OXFORD

Cerebral Cortex, 2014, 1–15

doi: 10.1093/cercor/bhu284
Original Article

ORIGINAL ARTICLE

Neural Differentiation Tracks Improved Recall of Competing Memories Following Interleaved Study and Retrieval Practice

J. C. Hulbert[1] and K. A. Norman[1,2]

[1]Princeton Neuroscience Institute, and [2]Department of Psychology, Princeton University, Princeton, NJ 08544, USA

Address correspondence to Kenneth A. Norman, Princeton Neuroscience Institute, Princeton University, Washington Road, Princeton, NJ 08544, USA. Email: knorman@princeton.edu

Abstract

Selective retrieval of overlapping memories can generate competition. How does the brain adaptively resolve this competition? One possibility is that competing memories are inhibited; in support of this view, numerous studies have found that selective retrieval leads to forgetting of memories that are related to the just-retrieved memory. However, this retrieval-induced forgetting (RIF) effect can be eliminated or even reversed if participants are given opportunities to restudy the materials between retrieval attempts. Here, we outline an explanation for such a reversal, rooted in a neural network model of RIF that predicts representational differentiation when restudy is interleaved with selective retrieval. To test this hypothesis, we measured changes in pattern similarity of the BOLD fMRI signal elicited by related memories after undergoing interleaved competitive retrieval and restudy. Reduced pattern similarity within the hippocampus positively correlated with retrieval-induced facilitation of competing memories. This result is consistent with an adaptive differentiation process that allows individuals to learn to distinguish between once-confusable memories.

Key words: differentiation, hippocampus, memory, neural network model, pattern similarity, retrieval-induced forgetting

Introduction

Selective retrieval of a target memory reliably improves the later accessibility of that memory (Roediger and Butler 2011), but memory for related items is sometimes impaired. This latter phenomenon, termed retrieval-induced forgetting (RIF), has been observed under a wide range of conditions (Anderson et al. 1994; Anderson 2003). RIF is beneficial so long as the weakened competitor remains irrelevant. However, items that are irrelevant in one situation can become relevant later. When this happens, RIF can be harmful to future retrieval success. A central question for theories of learning is how the brain mitigates these potentially harmful effects of RIF (MacLeod and Hulbert 2011).

In this paper, we describe a potential solution to this problem: In situations where participants are allowed to restudy the previously irrelevant item, the brain may differentiate the neural representation of this memory from other, competing memories, thereby reducing competition on subsequent retrieval attempts and improving recall of the full set of (previously competitive) memories. We motivate this hypothesis using our previous neural network modeling work, and we provide novel empirical support for this hypothesis using fMRI.

The typical retrieval-practice paradigm employed to investigate RIF begins with a study phase, in which participants learn category–exemplar pairings. In the retrieval-practice phase that follows, half of the exemplars from half of the categories are cued for retrieval. After multiple rounds of retrieval practice, a final cued-recall test is administered for all of the originally studied items. Although recall of items that underwent retrieval practice (Rp+) is typically facilitated compared with recall of items

from unpracticed baseline categories (Nrp), recall of the unpracticed competitors from practiced categories (Rp–) is characteristically impaired relative to Nrp (Anderson 2003). This latter finding indicates RIF.

Certain variations of the above-mentioned procedure have been found to reverse RIF. Notably, interleaving opportunities to restudy Rp–, Nrp, and Rp+ items between retrieval practice attempts leads to better recall of Rp– items, relative to Nrp items (Storm et al. 2008). This result is puzzling at first: Rp– and Nrp items were studied the same number of times, in the same way; they only differed in that participants practiced generating exemplars related to the Rp– items. In the standard retrieval-practice paradigm, this leads to impaired recall of Rp– items—why did the addition of restudy trials in the Storm et al. (2008) study lead to better recall of Rp– items than Nrp items?

We hypothesize that this improvement in recall—"reverse RIF" or *revRIF*—occurs because interleaved retrieval practice and restudy lead to differentiation of the neural representations of items from practiced categories. This hypothesis stems from our prior neural network modeling work exploring competition-dependent learning and RIF (Norman et al. 2006, 2007). Note that neither of these prior modeling papers directly addressed how differentiation could occur during a retrieval-practice paradigm: Norman et al.'s (2006) paper looked at competition-dependent learning and differentiation in general terms, but did not simulate a retrieval-practice paradigm; and Norman et al.'s (2007) model looked at cortico-hippocampal interactions during retrieval practice but used a simplified architecture that did not allow for differentiation. The predictions described here were derived by extrapolating outward from the basic principles outlined in the 2 papers.

Our predictions are outlined in Figure 1. When memories compete on retrieval-practice trials, the "winning" item (i.e., the item with the strongest activation; typically the Rp+ item) has its connections strengthened. Crucially, if related Rp– items begin to interfere during retrieval practice, connections between the (weakly active) features of these competing memories and other active features are weakened. As shown in Figure 1c, this weakening process results in the "shearing away" of the Rp – memory from features shared with the Rp+ memory, as well as a more general weakening of interconnectivity between the features of the Rp– representation. Because of this decreased

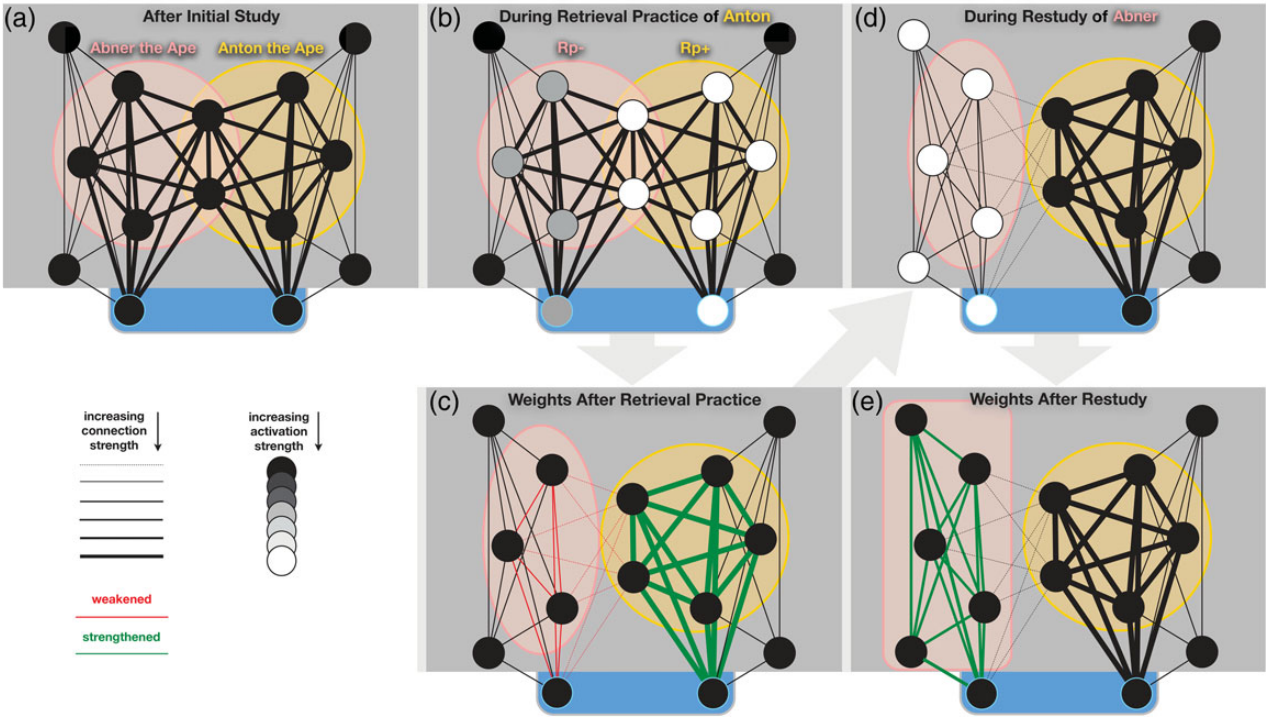

Figure 1. Predictions of our neural network model. (*a*) Here, we depict 2 partially overlapping memory representations—Abner the Ape (large pink circle) and Anton the Ape (large orange circle)—after their initial study. Inhibitory interneurons (not shown) enforce an approximate "set point" on the amount of neural activity in the network (O'Reilly and Munakata 2000). For the purpose of this diagram, we assume that the 5 units receiving the most excitatory input within the upper part of the network (i.e., within the large gray rectangle) are allowed to be strongly active; additional units are allowed to be weakly active. We also assume that there exist sensory units (in the blue zone) associated with each animal that represent the animal's sensory features. These sensory units are activated whenever the relevant animal is presented; activation spreads from the sensory units to the rest of the animal's representation. (*b*) During selective retrieval practice of Anton (the Rp+ item), the units associated with Anton's representation are strongly activated in memory. Because of the overlap, Abner's (the Rp– item's) representation is also partially activated. (*c*) In the model, strong activation of Anton's representation triggers further strengthening of connections between these units. Also, weak activation of Abner's representation triggers weakening of connections into Abner's units (from other active units). This competition-dependent weakening of connections (highlighted in red) leaves the Abner representation in a degraded (less-fully-interconnected) state. The decrease in interconnectivity is assumed to make the Abner memory harder to recall if memory were tested at this point in time (i.e., RIF). (*d*) When Abner is restudied, the 2 units that formerly were shared between Abner and Anton are no longer in the "top 5 most excited units" because of the weakening that took place earlier. As such, they drop out of the representation of Abner. Other units that were not previously activated then take their place via spreading activation, leaving Abner with a full complement of 5 activated units. (*e*) Learning in the model strengthens the connections (highlighted in green) between these new units and the other Abner units. In the final state of the network, the Abner representation is nearly as strong (i.e., its features are roughly as densely interconnected) as it was at the outset, and now it overlaps less with the representation of Anton. This neural differentiation will result in less competition at retrieval, which should boost recall of Anton above baseline (i.e., revRIF).

interconnectivity, the Rp– pattern is a less "attractive" state of the network (i.e., it is more difficult to coherently reactivate the pattern), so the model predicts decreased recall of the Rp– pattern (i.e., RIF) if memory were tested at this point in the experiment.

If the Rp– item is subsequently restudied (Fig. 1*d*), the unique features of the Rp– item will be activated, but activation may not spread to the features that were formerly shared by the Rp– and Rp+ item (because of the weakening of connections that occurred during retrieval practice). Crucially, Norman et al.'s (2006) model includes inhibitory interneurons (not shown in Fig. 1) that enforce an approximate "set point" on the amount of excitatory neural activity in the network (O'Reilly and Munakata 2000). If the Rp– item fails to reactivate some of the previously shared features, then this "set point" property will result in other features (not shared with the Rp+ item) joining the representation, and the network will strengthen the connections between the full set of activated features. In so doing, the previously degraded representation recovers its initial integrity (insofar as it now has a full complement of tightly integrated features), which should boost recall. The process of swapping out shared for unshared features results in an overall reduction of overlap between the representations of the Rp+ and Rp– items. This neural differentiation leads to reduced competition between these items and a further improvement in recall on the final test (above and beyond what is gained from strengthening the Rp– memory), thereby accounting for the finding of revRIF.

