Behavioral/Cognitive

Neural Differentiation of Incorrectly Predicted Memories

Ghootae Kim, Kenneth A. Norman, and Nicholas B. Turk-Browne
Department of Psychology and Neuroscience Institute, Princeton University, Princeton, New Jersey 08544

When an item is predicted in a particular context but the prediction is violated, memory for that item is weakened (Kim et al., 2014). Here, we explore what happens when such previously mispredicted items are later reencountered. According to prior neural network simulations, this sequence of events—misprediction and subsequent restudy—should lead to differentiation of the item's neural representation from the previous context (on which the misprediction was based). Specifically, misprediction weakens connections in the representation to features shared with the previous context and restudy allows new features to be incorporated into the representation that are not shared with the previous context. This cycle of misprediction and restudy should have the net effect of moving the item's neural representation away from the neural representation of the previous context. We tested this hypothesis using human fMRI by tracking changes in item-specific BOLD activity patterns in the hippocampus, a key structure for representing memories and generating predictions. In left CA2/3/DG, we found greater neural differentiation for items that were repeatedly mispredicted and restudied compared with items from a control condition that was identical except without misprediction. We also measured prediction strength in a trial-by-trial fashion and found that greater misprediction for an item led to more differentiation, further supporting our hypothesis. Therefore, the consequences of prediction error go beyond memory weakening. If the mispredicted item is restudied, the brain adaptively differentiates its memory representation to improve the accuracy of subsequent predictions and to shield it from further weakening.

Key words: episodic memory; fMRI; neural network

Significance Statement

Competition between overlapping memories leads to weakening of nontarget memories over time, making it easier to access target memories. However, a nontarget memory in one context might become a target memory in another context. How do such memories get restrengthened without increasing competition again? Computational models suggest that the brain handles this by reducing neural connections to the previous context and adding connections to new features that were not part of the previous context. The result is neural differentiation away from the previous context. Here, we provide support for this theory, using fMRI to track neural representations of individual memories in the hippocampus and how they change based on learning.

Introduction

When faced with a familiar situation, we can often predict who or what will appear. What happens to the memories supporting these predictions when they turn out to be wrong? We previously found that incorrect prediction of an item in a familiar context leads to worse subsequent memory for that item (Kim et al., 2014). Through this process, the brain might prune memories that correspond to changed or unstable aspects of the environment. However, an item that is irrelevant in one situation might later become relevant in another situation. In this case, the previously weakened memory needs to regain its mnemonic strength. How does the brain accomplish this goal? Based on our previous network modeling work (Norman et al., 2006, 2007) and empirical findings (Hulbert and Norman, 2015), we propose that the brain resolves this by adaptively differentiating the memory from its previous context.

The model's predictions are illustrated in Figure 1. Consider two items, A and B, that have been paired with each other previously such that the appearance of A leads to a prediction of B. However, on this particular trial, B does not appear. Memory A is strongly activated (because it was just shown) and memory B is moderately activated (because it is being predicted from memory). A key postulate in the model is that moderate activation weakens connections from other, strongly activated features (Norman et al., 2006, 2007). Therefore, connections from the strongly activated features of A to the moderately activated features of B (those that are not shared with A) are weakened, effectively "shearing" these unique features of B from features shared

Received Oct. 22, 2016; revised Dec. 19, 2016; accepted Jan. 13, 2017.

Author contributions: G.K., K.A.N., and N.B.T.-B. designed research; G.K. performed research; G.K. analyzed data; G.K., K.A.N., and N.B.T.-B. wrote the paper.

This work was supported by the National Institutes of Health (Grant R01 MH069456 to K.A.N. and N.B.T.-B. and Grant R01 EY021755 to N.B.T.-B.

The authors declare no competing financial interests.

Correspondence should be addressed to Dr. Ghootae Kim, Department of Psychology and Neuroscience Institute, Princeton University, Peretsman-Scully Hall, Princeton, NJ 08544. E-mail: kimghootae@gmail.com.

DOI:10.1523/JNEUROSCI.3272-16.2017

 0270-6474/17/372022-10$15.00/0

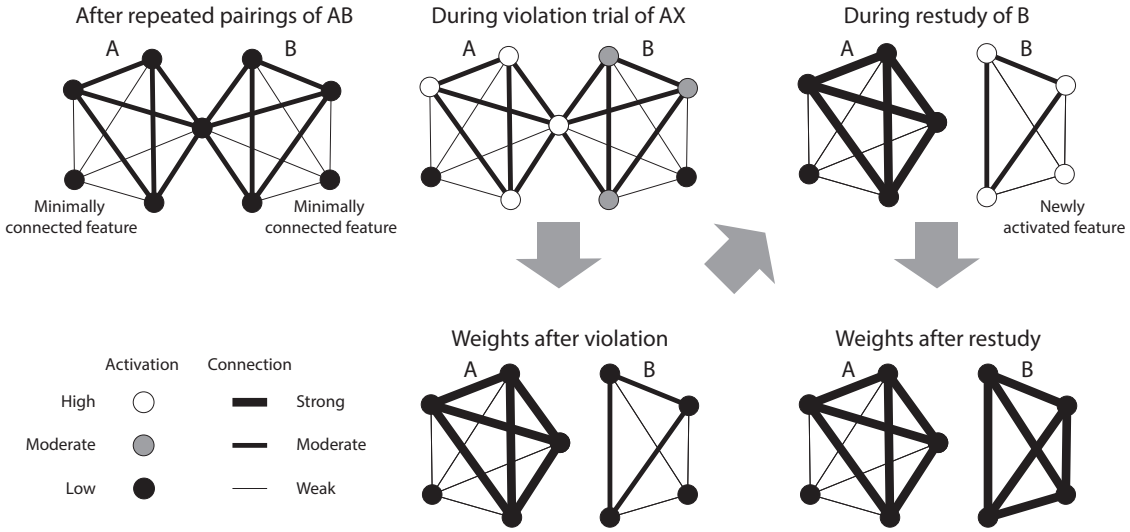

Figure 1. How interleaved misprediction and restudy lead to differentiation. A and B have been paired previously (AB), but in this instance of A, B does not appear (AX). This unconfirmed prediction leads to moderate activation of the features of B. According to our theory (Norman et al., 2006, 2007; Hulbert and Norman, 2015), this leads to weakening of connections into these moderately active features from other, strongly activated features (including features shared with A). If the B item is restudied later after a novel item, activation spreads to new features that were not formerly activated by A, resulting in strengthening of these connections to new features. This cycle, whereby misprediction of B causes shearing of B features from A features and restudy leads to acquisition of new features, has the overall effect of differentiating B's neural representation from A's neural representation.

with A. This weakening of connections within the B representation is a possible explanation for our previous findings of impaired recognition of mispredicted items (Kim et al., 2014). Crucially, if B is restudied later, then the unique features of B will be activated, but not features previously shared with A (because connections to these features were weakened). Instead, activation will spread to other new features (not previously shared with A) and connections to these features will be strengthened. This process of swapping out shared for unshared features decreases overlap between the A and B memories.

Other studies have demonstrated neural differentiation from learning of interrelated materials (Schapiro et al., 2012; Schlichting et al., 2015; Favila et al., 2016). The key contribution of the present work is that we provide a mechanism for differentiation (described above). Our proposed mechanism leads to two specific, testable claims that go beyond basic differentiation: First, the mispredicted item should specifically differentiate from its prior context (as opposed to becoming generally more distinct from other items). Second, across items, the degree of misprediction should relate to the degree of differentiation (insofar as misprediction leads to shearing off of shared features, opening the door for new features to be acquired).

