionally useful for extracting information about cognitive processes from neuroimaging data. Models provide a basis for identifying the multiple neural processes that contribute to behavior, and neuroimaging data can provide a powerful test bed for constraining and testing model predictions.

Keywords: category learning, model-based imaging, medial temporal lobe, entropy, recognition

Categorization is a fundamental process that supports numerous behaviors in many organisms. Categories help organisms make sense of a complex world by grouping objects that share behaviorally relevant properties together to facilitate generalization and inference. For example, the category *wombats* allows people to recognize multiple instances of wombat as members of a kind and to generalize properties, such as *likes trees,* from one wombat to another.

Categorization itself is not a unitary process but, like many psychological phenomena, is made up of a number of component processes. Many of the processes associated with categorization occur simultaneously or within close temporal proximity to one another. For example, a person watching a wombat outside the window of his or her Peugeot may retrieve memories of past experiences with wombats and facts about wombats, all while visually processing the wombat and (one hopes) the road ahead. A central problem in categorization research is how to identify multiple, simultaneously occurring processes both behaviorally and neurally.

The problem of how to separate multiple cognitive processes that underlie a behavior is particularly pertinent in the neuroimaging literature. Standard functional magnetic resonance imaging (fMRI) analysis techniques typically compare brain activation across two or more task conditions (e.g., correct categorization relative to incorrect categorization) to identify brain regions that are more engaged for one condition than for another. These contrast-based techniques assume that some cognitive process of interest differs between conditions. However, in reality, most cognitive processes are not neatly separated between conditions but are simultaneously engaged to varying degrees throughout a task. For example, when encountering a novel object, a person may simultaneously engage recognition processes to determine the object's fit to current knowledge structures and motivational processes promoting exploration of the object to acquire additional

Tyler Davis, Imaging Research Center, The University of Texas at Austin; Bradley C. Love, Department of Psychology, The University of Texas at Austin; Alison R. Preston, Department of Psychology, Center for Learning and Memory, and Institute of Neuroscience, The University of Texas at Austin.

Bradley C. Love is now at the Department of Cognitive, Perceptual, and Brain Sciences, University College London, London, United Kingdom.

This work was made possible by Army Research Office Grant 55830-LS-YIP and National Science Foundation CAREER Award 1056019 to Alison R. Preston; Air Force Office of Scientific Research Grant FA9550-10-1-0268, Army Research Laboratory Grant W911NF-09-2-0038, and National Science Foundation Grant 0927315 to Bradley C. Love; and National Institute of Mental Health Grant MH091523 to Alison R. Preston and Bradley C. Love. Thanks to April Dominick, Jackson Liang, and Sasha Wolosin for help with data collection. Thanks to Manoj Doss for assistance with segmentation of the regions of interest.

Correspondence concerning this article should be addressed to Tyler Davis, Imaging Research Center, The University of Texas at Austin, 1 University Station, A8000, Austin, TX 78712. E-mail: thdavis@mail.utexas.edu

information. Accordingly, condition-based contrasts typically identify a wide variety of brain regions for a given comparison. While the common assumption is that different regions active for a particular condition support different cognitive processes, the ascription of a cognitive process onto any given region is vastly underdetermined by the data and relies unduly on "reverse inferences" from past research and theory (Poldrack, 2006).

Model-based fMRI offers a potential solution to the problem of localizing cognitive function in the brain (Daw, 2011; O'Doherty, Hampton, & Kim, 2007). In model-based imaging, quantities linked to processes in mathematical models are used to isolate and interpret patterns of brain activation. These model measures can be used to interrogate neuroimaging data and provide a more precise description of the cognitive functions mediated by different brain structures. Importantly, unlike most standard condition-based neuroimaging approaches, models can define distinct processes that are engaged at the same moment in time. Thus, combining fMRI with mathematical models of cognition offers an extremely powerful technique for understanding the neural basis of cognitive processes that govern behaviors like categorization.

Here, we combine computational modeling with high-resolution fMRI of the medial temporal lobes (MTL) and striatum, two neural systems that have played a central role in recent, neurobiologically inspired category-learning research. By combining the strengths of both techniques, we are able to identify separable computational processes related to category learning that occur simultaneously within subregions of the MTL and striatum. Specifically, we use a category-learning model to interrogate the brain basis of concurrent processes associated with item recognition and the uncertainty (i.e., entropy) of an item's assignment to a learned knowledge structure (i.e., cluster).

Neurobiological Accounts of MTL and Striatal Function

The MTL is one of the most frequently studied systems in neurobiological research on category learning, but current theories provide conflicting accounts of its computational role in category learning. Different theories have described the MTL's role in category learning a variety of different ways, including an explicit long-term memory-based system (Smith & Grossman, 2008), an exemplar-based system (Ashby & Maddox, 2005; Ashby & O'Brien, 2005; Pickering, 1997), a locus for storage and/or retrieval of rules in a rule-based system (Nomura et al., 2007; Seger & Cincotta, 2006), and a prototype-based system (Aizenstein et al., 2000; Glass, Chotibut, Pacheco, Schnyer, & Maddox, 2012; Reber, Gitelman, Parish, & Mesulam, 2003; Zeithamova, Maddox, & Schnyer, 2008). Recently, we put forward a more general, model-based account, which proposes that the MTL forms cluster-based representations that are tailored to meet the demands of the learning context (Davis, Love, & Preston, 2011). When categories can be distinguished by a regularity that can be captured by a prototype or a simple rule, a single cluster represents each category. If a context requires more fine-grained discriminations, multiple clusters are stored. That is, opposed to assuming a fixed representational form across learning contexts (e.g., rule, prototype, or exemplar), our approach suggests that the MTL can flexibly tailor representations to a given task.

Consistent with this theory, in a previous whole-brain fMRI study, we found that predictions generated from a clustering model, supervised and unsupervised stratified adaptive incremental network (SUSTAIN; Love, Medin, & Gureckis, 2004), tracked MTL activation during a category-learning task. One measure, recognition strength, indexed model-based processes related to retrieving stored representations from memory. The recognition strength measure predicted MTL activation during a stimulus presentation period when participants were trying to determine category membership. Another measure, error-correction, indexed processes related to updating memory in response to errors; this measure predicted MTL engagement during feedback when participants could update current category representations in response to decision outcomes.

The striatum is another region that has received widespread attention in the neurobiological category-learning literature. The striatum is believed to have a role in connecting category representations to behavioral responses via associative learning mechanisms (Ashby, Alfonso-Reese, Turken, & Waldron, 1998; Maddox & Ashby, 2004; Seger, 2008; Shohamy, Myers, Kalanithi, Gluck, 2008). The striatum is thought to support a procedural learning system that is functionally separate from and competitive with an explicit memory system comprising the frontal lobes and the MTL (Ashby et al., 1998; Poldrack & Packard, 2003). Although not the focus of our previous study, we found widespread activation in the striatum that correlated with both the recognition strength and the error correction measures, suggesting that the psychological processes indexed by our model-based measures may also depend on computations occurring in the striatum (Davis, Love, & Preston, 2011).

In the category-learning literature, the MTL and striatum are often treated as unitary structures that engage a single computational process at any given point in time. Anatomically, however, the MTL and striatum have a number of distinct subregions that may simultaneously contribute different cognitive processes in support of category-learning behavior. Anatomical differences between regions of the MTL provide possible clues to differences in their underlying function. While the MTL as a whole is often thought of as a region critical for the storage and retrieval of information in memory (for reviews, see Eichenbaum, Yonelinas, & Ranganath, 2007; Preston & Wagner, 2007; Squire, 1992) because of its increased neuromodulatory input and connectivity with putatively emotional regions of the brain (Canteras & Swanson, 1992; Cavada, Company, Tejedor, Cruz-Rizzolo, & Reinoso-Suarez, 2000; Witter, Wouterlood, Naber, & Van Haeften, 2006), the anterior hippocampus may be more sensitive to motivational factors than are other parts of the MTL (Fanselow & Dong, 2010; Moser & Moser, 1998; Strange & Dolan, 2006). Accordingly, in category learning and the broader memory literature, the anterior hippocampus is particularly engaged by novel and uncertain stimuli (Daselaar, Fleck, & Cabeza, 2006; Henson, Cansino, Herron, Robb, & Rugg, 2003; Schott et al., 2004; Seger, Dennison, Lopez-Paniaqua, Peterson, & Roark, 2011; Strange, Duggins, Penney, Dolan, & Friston, 2005; Strange, Fletcher, Henson, Friston, & Dolan, 1999; Strange, Hurlemann, Duggins, Heinze, & Dolan, 2005). Such novelty signals may relate to motivational signals that are present during category learning and may serve to orient attention to uncertain or behaviorally salient events (Davis, Love,

& Maddox, 2009) and to guide memory formation (Lisman & Grace, 2005).

Like the MTL, anatomical diversity within the striatum provides clues to functional differences between striatal subregions in terms of knowledge retrieval and uncertainty processing. The striatum interacts with cortical regions via a number of corticostriatal loops (Alexander, Delong, & Strick, 1986). Regions of the striatum that instantiate visual and motor loops, the dorsal tail and body of the caudate and putamen, form connections with cortical regions involved in visual perception and motor behavior and may have a role in guiding categorization choice by associating category representations to behavioral responses (Ashby et al., 1998; Cincotta & Seger, 2007; Seger & Cincotta, 2005; Seger, 2008). The ventral striatum is thought to engage a motivational loop (Seger & Miller, 2010) that connects the striatum to motivational processing centers in the ventromedial frontal cortex, midbrain, and amygdala. Like the anterior hippocampus, motivational loops in the ventral striatum may be involved with aspects of category learning related to motivational salience (Seger & Miller, 2010) and reinforcement learning (Seger et al., 2010). While the ventral striatum has received less attention in the category-learning literature than have other parts of the striatum, in the broader reinforcement learning literature, it is associated with signaling unexpected rewards from feedback (Berns, McClure, Pagnoni, & Montague, 2001; O'Doherty, Dayan, Friston, Critchley, & Dolan, 2003; Shultz et al., 1997) and may be a source for a "novelty exploration bonus" or uncertainty signal when people are confronted with uncertain but motivationally salient events (Krebs, Schott, Schutze, & Duzel, 2009; Wittman, Daw, Seymour, & Dolan, 2008).

Model-Based Predictions

We use the rational model of categorization (RMC; Anderson, 1991; Sanborn, Griffiths, & Navarro, 2010) to generate predictions about the simultaneous engagement of cognitive processes in subregions of the striatum and the MTL during category learning. The RMC was originally proposed as a computational-level model that describes, from a Bayesian perspective, what the basic categorization problem is and how it can be solved rationally. Here, we take a different approach, adopting a mechanistic interpretation of the RMC (Jones & Love, 2011; Sanborn, Griffiths, & Navarro, 2010) and using it to predict the processes that are occurring in different brain regions as participants learn novel categories. The RMC embodies many of the same processes as SUSTAIN, a different clustering model that we previously used to predict patterns of activation related to error correction and recognition strength in a similar task (Davis, Love, & Preston, 2011). Indeed, because of the high degree of similarity in the RMC and SUSTAIN's functional architecture, predictions for error correction and recognition strength from the two models share a high degree of overlap. Here, we use the RMC, instead of SUSTAIN, because the probabilistic formulation of the RMC makes it straightforward to define an additional measure, entropy, which may index motivational processes related to uncertainty processing in the striatum and the MTL.

We explore how the measures derived from the RMC relate to activation in the MTL and the striatum as participants learn a rule-plus-exception category-learning task (Davis, Love, & Preston, 2011; Love & Gureckis, 2007). In this task, participants learn to sort schematic beetles into one of two categories based on perceptual features (see Figure 1A, Table 1). Each trial contains a *stimulus presentation* period, during which participants make judgments about the category membership of a stimulus, and a *feedback* period, during which they receive corrective feedback (see Figure 1B). Participants are informed prior to beginning the task that most beetles can be accurately categorized by using a simple rule (e.g., if it has thick legs it is a Category A beetle) and are explicitly provided with directions to attend to the dimension (e.g., legs) that the rule will be based on. Participants are also informed that each category will contain an exception item that violates the rule and appears as if it should belong in the opposing category.