In the present study, we set out to test this account by relating a measure of neural differentiation (after interleaved retrieval practice and restudy) to individual differences in the expression of revRIF. If revRIF is driven by neural differentiation, then greater levels of differentiation should be associated with greater levels of revRIF. To this end, we leveraged a multi-voxel pattern analysis technique that allowed us to compare how neural similarity structure changed as a consequence of our behavioral manipulation (Kriegeskorte et al. 2008). A priori, we expected that the hippocampus might play a key role in this process, given its noted involvement in learning novel episodic associations (see, e.g., McClelland et al. 1995). Thus, we focus our analyses on hippocampal activity obtained using an fMRI sequence optimized to recover signal from the medial temporal lobe.

Materials and Methods

Participants

Thirty-five right-handed, fluent English speakers recruited from the Princeton University community participated in the main neuroimaging experiment. Following a protocol approved by Princeton University's Institutional Review Board, these participants were compensated at a rate of $20/h. Of these individuals, 2 participants were excluded from analysis due to excessive head motion and another 2 participants were excluded on account of their chance performance on the baseline parity task (see General Procedure). Additionally, 7 of the first 20 participants were excluded because they admitted, on a post-experiment questionnaire, to violating the instructions by covertly "quizzing themselves" during periods when they were supposed to be studying the image–name pairs (without attempting retrieval). The high incidence of this behavior led us to adjust the instructions to further underscore the importance of only attempting retrieval during designated intervals (see General Procedure). The 15 participants who received the elaborated instructions all reported that they were able to successfully follow the instructions. As such, data from these 15 participants were pooled together with the 9 usable participants from the first batch of 20, leaving a final sample of 24 participants, ranging in age from 18 to 30 years (mean = 20.83; SD = 2.73; 16 female).

An additional 18 individuals (age range: 19–30 years [mean = 21.56; SD = 3.03]; 9 female; 1 left handed) were recruited to participate in a behavioral control study following the same general protocol, described later. They were compensated at a rate of $12/h.

Materials

Eight pictorial exemplars from each of 6 different mammalian categories (apes, elephants, giraffes, lions, otters, and pigs) were sourced from the Internet, totaling 48 images (see Fig. 2). We selected exemplars that would be visually confusable (within-category) to the untrained eye but varied across multiple dimensions (e.g., facial expressions and markings). After digitally removing the original background, cropping, and scaling the images to fill as much of the 500 × 500 pixel white canvas as possible, the grayscale images underwent further preprocessing to control low-level image properties using the SHINE toolbox (Willenbockel et al. 2010). Specifically, we sought to equate the luminance histograms across the images. Image-specific scrambled background fills were generated by retaining each image's original power spectrum but adding a random phase.

Retrieval status of the 6 categories was counterbalanced across participants, such that for any 1 participant, 3 categories would be subject to competitive retrieval practice (Rp) and 3 would not (Nrp). There were 6 levels of this counterbalancing factor. Within each of the 3 Rp categories, 4 exemplars were randomly assigned to the Rp+ condition and the other 4 to the Rp– condition. An analogous randomization process was performed on the items in the 3 Nrp categories, with items assigned to either Nrp_a or Nrp_b conditions, though this particular distinction only became relevant at the analysis stage. Randomization was conducted independently for each participant.

A matching number of low-frequency two-syllable names were selected, with the additional constraint that the first letter of each name corresponded to the first letter of the name of its superordinate category (e.g., "Abner the Ape" and "Egan the Elephant"). Names were randomly paired with category-appropriate images for each participant.

General Procedure

The experiment began by informing participants that they would be introduced to various "animal characters," such as Hartley the Horse (a filler item). Participants were then familiarized with the entire set of proper names that would appear in the experiment. To this end, a Matlab script, utilizing Psychophysics Toolbox extensions (Brainard 1997), presented the printed form of the 48 names in isolation (e.g., "Abner"). Presentation of the proper names occurred in a block-randomized order, such that every block of 4 items contained a randomly selected representative from each of the 4 conditions (Rp+, Rp–, Nrp_a, and Nrp_b). The ordering of the 4 conditions within any given block was also randomized. Each printed name was presented centrally for 500 ms before it was replaced with "???" for 750 ms. Participants were instructed to say the just-presented name aloud when the question marks appeared. This process was repeated, in a new block-randomized order, so each name was practiced exactly twice.

A visual outline of the remaining parts of the experiment can be found in Figure 3. MRI acquisition commenced with an initial study phase (S1). In this phase, participants encountered the

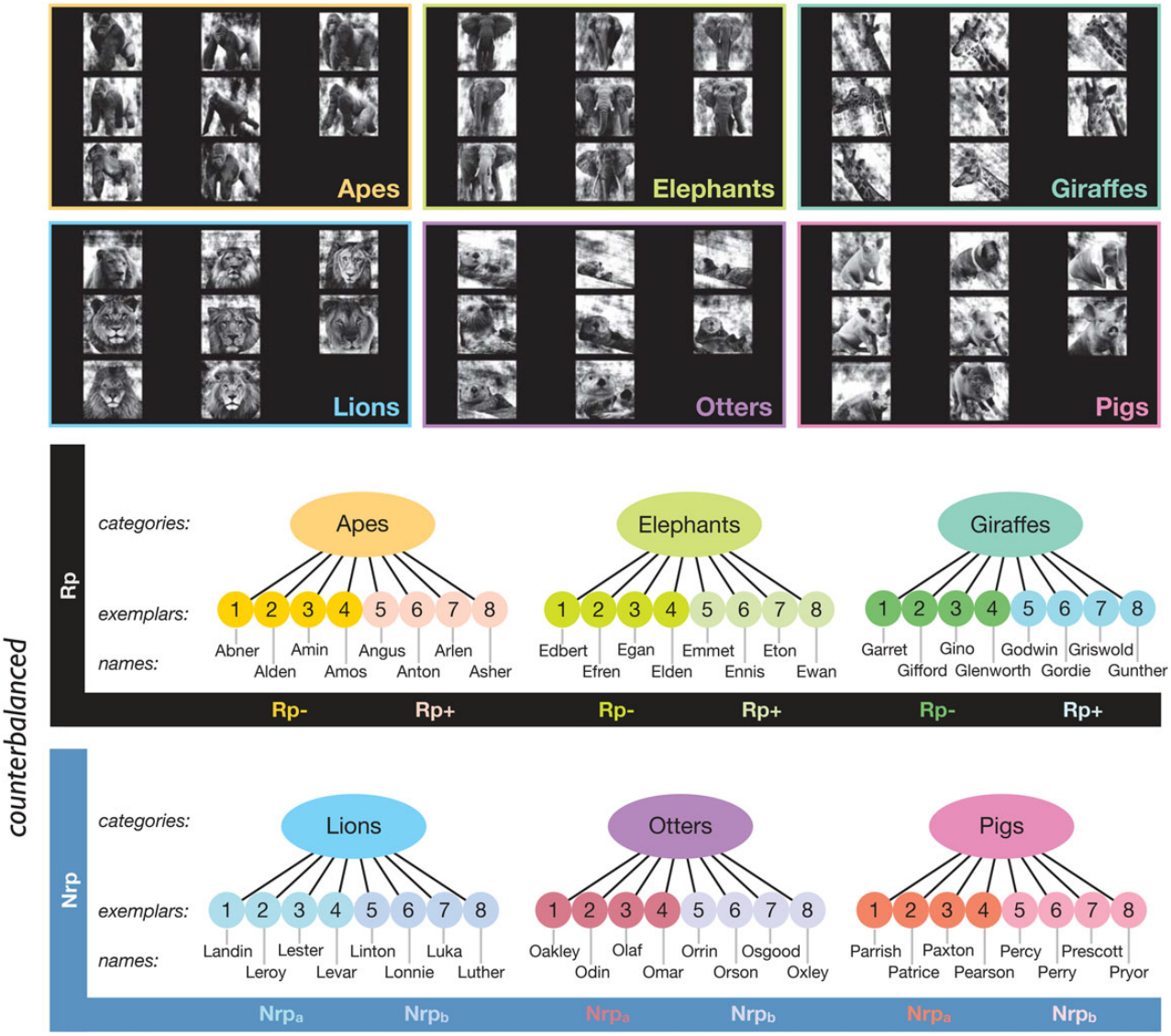

Figure 2. Stimuli. Our 6 animal categories, each containing 8 pictorial exemplars, were assigned to either the Rp (retrieval practice) or Nrp (baseline) condition in a counterbalanced fashion across participants. Proper names beginning with the first letter of the relevant category were randomly assigned to the individual pictures, which were also randomly divided into an A-set and a B-set. Image–name pairs randomly assigned to the A-set of Rp categories populated the Rp– condition, and the other half of the items populated the Rp+ condition. Items in the Nrp categories were also randomly split between the Nrp_a and Nrp_b conditions.

animal images for the first time, each paired with a name they had practiced previously. For example, a participant might have seen a picture of 1 of the 8 apes with the text, "Angus the Ape" printed below it in white, against a black background. The purpose of this phase was to record patterns of fMRI activity elicited by the items, so we could obtain an initial measurement (prior to the retrieval-practice/restudy phase) of the similarity structure of these patterns.

Each item was presented once during the S1 phase, in a randomized order. Participants were encouraged to attend to and study each image–name pairing when it appeared on the screen for 2 s, as they would be tested for all the animals' names at the end of the scan session. Participants were also asked to spend the active 6-s inter-trial interval (ITI) entirely focused on completing —as quickly and as accurately as possible—a series of parity judgments. This task prompted participants to indicate whether the sum of 2 centrally presented positive integers (1–9) was even or odd by pressing a button with their right index finger or middle finger, respectively. They had 1 s to respond, and then, regardless of whether they responded, a fixation cross was presented for 100 ms, followed by a new addition problem. This cycle continued until the 6-s ITI elapsed, or there was insufficient time to perform another parity judgment at its maximum allowable duration, in which case the remainder of the ITI was filled with fixation.