We tested these two hypotheses in an fMRI study. Observers were exposed to a continuous sequence of scenes and faces while performing a cover task. Unbeknownst to them, this sequence was generated from pairs (e.g., scene A–scene B), creating an expectation that B will follow A. For some pairs, these expectations were violated (A was followed by X instead of B). All B items were subsequently restudied. We hypothesized that misprediction of B followed by its restudy would lead to differentiation of B from A compared with a control condition consisting of pairs in which expectations were not violated. To test this, we used fMRI and pattern similarity analysis to track changes in neural overlap between A and B and to track how strongly B was predicted on violation trials. Our results showed a direct relationship between competitive dynamics (i.e., misprediction) during learning and representational change, thereby supporting for our mechanistic model of memory updating.

Materials and Methods

Participants. Thirty-two adults (19 women, 27 right-handed, mean age 20.1 years) participated for monetary compensation. All had normal or corrected-to-normal vision and provided informed consent. The Princeton University Institutional Review Board approved the study protocol.

Stimuli. Participants were shown color photographs of indoor and outdoor scenes (including some from http://cvcl.mit.edu/MM/sceneCategories.html), male and female faces (including some from www.macbrain.org/resources.htm), and natural and manmade objects. Stimuli were projected on a screen behind the scanner and viewed with a mirror on the head coil (subtending 8.8 × 8.8°). Participants fixated a black central dot that changed to white when a response was recorded.

Procedure. The experimental procedure unfolded over 2 d (Fig. 2*A*). All phases of the experiment were scanned with fMRI. The first session consisted of six runs of an incidental encoding task. Participants made indoor/outdoor judgments for scenes that were presented for most of trials (92%) and female/male judgments for faces that were presented occasionally (8%). Unbeknownst to participants, the stimulus sequence contained scene image pairs (e.g., scene A–scene B). There were 8 new pairs for each of the 2 conditions (violation and nonviolation) within each run (8 pairs × 6 runs = 48 pairs in total per condition). The first and second members of each pair were presented once separately (randomly intermixed with items from other pairs) before they were ever shown together in a pair. This allowed us to measure the neural representation of each item on its own before learning ("prelearning snapshot") and uncontaminated by the pair-mate (Schapiro et al., 2012).

There was no explicit distinction between the prelearning snapshots and the presentation of pairs. After the items in a given pair were shown separately (to acquire prelearning snapshots), the items were shown as a pair three times, interleaved with repetitions of other pairs. For pairs assigned to the violation condition, the three repetitions were followed by a violation event where the first scene in the pair was followed by a novel face instead of the paired scene (e.g., scene A–face X); this event was omitted for nonviolation pairs. Crucially, in both conditions, the B item was subsequently presented ("restudied") on its own, following a novel item in the sequence rather than its previous context A. This cycle of violation and restudy was repeated once more for the violation pairs—our modeling work suggested that two cycles would produce more differentiation than one—leading to the following overall event sequence per pair: AB–AB–AB–AX–B–AY–B. There was also a second restudy for

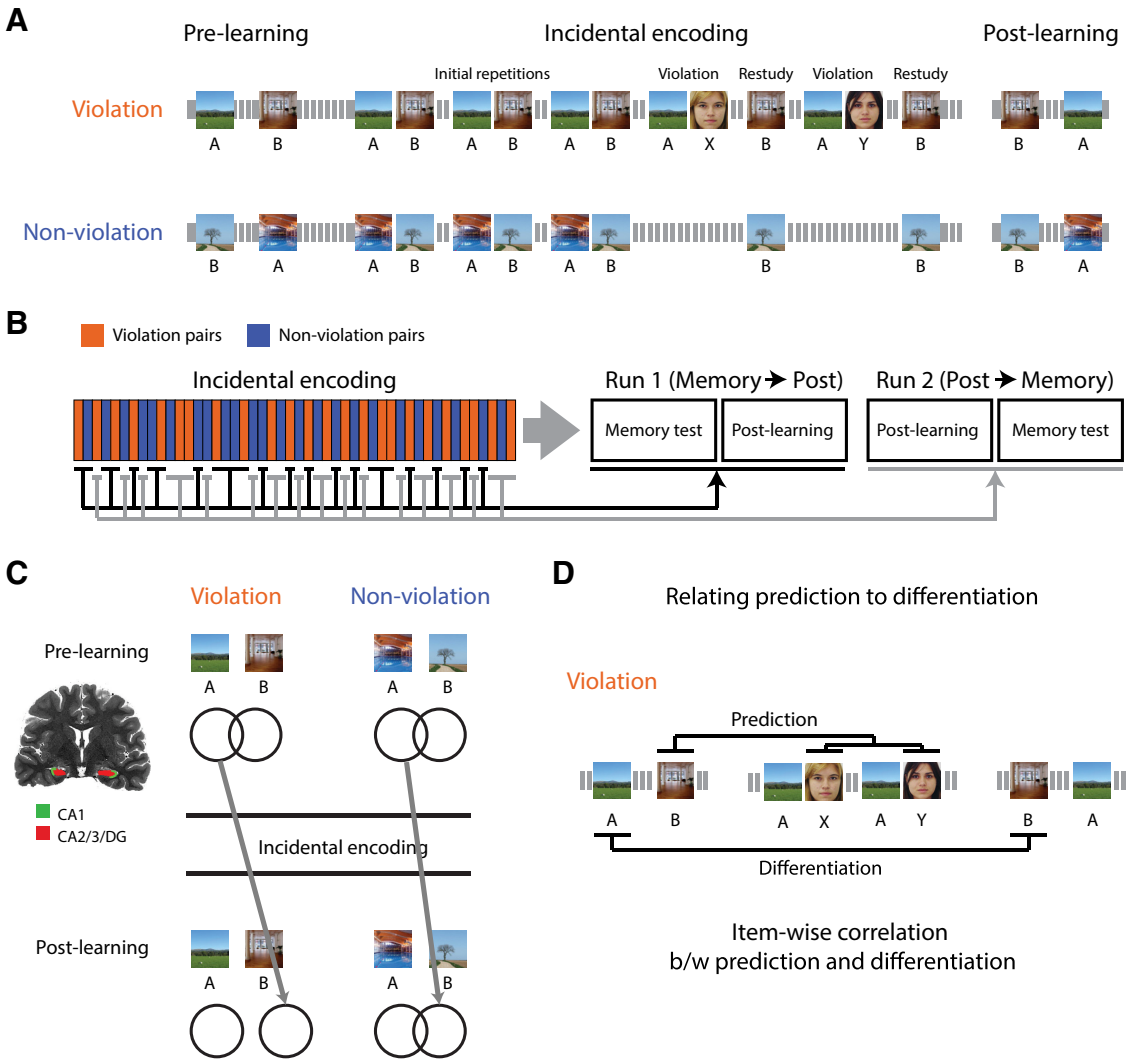

Figure 2. Experimental design and main analyses. ***A***, During incidental encoding, participants performed a categorization task for scenes (indoor/outdoor) and faces (male/female). Before the main incidental encoding phase, all scene images were presented in a random order so that we could take prelearning snapshots of the neural representations of these images. During the incidental encoding phase, the trial sequence for the violation condition was constructed from three initial repetitions of AB pairs (AB–AB–AB) and two cycles of violation and restudy trials (AX–B–AY–B), whereas the violation trials were omitted for the control (nonviolation) condition (AB–AB–AB–B–B). Each individual trial in these sequences was interleaved with trials of other pairs at different points in their sequences. After the main incidental encoding phase, the same scene images were presented again in a random order, allowing us to take postlearning snapshots of their neural representations. ***B***, Half of the pairs in the incidental encoding phase were randomly assigned to memory and postlearning run 1 (memory run preceded postlearning run) and the other half to memory and postlearning run 2 (postlearning run preceded memory run). The order of run 1 and 2 was counterbalanced across participants. ***C***, In left CA2/3/DG, we measured neural differentiation by computing the similarity of the prelearning snapshot of A and the postlearning snapshot of the corresponding B item in the violation condition versus the control condition. ***D***, We measured prediction strength in the violation condition by computing the similarity of the prelearning snapshot of B to neural patterns present at the moment of the violation (when X/Y items were presented). We then computed the correlation between prediction strength and neural differentiation in a trial-by-trial fashion.