Behaviorally, exceptions tend to be associated with a higher frequency of errors during learning and lead to greater recognition success in postlearning recognition memory tests (Davis, Love, & Preston, 2011; Palmeri & Nosofsky, 1995; Sakamoto & Love, 2004, 2006). Neurobiologically, rule-plus-exception tasks are thought to engage clustering mechanisms in the MTL that form task appropriate groupings of the items by separating exceptions and rule-following items into their own clusters (Davis, Love, & Preston, 2011). Regions of the striatum that instantiate visual and motor loops may be involved with associating these category representations to behavioral responses (Meeter, Radics, Myers, Gluck, & Hopkins, 2008). The RMC is able to account for basic behavioral effects in rule-plus-exception tasks because, like SUSTAIN, it tends to form clusters or groupings of items that are appropriate for the task (see Figure 2). Exceptions tend to be represented by their own separate clusters, whereas rule-following items are more likely to be grouped in shared clusters.

Given that the RMC, like SUSTAIN, is a valid behavioral model for rule-plus-exception tasks, it has the potential to also provide an accurate account of the neural processes that participants engage while they learn the task. We examine three quantitative measures derived from the RMC: two that are designed to identify different computations that are present during the stimulus presentation period of the trial, recognition strength and entropy, and one that is used to account for activation during the feedback portion of the trial, error. The recognition strength and error measures predicted by the RMC overlap highly with predictions from SUSTAIN, and the entropy measure is a completely novel measure that captures processes related to motivated learning under uncertainty. Here, we give a brief algorithmic description and psychological interpretation of these measures (see the Appendix for model formalism).

The first measure that we examine in relation to activation during stimulus presentation is recognition strength. Recognition strength indexes the degree to which a stimulus is likely or expected given the RMC's probabilistic representation of the task (i.e., the probability of an item given the model). Recognition strength strongly relates to familiarity measures used to predict recognition performance following learning in rule-plus-exception tasks (e.g., Palmeri & Nosofsky, 1995; Sakamoto & Love, 2004). Psychologically, the recognition strength measure relates to the extent to which an item matches the RMC's stored category representations. Recognition strength tends to be similar for exceptions and rule-following items early in the task but differentiates the item types as learning progresses. The exception and rule-following items are differentiated because exception items tend to be stored in their own clusters, which provide a perfect

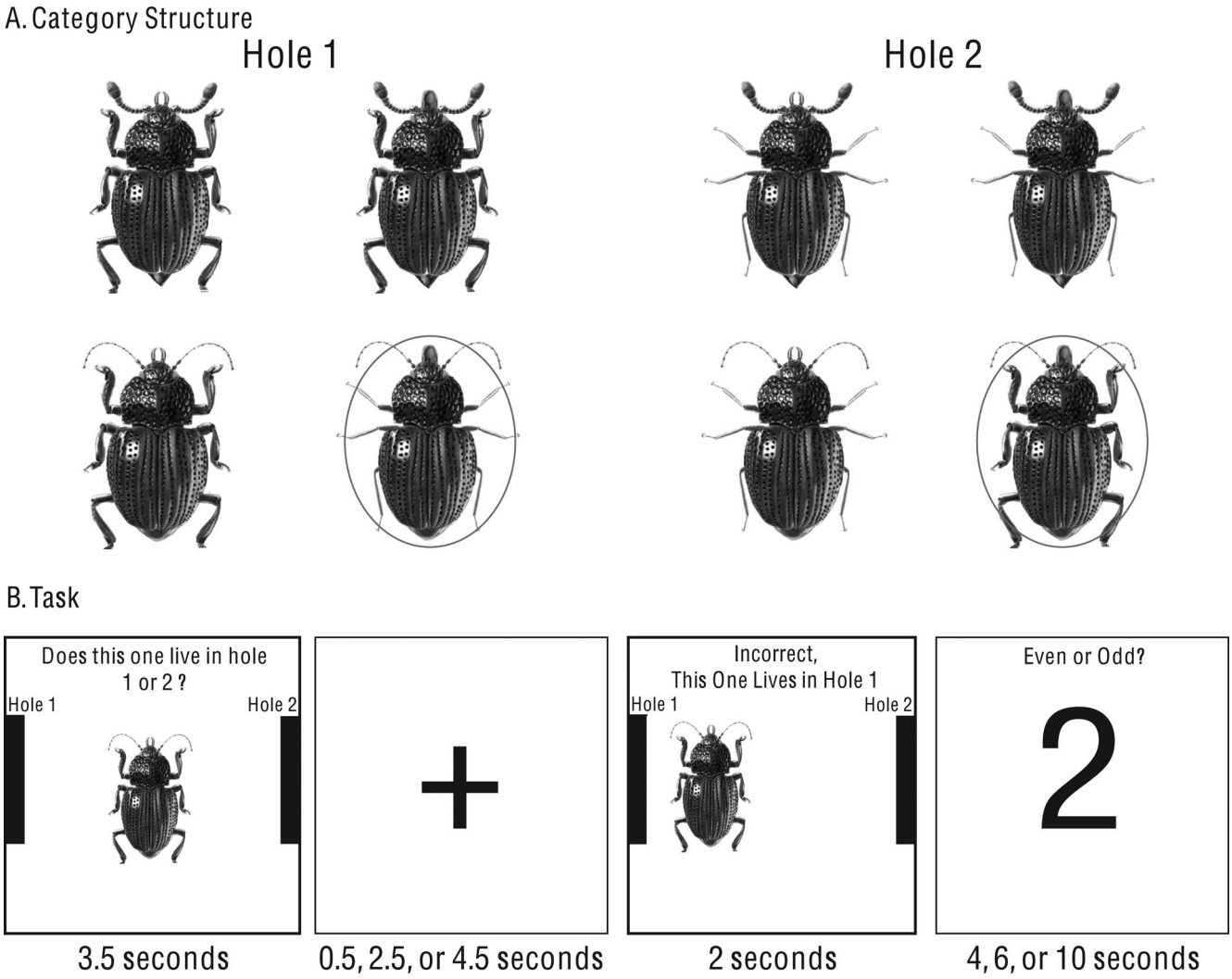

Figure 1. A: An example of category structure. The beetles vary on four of the following five perceptual dimensions when the fifth dimension is held fixed: eyes (green or red), tail (oval or triangular), legs (thin or thick), antennae (spindly or fuzzy), and fangs (pointy or round). The rule-relevant dimension in this example is legs. Most (3/4) of Hole 1 beetles have thick legs, whereas most (3/4) of Hole 2 beetles have thin legs. The two 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 are held constant in this example. B: Trial structure. During the stimulus presentation, a beetle was presented, and participants were asked to classify it as a Hole 1 or a Hole 2 beetle. Following a variable fixation, participants received feedback about whether they were correct or incorrect and about the correct category assignment. Feedback was followed by a variable number of even–odd digit trials that served as baseline.

match to exception features, leading to high likelihood. The rule-following items tend to be less differentiated than exception items because they are stored in shared clusters that do not tend to match any particular rule-following item perfectly, leading to lower likelihood for rule-following items on average.

The recognition strength measure (see Figure 3) is predicted to correlate with regions of the MTL involved with retrieving stored category representations from memory, as found in previous studies with SUSTAIN. We also predict that regions of the striatum that instantiate visual and motor loops, such as the tail of the caudate and posterior striatum, will correlate with the recognition strength measure because they are thought to be recruited for associating category representations to behavioral responses.

A second measure that we use to examine activation during stimulus presentation, entropy (see Figure 3), indexes the extent to which the RMC is uncertain about which cluster a stimulus belongs to, given the model's current probabilistic representation of the task. Entropy is highest when all potential outcomes (i.e., cluster assignments) are equally probable and lowest when only one potential outcome is likely. Events in the real world can be measured in terms of entropy as well. For example, on any normal day, a woman might expect to get a phone call only from her life

Table 1
Abstract Category Structure

Hole 1 beetles	Hole 2 beetles
2 2 2 2[a]	**1** 2 2 2[a]
1 1 1 2	**2** 1 1 2
1 1 2 1	**2** 1 2 1
1 2 1 1	**2** 2 1 1

Recognition test foils
1 1 1 1
1 1 2 2
1 2 1 2
1 2 2 1
2 2 2 2
2 2 1 1
2 1 2 1
2 1 1 2

Note. Each row represents a unique stimulus (i.e., beetle). The four values assigned to a stimulus denote the four stimulus dimensions (e.g., legs, antennae) assigned to a beetle. Each numeric value (1 or 2) represents a specific feature instantiation (e.g., red or green eyes). The first dimension (in bold) indicates the rule-relevant dimension. Most Hole 1 beetles have a 1 on the first dimension (e.g., thick legs) whereas most Hole 2 beetles have a 2 (e.g., thin legs). The first stimulus in each of the columns is therefore an exception.
[a] Exception item.

partner, making the identity of a caller highly predictable and entropy low. On a birthday or holiday, however, the number of people who are likely to call may increase, making the identity of a caller less predictable and entropy high.

In the task, entropy is higher initially for both item types because the model's representations of the category structure are still being formed and updated by new stimuli. As learning progresses, entropy decreases because the assignment of items to clusters becomes more certain. In addition to the primary temporal component whereby entropy decreases over the course of the experiment, the entropy measure also predicts an item-based component whereby entropy is higher for the exception items, compared with rule-following items, throughout the task. Exception items are associated with higher entropy rather than rule-following items because exceptions are more confusable with stimuli in the opposing category and, thus, partially match clusters for the opposing category. When there are partial matches to multiple clusters, as in the case of exceptions, entropy is higher because the RMC is more uncertain about which cluster to choose. Eventually, if the model were to run on the same set of trials indefinitely, entropy would approach zero for both item types. In short, the RMC, like people, tends to be more uncertain about which clusters to assign items to early in the experiment and remains more uncertain for exception items throughout the number of trials used in the present experiment. These two factors, greater uncertainty for exceptions and decreasing uncertainty with more training, are both reflected in the entropy measure.

Because uncertainty is a powerful motivating force for learning, the entropy measure could potentially relate to activation in widespread regions of the MTL and striatum. Two subregions that may be particularly sensitive to the entropy measure are the anterior hippocampus and ventral striatum. These regions have been associated with novelty and uncertainty processing (Strange, Duggins, et al., 2005; Strange et al., 1999; Strange, Hurlemann, et al., 2005) and with directing encoding toward sources of novel or unexpected information (Wittman, Bunzeck, Dolan, & Duzel, 2007; Wittman et al., 2008). Such "information readiness" signals are conceptually related to the orienting response (Bradley, 2009; Sokolov, 1966; Sokolov, Spinks, Naatanen, & Lyytinen, 2002), a complex of neurophysiological motivational responses that have been related to exception learning in a similar task (Davis, Love, & Maddox, 2009).

Error is a final measure (see Figure 4) that we use to predict activation during the portion of the trial in which participants receive feedback (see Figure 1B). The error measure is the probability that the RMC will make an error on a given trial and is thought to index psychological processes associated with updating representations in memory in response to prediction error. Because rule-following items look like many of the other members of their category, they are associated with fewer errors than are the exception items, which look like members of the opposing category. The resemblance that rule-following items share with other members of their category also leads them to be grouped in shared clusters with other rule-following items, whereas exceptions will be stored in individual clusters. Because multiple rule-following items are stored in the rule-following clusters, these clusters also have a higher base rate than exception clusters, which makes them more probable and contributes to higher error. MTL regions that store category representations should correlate with the error measure when participants receive feedback because an error provides a signal to update these representations. Likewise, an error should provide a signal to regions of the striatum that instantiate visual and motor loops to adjust the strength of their associations between category representation and behavioral responses.

We use high-resolution fMRI of the MTL and the striatum to investigate predictions based on the model-based measures. The greater spatial resolution and higher signal-to-noise ratio of high-resolution fMRI (Carr, Rissman, & Wagner, 2010) enhances our ability to isolate simultaneous computational processes within the MTL and the striatum, compared with conventional whole-brain imaging techniques employed in previous studies (i.e., Davis, Love, & Preston, 2011). For example, processes within a single subregion of the MTL or the striatum may reflect both entropy and recognition. In conventional whole-brain imaging, such signals are more likely to be blurred together. By using high-resolution fMRI, we thus have a greater ability to disentangle patterns reflecting different processes.