Four functional runs of retrieval practice/restudy followed the initial study phase, with each of the learned image–name pairs presented twice in every run. Four presentation schedules—1 for each run—were generated using OptSeq2 (Dale 1999), which also determined the duration of the ITI (jitter range: 1–7.5 s; mean = 2.86; SD = 1.33 across runs). While all participants shared the same 4 abstract schedules, their ordering (runs 1–4) and the assignment of items to the condition placeholders within each run was randomized for every participant. Items from a given

Figure 3. Behavioral paradigm. Initially, participants studied each one of the animal–name pairs in isolation once, with a parity judgment baseline task separating each presentation. Participants were then given the opportunity to retrieve the name of each exemplar out loud (yellow prompt) and restudy the correct pairing (blue prompt) twice in each of the following 4 rounds of interleaved retrieval practice and restudy. We manipulated the order of these 2 constituent tasks, such that Rp+ items were subjected to Competitive retrieval practice attempts followed by feedback (i.e., the yellow preceded the blue prompt, as in the case of Abner). Items in the Rp– and $Nrp_{a/b}$ conditions (e.g., Odin), in contrast, underwent retrieval practice only after first receiving the correct answer in blue, making the retrieval task relatively Non-Competitive. A final opportunity to study all of the intact pairings once, without retrieval, was provided in the same manner as the initial study period, before the final cued-recall test. Note that Rp– and Nrp items were seen the exact same number of times in exactly the same fashion; the only difference between the conditions was that participants performed Competitive retrieval practice on other items from the Rp categories (but not from the Nrp categories).

condition were all presented once before any would appear a second time in any run.

The item's condition dictated the ordering of the 2 constituent tasks to be performed in the retrieval-practice/restudy phase (see Fig. 3). For Rp+ items, participants were asked first to verbally retrieve the animal's proper name, given the associated image and a category cue. They had 2 s to do so and were encouraged to guess if unsure of the correct response. Vocal responses were recorded with an MR-safe noise-canceling microphone and later coded for accuracy offline. The correct name then appeared on the screen along with the corresponding visual image for 2 s, during which time participants were asked to passively restudy the intact pairing. Because the correct answer only appeared after a selective recall attempt in this condition, we will refer to this task as "Competitive recall."

The order of the 2 components was simply reversed whenever an Rp– or Nrp item was presented: Participants received the correct answer immediately prior to having to repeat the name aloud. We expected retrieval competition to be relatively low in this condition, so we will refer to this task as "Non-Competitive recall." Notably, our stimuli and procedure were designed to minimize the degree of overlap across categories. If items from different categories compete with one another at retrieval, this would make it more difficult to observe the predicted differences between Rp categories (which were subject to Competitive recall) and Nrp categories (which were not).

During both Competitive and Non-Competitive recall attempts, participants were asked to fixate on a central cross and "clear their minds" during the variable ITI, rather than thinking about any animals or their names. Moreover, the experimenter emphasized that participants should not attempt to covertly retrieve the name when shown the correct answer. Instead, they were instructed to passively review the presented information. To further encourage compliance, the final 15 participants received elaborated instructions to explain that any attempts to "quiz themselves" during the restudy period would render useless the intended contrast between retrieval and restudy conditions. Their compliance with this special instruction was assessed verbally in between scanner runs and post-experimentally on a written questionnaire.

At the end of the retrieval-practice/restudy phase, participants were given 1 last opportunity to study all of the intact image–name pairings. The purpose of this final study phase (S2) was to collect measurements of neural similarity that we could compare with the measurements obtained during the matched initial round of study (S1). During S2, participants were reminded that they would receive a test on all the items at the end of the scan session. However, they were instructed to avoid overt (and covert) retrieval practice during this phase and to focus exclusively on the single animal presented on the screen. This phase paralleled that of the initial study period, albeit with a freshly randomized presentation schedule.

In the final recall test, participants were presented with each of the 48 images, along with a category cue, one at a time. They were instructed to verbally recall the associated proper name as quickly as possible. Participants were given 4 s to respond aloud and encouraged to guess if unsure of the response. A 2-s ITI separated each test trial. The order of the test trials was block-randomized, such that every 4 test trials contained 1 representative from each of the 4 conditions. No scanning occurred during this phase.

Control Experiment

In addition to the imaging study, we conducted a behavioral control experiment ($N = 18$). The goal of the control experiment was to assess whether we would get RIF (instead of revRIF) if we removed opportunities for restudy of Rp– and Nrp items, but otherwise kept everything the same. The procedure of this control experiment was identical to the imaging experiment, up until the start of the retrieval-practice phase. At that point, participants in the control experiment practiced retrieving the Rp+ items with feedback (the number and structure of these Rp+ trials exactly matched the Rp+ trials from the corresponding phase of the imaging experiment, though the ITI was fixed at 2 s during this period of the control experiment). However, unlike the imaging study, participants were not given the opportunity to restudy or retrieve (even non-competitively) Rp– or Nrp items during this phase. Immediately following the retrieval-practice phase, participants were given the instructions for the final cued-recall test, which they then completed.

fMRI Acquisition and Preprocessing

Data were acquired on a 3T Siemens Skyra scanner with a 16-channel phased array head coil. Data from the 6 functional runs were acquired using a T2*-weighted gradient-echo echo-planar imaging sequence composed of 30 interleaved slices oriented parallel to the long axis of the hippocampus (TR = 2000 ms; TE = 30 ms; flip angle = 71°; FoV read = 256 mm; FoV phase = 90.6%; base resolution = 128; voxel size = 2 × 2 × 3 mm; acceleration = 2 × GRAPPA). As the sequence (based on LaRocque et al. 2013) was designed to optimize signal recovery from the medial temporal lobe, coverage of the dorsal parietal and frontal lobes, as well as portions of the ventral occipital lobes, was sometimes sacrificed (see Supplementary Fig. 1 for a coverage map). The first 5 recorded brain volumes (and 3 hidden dummy volumes) of each functional run were ignored to allow for T1 stabilization. We acquired a coplanar T1-weighted FLASH sequence at the end of each scan session to assist in functional coregistration with the high-resolution 3D T1-weighted MPRAGE image collected prior to the functional runs (176 sagittal slices; TR = 2530 ms; TI = 1100 ms; TE = 3.3 ms; flip angle = 7°; FoV = 256 mm; voxel size = 1 mm isotropic; acceleration = 2 × GRAPPA). Field-mapping scans were also acquired after the final recall test.

Preprocessing of the functional data was conducted via FEAT version 5.98 in FSL version 4.1.9 (www.fmrib.ox.ac.uk/fsl). Each run was subjected to the following preprocessing steps: motion correction using MCFLIRT; field map-based EPI unwarping using PRELUDE+FUGUE; slice-timing correction using Fourier-space time-series phase-shifting; non-brain removal using BET; grand-mean intensity normalization of the entire 4D data set by a single multiplicative factor; and high-pass temporal filtering with a 64s-sigma Gaussian kernel.

Imaging data from the initial study and final restudy periods were not smoothed, in order to retain the highest possible spatial resolution for the multi-voxel pattern analyses that would follow (Zeineh et al. 2003; Carr et al. 2010).

FLIRT was used to carry out registration of the functional runs to the FLASH, MPRAGE, and standard brain (MNI152 with 2-mm isotropic voxels). The pattern similarity analyses of primary interest were conducted on hippocampal ROIs defined in native space and realigned to the first volume of the initial study phase.

Hippocampal ROIs were defined anatomically via FreeSurfer's (http://surfer.nmr.mgh.harvard.edu) automated segmentation of each participant's high-resolution structural image (for details, see Fischl et al. 2004). Individual left and right hippocampal masks were extracted from the output. The binary masks were then registered to functional space using FLIRT. We created a bilateral ROI by merging the left and right hippocampal masks for each participant.

Analysis of Neural Pattern Similarity

For each of our hippocampal ROIs, we constructed similarity matrices, representing the extent to which the spatial pattern of BOLD activity associated with the presentation of each item correlated with the patterns associated with other items from the same category. By looking at the average similarity separately for Rp and Nrp categories across time, we aimed to quantify how differences between these 2 conditions related to behavioral performance on the final recall test.

Figure 4 illustrates our analysis procedure. Each participant's preprocessed data were z-scored on a voxel-wise basis, separately for the initial study and final restudy runs. Custom Matlab routines calling functions from the Simitar toolbox (Pereira and Botvinick 2013) were used to label and group volumes by the category (type of animal) and condition (Rp–, Rp+, Nrp_a, or Nrp_b) of the associated stimulus at each time point. To account for the hemodynamic lag and reduce noise, we averaged across the second, third, and fourth brain volumes acquired after the onset of the image–name displays. We next computed the Pearson correlation between the patterns of BOLD activity associated with items within each category. For our primary analysis, Rp– items were compared with Rp+ of the same animal category, just as Nrp_a items were compared with Nrp_b items from the same animal category. The dummy coding in the case of Nrp served to match the number of items within each of the 4 conditions, which would later be compared. As shown in Figure 1, our theory specifically predicts that Rp– items should be repelled away from Rp+ items. Computing within-category similarity based on the neural distance between Rp– and Rp+ items was assumed to maximize our sensitivity to this effect (see "Subsidiary analyses" section below for additional variants of this analysis).

After Fisher z-transforming the correlation coefficients, we computed the average similarity for the Rp and Nrp categories.