the nonviolation pairs, leading to a sequence that was matched for the number of exposures to B, but without violation events: AB–AB–AB–B–B. Importantly, these manipulations were incidental to the primary task of making categorical judgments and the interleaving of multiple pairs obscured the pair structure. In a separate behavioral pilot, we encouraged participants to report any regularity and none noticed that the sequence was constructed from repeated pairs. Each trial began with a blink of the fixation cross to signal an upcoming stimulus, followed by the stimulus presentation for 1 s and a blank interval of 2 s. There were a total of 192 trials (32 prelearning, 160 pair sequences) within each run, which lasted 10 min and 6 s.

The second session occurred the day after the first session (at least 12 h) and consisted of three tasks: postlearning snapshot (two runs), surprise memory test (two runs), and functional localizer (three runs). The postlearning snapshots were collected in the same manner as the prelearning snapshots: all scenes from the first session were shown again, one at a time, in a random order and with indoor/outdoor judgments. In the recognition memory test, we presented each participant with all B scenes from both conditions (48 violation and 48 nonviolation) randomly intermixed with 48 novel scenes. A two-step memory response was made for each scene ("old" or "new") and then confidence level ("sure" or "unsure"). Images remained on the screen until the responses were made and the next trial began on the first subsequent TR.

The four postlearning and memory runs were interleaved and their order was counterbalanced across participants (odd participants: memory run 1, postlearning run 1, postlearning run 2, memory run 2; even participants: postlearning run 1, memory run 1, memory run 2, postlearning run 2). Half of the studied pairs were randomly assigned to memory run 1 and postlearning run 1 and the other half to memory run 2 and postlearning run 2. As a result of this design, for half of the B items, memory was tested before the postlearning snapshot was taken and, for the other half of B items, the postlearning snapshot was taken before memory was tested (Fig. 2*B*). We designed the procedure this way because we were concerned that testing memory for an item could contam-

inate the subsequent postlearning snapshot for that item and vice versa. We wanted some items to get an uncontaminated postlearning snapshot (before memory was tested). Although behavioral recognition memory performance was not the focus of this study, we also wanted some items to get an uncontaminated memory judgment (before the snapshot was taken). For the analyses focusing on neural differentiation (and related follow-up analyses), we used only the pairs for which the postlearning snapshot preceded the recognition memory test, thereby ensuring that the postlearning snapshot would not be affected by any learning that occurred as a result of recognition memory test.

After the postlearning and memory runs, participants completed three runs of a functional localizer. Each run contained 15 blocks, with five blocks from each of three categories: faces, scenes, and objects. Participants judged faces as male or female, scenes as indoor or outdoor, and objects as manmade or natural. Each stimulus was presented for 500 ms, followed by a blank interval of 1000 ms. There were 10 trials per block and each 15 s block was followed by 15 s of fixation, treated as a rest category. The duration of each run was 465 s. For the analyses below, we used only the scene blocks. Specifically, we calculated a "template" activity pattern for the scene category from the scene blocks; this template was later used to evaluate whether our results were item specific.

Data acquisition. The experiment was programmed in MATLAB using the Psychophysics Toolbox (http://psychtoolbox.org). MRI data were acquired with a 3T MRI scanner (Siemens Prisma) with a 64-channel head coil. Functional scans came from a T2*-weighted multiband EPI sequence (TR = 1.5 s, TE = 39 ms, 128 × 128 matrix, voxel size = 1.5 mm iso, 192 mm field of view, flip angle = 50°, acceleration factor 4, shift 3), with 52 oblique axial slices (transverse to the long axis of the hippocampus) acquired in an interleaved order. These slices covered a partial volume encompassing the temporal and occipital lobes, which allowed us to maximize spatial and temporal resolution over hippocampal subfields. A whole-brain T1 MPRAGE image and a coplanar T1 FLASH image were acquired for registration to other participants. Two T2 TSE images were acquired for probabilistic segmentation of hippocampal subfields (54 slices perpendicular to the long axis of the hippocampus; 0.44 × 0.44 mm in-plane, 1.8 mm thick). A field map was acquired to correct for B0 inhomogeneity.

Regions of interest (ROIs). This study involves rapidly learning new, arbitrary associations between stimuli. Because the hippocampus supports the encoding and retrieval of such memories (Davachi, 2006; Norman and O'Reilly, 2003; Schapiro et al., 2016), we expected that representational changes would occur in the hippocampus. Based on a previous study of neural differentiation (Hulbert and Norman, 2015), we focused specifically on the left hippocampus. Within the hippocampus, we were primarily interested in the CA3 and dentate gyrus (DG) subfields because they are core storage sites for episodic memories and generate predictions via pattern completion (Hindy et al., 2016).

Hippocampal segmentation. Subfields of the hippocampus (CA1 and CA2/3/DG) were defined probabilistically in MNI space based on a database of manual hippocampal segmentations from a separate set of 24 participants. Manual segmentations were created on T2-weighted TSE images using anatomical landmarks (Duvernoy, 2005; Carr et al., 2010; Aly and Turk-Browne, 2016) and then registered to an MNI template. Voxels in the MNI template were assigned subfield labels if the probability was >0.5. Nonlinear registration (FNIRT; Andersson et al., 2007) was used to register each participant's high-resolution MPRAGE to this probabilistic label atlas. Subfields were extracted separately from each hemisphere and merged for bilateral analyses.

Preprocessing. fMRI data were preprocessed with FSL (http://fsl.fmrib.ox.ac.uk). Functional scans were corrected for slice acquisition time and head motion, high-pass filtered (128 s period cutoff), and aligned to the middle volume within each run. As a second alignment step, all preprocessed images in the first session were aligned to the first volume of the first functional run. Functional scans from the second session were aligned to the same volume based on first aligning the MPRAGE scans across sessions.

Measuring differentiation. For our initial analysis of overall differentiation, we measured how much B's neural representation after learning moved away from the original representation of A and whether this differed for violation and nonviolation pairs. Specifically, we measured the Pearson correlation between the prelearning snapshot of A and the postlearning snapshot of B and then transformed it to Fisher's *z* (Fig. 2*B*). These snapshots were defined by the spatial pattern of activity elicited by each item in a particular ROI (e.g., left CA2/3/DG) at the peak of the hemodynamic response (4.5 s after image onset). This approach differs slightly from previous studies of representational change that used similar neural snapshot methods (Schapiro et al., 2012; Favila et al., 2016). For example, Favila et al. (2016) measured neural overlap between competing items within a postlearning phase relative to noncompeting control items. Such an approach was possible because there was no difference between items in that phase other than their learning history. However, in our study, A in the violation condition was presented two times more than A in the nonviolation condition (because of AX and AY violation trials), so any comparison between violation and nonviolation conditions that includes the postlearning snapshot of A is confounded by item frequency. Therefore, we used the prelearning snapshot of A (before any difference between conditions) as the baseline for representational change.