Materials and Method

Participants

Thirty-three healthy, right-handed volunteers participated in the experiment after giving informed consent in accordance with a protocol approved by the Stanford and the University of Texas institutional review boards. Participants received $20 per hr for their involvement. Seven participants were excluded for failing to achieve greater than 50% performance on exception items in the final (6th) run.

Materials

Participants completed a rule-plus-exception category-learning task (Love & Gureckis, 2007) during fMRI scanning. The stimuli

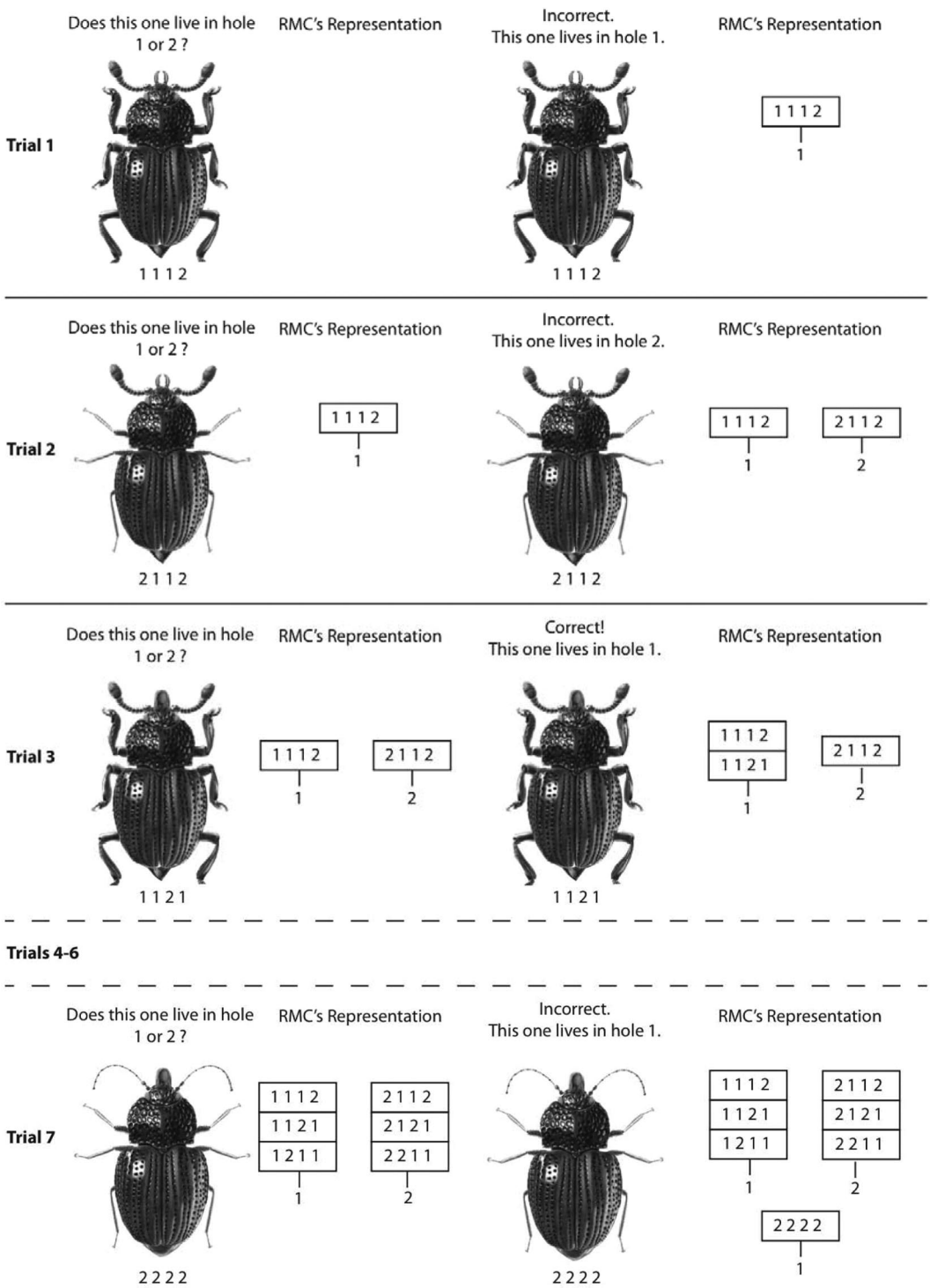

Figure 2 (opposite).

used in the task were schematic beetles that varied along four perceptual dimensions (see Figure 1A, Table 1) and were assigned to categories (Hole 1 or Hole 2) based on their combinations of feature values. For each stimulus, four of five 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 1A, the rule-relevant dimension was the legs; all but one of the beetles in Hole 1 had thick legs, and all but one of the beetles in Hole 2 had thin legs. The other two beetles (circled beetles in Figure 1A) 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 the Table 1. In order to minimize the effects of feature salience, the mapping of each abstract dimension to a physical dimension was randomized for each participant.

Procedures

On each trial of the category-learning task, a single beetle was presented in the center of the screen, and participants were asked to decide whether it was a Hole 1 or Hole 2 beetle (see Figure 1B). Each stimulus was presented for 3.5 s, during which time participants had to indicate category membership via a button box held in their right hand. After a brief fixation (0.5, 2.5, or 4.5 s; M = 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). At the end of the trial, participants were informed about whether their response on that trial was correct or incorrect. An even–odd digit task (Stark & Squire, 2001) served as a baseline between trials (mean baseline time per trial = 6 s). We chose an active baseline task over a passive baseline task (e.g., rest or fixation) because the MTL tends to have more variable and higher magnitude activation during passive baseline tasks. Active-baseline tasks, which have become standard in the literature on MTL function, provide enhanced sensitivity for measuring task-related activation in the MTL (Stark & Squire, 2001). No feedback was given during the even–odd digit task.

Participants were trained with the rule-plus-exception procedure for six functional runs, each lasting 9 min and 53 s. During each run, the eight stimuli (beetles) were presented five times sequentially in a pseudorandom order. We manipulated the order of stimuli within each run to facilitate exception learning and reduce the number of nonlearners. In contrast to previous studies in which stimulus order was random in every block, in the first block of each run, each of the rule-following items were presented one time before any exception items. Trial order and duration were optimized for each of the six 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 six functional runs across participants. The first 12 s of each run, consisting of fixation, were discarded. Prior to beginning the task, participants were given explicit instructions indicating the rule-relevant dimension for category membership and were encouraged to memorize the exceptions to the rule (Davis, Love, & Preston, 2011; Love & Gureckis, 2007).

Following the category-learning task, participants completed a self-paced, two-alternative forced-choice recognition memory task outside of the scanner. Recognition memory tests are often used as a measure of exception processing in rule-plus-exception tasks (e.g., Palmeri & Nosofsky, 1995; Sakamoto & Love, 2004, 2006; Davis, Love, & Maddox, 2009). Testing recognition also serves as a behavioral check for predictions of the model-based recognition strength measure, which predicts that exceptions will be recognized more strongly than will rule-following items. On each trial of the recognition task, participants were presented with two beetles: one that was presented during the category-learning phase and a foil that was not presented during the category-learning phase. Participants were asked to identify the old item presented during the scanned rule-plus-exception task.

fMRI Acquisition

Imaging data were acquired on a 3.0 T General Electric Signa whole-body magnetic resonance imaging (MRI) system (GE Medical Systems, Milwaukee, WI) with an eight-channel head coil array. High-resolution structural images using a T2-weighted, flow-compensated spin-echo sequence (TR [repetition time] = 3 s, TE [echo time] = 68 ms, 0.43 × 0.43 inplane resolution) were acquired with 20 3-mm thick slices that were perpendicular to the main axis of hippocampus, to optimize visualization of hippocampal subfields, MTL cortical subregions, and striatum. Functional images were acquired with a high-resolution T2*-sensitive, gradient-echo spiral in/out pulse sequence with the same slice prescription as the structural images (TR = 4 s, TE = 34 ms, flip angle = 80°, FOV [field of view] = 22 cm, 1.7 × 1.7 × 3.0 mm resolution). A high-order shimming procedure was used prior to scanning to reduce B_0 heterogeneity.

Figure 2 (opposite). A trial-by-trial schematic of the task and the rational model of categorization's (RMC's) clustering behavior over the first seven trials. On each trial, a beetle stimulus is presented, and the RMC is asked to guess the category membership. Each beetle corresponds to a four binary-digit code listed in Table 1 that corresponds to the beetle's feature instantiations on each of the perceptual dimensions. These codes represent how the beetles' are coded in the RMC. The RMC compares the presented beetle to cluster representations stored in memory and makes a choice. After a choice is made, the RMC is given feedback about the correct category membership. The RMC uses this feedback to update its cluster representations. On each trial, a stimulus is added to either a current cluster or a new cluster. In this example, the RMC sees all of the rule-following items for each category before the first exception (on Trial 7). In the first two trials, it forms two clusters that represent the rule-following items in each category. In the next four trials, it adds subsequent rule-following items to these clusters because they do not violate the regularity in the cluster. In the first 7 trials, all stimuli in a category cluster (Category 1 or 2) have the same value on the first, rule-following dimension. The exception (Trial 7) violates this regularity and thus, when encountered, requires a new cluster to be formed.

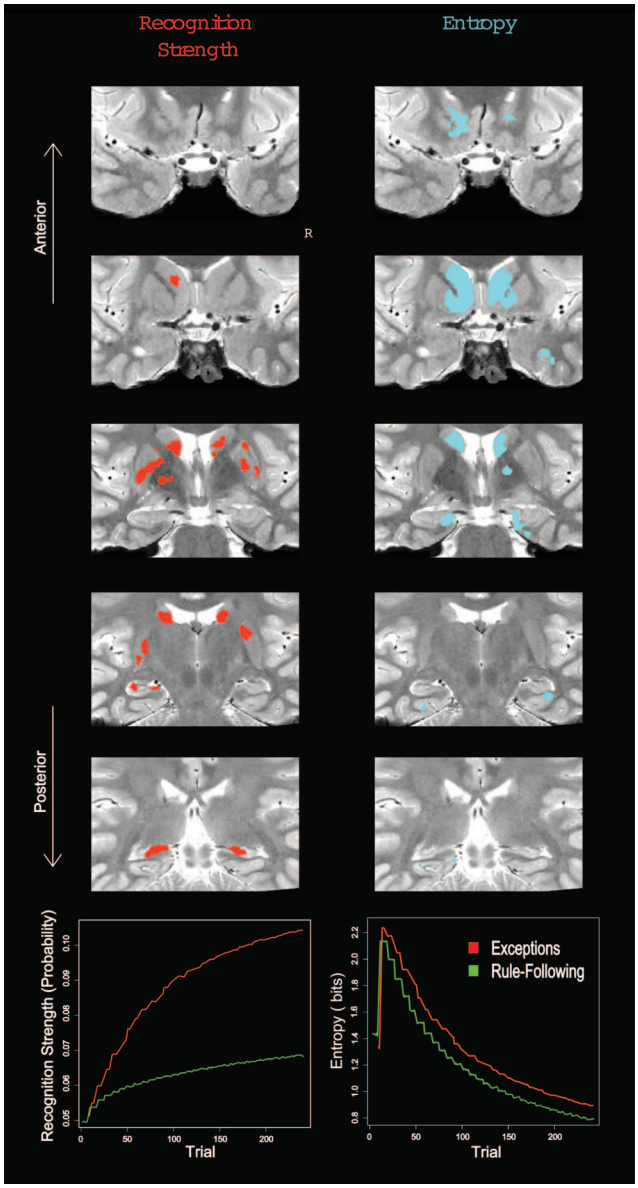

Figure 3. Illustrations of the model-based measures used to predict activation during the stimulus presentation period and corresponding functional magnetic resonance imaging (fMRI) results. Brain regions associated with the rational model of categorization's (RMC's) recognition strength measure are depicted in red, and brain regions associated with the entropy measure are depicted in cyan. The bottom panel represents the predicted shape of each model-based regressor for the two item-types over the course of the experiment. For the model-based measures, the predicted pattern for exception trials is given in red, and the predicted pattern for rule-following trials is given in green. R = right.

