

**From the holdings of the Department of Special Collections , Princeton University Library,
Princeton, NJ**

These documents are either in the public domain, or can only be used for educational and research purposes (Fair use) as per U.S. Copyright law (text below). By accessing this file, all users agree that their use falls in the public domain or within fair use as defined by the copyright law. We do not charge any permission or usage fees for the publication of images of material in our collections, including those provided by Princeton University Library, via our website, catalog, or directly from staff. For more information regarding copyright, credit, and citation please review the information on the Copyright, Credit and Citation Guidelines page of our website.

<https://library.princeton.edu/special-collections/policies/copyright-credit-and-citation-guidelines>. If you have any questions, including how to order higher-quality images from the collection, contact Special Collections Public Services via our Ask Us! Form. <https://library.princeton.edu/special-collections/ask-us>

Doctoral Dissertations, 2011-2026

Psychology

2016

RESOLVING MEMORY CLUTTER BASED ON COMPETITION: BEHAVIORAL, AND NEUROIMAGING INVESTIGATION

Kim, Ghootae

Princeton, NJ : Princeton University

<http://arks.princeton.edu/ark:/88435/dsp010z708z920>

U.S. Copyright law: The copyright law of the United States (Title 17, §108, United States Code) governs the making of photocopies or other reproductions of copyrighted material. Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be used for any purpose other than private study, scholarship, or research. If a user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of fair use that user may be liable for copyright infringement. This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve a violation of copyright law. <https://www.copyright.gov/title17/title17.pdf>

RESOLVING MEMORY CLUTTER
BASED ON COMPETITION: BEHAVIORAL,
AND NEUROIMAGING INVESTIGATIONS

Ghootae Kim

A DISSERTATION PRESENTED TO THE FACULTY
OF PRINCETON UNIVERSITY
IN CANDIDACY FOR THE DEGREE
OF DOCTOR OF PHILOSOPHY

RECOMMENDED FOR ACCEPTANCE
BY THE DEPARTMENT OF PSYCHOLOGY

Advisers: Nicholas Turk-Browne and Kenneth Norman

November 2016

Abstract

We quickly memorize relationships between items and contexts within episodic experiences. While this ability of relational memory enables us to use past experiences to guide our future behaviors, it can also bring about competitions between overlapping memories. How does the brain resolve the potential memory clutter? Previous studies have demonstrated that the brain reduces the competition by forming separated memory representations across overlapping experiences, and that the memory clutter can be attenuated based on effortful inhibition of distracting information. Much of this dissertation work provides evidence of a novel neural mechanism whereby the brain minimizes the competition by *re-organizing* overlapping memory representations based on context-based prediction signal. First, we demonstrated that moderate levels of incorrect prediction of an item within a familiar context weaken memory of the mispredicted item. This finding suggests that, through this process, the brain determines which memory is an irrelevant aspect of the environment, and prunes it from long-term memory storage. Second, we investigated how the previously weakened memory regains its memory strength when re-experienced later. We found that the weakened item representation becomes incrementally differentiated from its initial context after subsequent restudy, and we also observed that the neural differentiation effect is mediated by the amount of incorrect prediction. This set of findings sheds light on how the brain reorganizes existing overlapping memories based on prediction signal to minimize potential memory clutter. Finally, we investigated a novel online mechanism whereby the brain prevents interference between repeated experiences during encoding a later one. We observed that the degree of representational overlap across repeated experiences regulates subsequent source memory. If an initial representation of an item is less reinstated when it is re-experienced later, the initial representation suffers less from competition

from the novel context of the later encounter compared to when it was highly reinstated. To summarize, this work characterizes the novel neural mechanisms wherein the brain minimizes memory competition by regulating representations of overlapping experiences.

Acknowledgements

I have been highly privileged to have incredibly great advisors Nick Turk-Browne and Ken Norman. I would like to thank them for their encouragement and support. These studies could not have been completed without their expertise, knowledge, and guides. Above all, I am greatly indebted to them for cultivating my way of thinking, for fostering my excitement for doing research. I am so grateful to my previous advisor Do-Joon Yi, who built up the basic scientific-mind in me. I have also benefited greatly from the mentorship and collaboration of Jarrod Lewis-Peacock, who contributed to much of this work. I thank my labmates and friends in the department for their enormous support. Finally, ELEPHANTIC thanks to my family for their endless love.

General Introduction

Every time we experience an episode, we quickly learn relationships between the item and context within the episode. This associative nature of human episodic memory guides our future behavior (Bar, 2007; Schacter, Addis, & Buckner, 2008). For example, imagine a situation where you are visiting a town for the first time. You meet a barista at a coffee shop whose brewing technique is exceptional. The next morning, encountering the coffee shop again while you are jogging, you immediately retrieve the experiences of the day before, and open the door expecting to see the barista again.