Our hypothesis posits that differentiation effects should be item specific. Weakening of connections between the shared features of A/B and the unique features of B (as a result of misprediction) allows for the subsequent addition of new features to B when it is restudied and this leads to an overall decrease in neural overlap between A and B. In other words, it is important for our hypothesis that B become more distinct from A specifically, not just generically more distinct from other items. The basic measure of differentiation above is consistent with both possibilities, so we performed an additional randomization analysis to verify item specificity: We scrambled the pair assignments of A and B 1000 times within each condition and recalculated neural differentiation. That is, if A_i and B_i indicate that these items were from the same pair (i), the original analysis involves calculating differentiation between A_1 and B_1, A_2 and B_2, A_3 and B_3, etc., whereas a given permutation in the randomization test might compare A_1 and B_7, A_2 and B_4, A_3 and B_2, etc. If differentiation occurs in a generic sense, then the A items are interchangeable and the original effect will not differ from the permuted distribution. If, as predicted by our model, differentiation is item specific, the original magnitude of differentiation should be larger than the permuted distribution. For statistical analysis, within each participant, we measured the difference in pattern similarity of prelearning A to postlearning B between the violation and nonviolation conditions for both the correct pairing and the permuted pairings and computed the *z*-score of the true difference relative to the mean and SD of 1000 permuted differences. We then examined the reliability of these *z*-scores across participants with a one-sample *t* test against zero.

Relating prediction to differentiation. Beyond showing item-specific neural differentiation, a key goal of our study is to provide an explanation for how it arises, as a result of misprediction. The violation and nonviolation conditions only differed in the presence of violation trials that allowed for misprediction; therefore, any difference in overlap between these conditions (as captured in the differentiation measure described above) can be attributed to misprediction. Nevertheless, we sought stronger and more continuous evidence by attempting to relate, on a trial-by-trial basis, the amount of prediction on violation trials to the amount of subsequent differentiation. This analysis was performed only for the violation condition because there were no violation trials in the nonviolation condition. To measure prediction on violation trials, we calculated the amount of evidence for B during the presentation of unexpected items X and Y (i.e., the items that appeared after A, when B should have been presented). Specifically, we measured pattern similarity between the prelearning snapshot of B (not used to calculate differentiation) and the pattern of activity evoked by both X and Y events and then averaged these two similarities to provide a single index of prediction for each AB pair. Across pairs, we then calculated the correlation of this prediction score with the pair-specific neural differentiation effect within participant and then examined its reliability at the group level using a one-sample *t* test.

Again, our hypothesis about a relationship between prediction strength and differentiation is item-specific; the activation of B in partic-

ular is what induces competition with, and differentiation from, A. However, the analysis above does not guarantee that the correlation is item-specific. Pattern similarity between the prelearning snapshot of B and X/Y could reflect prediction of item-specific features of B or more generic categorical features of B shared across items (i.e., predicting that a scene will appear, without specifying the scene). We conducted two further analyses to address this. First, we performed the same type of randomization test used above for the main effect of differentiation. Specifically, if prediction reflects generic scene activation, then the prelearning snapshots of different B items should be interchangeable when calculating pattern similarity with X and Y. Therefore, we scrambled the original pairings of presnapshot B and X/Y 1000 times, recalculated their pattern similarity, and then recomputed the prediction–differentiation relationship. For example, if the original prediction scores were derived by correlating pre-B_1 with X_1/Y_1 (where X_1 and Y_1 followed A_1) and the original differentiation scores were derived by correlating pre-A_1 with post-B_1, a permutation might involve recomputing the prediction scores (comparing pre-B_7 to X_1/Y_1, etc.) but keeping the differentiation scores the same; the recomputed prediction scores were then correlated with the differentiation scores, yielding a new (null) value for the prediction–differentiation relationship. As before, a *z*-score for the original prediction–differentiation relationship was calculated relative to the permuted distribution for this relationship within participant and these *z*-scores were assessed for reliability with a one-sample *t* test across participants. According to our hypothesis (but not a generic scene prediction account), permuting the items in this way should abolish the relationship between prediction and differentiation.

Second, we used regression to remove generic category-level information from the activity patterns before calculating pattern similarity. Specifically, we defined a template activity pattern for the scene category by averaging over the many scene images in the localizer; we then regressed this template separately from each of the patterns used for this analysis (i.e., we scaled the scene template to maximize fit to the observed pattern and then took the residuals). By definition, the residual patterns after this regression are orthogonal to the scene template, thereby reducing the possibility that generic scene prediction drove the prediction score. As the final step, we repeated the original prediction–differentiation analysis with these residuals. According to our hypothesis, the relationship should be preserved.

Searchlight analysis. Given the nature of learning in our study, we expected that representational change would occur in the hippocampus, so we performed the above analyses in hippocampal subfields. However, it is still possible that other brain regions might show these effects. To address this issue, we performed an exploratory searchlight analysis.

First, for every subject, we swept a 10.5-mm-voxel cubic searchlight (radius = 3 voxels, volume = 343 voxels, except along boundaries) throughout the EPI volume, which covered the temporal and occipital lobes (see "Data acquisition" section). In each of the searchlights, we computed neural overlap between the prelearning snapshot of A and the postlearning snapshot of B for the violation versus nonviolation conditions and then measured the trial-by-trial relationship between the amount of prediction (on violation trials) and subsequent neural overlap scores in the violation condition. We assigned the final two outcomes (i.e., the average neural overlap score and the correlation coefficient between prediction and neural overlap) to the center voxel of each searchlight. The two resulting maps were then transformed to the MNI standard template (2 mm isotropic) and masked to exclude white matter. We then examined the reliability of each of the two analyses across participants by applying a one-sample *t* test (against zero) for every voxel. Finally, we selected voxels for which the *p*-values of both analyses were <0.05 (uncorrected).

Results

Behavioral performance

Judgments in the categorization cover task (indoor/outdoor for scenes, male/female for faces) were accurate in general (mean = 90.51%, SD = 10.91; vs chance: $t_{(31)} = 21.00$, $p < 0.001$) and did not differ across the prelearning snapshots, pair sequences, or postlearning snapshots ($F < 1$). For the recognition memory test, we restricted analysis to the B items that were tested before the postlearning snapshot was taken. Performance was good in general, with B items more likely to be endorsed as "old" than new items (mean $A' = 0.84$, $t_{(31)} = 26.17$, $p < 0.001$). We did not have a specific expectation for a difference in memory between violation and nonviolation conditions because neural differentiation can have opposite effects on the underlying processes that support recognition memory (see Discussion). Indeed, there was no difference between conditions ($t_{(31)} = -0.24$, $p = 0.81$). Neural analyses were restricted to the other B items, the postlearning snapshots of which were collected before the (potentially contaminating) memory test.

Neural differentiation

We examined how much the neural representation of the B items moved away from their initial A context items by measuring pattern similarity between the postlearning snapshot of B and the prelearning snapshot of A. We focused initially on left CA2/3/DG ROI, given prior findings of differentiation (Hulbert and Norman, 2015). We hypothesized that misprediction of B items in the violation condition would reduce subsequent neural overlap with A after restudy. Indeed, pattern similarity was lower for pairs from the violation versus nonviolation conditions ($t_{(31)} = -2.82$, $p = 0.008$; Fig. 3*A*). We conducted additional exploratory analyses outside of the main ROI, in other potentially relevant hippocampal subfields, and in both a bilateral and unilateral manner: right and bilateral CA2/3/DG and left, right, and bilateral CA1. None of these regions showed a reliable difference between conditions (right CA2/3/DG: $t_{(31)} = 0.69$, $p = 0.50$; bilateral CA2/3/DG: $t_{(31)} = -1.39$, $p = 0.17$; left CA1: $t_{(31)} = 0.31$, $p = 0.76$; bilateral CA1: $t_{(31)} = 1.22$, $p = 0.23$) and right CA1 actually showed a trend in the opposite direction from left CA2/3/DG ($t_{(31)} = 1.88$, $p = 0.07$).