A total of 858 volumes were acquired for each participant. The first volume of each functional run was collected with an echo time 2 ms longer than all subsequent volumes to create a field map for the correction of magnetic field heterogeneity. For each slice, the map was calculated from the phase of the first two time frames and applied as a first-order correction during reconstruction of the functional images to minimize blurring and geometric distortion on a per-slice basis. Correction for off-resonance due to breathing was applied on a per-time-frame basis, using phase navigation. The initial volume and the following two volumes of each functional run (a total of 12 s) were discarded to allow for T1 stabilization.

fMRI Analyses

The fMRI data were analyzed using SPM5 and custom MATLAB routines. Slice timing and motion correction were applied to all images. A mean functional image was computed and used to coregister functional and the high-resolution anatomical images.

Voxel-based statistical analyses were conducted at the individual participant level according to the general linear model. Regressor functions were constructed by modeling condition related activation as an impulse function convolved with a canonical hemodynamic response. For the model-based analysis, model-based measures of recognition strength and entropy were fit, in separate models, as parametric modulators of the stimulus presentation period. The error measure was fit as a parametric modulator of the feedback period in each model.

Group-level analyses were conducted using a nonlinear diffeomorphic transformation method (Vercauteren, Pennec, Perchant, & Ayache, 2009). Each participant's anatomically defined MTL and striatal regions-of-interest (ROIs) were aligned with those of a representative "target" participant, using a diffeomorphic deformation algorithm that implements a biologically plausible transformation. First, anatomically defined ROIs were demarcated on the T2-weighted, high-resolution inplane structural images for each individual participant. Eight MTL subregions were defined in each hemisphere: the hippocampal subfields (dentate gyrus/$CA_{2/3}$, CA_1, and subiculum) within the body of the hippocampus, surrounding MTL cortices (perirhinal, parahippocampal, and entorhinal cortex), and the most anterior and posterior slices of the hippocampus in which the subfields cannot be delineated at the resolution employed (Preston et al., 2007; Zeineh, Engel, Thompson, & Bookheimer, 2003). Three striatal subregions were defined in each hemisphere: putamen, caudate, and pallidum.

All participants' anatomical images were warped into the target participants' space in a manner that maintained the between-regions boundaries. To enhance the accuracy of the registration processes, registrations were performed separately for each hemisphere (left–right) for each of the primary ROI groups: hippocampus, MTL cortex, and striatum. Compared with conventional whole-brain normalization, the use of separate ROIs results in more accurate correspondence of MTL subregions across participants and higher statistical sensitivity (e.g., Kirwan, Jones, Miller, & Stark, 2007; Yassa & Stark, 2009).

Participant-level statistical contrast maps for each region were transformed with the anatomical normalization matrix and combined across participants for group statistical analyses. Group-level statistical maps were created with an uncorrected voxel-wise threshold of $p < .025$ and a multiple comparisons corrected cluster-level threshold of $p < .05$. The minimum cluster sizes were determined separately for the hippocampus (32 voxels), MTL cortex (37 voxels), and striatal (43 voxels) ROIs with Monte Carlo simulations implemented in AFNI's AlphaSim, which takes into account the size and shape of each region, as well as the height threshold p value and the smoothness of actual data. For each

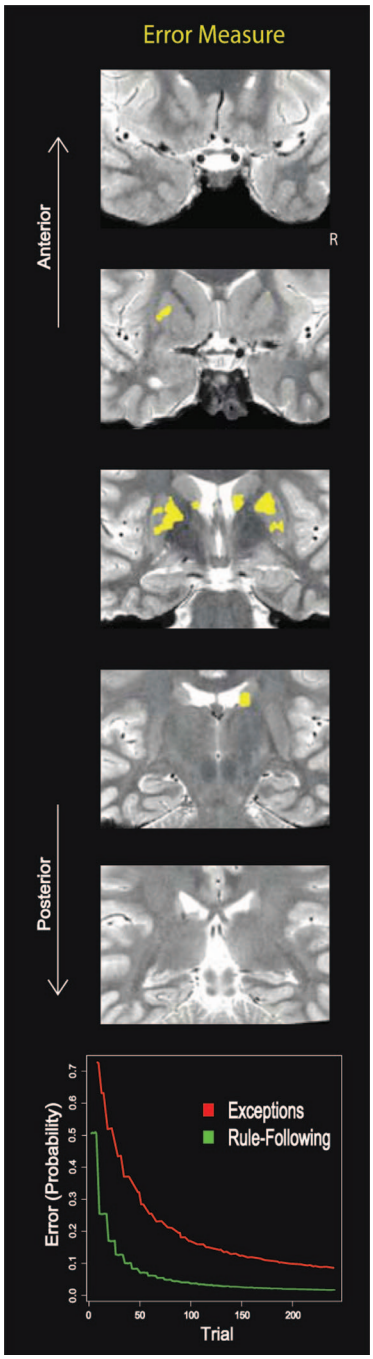

Figure 4. Illustration of the model-based error measure used to predict activation during the feedback period and corresponding functional magnetic resonance imaging (fMRI) results. The top panels show the brain regions associated with the rational model of categorization's (RMC's) error measure in yellow. Bottom panels represent the predicted shape of the model-based error regressor for the two item-types (exceptions in red, rule-following items in green) over the course of the experiment. R = right.

significant cluster, we report the cluster size, *k*, and the uncorrected voxel-wise *z* value at the peak.

Model-Based Analysis

Recognition strength, error, and entropy measures were formulated within the single particle filter version of the RMC. Here, we give a conceptual overview of how the model works and how each of the measures was generated. A formal mathematical description of the RMC is detailed in the Appendix (see also, Anderson, 1991; Sanborn et al., 2010).

Model operation. Figure 2 details a trial-by-trial progression of how the RMC forms clusters and updates clusters over the first seven trials of the experiment. On the first trial, the RMC begins the task with no cluster representations stored in memory and recruits a new cluster to represent the first stimulus. On each subsequent trial of the task, the RMC is presented with a stimulus, and the RMC compares it with cluster representations stored in memory to try to assign the unobserved category label. The clusters code the conjunctions of stimulus features and category labels observed on previous trials. The RMC uses the match between the current stimulus' features and the features of the stored clusters to compute a probability of membership for each cluster and an entirely new cluster. Each of the clusters makes a prediction about the category to which the stimulus belongs. The RMC combines the category label predictions across clusters by weighting each cluster's prediction by the cluster's probability, with proper Bayesian aggregation. The model uses the category probabilities that are aggregated across clusters to choose a response. After choosing a category, the model, like the participants, is given feedback about the correct category label. The RMC uses this feedback to update its representations by recomputing the probability that the stimulus belongs to each of the clusters, now with information about the stimulus' category label. The RMC assigns the stimulus to one of the current clusters or a new cluster, with a probability proportional to the cluster's probability, given the stimulus.

Model-based measures. The model-based measures are computed during different stages of the RMC's trial-by-trial computations. Recognition strength is computed during the stage at which the model is computing the cluster probabilities (see Appendix), and recognition strength reflects the strength of the match between the current stimulus and all clusters stored in memory. Recognition strength is the model's expected probability of seeing the current stimulus, given its current cluster representations (see Figure 3). Recognition strength has a general increasing trend because the stimulus set repeats over blocks, and thus, the model comes to anticipate each of the items. Recognition strength is higher for exceptions because they tend to be stored in individual clusters, which provide an exact match to each exception item. Rule-following items tend to be stored in shared clusters that do not perfectly match any given rule-following item (see Figure 2).

Entropy is computed over the uncertainty of cluster membership (see Appendix for formula for entropy). Entropy is high when the stimulus is likely to be a member of multiple clusters and low when only one cluster is likely. The RMC predicts that entropy is high early in learning but decreases throughout the experiment as the model becomes more certain in its cluster assignments. Entropy is also higher for exception items because they tend to partially match clusters of opposing categories, leading to more cluster uncertainty.

Finally, the error measure is computed after the RMC has computed a response probability, and the error measure is the probability that the category assignment chosen by the model is incorrect. Because exceptions tend to match clusters that contain members of the opposing category, this confusion continues into the decision stage and leads to more errors.

Model fitting. Each of the model-based measures was generated from fits of the RMC to participants' average performance over each scanning run (see Figure 5). For each scanning run, the model's predicted performance for exception and rule-following items was computed and compared with the participants' actual behavioral performance. The parameters that minimized the discrepancy between the predicted performance and actual performance across runs were selected (see Appendix). We followed standard practices in cognitive modeling and model-based fMRI and used predictions based on fits to group data (Daw, Gershman, Seymour, Dayan, & Dolan, 2011; Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006; Davis, Love, & Preston, 2011). Predictions from the models' fit to group data are frequently used in model-based imaging to overcome the noise inherent in single-participant data (see Daw, 2011). Because individual participant data are often noisy, parameter estimates and measures obtained from fits to individual participant data can be unreliable (for related, broader arguments for group-averaged data, see Cohen, Sanborn, & Shiffrin, 2008).

Measures averaging over items. In addition to fitting each of the measures as formulated by the model, we also examined measures that average over the differences between exceptions and rule-following items predicted by the entropy, recognition strength, and error measures to test the role of each measure's general time course in driving the results obtained. For example, averaging over exceptions and rule-following items in the entropy measure creates a single time course that is a weighted average of the rule-following and exception lines depicted for entropy in Figure 3. Coupled with the standard item-based analysis, the temporal measures that average over item effects provide auxiliary information about model successes and failures; they allow us to assess the general role of time course in driving the results observed for model-based analysis, whereas the item-based analysis allows us to assess the role of each of the measures' predicted item-based effects.

Results

Behavioral Results

Participants remembered exception items ($M = 0.69$, $SD = 0.28$) more accurately than rule-following items ($M = 0.52$, $SD = 0.16$) during the postscan recognition memory test, $t(25) = 2.91$, $p < .001$ (see Figure 6). In contrast, during the learning phase, participants categorized exception items ($M = 0.77$, $SD = 0.14$) less accurately than rule-following items ($M = 0.91$, $SD = 0.05$), $t(25) = 7.50$, $p < .001$. Although participants were less accurate in categorizing exception items in the current task, they were considerably more accurate than in previous related tasks (Davis, Love, & Preston, 2011; Love & Gureckis, 2007), suggesting that presenting all rule-following items in the beginning of each run

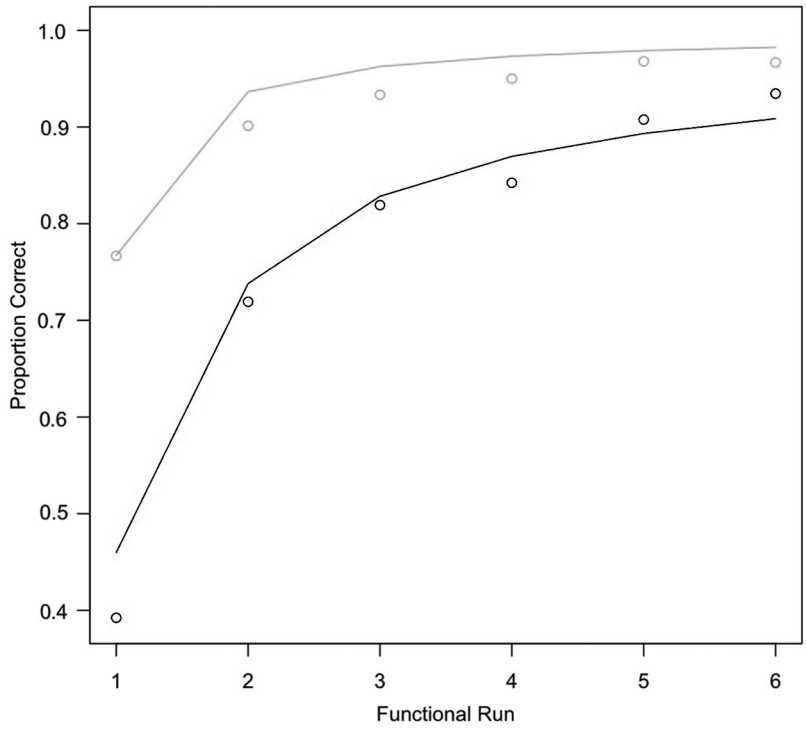

Figure 5. Illustration of model fit to behavior. Circles depict the observed proportion correct for each item type (black line = exception; gray line = rule following). Lines depict the proportion correct predicted by the model for each item type. Observed and predicted results are averaged over the functional runs.