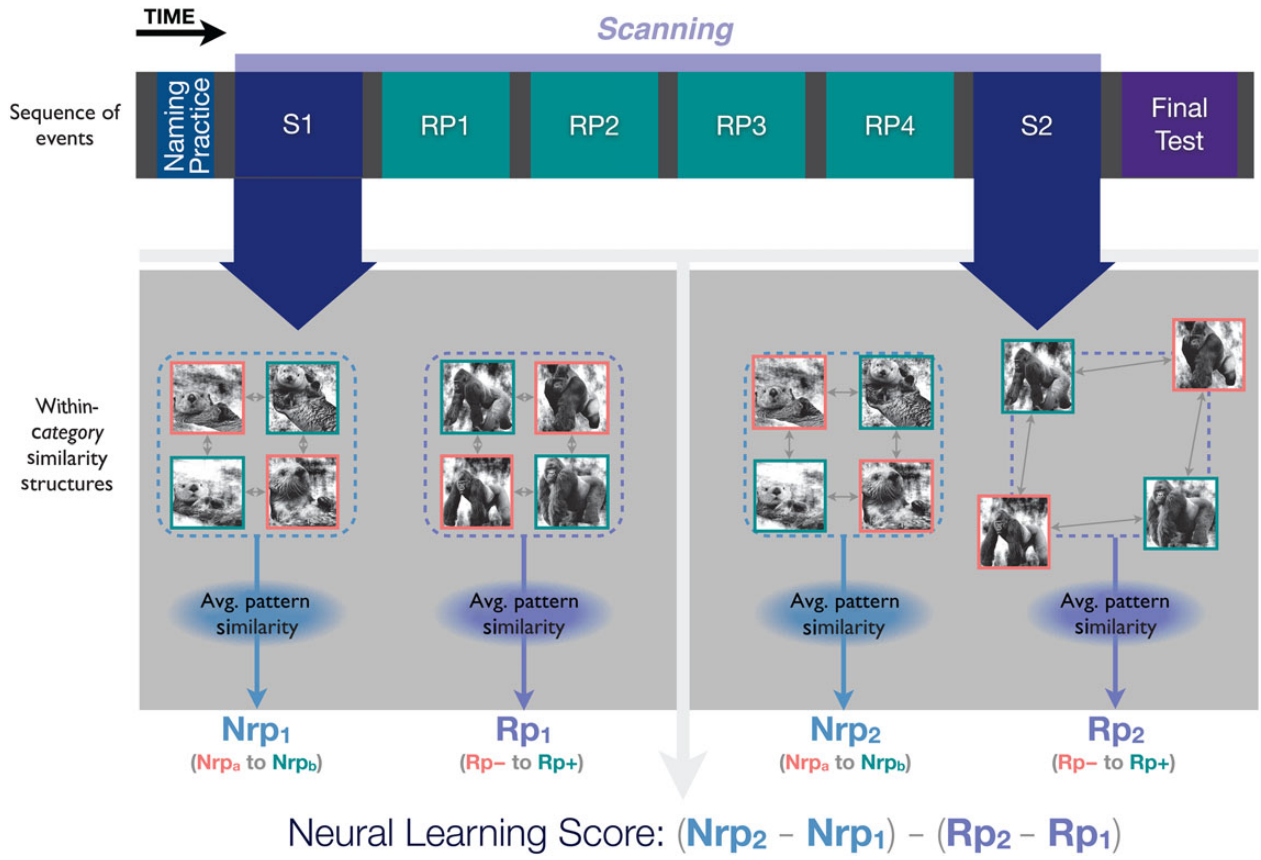

Figure 4. Derivation of the neural learning score. For simplicity, in this diagram, we only consider 1 animal category per Nrp/Rp condition, each with 4 exemplars (the outline colors indicate the sub-condition within that category, with pink representing Rp– and Nrp_a items and green representing Rp+ and Nrp_b items). For our primary analyses, we computed similarity matrices separately for items within each Rp category (Rp– to Rp+ similarity) and Nrp category (Nrp_a to Nrp_b similarity) based on data collected during the initial study period (S1). The within-category item similarity was expected to be comparable, on average, across categories prior to our behavioral intervention. Thus, items from S1 are represented as being equidistant from each other in the within-category similarity structures. In the figure, differentiation is shown as reduced similarity for items in the Rp category during the S2 phase, relative to the level of similarity that was present during S1. Our "neural learning score" summarizes these effects using a single number that reflects the change in similarity (from S1 to S2) for Rp items, relative to Nrp items. Because the values entering into the score are similarity (rather than distance) measures, we interpreted larger, positive scores as reflecting greater differentiation of the Rp items compared with Nrp items over time. Of primary interest was the relationship, across participants, between the magnitude of the neural learning score and the size of the revRIF effect, measured on the final recall test.

In order to determine whether there was more differentiation within our Rp condition than our Nrp condition, we computed a "neural learning score" defined as a double difference of the relevant similarity scores obtained for the initial study (S1) and final restudy (S2) runs:

$$(Nrp_2 - Nrp_1) - (Rp_2 - Rp_1)$$

Thus, for each participant, we arrived at a score for which positive values indicated greater differentiation within Rp categories relative to Nrp categories. Critically, we then tested for a relationship between individual differences in the neural learning score and behavioral revRIF, hypothesizing that greater differentiation would predict more revRIF. While we had a strong, a priori prediction about the direction of this effect, we report the results of two-tailed tests throughout.

Subsidiary Analyses

In addition to predicting that the change in Rp+ to Rp– pattern similarity (relative to baseline) should correlate with revRIF, our model also predicts that initial and final similarity should relate to revRIF. High levels of initial similarity (during S1) should lead to higher levels of competition during retrieval that, in turn, should yield greater differentiation and revRIF. The predicted relationship between final similarity (during S2) and revRIF goes in the opposite direction: Lower levels of pattern similarity for Rp items (vs. Nrp items) during S2 should be associated with less interference on the final recall test, resulting in better recall of Rp– items on the final test and more revRIF.

These ancillary effects may be harder to detect than the relationship between the change in pattern similarity (from S1 to S2) and revRIF. Consider that some participants may show a higher level of neural similarity than others for reasons unrelated to learning (e.g., they might have a more stable BOLD signal). Computing the change in similarity across the 2 time points helps to cancel out nuisance factors of this sort, provided they affect S1 and S2 equally. In contrast, when analyzing S1 or S2 on their own, we do not benefit from this noise cancelation. For this reason, we focused on the more powerful test afforded by the neural learning score (which looks at the change between S1 and S2) in our main analysis, though we also report findings from S1 and S2 individually.

Our model also predicts some degree of differentiation between items within each of the 2 Rp conditions (i.e., between

Rp– items and other Rp– items; and between Rp+ items and other Rp+ items). In the case of Rp– items, we might expect some incidental differentiation as a byproduct of those items being repelled from their Rp+ counterparts. As Rp– items are pushed away from Rp+ items, they should also—on average—be (indirectly) pushed away from each other. That is, the Rp items should start out tightly clustered in representational space, and the Rp– items should be forced outward as they are repelled from the Rp+ items. If they are forced outward in different directions, the distance between them should increase. However, because this effect is indirect (and we cannot guarantee that they will always be forced outward in different directions), we expect that this will be much less reliable than the Rp+ to Rp– differentiation effect, and thus less predictive of behavioral revRIF.

In the case of Rp+ items, we would expect differentiation to the degree that these items compete with other Rp+ items during retrieval practice; considered on its own, this differentiation should boost memory for Rp+ items, leading to a correlation between Rp+ differentiation (measured relative to an Nrp baseline) and Rp+ recall on the final test. Having said this, differentiation effects for Rp+ items could be offset by processes that increase Rp+ to Rp+ similarity (e.g., if there are parts of the hippocampus that track memory strength, Rp+ items may increasingly come to engage these regions, boosting similarity between Rp+ items; we talk about this possibility in the "Univariate confounds" section of the discussion). Competitive retrieval practice may also boost recall of Rp+ items through strengthening, even if no differentiation takes place. To the extent that differentiation is not the only determinant of Rp+ recall, this will make it more difficult to observe the predicted relationship between differentiation and Rp+ facilitation.

These concerns (about the reliability of the Rp– to Rp– similarity and Rp+ to Rp+ similarity as predictors of behavior) led us to focus on analyses relating Rp+ to Rp– similarity to behavior—as noted earlier, this is where our model makes its strongest predictions. However, for completeness, we also report results of analyses relating Rp– to Rp– similarity and Rp+ to Rp+ similarity to recall behavior.

To assess whether multivariate methods were needed to detect a relationship between changes in hippocampal activity and revRIF, we conducted a univariate version of our primary analysis. For this analysis, we computed the difference in univariate hippocampal activation elicited by Rp+ and Rp– items, and we measured how this difference changed from S1 to S2 (relative to baseline). Specifically, we extracted the univariate contrast of parameter estimates from a spatially smoothed (using a 5-mm FWHM Gaussian kernel) version of the data obtained during S1 and S2 for Rp+ versus Rp– items, and for Nrp_a versus Nrp_b items. This enabled us to look at changes in the Rp+ to Rp– contrast from S1 to S2, relative to changes in the corresponding Nrp_a to Nrp_b contrast from S1 to S2.

Lastly, we also conducted an exploratory whole-brain searchlight analysis, designed to identify extra-hippocampal regions that might also exhibit a positive relationship between the neural learning score and behavioral revRIF. This analysis proceeded in manner analogous to the ROI approach described earlier, except that, rather than a single ROI, the neural learning score was computed for a roving 3 × 3 × 3 voxel ROI (or smaller, in the case of searchlights on the edge of the brain mask) and assigned to each searchlight's centroid (Kriegeskorte et al. 2008). For these purposes, the subject-level results of the exploratory whole-brain searchlight were warped into standard space using FLIRT with further refinements made by FNIRT nonlinear registration, in order to facilitate group-level statistical analyses. To assess the statistical significance of the relationship between these (per-searchlight, per-subject) neural learning scores and revRIF, we ran a regression analysis using FSL's Randomise (version 2.9). In this analysis, the strength of the observed across-subjects relationship (at a particular brain location) between neural learning scores and revRIF was compared with an empirical null distribution, which was generated by randomly permuting participants' neural learning scores with respect to their revRIF scores 5000 times. A voxel-wise threshold of $P < 0.0001$, uncorrected for multiple comparisons, was adopted for this exploratory regression analysis.

Results

Behavioral Results

Neuroimaging Participants

Separate ANOVAs were conducted on retrieval-practice success (measured during the retrieval-practice/restudy phase) and on final recall accuracy. As expected, given the trivial nature of the Non-Competitive retrieval-practice task that was used for Rp– and Nrp items, retrieval-practice success for Rp– and Nrp items was at ceiling (mean across conditions and runs, 0.99). Therefore, we focused on retrieval-practice success for the Rp+ items, which involved Competitive retrieval from long-term memory. The data presented in the left-hand panel of Figure 5 reveal a significant linear increase in Rp+ retrieval-practice success ($F_{(1,23)} = 101.99$, $P < 0.001$) as participants gained more experience retrieving the targets from the first to the fourth run of retrieval practice/restudy.

Final cued-recall accuracy for the proper names of animals that were subjected to Competitive retrieval practice (Rp+ mean = 0.65; SD = 0.22) was higher than that for baseline items (Nrp mean = 0.47; SD = 0.26; $F_{(1,23)} = 13.87$, $P = 0.001$). This facilitation is a common feature of the RIF literature (see Anderson 2003 for a review). We also observed a marginal trend toward reverse RIF: improved recall of Rp– items relative to Nrp items (Rp– mean = 0.57; SD = 0.22; $F_{(1,23)} = 4.25$, $P = 0.05$). See the right-hand panel of Figure 5 for a graphical representation of these results, which show the same qualitative pattern as Storm et al.'s (2008) results, despite numerous differences in the materials and procedures. Failures to recall the correct name can be divided into occasions on which the participant did not recall anything and those on which participants provided an erroneous name (commission errors). To quantify the relative incidence of these events, we divided the number of commission errors by the total number of test trials marked as incorrect. Overall, our sample of 24 participants had an average ratio of 0.72 (SD = 0.21) across conditions, indicating a high incidence of guessing/competition on trials that posed a challenge for them.

We used the difference between each participant's final Rp– recall accuracy and his or her final Nrp recall accuracy—dubbed reverse RIF or revRIF because positive values indicate better recall of Rp– items than Nrp items—as our dependent measure in the individual differences analyses described below.