Our hypothesis explains this neural differentiation in terms of swapping out shared features with A for new features of B. Therefore, the representational change should be specific to B's relationship with A and not reflective of increased distinctiveness from other items in general. We evaluated this possibility by permuting the relationship between the prelearning snapshots of A and the postlearning snapshots of B (Fig. 3*B*). If neural differentiation is item specific, then this scrambling should eliminate the difference in pattern similarity between violation and nonviolation conditions. Indeed, consistent with our model, the differentiation effect in left CA2/3/DG was stronger for the correct AB pairings than the null distribution of permuted pairings ($t_{(31)} = -2.84$, $p = 0.008$; Fig. 3*C*).

Relationship between prediction and differentiation

The analysis above relies on the categorical manipulation of condition (violation vs nonviolation) to establish that misprediction is responsible for differentiation upon restudy. To provide further support for this hypothesized mechanism, we conducted continuous analyses within the violation condition. According to our model, greater prediction of B during the AX/AY violation trials induces more weakening of its shared features with A, which in turn allows for better subsequent acquisition of new features during restudy of B in a novel context. Accordingly, in left CA2/3/DG, there should be a negative relationship between the amount of prediction of B on violation trials (pattern similarity of X/Y with the prelearning B snapshot) and the strength of neural overlap for the corresponding pair (pattern similarity between prelearning A and postlearning B snapshots). Indeed, as shown in

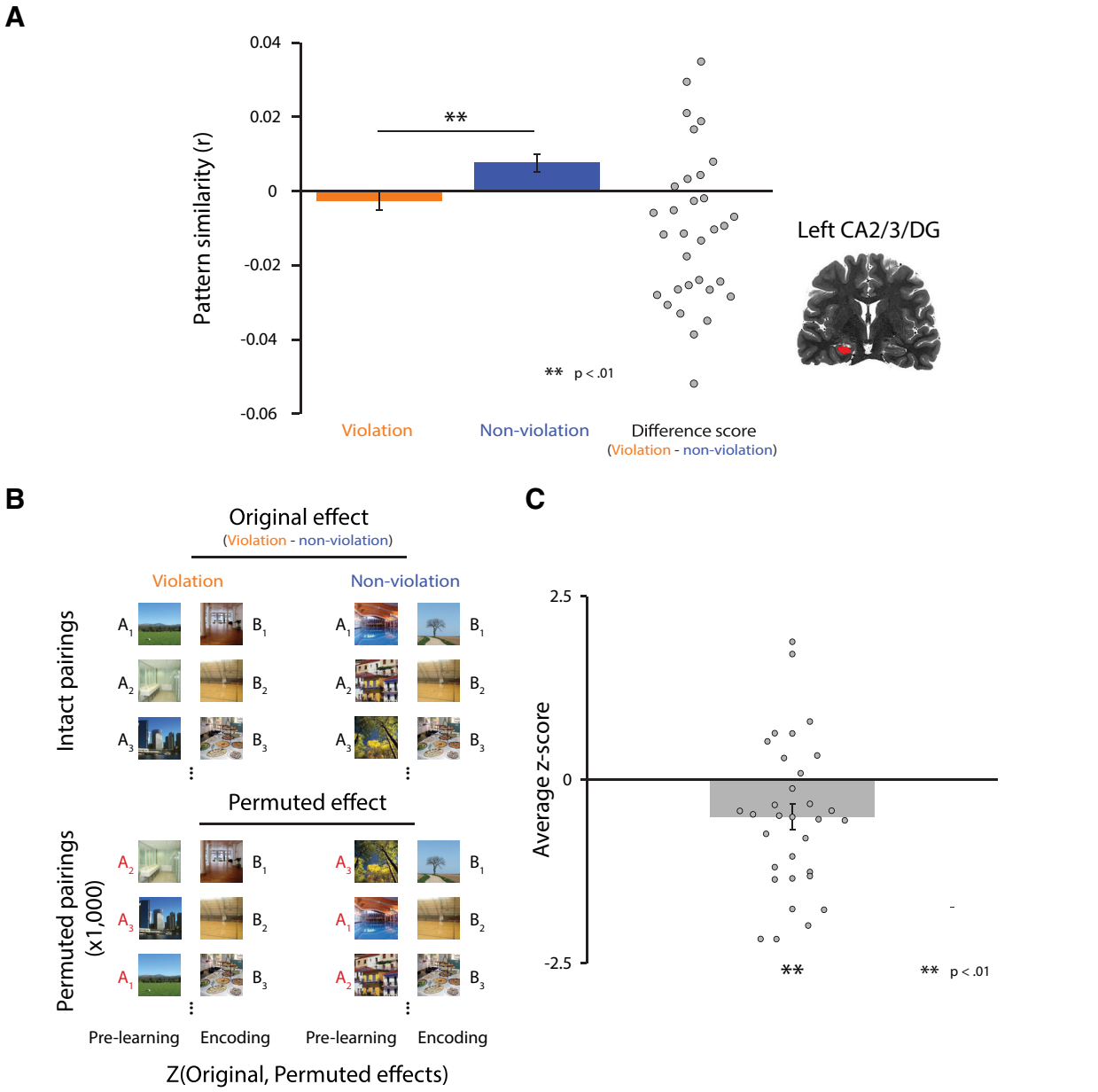

Figure 3. Neural differentiation effect and item specificity test in left CA2/3/DG. ***A***, Pattern similarity between the prelearning snapshot of A and the postlearning snapshot of B was significantly lower for the violation versus nonviolation conditions. ***B***, Schematic of procedure for randomization analysis. For each participant, we shuffled the AB pairings 1000 times within each condition. In each iteration, we calculated the same prelearning A to postlearning B pattern similarity for each condition and stored the difference between conditions. This produced a null distribution of differences and we calculated the *z*-score of the original effect with respect to this distribution. The reliability of the *z*-scores was assessed across participants. ***C***, Original differentiation effect based on the intact AB pairings was reliably lower compared with the null distribution acquired from permuted AB pairings. This result is consistent with the neural differentiation effect being item specific.

Figure 4*A*, there was a reliable negative correlation between these measures ($t_{(31)} = -2.61$, $p = 0.01$).

As with the differentiation effect, our model also posits that prediction of B *per se* is critical because more generic prediction (e.g., of a scene) would not specifically weaken the features of B. We evaluated this possibility using two different analyses: First, we performed a randomization test by permuting the prelearning B snapshots when calculating prediction during violation trials (under the null hypothesis that the B items are interchangeable) and then recomputed the trial-by-trial relationship with differentiation (Fig. 4*B*). Consistent with item-specific prediction being the critical ingredient, the relationship between prediction strength and differentiation was stronger for the original pairings of B items and violation trials relative to the null distribution from permuted pairings ($t_{(31)} = -2.42$, $p = 0.02$; Fig. 4*C*). Second, we regressed generic category-level information out of the prelearning B snapshots and violation trials before calculating prediction scores (thus attenuating the contribution of non-item-specific features of B to pattern similarity) and then reran the analysis. As expected, the negative relationship between prediction and differentiation persisted ($t_{(31)} = -2.47$, $p = 0.02$).