It is widely known that this rapid relational learning is implemented by a mechanism in the hippocampus (O'Keefe & Nadel, 1978; Aggleton & Brown, 1999; Lila, 2006). According to the complementary learning systems framework (McClelland, McNaughton, & O'Reilly, 1995; Norman & O'Reilly, 2003), the hippocampus binds together sets of co-occurring features (e.g., the barista and coffee shop) by linking those features into a hippocampal representation. A subsequent encounter of a familiar contextual cue (e.g., coffee shop) reactivates the entire hippocampal representation, which leads to full retrieval of the previous episode (e.g., your experiences at the coffee shop the day before).

Although this rapid relational learning mechanism enables us to use past experiences flexibly, it can also cause competition between related memories (Norman & O'Reilly, 2003). In the situation where you saw a new barista at the coffee shop, the reactivated memory of the old barista might interfere with forming a representation of the new experience. Further, if you are later trying to remember the old barista, the overlapping, recent experience may interfere with retrieval of the sought-after memory.

This competition-driven memory interference has long been investigated using the AB/AC paradigm (McGovern, 1964; Postman & Underwood, 1973; Richter et al., 2016). In the standard design, participants study a list of cue-associate word pairs (e.g., cream – laptop; AB pairs), and then are subsequently required to learn a second list of pairs using the same cues but with new associates (e.g., cream – sedan; AC pairs). In a control condition, the later second list study is omitted (i.e., study AB list only). It has been typically observed that recall of the first AB pairs (recalling B given A as a memory cue) is impaired by later learning of AC pairs compared to the control condition. In short, subsequent learning has a negative effect on prior memories, which was termed retroactive interference.

In addition, it has also been observed that the interference effect works in a proactive direction (Watkins & Watkins, 1975). In a typical proactive interference paradigm, the list conditions are the same as the retroactive interference design (i.e., AB learning is followed by AC learning). However, for the control condition, AB learning, instead of AC learning, is omitted (i.e., study AC list only). When tested for AC pairs, associative memory for the interference condition is worse compared to the control, which indicates the cost of prior experiences on later, new (but overlapping) learning (citation).

These two types of interference effects clearly demonstrate the dark side of rapid relational learning. That is, if A becomes connected to both B and C associates, they then compete with each other to be retrieved given A as a retrieval cue, which leads to impaired memory of both associates (Watkins & Watkins, 1975; Anderson & Neely, 1996). In addition to interference during a memory test, previous studies have found that competition can exert its negative influence on both retaining prior memories (i.e., memories of AB) and encoding new experiences (i.e., forming new memories of AC; Anderson, 2003; Koen & Rugg, 2016; Norman,

Newman, & Detre, 2007). Therefore, it is a highly important question to ask how the brain resolves competition between overlapping memories.

The complementary learning systems model (McClelland, McNaughton, & O'Reilly, 1995; Norman & O'Reilly, 2003) posits that sparse coding in the hippocampus provides a *default* level of resolution. That is, the hippocampus assigns distinct (pattern-separated) representations to episodes, which are less dependent on their similarities (Yassa and Stark, 2011). This property allows even similar episodes to be memorized in the hippocampus with less interference. Nevertheless, this pattern separation mechanism does not provide a complete resolution to interference. That is, despite the fact that pattern separation reduces representational overlap of similar memories stored in the hippocampus, there is still some residual overlap in the representations between those memories, which may lead to some level of interference (Norman and O'Reilly, 2003).

Previous studies in the area of memory control have reported that people are able to strategically select a sought-after memory while inhibiting competing non-target memories (Anderson & Green, 2001; Anderson, 2003). Retrieval induced forgetting (RIF) nicely demonstrates the role of memory control over resolving memory interference (Anderson, Bjork, & Bjork, 1994, 2000). A typical RIF paradigm consists of three phases. In the first study phase, participants learn category-exemplar pairings (e.g., "Fruit – apple", "Fruit – banana", "Tree – birch", and "Tree – oak" etc.). In the following practice phase, half of the exemplars from half of the categories are cued multiple times for retrieval ("Fruit – a__"). In the final phase, cued-recall test is performed for all of the originally studied items (e.g., "Fruit – a__", "Fruit – b__", "Tree – b__", and "Tree – o__"). Not surprisingly, memory performance for the practiced items ("apple"; Rp+) is higher compared to items from unpracticed baseline categories ("birch";