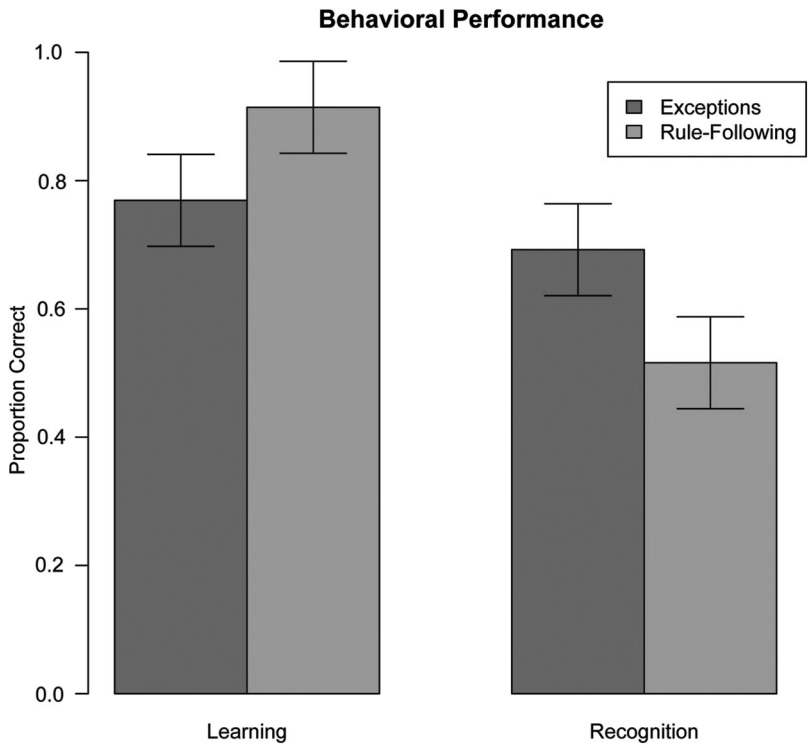

Figure 6. Behavioral results for the category-learning phase and postscanning two-alternative forced-choice (2AFC) recognition phase. Error bars give 95% within-subjects confidence intervals.

prior to introducing exception items succeeded in increasing performance.

fMRI Results

Activation during stimulus presentation. The primary goal of the fMRI analysis was to identify simultaneous processes occurring during the stimulus presentation period, when participants are trying to determine category membership. Two types of processes are hypothesized to occur during this decision phase: (a) retrieval of category representations from memory to drive choice and (b) cognitive and motivational processes related to detecting and processing uncertainty (e.g., Davis, Love, & Maddox, 2009). The recognition strength measure is used to localize processes related to memory retrieval, and the entropy measure is used to localize processes related to uncertainty.

Recognition strength measure. The recognition strength measure indexes the degree of match between a stimulus and stored category representations (see Figure 3). Consistent with previous work (Davis, Love, & Preston, 2011) and a widely supported role for the hippocampus in the storage and retrieval of memories, we observe bilateral clusters of activation with peaks in the posterior hippocampus that correlated with the recognition strength measure (right: $k = 227$, $z = 3.27$; left: $k = 66$, $z = 3.16$).

In addition, there was also significant activation in the striatum with bilateral peaks in the posterior/tail of the caudate (right: $k = 390$, $z = 4.83$; left: $k = 348$, $z = 4.33$), the location of which is consistent with the anatomical localization of the visual loop. This activation extended into the posterior slices of the anterior/head of the caudate and into the putamen, consistent with the anatomical localization of the motor loop (right: $k = 556$, $z = 3.95$; left: $k = 318$, $z = 3.58$). This pattern of activation is consistent with theories suggesting that the striatum has a role in associating category representations to behavioral responses.

Entropy measure. The entropy measure indexes the degree to which the model is uncertain about the cluster assignment for a stimulus and should relate to cognitive and motivational processes engaged in the anterior hippocampus and ventral striatum that detect uncertainty and facilitate learning under uncertain conditions (see Figure 3). Consistent with these predictions, activation that was correlated with the entropy measure was observed in both the anterior hippocampus (right: $k = 69$, $z = 4.13$; left: $k = 70$, $z = 2.98$) and anterior striatum, with the anterior striatal activation also extending into the ventral striatum (right: $k = 901$, $z = 4.18$; left: $k = 868$, $z = 4.90$) and posterior into the body/tail of the caudate. There were also clusters of activation in the right posterior parahippocampal cortex ($k = 144$, $z = 2.89$), left entorhinal cortex, and left perirhinal cortex ($k = 55$, $z = 3.69$; $k = 108$, $z = 3.49$; $k = 45$, $z = 3.19$). The MTL clusters were completely nonoverlapping with the recognition strength measure, and in the striatum, the entropy measure was the only model-based measure for which the clusters extended into the ventral striatum.

Auxiliary measures. In addition to the full model-based measures, we also examined individual components of the model-based measures to assess their contribution to the overall fit. Each model-based measure predicts an item-based component (exception $>$ rule-following items) as well as a temporal

component or the measure's general trajectory across time. We assessed the item-based component of the model-based measures using a standard general linear model based contrast comparing activation associated with correct exception items to correct rule-following items. This item-based analysis assumes that the difference between item types is constant throughout the task. To assess the general temporal component, we created measures that averaged over the item effects in the original model-based measures, thereby predicting a single, average time course for exceptions and rule-following items (see Materials and Method section). It is important to note that these auxiliary measures are not alternative hypotheses to the full model-based measures. Indeed, they are components of the model-based measures themselves.

Both the entropy and recognition strength measures predict an item-based component, and thus the contrast of correct exceptions > correct-rule-following items revealed clusters that approach a spatial mixture of the two model-based measures (see Figure 7). However, because the item-based component was a much larger part of the recognition strength measure, the results most strongly matched those of the recognition strength measure. In the MTL, activation for this contrast was primarily localized in the posterior hippocampus (right: $k = 432$, $z = 4.06$; left: $k = 267$, $z = 4.17$). The anterior hippocampal region identified with the model-based entropy measure was not observed with this condition-based comparison. In the striatum, this contrast revealed clusters of activation in the caudate that were separately identified as being associated with model-based measures of recognition and entropy (right tail: $k = 553$, $z = 4.58$; right anterior: $k = 55$, $z = 2.80$; left tail: $k = 514$, $z = 4.71$).

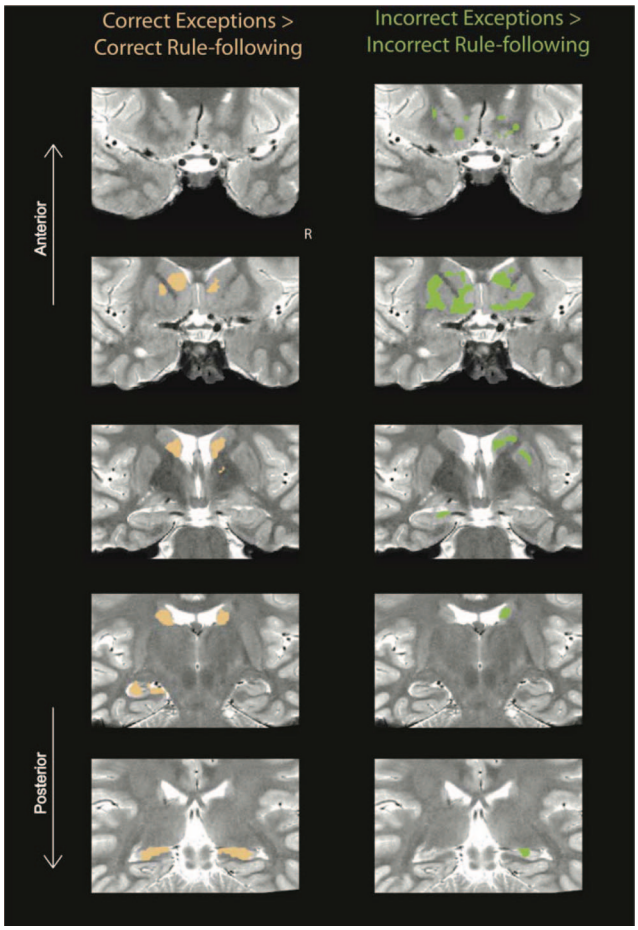

Figure 7. The fMRI results for the standard condition-based contrasts. Copper depicts activation during stimulus presentation that is significantly greater for correct exceptions, compared with correct rule-following items. Green depicts activation during feedback that is significantly greater for incorrect exceptions, compared with incorrect rule-following items. R = right.

The general time courses of the model-based measures, averaging over items, were also able to recover aspects of the results revealed by the full model-based measures. The amount of overlap between these results and those for the full model-based measures was largely related to the amount of variance in the full model-based measures that was accounted for by the general time course. The general change across time, averaged over items, accounted for approximately 97% of the variance in the full model-based entropy measure, but in the recognition strength measure, the general changes across time only accounted for approximately 28% of the variance. Accordingly, the results revealed by the general time course in the entropy measure highly overlapped those predicted by the original entropy measure. Activation was observed in the anterior hippocampus (right: $k = 76$, $z = 3.84$; left: $k = 79$, $z = 3.04$), anterior and ventral striatum (right: $k = 775$, $z = 4.51$; left: $k = 779$, $z = 4.49$), and MTL cortex (right: $k = 135$, $z = 2.85$; left: $k = 48$, $z = 3.57$; $k = 47$, $z = 3.20$; $k = 97$, $z = 3.60$). The general time course of the recognition strength measure, averaging over items, had less success in accounting for the results of the original recognition strength measure, with clusters restricted to the putamen (right: k = 673, z = 4.53; l‘9eft: k = 782, z = 4.02).

Taken together, the item-based and temporal aspects of the model-based measures are able to predict activation in many of the regions observed for the full model-based measures. This pattern is expected, as both are components of the full model-based measures. However, because fMRI data are noisy and likely do not correspond perfectly with the model-based measures, tests for the individual temporal or item-based components can be underpowered, particularly for effects that account for a small amount of the variance in the original measures (e.g., the item-based effect in entropy). Nevertheless, these auxiliary measures derived from the model help to illustrate the aspects of the full model-based measures to which the subregions of the MTL and striatum are particularly sensitive.

Activation during feedback. In addition to the primary objective of dissociating the computational processes engaged during the stimulus presentation period, we examined how the model-based predictions for error updating engaged MTL and striatal regions during feedback. The reason for modeling feedback is twofold: first, it is critical to model feedback separately from the stimulus presentation because the psychological processes that participants engage during these time periods have very different time courses. Second, modeling the feedback period can provide convergent evidence for conclusions drawn using SUSTAIN in previous studies (Davis, Love, & Preston, 2011).

Error measure. The error measure indexes processes that occur during the feedback portion of the trial (see Figure 1B)

associated with updating representations in response to prediction error (see Figure 4). The striatum is one of the key structures thought to update associations between visual stimuli or category representations and motor outputs (Ashby et al., 1998; Cincotta & Seger, 2007; Seger, 2008; Seger & Cincotta, 2005). Specifically, connections between the visual and motor loops are thought to be strengthened on the basis of feedback. Accordingly, we found peaks of activation that correlated with the error function bilaterally in the tail of the caudate (right: $k = 125$, $z = 3.87$; left: $k = 217$, $z = 5.05$) and putamen (right: $k = 432$, $z = 3.58$; left: $k = 343$, $z = 4.38$). These regions overlapped with those from the recognition strength measure, suggesting that regions of the striatum that are engaged during categorization are also engaged during feedback. That is, the same regions that exhibit one time course during stimulus presentation exhibit a very different time course during feedback; however, both time courses are consistent with the general role of the striatum in associating category representations and responses.