Searchlight results

The above analyses were performed within hippocampal subfields. However, representational changes may have also occurred in other brain regions. To examine this possibility, we swept a cubic searchlight through a functional volume covering the temporal and occipital lobes. For each of the searchlights, we

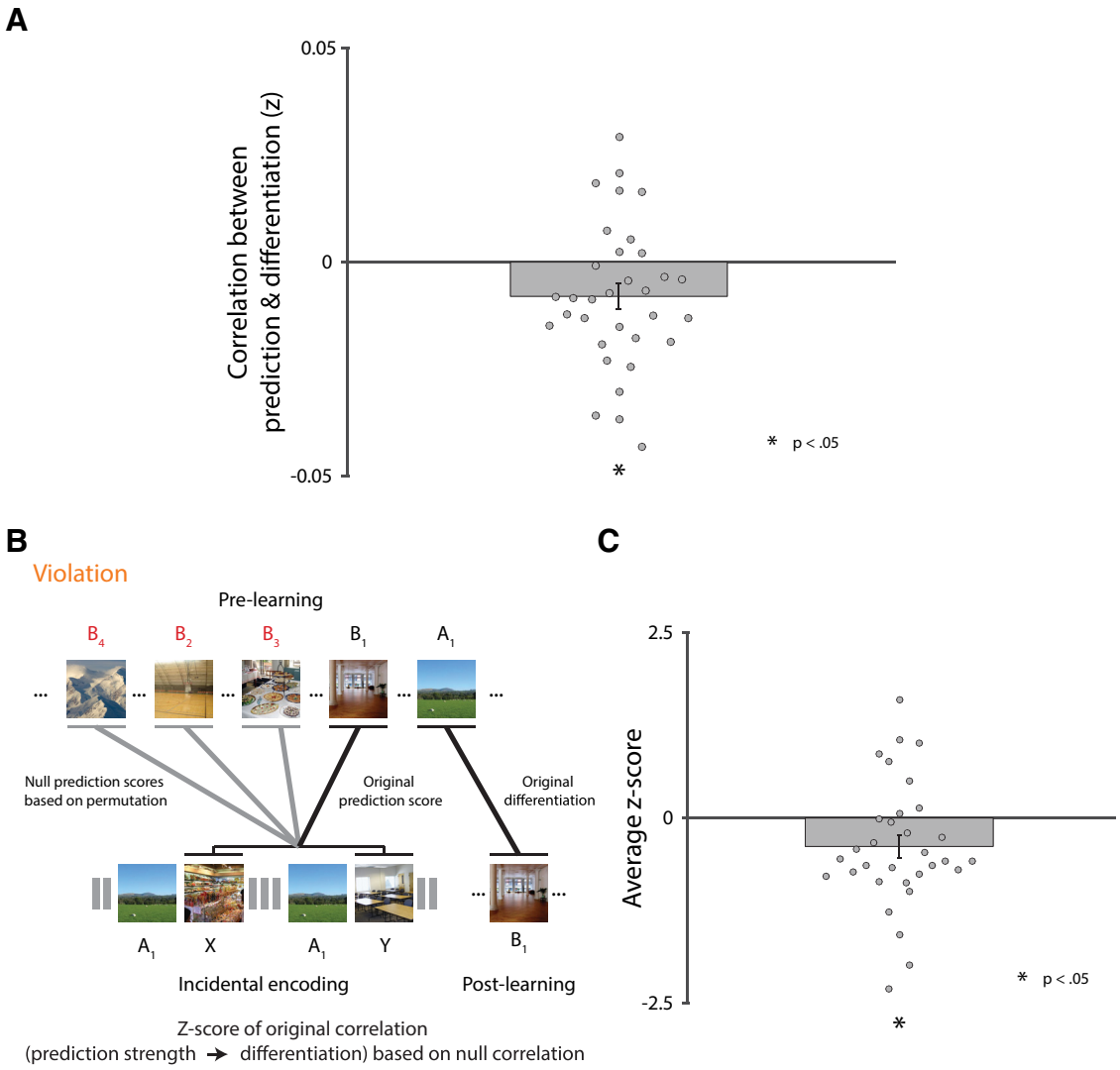

Figure 4. Relationship between prediction strength and differentiation and item specificity test for this relationship. ***A***, Relationship between prediction strength and differentiation was significantly negative. ***B***, Schematic of procedure for permutation analysis. For each participant, within the violation condition, we permuted the pairings of the prelearning B snapshots and the violation (X/Y) trials for the associated A items 1000 times. We then recalculated the prediction scores and correlated them with the actual differentiation scores. Finally, we calculated a *z*-score for the original relationship with respect to the null distribution of relationships and assessed its reliability across participants. ***C***, Original relationship between prediction strength and differentiation was significantly more negative than the relationships based on permuted pairs.

performed the two main analyses. First, we computed neural overlap between the prelearning snapshot of A and the postlearning snapshot of B for the violation versus nonviolation conditions. Second, we measured the trial-by-trial relationship between prediction strength (on violation trials) and subsequent neural overlap scores in the violation condition. We focus on the two largest clusters (4 and 5 voxels, respectively) that survived statistical tests for both analyses ($p < 0.05$ uncorrected; see caption of Fig. 5 for other smaller clusters). Replicating the main findings above, we found both a neural differentiation effect (violation < nonviolation) and a negative relationship between the amount of prediction and neural overlap in the left hippocampus (Fig. 5*A*). Interestingly, the left intracalcarine cortex showed the opposite pattern: neural overlap was greater for the violation versus nonviolation condition and prediction strength was positively correlated with this neural "integration" effect (Fig. 5*B*).

Ruling out univariate confounds

We have assumed that pattern similarity between A and B reflects a change in the relationship between the distributed representations of these items. However, univariate activation can affect pattern similarity (Coutanche, 2013; Davis and Poldrack, 2013; Davis et al., 2014; Aly and Turk-Browne, 2016a, 2016b), which could in principle explain some of our results. For example, weakened memory of B items in the violation condition might be expressed as lower activation in the postlearning phase, which could reduce pattern similarity for these items. This could also potentially explain the observed negative relationship between prediction strength and differentiation: greater misprediction could lead to more weakening, which (due to univariate confounds) could show up as lower pattern similarity. These scenarios are unlikely, however, in light of the item specificity of the differentiation effect. If the differentiation effect were merely due to reduced activation of B items in the violation condition, then the same pattern of results should have persisted even after permuting the AB pairings.