Control Participants

The behavioral control experiment was intended to establish whether our materials and general paradigm were sufficient to yield RIF when interwoven restudy episodes—our key manipulation—were excised. Indeed, the final test results from this control experiment showed significant RIF: Recall for the Rp– items

(mean recall = 0.01, SD = 0.02) was impaired compared with recall of baseline (Nrp) items (mean recall = 0.06, SD = 0.05), $F_{(1,17)} = 14.64$, $P = 0.001$. Of our 18 participants, only a single one showed an above-baseline (revRIF) effect (8%). Unsurprisingly, Rp+ recall was high across participants (mean = 0.74, SD = 0.21) on account of the 8 study/feedback cycles throughout the retrieval-practice phase.

fMRI Results

In an attempt to quantify and track the overlap between competitors across runs, we computed the change in similarity between hippocampal activity patterns associated with exemplars from each of our 6 categories. Averaged across participants, the neural learning score was not significantly different from 0 in the left hippocampus (mean, −0.01; SD, 0.16; $t_{(23)} = -0.31$, $P = 0.76$), the right hippocampus (mean, −0.01; SD, 0.20; $t_{(23)} = -0.27$, $P = 0.79$), or the bilateral hippocampal ROI (mean, −0.01; SD, 0.18; $t_{(23)} = -0.32$, $P = 0.75$). This finding indicates that, at the group level, the change in similarity of hippocampal BOLD patterns (within category) was comparable across the Rp and Nrp categories. However, this approach does not account for the vast individual differences observed in the data. Critically, we predicted that individual differences on our measure of behavioral differentiation (revRIF) should correlate with the degree of neural differentiation that took place between the initial study and restudy periods (as measured by our neural learning score).

A correlation analysis (all two-tailed tests) revealed a significant relationship in the expected (positive) direction between revRIF and the neural learning score extracted from the left hippocampus ($r_{(22)} = 0.43$, $P = 0.03$). The same relationship was marginal in the bilateral hippocampal ROI ($r_{(22)} = 0.34$, $P = 0.10$), and it was not reliable in the right hippocampus ($r_{(22)} = 0.26$, $P = 0.22$). See Figure 6 for the associated scatterplots.

A positive neural learning score, as previously defined, could arise from an increase in pattern similarity within Nrp categories, a decrease in pattern similarity within Rp categories, or some combination of the 2. However, our theory holds that revRIF should depend on the degree to which items from Rp categories differentiate. To examine this prediction further, we separated out the measured change in similarity for Nrp and Rp categories and tested to see whether either was significantly correlated with revRIF. For these analyses, we subtracted the final restudy period's neural similarity scores from those of the initial study period; positive scores on this measure indicate greater differentiation.

As expected, there was a significant positive relationship between changes in neural similarity within Rp categories (i.e., between Rp+ and Rp− items) and revRIF; no significant relationship emerged between changes in neural similarity within Nrp categories (i.e., between Nrp_a and Nrp_b items) and revRIF. Importantly, these correlations (for Rp items and Nrp items) were

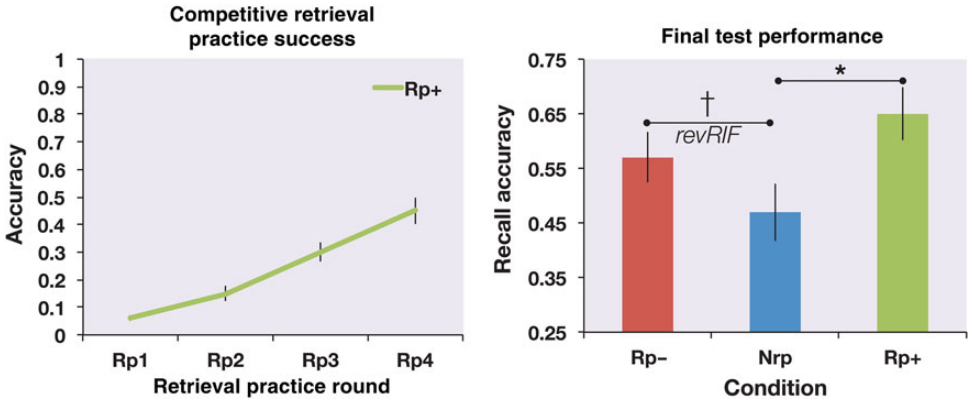

Figure 5. Behavioral results. The left panel depicts Competitive retrieval-practice success for Rp+ items across the 4 intervening rounds between the initial study phase and the final restudy opportunity. While participants initially struggled to name the Rp+ animals in the Competitive retrieval condition, they managed to do so with greater success on subsequent retrieval practice attempts. The right panel depicts the final recall accuracy for all 3 conditions. Competitive retrieval practice facilitated Rp+ items above the Nrp baseline. There was a marginally significant trend for the Rp− competitors to be facilitated above the Nrp baseline, as well, indicating numeric revRIF across participants. Error bars represent SE of the mean across participants. *$P < 0.05$, †$P = 0.05$.

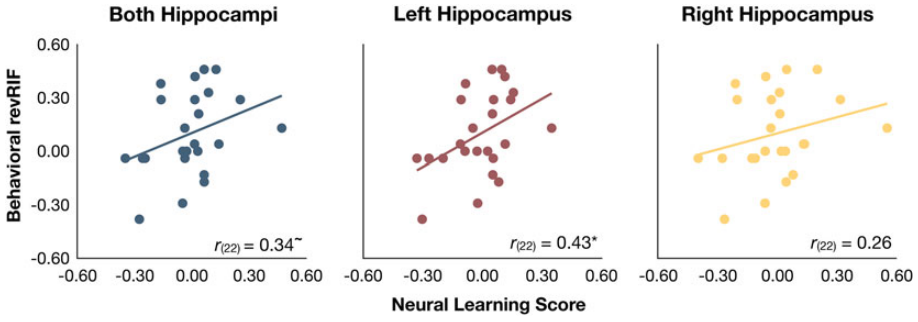

Figure 6. Individual differences analysis linking brain and behavior. We found a significant positive correlation between our neural learning score derived from a left hippocampal ROI (in red) and the degree to which the Rp− items were facilitated above baseline on the final test. A similar trend was found when considering a bilateral hippocampal ROI (in blue). The right hippocampal ROI (yellow) showed a far less reliable trend in the same direction. *$P < 0.05$, ~$P = 0.10$.

significantly different according to Meng's z-test (Meng et al. 1992). This pattern of findings held both for similarity scores extracted from the left hippocampal ROI ($r_{Rp(22)} = 0.58$, $P = 0.003$; $r_{Nrp(22)} = -0.09$, $P = 0.68$; Meng's $z_{(23)} = 2.20$, $P = 0.01$) and from the bilateral hippocampal ROI ($r_{Rp(22)} = 0.49$, $P = 0.02$; $r_{Nrp(22)} = -0.11$, $P = 0.63$; Meng's $z_{(23)} = 2.22$, $P = 0.01$). See Figure 7 for the associated scatterplots. Numerically, the data from the right hippocampal ROI followed the same pattern ($r_{Rp(22)} = 0.31$, $P = 0.14$; $r_{Nrp(22)} = -0.11$, $P = 0.61$; Meng's $z_{(23)} = 1.51$, $P = 0.07$). However, we focused our subsequent analyses on the 2 hippocampal ROIs (left and bilateral) that reliably exhibited the basic pattern of results.

The strongest predictions of our model pertained to changes in Rp– to Rp+ similarity (relative to baseline). Our neural learning score was designed to capture this type of differentiation. As previously noted, we also might expect some indirect (and, thereby, weaker) differentiation of Rp– items from other Rp– items within a given category. In our study, the degree to which Rp– items differentiated from each other (relative to baseline) did not reliably predict behavioral revRIF in the left ($r_{(22)} = 0.09$, $P = 0.69$) or bilateral hippocampal ROIs ($r_{(22)} = 0.03$, $P = 0.91$). To the extent that Rp+ items competed with each other, they too might be expected to differentiate, leading to improved recall of Rp+ items (relative to baseline). Yet, as we mentioned in "Materials and methods," there are several countervailing factors that could potentially work against this effect. As such, we did not have strong predictions as to the relationship between Rp+ differentiation and final Rp+ recall. In our study, we found a nonsignificant trend whereby greater similarity—as opposed to differentiation—of Rp+ items (relative to baseline) predicted better recall performance for Rp+ items (relative to baseline) in the left hippocampal ROI ($r_{(22)} = -0.33$, $P = 0.12$). The same trend was significant in the bilateral hippocampal ROI ($r_{(22)} = -0.46$, $P = 0.02$).

As noted earlier, focusing on a single time point reduces our ability to subtract out nuisance factors unrelated to learning. Nevertheless, our model predicts that initial and final similarity states should also relate to final recallability. An attempt was made to test these ancillary predictions with the available data. As predicted, greater similarity between Rp– and Rp+ items measured during initial study (S1) corresponded to greater revRIF in both the left hippocampal ROI ($r_{(22)} = 0.68$, $P < 0.001$) and the bilateral hippocampal ROI ($r_{(22)} = 0.57$, $P = 0.004$). In contrast, the analogous analysis involving Nrp_a and Nrp_b items failed to exhibit a reliable relationship in either the left hippocampal ROI ($r_{(22)} = 0.12$, $P = 0.56$) or the bilateral ROI ($r_{(22)} = 0.08$, $P = 0.72$). The relationship between initial similarity and revRIF was significantly larger for Rp categories than Nrp categories in both the left hippocampal ROI (Meng's $z_{(23)} = 2.39$, $P = 0.008$) and the bilateral ROI (Meng's $z_{(23)} = 2.04$, $P = 0.02$). This latter set of findings accords with our computational model, which predicts that higher levels of initial similarity would lead to higher levels of competition during Competitive retrieval that, in turn, would drive greater differentiation and revRIF within the Rp condition.

While our analyses relating initial (S1) similarity to revRIF came out as predicted, our analyses relating final (S2) similarity to revRIF did not: We failed to observe any reliable correlation between similarity measured in the left hippocampal ROI during S2 and revRIF, in the Rp condition ($r_{(22)} = -0.22$, $P = 0.31$), in the Nrp condition ($r_{(22)} = 0.04$, $P = 0.86$), or the subtraction of the 2 ($r_{(22)} = 0.15$, $P = 0.50$). A comparable pattern was observed within the bilateral hippocampal ROI: in the Rp condition ($r_{(22)} = -0.23$, $P = 0.28$), in the Nrp condition ($r_{(22)} = 0.04$, $P = 0.86$), or the subtraction of the 2 ($r_{(22)} = 0.14$, $P = 0.51$).

We also tried a univariate version of our neural learning score analysis: Instead of computing how the multivariate distance between Rp+ and Rp– items changed from S1 to S2 (relative to baseline), we computed the difference in univariate hippocampal activation elicited by Rp+ and Rp– items, and we measured how this difference changed from S1 to S2 (relative to baseline).