In contrast to previous research, the error measure did not correlate with activation in any region of the MTL. The MTL has also been hypothesized to update representations in response to prediction error (Davis, Love, & Preston, 2011), but its time course at feedback was not well accounted for by the error measure in the current task (see Auxiliary Measures). While this result is somewhat surprising, it is likely due to the accelerated exception learning observed in the present experiment; memory updating was likely faster than the gradual curve predicted by the RMC. Indeed, as described in the Materials and Method section, we deliberately presented all rule-following items before any exceptions at the beginning of each run to speed up learning and reduce the number of nonlearners from previous experiments.

Auxiliary measures. We also conducted auxiliary analyses to separately examine the temporal and item-based effects predicted by the error measure (see Figure 7). The general time course of the error measure, averaging over items, accounted for an intermediate amount of the variance in the original error measure (69%) and, accordingly, was able to capture some, but not all of the error measure's effects, with activation primarily localized in the putamen (right: $k = 411$, $z = 3.92$; left: $k = 262$, $z = 3.30$). An item-based contrast comparing incorrect exceptions to incorrect rule-following items predicted activation across the striatum similar to those revealed by the error measure (right: $k = 607$, $z = 3.88$; left: $k = 900$, $z = 4.42$). However, this item-based feedback contrast also revealed patterns of activation in the hippocampus that were missed by the model-based analyses (right anterior hippocampus: $k = 92$, $z = 2.89$; left posterior hippocampus: $k = 85$, $z = 4.14$), suggesting that the hippocampus was indeed engaged during feedback but that its time course was different from that predicted by the error measure. While this contrast does reveal that the hippocampus is engaged during feedback, it is important to remember that it does not speak to the specific computational processes that the hippocampus is supporting.

Discussion

Our goal in the present study was to use model-based analysis to identify multiple, simultaneous category-learning processes within the MTL and striatum. While traditionally the MTL and striatum have been treated as functionally separate and homogeneous systems that contribute a single process to category-learning behavior, our results suggest that there is functional heterogeneity within each system in terms of their underlying computational contributions to category learning. Notably, subregions within the MTL and striatum simultaneously contribute different computational processes to support category-learning behavior.

While recent research has begun to posit functional specificity within the MTL and striatum, progress has been hampered by the fact that many simple behavioral or condition-based contrasts exhibit activation in widespread regions of the MTL and striatum. By combining computational modeling's ability to predict fine-grained differences between simultaneously active computational processes with the enhanced spatial resolution and signal-to-noise ratio of high-resolution fMRI, we are able to isolate signals related to three different cognitive processes in the MTL and striatum: recognition strength, entropy, and error correction. These model-based measures, derived from the RMC, index psychological processes related to the retrieval of category representations from memory, cognitive and motivational processes associated with uncertainty, and processes associated with updating representations in response to error, respectively.

Consistent with previous results and the general theory that the hippocampus has a role in storing and retrieving information from memory, we found that a region of the posterior hippocampus correlated with the model-based recognition strength measure. A number of recent findings in the episodic memory literature suggest that retrieval processes are subserved primarily by the posterior hippocampus (Chua, Schacter, Rand-Givannetti, & Sperling, 2007; Daselaar et al., 2006; Prince, Daselaar, & Cabeza, 2005; Strange et al., 1999; Strange, Duggins, et al., 2005; Strange, Hurlemann, et al., 2005b); however, other studies have linked episodic retrieval to both anterior and posterior hippocampal activation (Kircher et al., 2008). Our finding of a recognition strength signal localized in the posterior hippocampus may result from the enhanced ability of model-based methods to separate retrieval-based processes from other simultaneously occurring processes or, perhaps, methodological differences between category learning and episodic memory experiments. Future researchers may wish to incorporate model-based predictions for retrieval into experiments in the episodic memory domain to provide further delineation between the mnemonic functions of the anterior and posterior hippocampus.

In addition to the posterior hippocampus, we also found activation in the tail of the caudate and posterior putamen that correlated with the recognition strength measure. These results are consistent with the hypothesis that the tail of the caudate and posterior putamen engage visual and motor loops that associate category representations to behavioral responses (Ashby et al., 1998; Cincotta & Seger, 2007; Seger, 2008; Seger & Cincotta, 2005). The simultaneous activation of the MTL and the striatum for the recognition strength measure is consistent with a cooperative role between the systems during categoriza-

tion (e.g., Meeter et al., 2008) but, as we discuss below, may be consistent with other explanations as well.

During stimulus presentation, anatomically distinct subregions within the MTL and the striatum were associated with recognition and entropy, providing evidence for simultaneously engaged cognitive processes during categorization decisions. Entropy was found to predict activation in the anterior hippocampus and ventral striatum, two regions that have been associated with processing novelty[1] (Daselaar et al., 2006; Strange et al., 1999).

While others have explored uncertainty processing in neurobiological category-learning research, the regions implicated in other paradigms differ from the present findings. For example, some studies have associated the ventral striatum with uncertainty processing (Grinband, Hirsch, & Ferrera, 2006), whereas others focus on the dorsal striatum (Daniel et al., 2011; Seger & Cincotta, 2005) or do not observe striatal activation related to uncertainty (Aron et al., 2004). Aside from the present study, we know of no category-learning studies relating the anterior hippocampus to uncertainty. Many previous tasks examining uncertainty in categorization have defined uncertainty in relation to the probabilistic structure of their task (Aron et al., 2004; Seger & Cincotta, 2005), in relation to the variability in the construction of their stimuli (Daniel et al., 2011), or in relation to participants' behavioral performance (Grinband et al., 2006). In contrast, by using a fully defined computational model of category learning, our approach takes into account the structure of the task, the participants' performance, and the dynamics of how psychological processes related to uncertainty change as participants learn. By combining all of this information in a psychologically plausible manner, our approach is better equipped to localize regions involved with processing uncertainty.

In terms of brain mechanisms, cluster uncertainty signals likely reflect a host of different cognitive, emotional, and motivational processes that are designed to direct learning resources to potentially significant sources of information. Indeed, one prominent model of biologically based reinforcement learning suggests that the anterior hippocampus and the ventral striatum form a circuit with midbrain dopamine neurons that strategically controls the impact of events on memory (Lisman & Grace, 2005). Events that are detected as novel, uncertain, or rewarding lead to higher midbrain dopamine release, which increases the impact that these events have on memory. Interestingly, the connection between the ventral striatum and the anterior hippocampus is thought to occur via the entorhinal cortex (Witter et al., 2006), one of the other MTL regions that correlated with the entropy measure. The notion that additional motivational processes are engaged for uncertain trials is consistent with findings that exception trials are associated with more orienting responses in a related paradigm (Davis, Love, & Maddox, 2009) and findings suggesting that uncertainty can act as an intrinsic reward or motivator during reinforcement learning (Krebs et al., 2009; Wittman et al., 2007, 2008).

As auxiliary measures and bases of comparison for our model-based results, we also separately examined predictions based on the temporal and item-based components of the original model-based measures. We examined predictions based on the general time course of the model-based measures by averaging over each measure's (entropy, recognition strength, and error) predicted item effect. To test the item-based component of the model-based measures, we employed standard condition-based analyses that predict that the difference between exception and rule-following items is constant throughout the task. Both analyses revealed subsets of the activation that were revealed by the model-based measures. The general temporal effect of the measures, averaging over items, was more suited toward revealing activation consistent with the entropy measure because the entropy measure has a strong time component. In contrast, while the item-based analyses revealed a spatial mixture of regions engaged for both recognition strength and entropy measures, the MTL results most closely matched the recognition strength measure because of the measure's strong item-based component. These auxiliary analyses are useful because they illustrate how components of the model-based measures are coded in the brain, but ultimately, the only way to recover the full pattern of results is to use the full model-based measures.

Beyond their superior ability to model the patterns of activations observed in the current task, the model-based measures we employ present a more continuous view of cognitive processes than do standard item-based methods. Many of our methods and interpretations were informed by previous findings in the neurobiological category-learning literature, as described above for the entropy measure. However, the methods employed in many of these seminal studies were not geared toward separating simultaneous processes that are present throughout a task, on every trial. For example, without a model-based approach, many previous studies had to rely on introducing different types of stimuli (e.g., probabilistic, random, deterministic: Seger & Cincotta, 2005; Seger et al., 2010) and assuming that different psychological processes would be engaged for only certain subsets of trials. The model-based perspective employed in the current study, along with the findings related to our own item-based contrasts, suggests that assuming that cognitive processes can be neatly separated between conditions is potentially misguided. It is not the case that particular cognitive processes are used during particular trials as opposed to others; it is that the extent to which cognitive processes are recruited varies depending on trial-by-trial task demands.

Multiple Systems?

There is a great deal of controversy across the category-learning literature on exactly how to characterize the roles of the different brain regions that support categorization behavior. Multiple-systems theories suggest that many of the brain regions that are engaged during category learning, such as the striatum, prefrontal cortex, and MTL, constitute functionally separate representational systems that learn categories in different ways and sometimes compete with one another. Single-system theories characterize different regions as contributing unique processes to a broader, more distributed category-learning system throughout the brain that operates on a single representational format (e.g., exemplars). For example, Palmeri and Flanery (2002) discussed how computations underlying different aspects of category learning, from object representation and selective attention to forming associations between category representations and responses, might all be

[1] It is important to note, however, that entropy is not synonymous with novelty; entropy is a measure of uncertainty, which critically depends on the model's representation of the task. Novelty contributes to uncertainty but is not synonymous with uncertainty.

mediated by separate brain regions, but only as a whole do they constitute a category-learning system.

While a multiple-systems characterization has historically been dominant in neurobiological approaches to category learning, recent trends have begun to lend support to single-system accounts. Indeed, the MTL and posterior striatum, characterized by multiple-systems theories as functionally separate representational systems, are often simultaneously activated and functionally correlated in category-learning studies (Dickerson, Li, & Delgado, 2011; Mattfeld & Stark, 2011; Sadeh, Shohamy, Levy, Reggev, & Maril, 2011; Seger et al., 2011), leading some researchers to suggest that the two regions work together to support category-learning behavior. For example, the posterior striatum may associate higher order category representations in the MTL to behavioral responses (Meeter et al., 2008; Seger & Miller, 2010).

The architecture of the RMC and our findings suggesting that the posterior striatum and MTL are both simultaneously correlated with the recognition strength measure parsimoniously lend themselves to a single-system characterization of the brain in which there is a single representational format (clusters) and in which there are multiple processes that operate on these representations. We point out, however, that dissociating single-system from multiple-systems characterizations is extremely difficult within a single experiment and that the MTL and striatum could potentially be simultaneously forming and retrieving functionally independent category representations in the present task.

In order to fully disentangle the contribution of various category-learning processes and neural regions, it will be critical to accumulate model-based imaging results over a range of experimental demands and develop richer methods for evaluating different computational models in terms of their ability to fit neural data. For example, advances in statistical methods that allow for stable estimation of the hemodynamic response on a trial-by-trial basis (e.g., Rissman, Gazzaley, & D'Esposito, 2004; Mumford, Turner, Ashby, & Poldrack, 2012), when coupled with appropriate trial spacing, may lead to better data visualization methods and allow for the rapid evaluation of a variety of multiple and single-system computational models and model-based measures in their ability explain patterns of activation within particular brain regions.

Conclusion

In conclusion, model-based fMRI presents a potentially powerful tool for understanding the neural basis of category learning. In the real world, cognitive processes associated with categorization occur simultaneously and unfold continuously over time. In contrast, standard fMRI techniques force a view by which cognitive processes are static and only occur on certain trials or in particular conditions. Model-based imaging offers category-learning researchers the opportunity to bring analysis strategies in line with theory and separate the dynamic signals recorded in fMRI tasks into distinct, simultaneously occurring cognitive processes. In the present study, by combining model-based measures of recognition strength, entropy, and error processing with fMRI, we were able to identify a number of distinct, simultaneously occurring computational processes that underlie function in subregions of the MTL and striatum. While the paths of cognitive neuroscience and computational modeling communities have occasionally diverged in the past, we believe that their futures would be stronger together.