Several additional results provide further evidence against this alternative account. First, univariate activation in left CA2/3/DG did not differ between the violation and nonviolation conditions during the postlearning phase ($t_{(31)} = -0.55$, $p = 0.58$). Second,

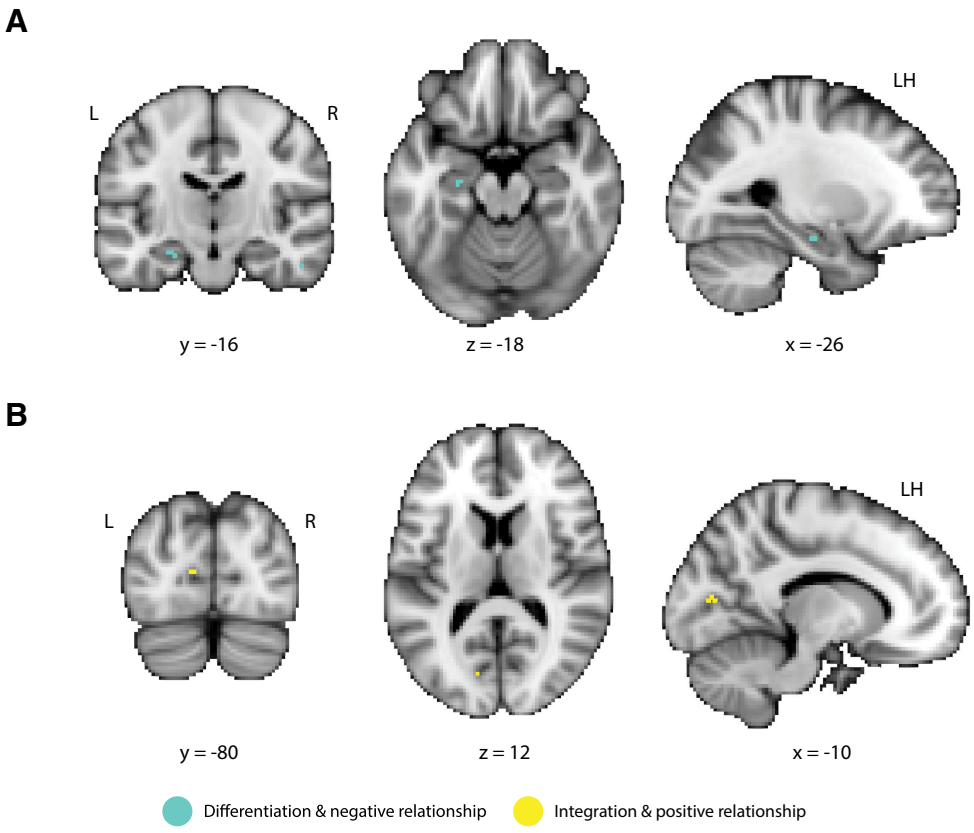

Figure 5. Searchlight results. ***A***, A cluster (4 voxels) in the left hippocampus showed both a main effect of differentiation and a negative relationship between prediction strength and neural overlap ($p < 0.05$). ***B***, A cluster (5 voxels) in the left intracalcarine cortex showed the opposite pattern: pattern similarity between the prelearning snapshot of A and the postlearning snapshot of B was greater for the violation versus nonviolation condition (neural integration) and prediction strength was positively correlated with neural overlap. Summary of smaller clusters (data not shown): (1) right intracalcarine cortex (6, −70, 16; 3 voxels), integration and positive relationship; (2) right precuneus (18, −60, 22; 2 voxels), differentiation and positive relationship; and (3) right inferior temporal gyrus (44, −50, −12; 2 voxels), integration and negative relationship.

there was no trial-by-trial relationship between univariate activation in the postlearning phase and differentiation ($t_{(31)} = 0.37$, $p = 0.71$). Third, the negative relationship between prediction strength and differentiation persisted after controlling for the univariate activation level (during the postlearning snapshot phase) with partial correlation ($t_{(31)} = -2.63$, $p = 0.01$). These observations are consistent with our interpretation that learning reflects differentiation of the underlying neural patterns rather than a change in overall activity.

Discussion

The results of this study extend our prior work showing that mispredicted memories are weakened (Kim et al., 2014). Here, we show that restudying a previously mispredicted item leads to differentiation of its hippocampal representation away from the prior context. We interpret this finding in terms of the nonmonotonic plasticity hypothesis (NMPH), which posits a U-shaped relationship between memory activation and learning. Low activation has no effect, moderate activation leads to memory weakening, and high activation leads to memory strengthening (Norman et al., 2006, 2007). Based on our prior study exploring effects of misprediction (Kim et al., 2014), we hypothesized that violation trials (where A was not followed by B) would elicit low-to-moderate levels of activation of the mispredicted B item, thereby weakening the synaptic connections between the (strongly activated) A item and the (moderately activated) B item. If memory is tested at this point, then the model predicts worse memory for the B item due to these weakened connections, as observed in Kim et al. (2014). Here, we explored what happens when the (weakened) B item is subsequently restudied. In this situation, our theory predicts that, due to the weakened connections to the shared features of A and B, activation spreads to new features that were not previously shared with A. These newly activated features at restudy are subsequently incorporated into the representation of B. This process of weakening connections to features (formerly) shared with A on the misprediction trial and strengthening connections to features not formerly shared with A (on the restudy trial) has the net result of moving B's representation away from A's representation, thereby differentiating these patterns.

Our theory generates two additional hypotheses that we were able to test in the current study. First, neural differentiation should be competition dependent, with stronger (but still moderate) prediction leading to more competition and greater subsequent differentiation. This hypothesis was supported by the observed negative relationship between prediction strength (at violation trials) and neural differentiation in left CA2/3/DG. Second, neural differentiation should occur with respect to the specific item that wins the competition and not other items. This hypothesis was supported by our randomization results, which showed that differentiation depends on the prediction of the specific B item that was previously paired with A and that differentiation reflects the neural representation of B specifically moving away from the neural representation of A, as opposed to becoming uniformly more distinct from other scenes.

The NMPH implies that there will be boundary conditions on our conclusions. In particular, the pattern of results observed here, an overall differentiation effect with greater prediction associated with more differentiation, will occur when activation for mispredicted items is in the low-to-moderate range. Because the NMPH posits that high activation leads to strengthening, much stronger predictions during violation trials may strengthen, rather than weaken, connections between A and unique features of B. This, in turn, will allow A to activate formerly unique features of B, leading to integration (i.e., increased neural overlap of A and B) instead of differentiation. We plan to test this prediction in future work by increasing associative strength between paired items (e.g., via more extensive exposure) or by adopting a more explicit prediction task (as opposed to the incidental approached used here).

A few other recent studies have used a similar pre–post "snapshot" approach to study differentiation (Schapiro et al., 2012; Hulbert and Norman, 2015; Schlichting et al., 2015; Favila et al., 2016). Schapiro et al. (2012) did so in a statistical learning paradigm, in which the transition probabilities between items varied: in the strong pair condition, A was always followed by B (transition probability = 1); in the weak pair condition, A was sometimes followed by B (0.33); and in the shuffled pair condition, A almost never was followed by B (~0). In CA2/3/DG, from the prelearning to postlearning phases (in which items were presented randomly), members of strong pairs showed increased neural overlap (integration) and members of weak pairs showed decreased overlap (differentiation), both relative to shuffled pairs. Hulbert and Norman (2015) used a retrieval practice (Rp) paradigm with highly similar pictures of animals: Rp^+ items were practiced, which should lead related Rp^- items to activate as competitors; then Rp^- items were restudied. The degree of left hippocampal differentiation predicted subsequent cued recall memory success. Favila et al. (2016) showed that linking two scene stimuli to a shared face associate leads to hippocampal differentiation of the scenes. Last, Schlichting et al. (2015) used a similar paradigm exploring the effect of linking two unrelated objects to a shared associate. They looked at a wide range of regions and showed differentiation in some regions and integration in others. They also manipulated whether linked pairs were trained in a blocked (all AB before any BC) or interleaved (intermixed AB and BC) manner and found that integration was more prevalent after blocked training.