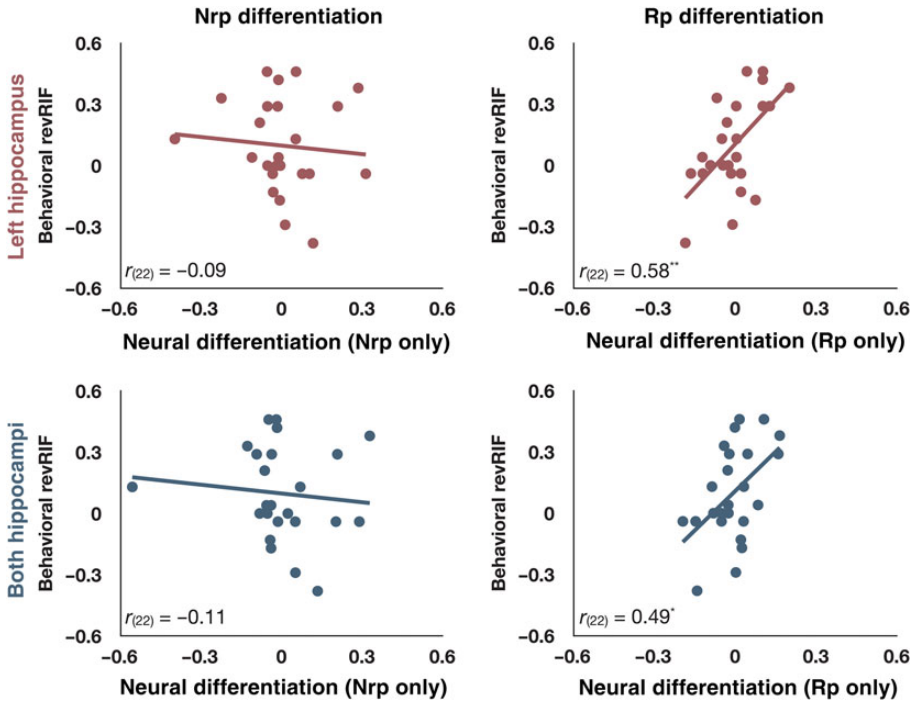

Figure 7. Within-condition differentiation and revRIF. By examining the relationships between revRIF and the constituent parts of the neural learning score derived from the left hippocampal ROI (top row, red) and the bilateral hippocampal ROI (bottom row, blue), we observed reliable correlations between revRIF and the neural differentiation of Rp categories, but not for Nrp categories. *$P < 0.05$, **$P < 0.01$.

Overall, we observed no evidence of a significant main effect (i.e., differential change in the Rp+ to Rp− contrast, relative to the Nrp_a to Nrp_b contrast) in either the left hippocampal ROI (mean, −1.42; SD, 93.36; $t_{(23)} = -0.07$, $P = 0.94$) or the bilateral ROI (mean, −1.82; SD, 100.81; $t_{(23)} = -0.09$, $P = 0.93$). Moreover, we failed to observe any reliable relationship between individual differences in this univariate measure and revRIF, considering either the left hippocampal ROI ($r_{(22)} = -0.05$, $P = 0.82$) or the bilateral ROI ($r_{(22)} = -0.03$, $P = 0.88$). This (null) result shows that—at least in this case—multivariate analysis was needed in order to observe a relationship between changes in hippocampal activity and revRIF.

Although we had a priori reasons to focus on the hippocampus and adjusted our imaging parameters to match, we also asked whether any other regions of the brain exhibited a reliable positive or negative relationship between revRIF and the neural learning score. To do so, we swept a 27-voxel searchlight across the entire brain volume. No searchlight centroids survived statistical thresholding at $P < 0.0001$ (uncorrected for multiple comparisons). It, of course, remains possible that other searchlight sizes/shapes would have identified hotspots.

Discussion

In this study, we followed up on a puzzling result from Storm et al. (2008), who reported that interleaved retrieval practice and restudy can lead to reverse RIF: improved recall of competing (Rp−) items relative to baseline (Nrp) items. Based on prior neural network simulations (Norman et al. 2006), we hypothesized that reverse RIF after interleaved retrieval practice and restudy was due to differentiation of competing representations. In support of this hypothesis, we found that—across participants—the degree of neural differentiation associated with Rp categories (relative to Nrp categories) in the left hippocampus predicted revRIF. Importantly, our behavioral control experiment showed that participants exhibit significant RIF when they were not given interleaved opportunities to restudy the Rp− and Nrp items.

Storm et al. (2008) accounted for accelerated relearning after RIF by appealing to Bjork and Bjork's (1992) "New Theory of Disuse," which holds that the long-term storage strength of a memory benefits from restudy attempts as a decreasing function of its current accessibility (called retrieval strength). According to this theory, retrieval practice reduces the retrieval strength of Rp− items. When Rp− items are restudied, they gain disproportionately in storage strength, owing to their diminished retrieval strength. Under some circumstances, this change in storage strength can outweigh the reduction in retrieval strength, resulting in a net increase in recall of Rp− items on the final test.

Crucially, theories that focus on memory strength generally fail to formalize another important dimension of memory: the representational overlap between competitors. Facilitation and inhibition may provide an adequate means of resolving retrieval-based competition in the short-term when targets and competitors remain fixed. However, different circumstances may call for the retrieval of both the previously inhibited and facilitated memories. Differentiation addresses this problem: Distinguishing the neural representations of the (previously) competing items ensures that all of items can be accessed in the future.

Factors Affecting the Neural Learning Score

Participants in this study varied extensively in the degree to which they showed neural differentiation (as indexed by the neural learning score) and behavioral revRIF. As predicted by our computational model, these neural and behavioral measures were correlated across participants (see Fig. 6). In addition to the correlation between the neural learning score and revRIF, there are 2 other features of Figure 6 scatterplot that stand out: 1) the high level of variability in neural learning scores across participants and 2) the lack of an overall trend toward differentiation—roughly equal numbers of participants showed negative versus positive neural learning scores. Here, we discuss possible factors that may have affected the neural learning score, leading to the observed pattern of results.

One possible source of individual differences is variance in how distinctively participants encoded the animal stimuli at the outset of the experiment. As discussed earlier, higher levels of initial similarity in the neural representations of these animals were associated with higher levels of revRIF on the final test.

Variability in participants' use of covert retrieval also may have contributed to individual differences in revRIF. Specifically, some participants may have consciously or unconsciously adopted a strategy of covertly retrieving Nrp items (e.g., by disregarding the provided correct answer and quizzing themselves during what are supposed to be Non-Competitive trials). Adopting such a strategy would have the effect of blurring the difference between the Nrp and Rp conditions, thereby reducing the neural learning score and the behavioral revRIF effect, which both rely on relative differences between the 2 conditions. Previous research indicates that RIF may be masked by covert attempts to retrieve putatively unpracticed items or categories (for a related discussion, see Weller et al. 2012). Indeed, such covert retrieval attempts—and their complicating influence on final recall—may be especially likely when retrieval practice is intermixed with restudy periods (Dobler and Bäuml 2013). As discussed in the "Methods" section, we excluded participants who admitted (on a post-test questionnaire) to deliberately quizzing themselves on Non-Competitive retrieval trials (see the "Strategy-based exclusions" section for further consideration of these participants). However, the remaining (non-excluded) participants may have unintentionally engaged in this strategy to varying degrees.

Importantly, while the aforementioned factors (variance in initial distinctiveness, covert retrieval) can explain variability in the neural learning score, they are not sufficient to explain why the distribution of neural learning scores was centered around 0. Taken at face value, this latter observation suggests that the differentiation process predicted by our model may not be occurring. However, another possibility is that strategic changes from S1 to S2 may have caused a net negative shift in the neural learning score that worked against the differentiation effect. For example, encoding variability may decrease as a function of participants' experience with stimuli from a particular category. During S1, participants may have been overwhelmed by the task of memorizing 48 animal–name pairings. Their attentional focus may have shifted rapidly between features of the presented animal and its given name, as they struggled to identify a useful strategy for the task. By S2, many participants may have settled on a more stable attentional strategy, yielding less noisy (and thus more similar) neural patterns across animals. Furthermore, we would expect this decrease in encoding variability to be larger for Rp categories, insofar as participants are forced to engage more deeply with Rp items (due to Rp+ Competitive retrieval practice) than Nrp items during the retrieval-practice phase. In summary, this encoding variability account posits a shift from relatively noisy/dissimilar patterns during S1 to less noisy/more similar patterns during S2, that is larger for Rp than

Nrp categories; because of the way that the neural learning score is computed, this shift would end up exerting a negative push on neural learning scores that works against the predicted differentiation effect.

Ruling Out Alternative Explanations

As noted earlier, our preferred account of the data is that revRIF is a consequence of neural differentiation. Here, we will address alternatives to this view.

The Role of Integration

Differentiation is one way to reduce interference; another way to reduce interference is to integrate the items (Anderson and McCulloch 1999). Through a series of instructional manipulations and reanalyses based on self-report data, Anderson and McCulloch (1999) found that attempts to integrate Rp+ and Rp– items reduced (though did not reverse) RIF. While integration could (in principle) have occurred in our study, we have numerous lines of evidence suggesting that integration was not a significant factor in driving the present findings. First, we explicitly instructed our participants not to engage in this behavior. Second, our own survey data revealed that only 6 participants indicated occasionally bringing multiple exemplars to mind for the purpose of comparison. Importantly, even in these cases, their stated intention was to highlight features that distinguished the exemplars, rather than to integrate over commonalities. Third, even if we were to take the extremely conservative approach and exclude these participants, the correlation between revRIF and the neural learning score remained significant within the left hippocampal ROI ($r_{(16)} = 0.53$, $P = 0.03$) and marginal within the bilateral hippocampal ROI ($r_{(16)} = 0.42$, $P = 0.08$). Most importantly, the hypothesis that integration is responsible for revRIF implies that there should be a positive relationship between neural similarity and revRIF (i.e., the more you integrate, the more neural similarity will increase for Rp items, and the more revRIF there should be), but we obtained the exact opposite pattern in our study—we found that decreased (not increased) neural similarity between Rp+ and Rp– items predicted revRIF. For all of the above-mentioned reasons, we argue that it is highly unlikely that integration gave rise to our basic brain–behavior correlation. Lastly, note that we *did* observe a positive relationship between the change in Rp+ to Rp+ item similarity and final recall of Rp+ items (relative to baseline). The directionality of this effect is consistent with the idea that participants are integrating Rp+ items during the retrieval-practice phase (leading to increased neural similarity and increased subsequent recall). However, for the reasons outlined earlier, we think that the effect may be due to other factors besides integration (e.g., the "Rp+ bump" hypothesis mentioned in the following section).

Univariate Confounds

Throughout the paper, we have interpreted positive values of the neural learning score as reflecting differentiation of multivariate patterns of activity in the hippocampus. However, the learning score could, in principle, be affected by univariate changes in neural activity. For example, consider what would happen if retrieval practice increased (or decreased) activity in a subset of hippocampal voxels for all Rp+ items. We will refer to this as the "Rp+ bump" hypothesis. Such an occurrence would reduce the neural similarity between Rp+ and Rp– items, insofar as the Rp+ patterns would contain the "bump" and the Rp– patterns would not. Consequently, it would show up as an increase in our neural learning score.