References

Aizenstein, H. J., MacDonald, A. W., Stenger, V. A., Nebes, R. D., Larson, J. K., Ursu, S., & Carter, C. S. (2000). Complementary category learning systems identified using event-related functional MRI. *Journal of Cognitive Neuroscience, 12,* 977–987. doi:10.1162/08989290051137512

Alexander, G. E., DeLong, M. R., & Strick, P. L. (1986). Parallel organization of functionally segregated circuits linking basal ganglia and cortex. *Annual Review of Neuroscience, 9,* 357–381. doi:10.1146/annurev.ne.09.030186.002041

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review, 98,* 409–429. doi:10.1037/0033-295X.98.3.409

Aron, A. R., Shohamy, D., Clark, J., Myers, C., Gluck, M. A., & Poldrack, R. A. (2004). Human midbrain sensitivity to cognitive feedback and uncertainty during classification learning. *Journal of Neurophysiology, 92,* 1144–1152. doi:10.1152/jn.01209.2003

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review, 105,* 442–481. doi:10.1037/0033-295X.105.3.442

Ashby, F. G., & Maddox, W. T. (1993). Relations between prototype, exemplar, and decision bound models of categorization. *Journal of Mathematical Psychology, 37,* 372–400. doi:10.1006/jmps.1993.1023

Ashby, F. G., & Maddox, W. T. (2005). Human category learning. *Annual Review of Psychology, 56,* 149–178. doi:10.1146/annurev.psych.56.091103.070217

Ashby, F. G., & O'Brien, J. B. (2005). Category learning and multiple memory systems. *Trends in Cognitive Sciences, 9,* 83–89. doi:10.1016/j.tics.2004.12.003

Berns, G. S., McClure, S. M., Pagnoni, G., & Montague, P. R. (2001). Predictability modulates human brain response to reward. *The Journal of Neuroscience, 21,* 2793–2798.

Bradley, M. M. (2009). Natural selective attention: Orienting and emotion. *Psychophysiology, 46,* 1–11. doi:10.1111/j.1469-8986.2008.00702.x

Canteras, N. S., & Swanson, L. W. (1992). Projections of the ventral subiculum to the amygdala, septum, and hypothalamus: A PHA-L anterograde tracing study in the rat. *Journal of Comparative Neurology, 324,* 180–194. doi:10.1002/cne.903240204

Carr, V. A., Rissman, J., & Wagner, A. D. (2010). Imaging the human medial temporal lobe with high-resolution fMRI. *Neuron, 65,* 298–308. doi:10.1016/j.neuron.2009.12.022

Cavada, C., Company, T., Tejedor, J., Cruz-Rizzolo, R. J., & Reinoso-Suarez, F. (2000). The anatomical connections of the macaque monkey orbitofrontal cortex. A review. *Cerebral Cortex, 10,* 220–242. doi:10.1093/cercor/10.3.220

Chua, E. F., Schacter, D. L., Rand-Giovannetti, E., & Sperling, R. A. (2007). Evidence for a specific role of the anterior hippocampal region in successful associative encoding. *Hippocampus, 17,* 1071–1080. doi:10.1002/hipo.20340

Cincotta, C. M., & Seger, C. A. (2007). Dissociation between striatal regions while learning to categorize via feedback and via observation. *Journal of Cognitive Neuroscience, 19,* 249–265. doi:10.1162/jocn.2007.19.2.249

Cohen, A. L., Sanborn, A. N., & Shiffrin, R. M. (2008). Model evaluation using grouped or individual data. *Psychonomic Bulletin and Review, 15,* 692–712.

Daniel, R., Wagner, G., Koch, K., Reichenbach, J. R., Sauer, H., & Schlösser, R. G. (2011). Assessing the neural basis of uncertainty in perceptual category learning through varying levels of distortion. *Journal of Cognitive Neuroscience, 23,* 1781–1793.

Daselaar, S. M., Fleck, M. S., & Cabeza, R. (2006). Triple dissociation in the medial temporal lobes: Recollection, familiarity, and novelty. *Journal of Neurophysiology, 96,* 1902–1911. doi:10.1152/jn.01029.2005

Davis, T., Love, B. C., & Maddox, T. (2009). Anticipatory emotions in

decision tasks: Covert markers of value or attentional processes? *Cognition, 112,* 195–200. doi:10.1016/j.cognition.2009.04.002

Davis, T., Love, B. C., & Preston, A. R. (2011). Learning the exception to the rule: Model-based fMRI reveals specialized representations for surprising category members. *Cerebral Cortex, 22,* 260–273. doi:10.1093/cercor/bhr036

Daw, N. D. (2011). Trial-by-trial data analysis using computational models. In M. R. Delgado, E. A. Phelps, & T. W. Robbins (Eds.), *Decision making, affect, and learning: Attention and performance XXIII* (Vol. 23, pp. 3–38). Oxford, England: Oxford University Press.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron, 69,* 1204–1215. doi:10.1016/j.neuron.2011.02.027

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature, 441,* 876–879. doi:10.1038/nature04766

Dickerson, K. C., Li, J., & Delgado, M. R. (2011). Parallel contributions of distinct human memory systems during probabilistic learning. *NeuroImage, 55,* 266–276. doi:10.1016/j.neuroimage.2010.10.080

Eichenbaum, H., Yonelinas, A. P., & Ranganath, C. (2007). The medial temporal lobe and recognition memory. *Annual Review of Neuroscience, 30,* 123–152. doi:10.1146/annurev.neuro.30.051606.094328

Fanselow, M. S., & Dong, H. W. (2010). Are the dorsal and ventral hippocampus functionally distinct structures? *Neuron, 65,* 7–19. doi:10.1016/j.neuron.2009.11.031

Glass, B. D., Chotibut, T., Pacheco, J., Schnyer, D. M., & Maddox, W. T. (2012). Normal aging and the dissociable prototype learning systems. *Psychology and Aging, 27,* 120–128.

Grinband, J., Hirsch, J., & Ferrera, V. P. (2006). A neural representation of categorization uncertainty in the human brain. *Neuron, 49,* 757–763. doi:10.1016/j.neuron.2006.01.032

Henson, R. N. A., Cansino, S., Herron, J. E., Robb, W. G. K., & Rugg, M. D. (2003). A familiarity signal in human anterior medial temporal cortex? *Hippocampus, 13,* 301–304.

Jones, M., & Love, B. C. (2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and Brain Sciences, 34,* 169–231.

Kircher, T., Weis, S., Leube, D., Freymann, K., Erb, M., Jessen, F., . . . Krach, S. (2008). Anterior hippocampus orchestrates successful encoding and retrieval of non-relational memory: An event-related fMRI study. *European Archives of Psychiatry and Clinical Neuroscience, 258,* 363–372. doi:10.1007/s00406-008-0805-z

Kirwan, C. B., Jones, C. K., Miller, M. I., & Stark, C. E. (2007). High-resolution fMRI investigation of the medial temporal lobe. *Human Brain Mapping, 28,* 959–966. doi:10.1002/hbm.20331

Krebs, R. M., Schott, B. H., Schutze, H., & Duzel, E. (2009). The novelty exploration bonus and its attentional modulation. *Neuropsychologia, 47,* 2272–2281. doi:10.1016/j.neuropsychologia.2009.01.015

Lisman, J. E., & Grace, A. A. (2005). The hippocampal-VTA loop: Controlling the entry of information into long-term memory. *Neuron, 46,* 703–713.

Love, B. C., & Gureckis, T. M. (2007). Models in search of a brain. *Cognitive, Affective, & Behavioral Neuroscience, 7,* 90–108. doi:10.3758/CABN.7.2.90

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review, 111,* 309–332. doi:10.1037/0033-295X.111.2.309

Maddox, W. T., & Ashby, F. G. (2004). Dissociating explicit and procedural-learning based systems of perceptual category learning. *Behavioural Processes, 66,* 309–332. doi:10.1016/j.beproc.2004.03.011

Mattfeld, A. T., & Stark, C. E. (2011). Striatal and medial temporal lobe functional interactions during visuomotor associative learning. *Cerebral Cortex, 21,* 647–658. doi:10.1093/cercor/bhq144

Meeter, M., Radics, G., Myers, C., Gluck, M., & Hopkins, R. (2008). Probabilistic categorization: How do normal participants and amnesic patients do it? *Neuroscience & Biobehavioral Reviews, 32,* 237–248. doi:10.1016/j.neubiorev.2007.11.001

Moser, M. B., & Moser, E. I. (1998). Functional differentiation in the hippocampus. *Hippocampus, 8,* 608–619. doi:10.1002/(SICI)1098-1063(1998)8:6<608::AID-HIPO3>3.0.CO;2-7

Mumford, J. A., Turner, B. O., Ashby, F. G., & Poldrack, R. A. (2012). Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. *NeuroImage, 59,* 2636–2643. doi:10.1016/j.neuroimage.2011.08.076

Nomura, E. M., Maddox, W. T., Filoteo, J. V., Ing, A. D., Gitelman, D. R., Parrish, T. B., . . . Reber, P. J. (2007). Neural correlates of rule-based and information-integration visual category learning. *Cerebral Cortex, 17,* 37–43. doi:10.1093/cercor/bhj122

Nosofsky, R. M., Gluck, M. A., Palmeri, T. J., & McKinley, S. C. (1994). Comparing models of rule-based classification learning: A replication and extension of Shepard, Hovland, and Jenkins (1961). *Memory & Cognition, 22,* 352–369. doi:10.3758/BF03200862

O'Doherty, J. P., Dayan, P., Friston, K., Critchley, H., & Dolan, R. J. (2003). Temporal difference models and reward-related learning in the human brain. *Neuron, 38,* 329–337. doi:10.1016/S0896-6273(03)00169-7

O'Doherty, J. P., Hampton, A., & Kim, H. (2007). Model-based fMRI and its application to reward learning and decision making. *Annals of the New York Academy of Sciences, 1104,* 35–53. doi:10.1196/annals.1390.022

Palmeri, T. J., & Flanery, M. A. (2002). *Memory systems and perceptual categorization: Psychology of learning and motivation* (Vol. 41, pp. 141–189). New York, NY: Academic Press. Retrieved from http://www.sciencedirect.com/science/article/pii/S0079742102800068

Palmeri, T. J., & Nosofsky, R. M. (1995). Recognition memory for exceptions to the category rule. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 548–568. doi:10.1037/0278-7393.21.3.548

Pickering, A. D. (1997). New approaches to the study of amnesic patients: What can a neurofunctional philosophy and neural network methods offer? *Memory, 5,* 255–300. doi:10.1080/741941146

Poldrack, R. A. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Sciences, 10,* 59–63. doi:10.1016/j.tics.2005.12.004

Poldrack, R. A., & Packard, M. G. (2003). Competition among multiple memory systems: Converging evidence from animal and human brain studies. *Neuropsychologia, 41,* 245–251. doi:10.1016/S0028-3932(02)00157-4

Preston, A. R., & Wagner, A. D. (2007). The medial temporal lobe and memory. In R. P. Kesner & J. L. Martinez (Eds.), *Neurobiology of learning and memory* (pp. 305–337). Burlington, MA: Academic Press. doi:10.1016/B978-012372540-0/50010-8

Prince, S. E., Daselaar, S. M., & Cabeza, R. (2005). Neural correlates of relational memory: Successful encoding and retrieval of semantic and perceptual associations. *The Journal of Neuroscience, 25,* 1203–1210. doi:10.1523/JNEUROSCI.2540-04.2005

Reber, P. J., Gitelman, D. R., Parish, T. B., & Mesulam, M. M. (2003). Dissociating explicit and implicit category knowledge with fMRI. *Journal of Cognitive Neuroscience, 15,* 574–583. doi:10.1162/089892903321662958

Rissman, J., Gazzaley, A., & D'Esposito, M. (2004). Measuring functional connectivity during distinct stages of a cognitive task. *NeuroImage, 23,* 752–763. doi:10.1016/j.neuroimage.2004.06.035

Sadeh, T., Shohamy, D., Levy, D. R., Reggev, N., & Maril, A. (2011). Cooperation between the hippocampus and the striatum during episodic encoding. *Journal of Cognitive Neuroscience, 23,* 1597–1608.