What is missing from these prior studies is a mechanistic explanation of why differentiation occurs, what determines the size of the effect, and when and where differentiation versus integration is observed. We hypothesize that the key mediating variable is the level of memory activation: moderate activation followed by restudy leads to differentiation and strong activation leads to integration. For example, in Schapiro et al. (2012), weak transition probabilities (0.33) during AB pair learning may induce moderate levels of prediction of B, leading to differentiation of A and B. In contrast, higher levels of prediction in the strong pair condition may lead to the opposite integration effect. Results from Schlichting et al. (2015) also provide hints regarding our hypothesis. Specifically, blocked AB study may increase competitor activation during BC learning, tilting the balance toward integration. Likewise, regional differences in differentiation versus integration may relate to how tightly activity is controlled: In regions such as the hippocampus that have sparse activation, it is harder for related memories to activate strongly, biasing learning toward differentiation; other regions with less sparse activation (e.g., in cortex) would be biased toward integration. Our searchlight results showing opposite representational changes in the left hippocampus (differentiation) and intracalcarine cortex (integration) are consistent with this speculation, although we plan to address this point in a more focused and systematic way in future research. Crucially, none of the above studies measured competitor activation, so they could not test our hypothesis. The main added value of our study is thus that we establish a link between competitive dynamics during learning and subsequent representational change.

Our finding of neural differentiation in left CA2/3/DG is distinct from the notion of hippocampal pattern separation (Hulbert and Norman, 2015). Pattern separation refers to the fact that the hippocampus automatically assigns distinct representations to stimuli due to sparse coding in DG and CA3 (Yassa and Stark, 2011). Although this pattern separation process reduces neural overlap in the hippocampus, there is still some residual overlap between similar items, which can lead to interference (Norman et al., 2003, 2005). Our differentiation mechanism operates on this residual overlap after standard pattern separation takes place. That is, the residual neural overlap between related memories of A and B leads to incorrect prediction of B when A is presented, which drives further reduction of the residual neural overlap and reduces subsequent interference.

This study was not designed to identify the behavioral consequences of neural differentiation, although that remains an important goal of future work. Here, we used an item recognition memory test to be consistent with our prior study (Kim et al., 2014), but this is not a sensitive way to measure differentiation. Prior modeling work (Norman and O'Reilly, 2003; Norman, 2010) and empirical studies (LaRocque et al., 2013) suggest that reduced neural overlap can have opposite effects on different components of memory: boosting recollection by reducing interference from other memories, but reducing familiarity by lowering global match. Because these effects go in opposite directions, they might cancel each other out in our recognition test, which is sensitive to both components. The aforementioned study by Favila et al. (2016) suggests a better way of behaviorally measuring differentiation. Those investigators found that neural differentiation between visually similar scenes (e.g., A and B) led to enhanced subsequent learning of new associations between these scenes and objects (A–X and B–Y), presumably because of reduced interference between the scenes.

Summary

We found that interleaved misprediction and restudy leads to neural differentiation. These findings are consistent with predictions from our neural network modeling work (Norman et al., 2006, 2007) and other recent studies in the field (Schapiro et al., 2012; Hulbert and Norman, 2015; Schlichting et al., 2015; Favila et al., 2016). In addition, our findings are consistent with recent evidence from the reconsolidation literature that misprediction is a prerequisite for subsequent memory modification (Sevenster et al., 2013; Merlo et al., 2015). Most importantly, by revealing a relationship between prediction strength and differentiation, this work suggests a key role for competition in driving representational change. This complements prior results showing that activation of nontarget memories can weaken them (Newman and Norman, 2010; Detre et al., 2013; Kim et al., 2014; Poppenk and Norman, 2014) and suggests that one function of such weakening is to prepare the memory to accept new features and associations. This adaptive optimization of memory may serve to increase the accuracy of memory with respect to the current environment, to

reduce subsequent interference, and to minimize prediction errors.

References

Aly M, Turk-Browne NB (2016) Attention stabilizes representations in the human hippocampus. Cereb Cortex 26:783–796. CrossRef Medline

Aly M, Turk-Browne NB (2016) Attention promotes episodic encoding by stabilizing hippocampal representations. Proc Natl Acad Sci U S A 113: E420–E429. CrossRef Medline

Andersson JL, Jenkinson M, and Smith S (2007) Non-linear registration, aka spatial normalisation. Technical Report TR07JA2, Oxford Centre for Functional Magnetic Resonance Imaging of the Brain, Department of Clinical Neurology. Oxford: OUP.

Carr VA, Rissman J, Wagner AD (2010) Imaging the human medial temporal lobe with high-resolution fMRI. Neuron 65:298–308. CrossRef Medline

Coutanche MN (2013) Distinguishing multi-voxel patterns and mean activation: why, how, and what does it tell us? Cogn Affect Behav Neurosci 13:667–673. CrossRef Medline

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

Davis T, Poldrack RA (2013) Measuring neural representations with fMRI: practices and pitfalls. Ann N Y Acad Sci 1296:108–134. CrossRef Medline

Davis T, LaRocque KF, Mumford JA, Norman KA, Wagner AD, Poldrack RA (2014) What do differences between multi-voxel and univariate analysis mean? How subject-, voxel-, and trial-level variance impact fMRI analysis. Neuroimage 97:271–283. CrossRef Medline

Detre GJ, Natarajan A, Gershman SJ, Norman KA (2013) Moderate levels of activation lead to forgetting in the think/no-think paradigm. Neuropsychologia 51:2371–2388. CrossRef Medline

Duvernoy, H. M (2005) The human hippocampus: functional anatomy, vascularization and serial sections with MRI. New York: Springer.

Favila SE, Chanales AJ, Kuhl BA (2016) Experience-dependent hippocampal pattern differentiation prevents interference during subsequent learning. Nat Commun 7:11066. CrossRef Medline

Hindy NC, Ng FY, Turk-Browne NB (2016) Linking pattern completion in the hippocampus to predictive coding in visual cortex. Nat Neurosci 19: 665–667. CrossRef Medline

Hulbert JC, Norman KA (2015) Neural differentiation tracks improved recall of competing memories following interleaved study and retrieval practice. Cereb Cortex 25:3994–4008. CrossRef Medline

Kim G, Lewis-Peacock JA, Norman KA, Turk-Browne NB (2014) Pruning of memories by context-based prediction error. Proc Natl Acad Sci U S A 111:8997–9002. CrossRef Medline

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. CrossRef Medline

Merlo E, Milton AL, Everitt BJ (2015) Enhancing cognition by affecting memory reconsolidation. Curr Opin Behav Sci 4:41–47. CrossRef

Newman EL, Norman KA (2010) Moderate excitation leads to weakening of perceptual representations. Cereb Cortex 20:2760–2770. CrossRef Medline

Norman KA (2010) How hippocampus and cortex contribute to recognition memory: revisiting the complementary learning systems model. Hippocampus 20:1217–1227. CrossRef Medline

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

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. CrossRef Medline

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

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

Poppenk J, Norman KA (2014) Briefly cuing memories leads to suppression of their neural representations. J Neurosci 34:8010–8020. CrossRef Medline

Schapiro AC, Kustner LV, Turk-Browne NB (2012) Shaping of object representations in the human medial temporal lobe based on temporal regularities. Curr Biol: CB 22:1622–1627. CrossRef Medline

Schapiro, A. C., Turk-Browne, N. B., Botvinick, M. M., and Norman, K. A (2016) Complementary learning systems within the hippocampus: a neural network modeling approach to reconciling episodic memory with statistical learning. Philos Trans R Soc. In press.

Schlichting ML, Mumford JA, Preston AR (2015) Learning-related representational changes reveal dissociable integration and separation signatures in the hippocampus and prefrontal cortex. Nat Commun 6:8151. CrossRef Medline

Sevenster D, Beckers T, Kindt M (2013) Prediction error governs pharmacologically induced amnesia for learned fear. Science 339:830–833. CrossRef Medline

Yassa MA, Stark CE (2011) Pattern separation in the hippocampus. Trends Neurosci 34:515–525. CrossRef Medline