To distinguish between these 2 accounts of what is driving the neural learning score (differentiation or an Rp+ bump), we can look at the similarity between Rp+ items and how this relates to the neural learning score. If the neural learning score is driven by an Rp+ bump, then high values of the learning score should be accompanied by increased similarity between same-category Rp+ items (insofar as all of these Rp+ items will contain the bump). In contrast, if the neural learning score is driven by differentiation, then high values of the learning score should be accompanied by reduced similarity between same-category Rp+ items (when an Rp+ item is practiced, other Rp+ items from that category may act as competitors, resulting in differentiation).

An examination of changes in Rp+ to Rp+ similarity within the left hippocampal ROI yielded no evidence for the Rp+ bump hypothesis. There was a numerical trend for the learning score to be negatively correlated with same-category Rp+ to Rp+ similarity ($r_{(22)} = -0.12$, $P = 0.59$). As noted earlier, this pattern is consistent with the differentiation hypothesis and directly contradicts the Rp+ bump hypothesis. For the sake of completeness, we also examined the complementary relationship between (baseline corrected) changes in Rp– to Rp– similarity and the neural learning score (i.e., an "Rp– bump"). No reliable effect was observed within the left hippocampal ROI ($r_{(22)} = -0.09$, $P = 0.68$).

Lastly, in addition to the correlational analyses presented here, there is 1 additional reason to discount the Rp+ bump hypothesis: Namely, it does not provide a mechanism for the observed correlation between revRIF and the neural learning score. If differentiation is driving the neural learning score, it is clear why—in terms of our theory—this would affect revRIF: Reduced overlap leads to reduced competition, which then leads to improved recall of Rp– items. In contrast, if an Rp+ bump were driving the neural learning score, it is unclear why this would affect revRIF, which is defined entirely based on the relationship between Rp– and Nrp items (i.e., it is unclear why a bump in activation for Rp+ items would make recall of Rp– items better, relative to Nrp items).

While current evidence suggests that the Rp+ bump is clearly not driving the neural learning score (or revRIF), this does not mean that an Rp+ bump does not exist. For instance, our finding that increased Rp+ to Rp+ similarity predicts an Rp+ recall benefit could be explained in terms of just such an Rp+ bump. That is, higher levels of memory strength could be reflected in a bump in hippocampal voxels that track memory strength (leading to increased Rp+ to Rp+ similarity) and also higher levels of Rp+ recall.

Process-of-Elimination Strategies

Another possible account of revRIF is that it reflects strategies at test, rather than representational differentiation. For example, if participants adopted a strategy of ruling out familiar (Rp+) animal names when confronted with a less familiar (Rp–) memory cues, this could artificially boost our measure of Rp– recall, relative to baseline. However, we have reason to believe that this type of strategy was not likely to have given rise to our results. To the extent that participants' familiarity with Rp+ items was driving a process-of-elimination strategy, leading to revRIF, the same revRIF effect should be present in the control study, but it was not—participants instead showed a robust RIF effect.

Strategy-Based Exclusions

A total of 7 participants were excluded on the basis of their self-reported failure to follow instructions and not quiz themselves during study episodes. As they represent a sizeable proportion of our overall sample, we examined how their strategy choice

may have influenced the results. In an attempt to characterize the behavior of the 7 exclusions in general terms, we first noted that the mean Nrp recall from these 7 exclusions (0.73, SD = 0.21) was higher than that for the 24 compliant participants (0.47, SD = 0.26). A similar pattern emerged for the Rp– condition within the excluded group (mean = 0.64, SD = 0.15) relative to our compliant participants (mean = 0.57, SD = 0.22). These increases in Nrp and Rp– recall are consistent with the suggestion that the excluded participants may have engaged in covert Competitive retrieval practice of the Nrp and Rp– items. As is apparent from the pattern of means listed earlier, the 7 excluded participants did not show revRIF—rather, they showed a numerical trend toward RIF. Including these 7 participants in our main analysis weakened the relationship between the neural learning score and revRIF without changing the direction of the effect within the left hippocampal ROI ($r_{(29)} = 0.23$, $P = 0.21$), the bilateral hippocampal ROI ($r_{(29)} = 0.20$, $P = 0.29$), or the right hippocampal ROI ($r_{(29)} = 0.16$, $P = 0.39$). The weakened results make sense in light of a separate descriptive analysis focusing only on the 7 excluded participants, which revealed negative brain–behavior correlation coefficients within our hippocampal ROIs (left: −0.86; bilateral: −0.90; right: −0.75). The coefficients are in the opposite direction of the reported relationship within our compliant group. Together, these results suggest that participants' strategies interacted strongly with (rev)RIF in this paradigm. Moreover, the fact that participants' self-reports (of strategy use) aligned with behavioral and neural evidence suggests that participants had some subjective insight into these strategies.

The above-mentioned results suggest a way to further explore the idea (mentioned earlier, in the "Factors affecting the neural learning score" section) that covert retrieval practice may have contributed to the low neural learning scores shown by some of the included participants. If this is the case, then included participants with low neural learning scores should show the same behavioral "signature" of covert retrieval practice that was shown by excluded participants (most diagnostically: enhanced recall for Nrp items). To investigate this, we did a median split on the neural learning scores (from the left hippocampus) among the 24 included subjects and compared the "high neural learning score" participants and "low neural learning score" participants on Nrp recall. We found that mean final Nrp recall for "low neural learning score" participants (0.56, SD = 0.26) was numerically better than that for "high neural learning score" participants (0.39, SD = 0.25) and numerically worse than that for the excluded participants (0.73, SD = 0.21). While this comparison is post hoc and based on small numbers of participants, the numeric ordering of these results fits with the idea that participants with low neural learning scores may have engaged in covert retrieval practice (albeit not to the same degree as excluded participants).

Related Neural Findings

The results of our study are consistent with recent results from Schapiro et al. (2012). Like our study, they measured pattern similarity before and after learning. Unlike our study, they used a statistical learning paradigm in which participants viewed a long sequence of fractal images. The fractal stream was composed of strong pairs, for which the second image in the pair followed the first image 100% of the time, mixed in with weak pairs, for which the second image in the pair followed the first image 33% of the time.

In the Schapiro et al. (2012) study, weak associates played a role analogous to Rp– items in our study. On trials in which the second item in a weak pair did not follow the first, we hypothesize that participants generated a prediction of the second item (leading to weak activation of that item's representation), at the same time that the sensory representation of the first item was strongly active. This combination of weak activity (for the second item) and strong activity (for the first item) is analogous to the combination of weak activity for the Rp– item and strong activity for the Rp+ item depicted in Figure 1*b*. As shown in Figure 1*c*, we would expect this to result in shared features being disconnected from the second item. The next time that the second item is presented in the sequence, we would expect this to result in the fractal's representation acquiring new features, thereby resulting in differentiation (see Fig. 1*d*,e). In keeping with this prediction, Schapiro et al. (2012) found that the hippocampal representations of items in weak pairs differentiated from one another. Although there have been other studies that have related hippocampal pattern similarity to memory (e.g., LaRocque et al. 2013), most of these studies have not compared pre- and post-learning pattern similarity. To our knowledge, the Schapiro et al. (2012) study is the only other study (besides ours) that has looked at differentiation of hippocampal representations as a function of learning.

Relationship between Differentiation and Pattern Separation

The differentiation process described in this paper should not be confused with the widely discussed notion of hippocampal pattern separation (see Yassa and Stark 2011 for a review). Pattern separation refers to the ability of the hippocampus to assign distinct representations to stimuli, regardless of their similarity (Marr 1971). This pattern separation bias, which results from sparse coding in the dentate gyrus and CA3, occurs automatically for all input patterns (O'Reilly and McClelland 1994; O'Reilly and Rudy 2001). Another key point is that pattern separation reduces overlap between neural representations, but there is still some residual overlap in the representations that are assigned to similar stimuli (Norman and O'Reilly 2003). This residual overlap is important for coding efficiency—if the hippocampus assigned completely non-overlapping representations to stimuli, it would quickly run out of neurons. As discussed by Norman and O'Reilly (2003), residual overlap in the hippocampus can lead to interference between memories, both at encoding (if new memories partially overwrite older ones) and at recall (if overlapping memories co-activate and compete with each other).

The differentiation process described in this paper operates on the residual levels of hippocampal overlap that are "left over" after (automatic) pattern separation processes take place. That is, differentiation is not automatic; rather, it is driven by competition at retrieval (resulting from overlap in representations), and it acts to reduce this competition (by further reducing representational overlap). The idea of adaptive differentiation has a longstanding history in the connectionist modeling literature: In these models, when similar stimuli are linked to distinct responses and are presented in an interleaved fashion, this can trigger learning processes that act to pull apart the internal representations of these stimuli, thereby supporting the network's ability to generate appropriately distinct responses to these stimuli (e.g., Gluck and Myers 1993; McClelland et al. 1995). O'Reilly and Rudy (2001) ascribed this adaptive differentiation process specifically to cortex, as opposed to hippocampus. A key claim here is that adaptive differentiation can happen in the hippocampus proper in addition to surrounding cortical regions (for a recent model of how adaptive, error-correcting learning can take place in the hippocampus, see Ketz et al. 2013; see also Gluck

and Myers 1993). We should also note that the differentiation process does not have to be complete in order to yield improvements in recall. Figure 1 (for expository convenience) shows a complete elimination of overlap between previously competing memories; however, our model and the brain data from differentiators in this study both show a more modest reduction in overlap, which we hypothesize causes a moderate reduction in competition, thereby leading to a moderate increase in recall accuracy.

One final issue relates to the representation of similarities between category members: If (automatic) hippocampal pattern separation reduces overlap between representations of same-category items, and (competition-driven) differentiation reduces hippocampal overlap even further, how does the brain still manage to represent all of the features that same-category items have in common with one another? A central claim of the Complementary Learning Systems (CLS) model (McClelland et al. 1995) is that cortex, not hippocampus, is responsible for representing this kind of semantic similarity structure. Whereas the hippocampus is biased to assign relatively distinct representations to stimuli, cortex is biased to assign overlapping, feature-based representations, such that stimuli with similar features are assigned similar internal representations. According to the CLS model, low overlap in the hippocampus explains how people can retrieve distinct names for similar-looking pictures, and high overlap in cortex explains how—at the same time—people can still recognize similarities between same-category items. As noted earlier, the CLS model also posits that differentiation-like processes occur in cortex (see O'Reilly and Rudy 2001), but cortical differentiation is hypothesized to occur on a different, more incremental, timescale. Moreover, it is done with the goal of refining rather than eliminating the cortical representation of similarity structure. The fMRI protocol that we used in the present study was optimized to study hippocampus, so we were not able to fully assess the degree to which differentiation also took place in cortical regions.