Sakamoto, Y., & Love, B. C. (2004). Schematic influences on category

learning and recognition memory. *Journal of Experimental Psychology: General, 133,* 534–553. doi:10.1037/0096-3445.133.4.534

Sakamoto, Y., & Love, B. C. (2006). Vancouver, Toronto, Montreal, Austin: Enhanced oddball memory through differentiation, not isolation. *Psychonomic Bulletin & Review, 13,* 474–479. doi:10.3758/BF03193872

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review, 117,* 1144. doi:10.1037/a0020511

Schott, B. H., Sellner, D. B., Lauer, C. J., Habib, R., Frey, J. U., Guderian, S., . . . Düzel, E. (2004). Activation of midbrain structures by associative novelty and the formation of explicit memory in humans. *Learning & Memory, 11,* 383–387. doi:10.1101/lm.75004

Schultz, W., Dayan, P., & Montague, P. R. (1997, March 14). A neural substrate of prediction and reward. *Science, 275,* 1593–1599. doi:10.1126/science.275.5306.1593

Seger, C. A. (2008). How do the basal ganglia contribute to categorization? Their roles in generalization, response selection, and learning via feedback. *Neuroscience & Biobehavioral Reviews, 32,* 265–278. doi:16/j.neubiorev.2007.07.010

Seger, C. A., & Cincotta, C. M. (2005). The roles of the caudate nucleus in human classification learning. *The Journal of Neuroscience, 25,* 2941. doi:10.1523/JNEUROSCI.3401-04.2005

Seger, C. A., & Cincotta, C. M. (2006). Dynamics of frontal, striatal, and hippocampal systems during rule learning. *Cerebral Cortex, 16,* 1546–1555. doi:10.1093/cercor/bhj092

Seger, C. A., Dennison, C. M., Lopez-Paniagua, D., Peterson, E. J., & Roark, A. A. (2011). Dissociating hippocampal and basal ganglia contributions to category learning using stimulus novelty and subjective judgments. *NeuroImage, 55,* 1739–1753.

Seger, C. A., & Miller, E. K. (2010). Category learning in the brain. *Annual Review of Neuroscience, 33,* 203–219. doi:10.1146/annurev.neuro.051508.135546

Seger, C. A., Peterson, E. J., Cincotta, C. M., Lopez-Paniagua, D., & Anderson, C. W. (2010). Dissociating the contributions of independent corticostriatal systems to visual categorization learning through the use of reinforcement learning modeling and Granger causality modeling. *NeuroImage, 50,* 644–656. doi:10.1016/j.neuroimage.2009.11.083

Shohamy, D., Myers, C., Kalanithi, J., & Gluck, M. (2008). Basal ganglia and dopamine contributions to probabilistic category learning. *Neuroscience & Biobehavioral Reviews, 32,* 219–236. doi:10.1016/j.neubiorev.2007.07.008

Smith, E. E., & Grossman, M. (2008). Multiple systems of category learning. *Neuroscience and Biobehavioral Reviews, 32,* 249–264. doi:10.1016/j.neubiorev.2007.07.009

Sokolov, E. N. (1966). Orienting reflex as information regulator. In A. Leontyev, A. Luria, & A. Smirnov (Eds.), *Psychological research in the USSR* (pp. 334–360). Moscow, Russia: Progress.

Sokolov, E. N., Spinks, J. A., Naatanen, R., & Lyytinen, H. (2002). *The orienting response in information processing.* London, England: Erlbaum.

Squire, L. R. (1992). Memory and the hippocampus: A synthesis from findings with rats, monkeys, and humans. *Psychological Review, 99,* 195–231.

Stark, C. E. L., & Squire, L. R. (2001). When zero is not zero: The problem of ambiguous baseline conditions in fMRI. *Proceedings of the National Academy of Sciences, USA, 98,* 12760–12766. doi:10.1073/pnas.221462998

Strange, B. A., Duggins, A., Penny, W., Dolan, R., & Friston, K. (2005). Information theory, novelty and hippocampal responses: Unpredicted or unpredictable? *Neural Networks, 18,* 225–230. doi:10.1016/j.neunet.2004.12.004

Strange, B. A., & Dolan, R. J. (2006). Anterior medial temporal lobe in human cognition: Memory for fear and the unexpected. *Cognitive Neuropsychiatry, 11,* 198–218. doi:10.1080/13546800500305096

Strange, B. A., Fletcher, P. C., Henson, R. N. A., Friston, K. J., & Dolan, R. J. (1999). Segregating the functions of human hippocampus. *Proceedings of the National Academy of Sciences, USA, 96,* 4034–4039. doi:10.1073/pnas.96.7.4034

Strange, B. A., Hurlemann, R., Duggins, A., Heinze, H. J., & Dolan, R. J. (2005). Dissociating intentional learning from relative novelty responses in the medial temporal lobe. *NeuroImage, 25,* 51–62. doi:10.1016/j.neuroimage.2004.12.014

Vercauteren, T., Pennec, X., Perchant, A., & Ayache, N. (2009). Diffeomorphic demons: Efficient non-parametric image registration. *NeuroImage, 45*(1, Suppl. 1), S61–S72. doi:10.1016/j.neuroimage.2008.10.040

Witter, M. P., Wouterlood, F. G., Naber, P. A., & Van Haeften, T. (2006). Anatomical organization of the parahippocampal-hippocampal network. *Annals of the New York Academy of Sciences, 911,* 1–24. doi:10.1111/j.1749-6632.2000.tb06716.x

Wittmann, B. C., Bunzeck, N., Dolan, R. J., & Duzel, E. (2007). Anticipation of novelty recruits reward system and hippocampus while promoting recollection. *NeuroImage, 38,* 194–202. doi:10.1016/j.neuroimage.2007.06.038

Wittmann, B. C., Daw, N. D., Seymour, B., & Dolan, R. J. (2008). Striatal activity underlies novelty-based choice in humans. *Neuron, 58,* 967–973. doi:10.1016/j.neuron.2008.04.027

Yassa, M. A., & Stark, C. E. (2009). A quantitative evaluation of cross-participant registration techniques for MRI studies of the medial temporal lobe. *NeuroImage, 44,* 319–327. doi:10.1016/j.neuroimage.2008.09.016

Zeineh, M. M., Engel, S. A., Thompson, P. M., & Bookheimer, S. Y. (2003). Dynamics of the hippocampus during encoding and retrieval of face-name pairs. *Science, 299,* 577. doi:10.1126/science.1077775

Zeithamova, D., Maddox, W. T., & Schnyer, D. M. (2008). Dissociable prototype learning systems: Evidence from brain imaging and behavior. *The Journal of Neuroscience, 28,* 13194–13201. doi:10.1523/JNEUROSCI.2915-08.2008

(*Appendix follows*)

Appendix

Modeling Procedure

The RMC models category-learning behavior by estimating, on each trial, the probability that a stimulus will belong to a given category based on its observed features and the features associated with cluster representations stored in memory. A cluster is a grouping of stimuli that code the features and category labels associated with stimuli in the task and can contain anywhere from a single stimulus to all of the stimuli observed in a task. The RMC assumes that one cluster was responsible for generating the observed stimulus and generates an overall probability of the category label by aggregating the category label predictions over clusters.

Formally, the probability that a stimulus i belongs to category j given its observed feature structure F and the features of k stored clusters is given by

$$P_i(j \mid F) = \sum_k P(k \mid F) P_i(j \mid k), \tag{1}$$

where $P(k \mid F)$ is the probability of the stimulus i coming from cluster k, given the observed features, and $P_i(j \mid k)$ is the probability of i having the category label j given the cluster. Each $P(k \mid F)$ is a posterior probability that combines the prior probability of cluster k and the likelihood of the stimulus features F, given cluster k:

$$P(k \mid F) = \frac{P(k)P(F \mid k)}{\sum_k P(k)P(F \mid k)}, \tag{2}$$

where $P(k)$ is the prior probability of the cluster over possible partitionings of the stimulus space. The denominator in Equation 2 normalizes the cluster probabilities to sum to one.

$P(k)$ is given by

$$P(k) = \begin{cases} \dfrac{cn_k}{(1-c)+cn} & \text{if } n_k > 0 (k \text{ is old}) \\ \dfrac{(1-c)}{(1-c)+cn} & \text{if } n_k = 0 (k \text{ is new}) \end{cases}, \tag{3}$$

where c is a coupling parameter that controls how likely objects are to be stored in the same cluster, n is the total number of stimuli presented so far, and n_k is the number of stimuli assigned to cluster k. $P(k)$ has a modulating effect on the posterior probabilities such that as the number of previous items stored in cluster k increases, the RMC judges the cluster as more probable to have generated a stimulus.

$P(F \mid k)$, in Equation 2, is the likelihood of the stimulus features F given cluster k, given by

$$P(F \mid k) = \prod_i P_i(j \mid k), \tag{4}$$

where js are the values on each i stimulus dimension in the feature set F. Equation 4 assumes that features within a cluster are independent of one another, so the individual contribution of each feature dimension can be multiplied to obtain the likelihood. In discrete dimension cases, like the present application, the contribution of matches–mismatches on an individual feature dimension i is given by

$$P_i(j \mid k) = \frac{n_j + \beta_i}{n_k + 2\beta_i}, \tag{5}$$

where n_j is the number of stimuli in cluster k with the same value as the to be classified stimulus on dimension i, n_k is the number of stimuli in cluster k (assuming all stimuli have a value on dimension i), and the β_is are symmetric beta priors that control the weight of matches (mismatches) on dimension i in the likelihood computation (i.e., dimensional salience). The β_is are constrained to be in the range (0,1), with lower values indicating higher salience. In the present application, we fit a separate beta prior for the rule-relevant dimension β_r and the last three other dimensions β_o to reflect the instructions given to participants to focus on the rule-relevant dimension. The beta prior for the category label dimension was fixed at .01, which reflects the assumption that stimuli with different category labels will belong to different clusters. All priors are symmetric such that for a given stimulus dimension/category label, matches and mismatches are weighted the same regardless of the particular feature instantiation (e.g., whether the category label is 1 or 2).

In the present application, the probability with which the RMC responds (resp) with category m given stimulus i is scaled using a probabilistic choice rule (Nosofsky, Gluck, Palmeri, & McKinley, 1994):

$$P(\text{resp} = m) = \frac{P_i(j = m \mid F)^\gamma}{\sum_{j=1}^{J} P_i(j \mid F)^\gamma}, \tag{6}$$

where γ is a decision parameter that scales the probabilities such that for $\gamma = 1$ (see also, Ashby & Maddox, 1993), the model probability matches and responds proportionally to the probability of the category label, given i calculated in Equation 1, and as γ approaches infinity, the responses become more deterministic, such that the model always chooses the most probable category. After a response is made and feedback is delivered, the cluster probabilities are recalculated, taking into account the now observed category label; as a result, the stimulus is stored in a given cluster with a probability proportional to the cluster's likelihood, given the stimulus. The probabilistic assignment of a stimulus to a

(*Appendix continues*)

cluster implements a single particle filter version of the RMC (Sanborn, Griffiths, & Navarro, 2010) and is a departure from the original RMC, which assigns stimuli to clusters deterministically using a local maximum aposterior probability (MAP) algorithm.

Values for β_r, β_o, c, and γ were estimated by fitting the model, using a combination of grid search and simplex algorithms, to participants' average performance for each trial type over each scanning run, minimizing the sum of squared error (see Figure 5). The best fitting parameter values were $\beta_r = .354$; $\beta_o = 1.0$; $c = 0.347$; and $\gamma = 1.556$. With these parameters, the model was run to generate the entropy, recognition, and error measures used in the model-based fMRI analyses.

Model-based measures for fMRI analysis were computed from the fitted version of the RMC as follows. Recognition strength R (see Figure 3) indexes the probability of stimulus i (prefeedback), given the model's current cluster representations:

$$R = \sum_k P(k)\, P(F \mid k). \tag{7}$$

Recognition strength is equivalent to the denominator in Equation 2 and is computed during the stage at which the RMC is comparing an observed stimulus to the current clusters that are stored in its memory.

Entropy E (see Figure 3) for stimulus i is defined using the prefeedback cluster probabilities from Equation 2:

$$E = -\sum_k P(k \mid F) \log_2[P(k \mid F)]. \tag{8}$$

Entropy is computed after the model has compared a presented stimulus with each of the clusters stored in memory and computed a probability of the stimulus belonging to each cluster.

The final measure, error, is the probability that the model's categorization choice (see Equation 6) is incorrect and is computed after the decision stage.

Received June 1, 2011
Revision received December 8, 2011
Accepted December 9, 2011 ■