Relationship to Other Accounts of Differentiation

The differentiation account presented here also differs from the account of memory differentiation presented by Shiffrin et al. (1990; see also Shiffrin and Steyvers 1997; Criss et al. 2013). Shiffrin et al. (1990) argued that—when a memory is strengthened—the representation of which features that memory does (and does not) contain becomes sharper, thereby making it less confusable with other memories. Applied to our study, this idea implies that strengthening Rp+ items via retrieval practice will make the representations of Rp+ items more distinct from all other items, thereby reducing competition and boosting recall of Rp– items. A key difference between Shiffrin et al.'s (1990) view of differentiation and ours is that—according to Shiffrin et al.—retrieval practice of Rp+ items can itself cause differentiation and, thus, revRIF. In contrast, our theory predicts that retrieval practice of Rp+ items, on its own, will harm memory for Rp– items and that interleaving retrieval practice of Rp+ items and extra study of Rp– items is needed in order to obtain differentiation and revRIF. While not definitive, results from Storm et al. (2008)—showing RIF after retrieval practice and revRIF only after interleaved learning of Rp+ and Rp– items—are more consistent with our view than the Shiffrin et al. (1990) view. The Storm et al. (2008) results suggest that the Shiffrin et al. (1990) hypothesis may be insufficient to account for extant data on RIF and revRIF on its own.

Future Directions

Subsequent investigations may further clarify our understanding of the neural learning score by probing the differentiation of individual items from their initial state, rather than relying on averages across entire categories. The results may speak to even more specific predictions arising from our neural network model. For instance, it may be possible to identify the neural signature of features Rp– items cede to Rp+ items, as well as any new features assumed by the former's representation. Additional research is also needed to investigate the relationship between the neural learning score and revRIF in extra-hippocampal regions. Our choice of MRI sequence optimized was guided by an a priori hypothesis based on evidence of the hippocampus's importance in episodic memory formation. This entailed sacrificing coverage of other areas, including some with known associations to the representation of animals, such as the lateral occipital complex (e.g., Weber et al. 2009).

Our findings may also shed light on previous results showing that periods involving sleep reduce RIF (MacLeod and Macrae 2001; Chan 2009; Baran et al. 2010; but see Racsmány et al. 2010; Abel and Bäuml 2012). Previously, we have argued that sleep presents an opportunity for interleaved replay of competing memories (Norman et al. 2005). More recently, we have argued that interleaved replay during sleep promotes differentiation, resulting in reduced competition and more accurate recall of individuating features (Schapiro et al. 2013). In future work, we will explore whether effects of sleep on RIF depend on the same processes that we manufactured during the waking state in this study (i.e., differentiation, caused by interleaved learning).

Conclusions

In summary, we found that neural differentiation of competing memories (as measured using fMRI pattern similarity analysis) was associated with improved recall of these memories. These results were in line with predictions derived from our prior neural network modeling work (Norman et al. 2006, 2007). The work highlights the contribution of differentiation—in combination with inhibition and facilitation—to reducing competition between memories. Without differentiation, memory would be a zero-sum process whereby strengthening one memory necessarily impairs recall of other, related memories. Differentiation provides a way out of this zero-sum trap: Our results provide initial evidence that neural differentiation can resolve competition in a way that enhances the accessibility of both practiced and unpracticed items.

Supplementary Material

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

Funding

This work was supported by the National Institutes of Health (R01 MH069456 to K.A.N.).

Notes

We thank Mariam Aly, Anna Schapiro, and Nick Turk-Browne for their comments on an earlier version of this manuscript. *Conflict of Interest:* None declared.

References

Abel M, Bäuml K-HT. 2012. Retrieval-induced forgetting, delay, and sleep. Memory. 20:420–428.

Anderson MC. 2003. Rethinking interference theory: executive control and the mechanisms of forgetting. J Mem Lang. 49:415–445.

Anderson MC, Bjork RA, Bjork EL. 1994. Remembering can cause forgetting: retrieval dynamics in long-term memory. J Exp Psychol Learn. 20:1063–1087.

Anderson MC, McCulloch KC. 1999. Integration as a general boundary condition on retrieval-induced forgetting. J Experim Psychol. 25:608–629.

Baran B, Wilson J, Spencer R. 2010. REM-dependent repair of competitive memory suppression. Exp Brain Res. 203:471–477.

Bjork RA, Bjork EL. 1992. A new theory of disuse and an old theory of stimulus fluctuation. In: Healy AF, Kosslyn S, Shiffrin RM, editors. From Learning Processes to Cognitive Processes: Essays in Honor of William K. Estes. Hillside, NJ: Erlbaum. p. 35–67.

Brainard DH. 1997. The psychophysics toolbox. Spat Vis. 10: 433–436.

Carr VA, Viskontas IV, Engel SA, Knowlton BJ. 2010. Neural activity in the hippocampus and perirhinal cortex during encoding is associated with the durability of episodic memory. J Cogn Neurosci. 22:2652–2662.

Chan JCK. 2009. When does retrieval induce forgetting and when does it induce facilitation? Implications for retrieval inhibition, testing effect, and text processing. J Mem Lang. 61:153–170.

Criss AH, Wheeler ME, McClelland JL. 2013. A differentiation account of recognition memory: evidence from fMRI. J Cogn Neurosci. 25:421–435.

Dale AM. 1999. Optimal experimental design for event-related fMRI. Hum Brain Mapp. 8:109–114.

Dobler IM, Bäuml KH. 2013. Retrieval-induced forgetting: dynamic effects between retrieval and restudy trials when practice is mixed. Mem Cogn. 41:547–557.

Fischl B, van der Kouwe A, Destrieux C, Halgren E, Ségonne F, Salat DH, Busa E, Seidman LJ, Goldstein J, Kennedy D, et al. 2004. Automatically parcellating the human cerebral cortex. Cereb Cortex. 14:11–22.

Gluck MA, Myers CE. 1993. Hippocampal mediation of stimulus representation: a computational theory. Hippocampus. 3:491–516.

Ketz N, Morkonda SG, O'Reilly RC. 2013. Theta coordinated error-driven learning in the hippocampus. PLoS Comput Biol. 9: e1003067.

Kriegeskorte N, Mur M, Bandettini PA. 2008. Representational similarity analysis—connecting the branches of systems neuroscience. Front Neurosci. 2:4.

LaRocque KF, Smith ME, Carr VA, Witthoft N, Grill-Spector K, Wagner AD. 2013. Global similarity and pattern separation in the human medial temporal lobe predict subsequent memory. J Neurosci. 33:5466–5474.

MacLeod MD, Hulbert JC. 2011. Retrieval inhibition, sleep, and the resolving power of human memory. In: Benjamin AS, editor. Successful Remembering and Successful Forgetting: Essays in Honor of Robert A. Bjork. North-Holland: Elsevier. p. 133–152.

MacLeod MD, Macrae CN. 2001. Gone but not forgotten: the transient nature of retrieval-induced forgetting. Psychol Sci. 12:148–152.

Marr D. 1971. Simple memory: A theory for archicortex. Philos T Roy Soc B. 262:23–81.

McClelland JL, McNaughton BL, O'Reilly RC. 1995. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. Psychol Rev. 102:419–457.

Meng X-L, Rosenthal R, Rubin DB. 1992. Comparing correlated correlation coefficients. Psychol Bull. 111:172–175.

Norman KA, Newman EL, Detre G. 2007. A neural network model of retrieval-induced forgetting. Psychol Rev. 114:887–953.

Norman KA, Newman E, Detre G, Polyn S. 2006. How inhibitory oscillations can train neural networks and punish competitors. Neural Comput. 18:1577–1610.

Norman KA, Newman EL, Perotte AJ. 2005. Methods for reducing interference in the complementary learning systems model: oscillating inhibition and autonomous memory rehearsal. Neural Netw. 18:1212–1228.

Norman KA, O'Reilly R. 2003. Modeling hippocampal and neocortical contributions to recognition memory: a complementary learning systems approach. Psychol Rev. 110:611–646.

O'Reilly RC, McClelland JL. 1994. Hippocampal conjunctive encoding, storage, and recall: avoiding a trade-off. Hippocampus. 4:661–682.

O'Reilly RC, Munakata Y. 2000. Computational Explorations in Cognitive Neuroscience: Understanding the Mind by Simulating the Brain. Cambridge, MA, US: The MIT Press.

O'Reilly RC, Rudy JW. 2001. Conjunctive representations in learning and memory: principles of cortical and hippocampal function. Psychol Rev. 108:311–345.

Pereira F, Botvinick M. 2013. Simitar: Simplified searching of statistically significant similarity structure. P Int Workshop Pattern Recogn Neuroimag. 1–4. doi:10.1109/PRNI.2013.10.

Racsmány M, Conway MA, Demeter G. 2010. Consolidation of episodic memories during sleep: long-term effects of retrieval practice. Psychol Sci. 21:80–85.

Roediger HL, Butler AC. 2011. The critical role of retrieval practice in long-term retention. Trends Cogn Sci. 15:20–27.

Schapiro AC, Kustner LV, Turk-Browne NB. 2012. Shaping of object representations in the human medial temporal lobe based on temporal regularities. Curr Biol. 17:1622–1627.

Schapiro AC, Rogers T, Norman KA, Chen L, McDevitt E, Mednick S. 2013. The role of sleep in consolidating semantic knowledge. J Vis. 13:666–666.

Shiffrin RM, Ratcliff R, Clark SE. 1990. List-strength effect: II. Theoretical mechanisms. J Exp Psychol Learn. 16:179–195.

Shiffrin RM, Steyvers M. 1997. A model for recognition memory: REM-retrieving effectively from memory. Psychon B Rev. 4:145–166.

Storm BC, Bjork EL, Bjork RA. 2008. Accelerated relearning after retrieval-induced forgetting: the benefit of being forgotten. J Exp Psychol Learn. 34:230–236.

Yassa MA, Stark CEL. 2011. Pattern separation in the hippocampus. Trends Neurosci. 34:515–525.

Weber M, Thompson-Schill SL, Osherson D, Haxby J, Parsons L. 2009. Predicting judged similarity of natural categories from their neural representations. Neuropsychologia. 47:859–868.

Weller PD, Anderson MC, Gómez-Ariza CJ, Bajo MT. 2012. On the status of cue independence as a criterion for memory inhibition: evidence against the covert blocking hypothesis. J Exp Psychol Learn. 39:1232–1245.

Willenbockel V, Sadr J, Fiset D, Horne G, Gosselin F, Tanaka J. 2010. Controlling low-level image properties: the SHINE toolbox. Behav Res Meth. 42:671–684.

Zeineh MM, Engel SA, Thompson PM, Bookheimer SY. 2003. Dynamics of the hippocampus during encoding and retrieval of face-name pairs. Science. 299:577–580.