Nrp). Importantly, however, recall of the unpracticed items (competitors) of practiced categories (“banana”; Rp-) is impaired compared to the baseline items (Nrp), and this impaired memory for the competing Rp- items indicates RIF. The inhibition theory of RIF (Anderson, 2003; Anderson & Neely, 1996) posits that memories of the competing Rp- items (“banana”) are actively suppressed during retrieval practice of Rp+ items (retrieving “apple” given “Fruit” as a cue). The cue-independence of the RIF effect supports this account: the memory impairment for Rp- items (“banana”) holds true even when a new and independent retrieval cue is used (e.g., “Monkey – b___”), rather than the original one (“Fruit”), which suggests that selecting target RP+ items during the retrieval practice creates inhibition of competing RP- memories. Subsequently, when the common cue is presented at test, this memory inhibition facilitates retrieval of the sought-after target while preventing competing (but inappropriate) items to be reactivated. A recent fMRI study (Kuhl, Dudukovic, Kahn, & Wagner, 2007) has provided neural evidence of the active memory control process using a RIF paradigm. They found that the activation level of a subset of prefrontal cortex (PFC) regions (dorsolateral prefrontal cortex, frontopolar cortex, and medial prefrontal cortex extending to anterior cingulate cortex), which engages in executive control processes (Aron, Robbins, & Poldrack, 2004; Cohen, Botvinick, & Carter, 2000; MacDonald, Cohen, Stenger, & Carter, 2000), decreased as the number of retrieval practices of RP+ items increased, which reflects the reduced demands on cognitive control over repeated retrieval practices. Most importantly, this repetition-related activation decrease in the PFC subregions predicted greater forgetting of RP- items, and these findings together suggest that competing RP- memories are incrementally inhibited through memory control processes (mediated by the PFC subregions) over repeated retrieval attempts of target items.

Although the generality of RIF has been demonstrated across a variety of stimulus types (Ciranni & Shimamura, 1999; Johnson & Anderson, 2004; Shivde & Anderson, 2001) and different forms of memory (Brown, Kouri, & Hirst, 2012; Coman, Manier, & Hirst, 2009; Fan & Turk-Browne, 2013), the underlying memory inhibition process might not work efficiently when executive control functioning is overloaded. For example, one behavioral study (Roman, Soriano, Gomez-Ariza & Bajo, 2009) observed that the RIF effect was eliminated when executive control was burdened by a concurrent task during retrieval practice.

The primary goal of our work is to provide empirical evidence of a novel mechanistic model of how the brain *self-organizes* overlapping memory representations based on the competitive neural dynamics of those memories. Our work was theoretically motivated by a recently proposed neural network model articulating how competition drives changes in memory representations (we refer to this as the nonmonotonic plasticity hypothesis, or NMPH; Norman, Newman, Detre, & Polyn, 2006; Norman et al., 2007).

The nonmonotonic plasticity hypothesis assumes that memories are composed of distributed populations of units (or neurons) and the strength of a memory is proportional to degree of interconnectivity between units in its ensemble. A neural network learning algorithm, which is the core of the model, minimizes competition between to-be-recalled target memories and related competing memories through inhibitory oscillations. First, the inhibitory oscillation algorithm identifies weak parts of the target memory: If a unit belonging to the target memory is receiving relatively little collateral support from other units, increasing inhibition will trigger a decrease in the activation of that unit. The learning algorithm then strengthens these weak units by increasing weights from other active units. Conversely, the algorithm identifies strong competing memories by lowering inhibition, and weakens units that pop up (i.e., units receiving

strong input, but not quite enough to become active given normal levels of inhibition) by attenuating connections from other active units. Because the relationship between activation and change in connection strength reverses as the level of activation goes up, this model predicts a U-shaped relationship between memory activation and change in memory strength. That is, moderate activation of a competing memory leadings to weakening of connections from other active units, and as the competing memory is a less “attractive” state of the network, its memory strength is impaired. In contrast, stronger activation strengthens connections between units of the memory, thereby increasing behavioral expression of the memory.

Chapter 1 describes a functional magnetic resonance imaging (fMRI) study (Kim, Lewis-Peacock, Norman, & Turk-Browne, 2014) that examined this hypothesized U-shaped relationship between memory activation and change in memory strength. In our study, participants were exposed to a continuous sequence of faces and scenes while performing a cover task. Unbeknownst to them, this sequence was generated from triplets (e.g., A-B-C). The first two “context” stimuli (i.e., A-B) in each triplet were later repeated, but a new stimulus appeared in the third position (e.g., A-B-D). We hypothesized that the original third C item would be moderately reactivated by repetition of the context AB items, leading to weakening of the mispredicted C item’s memory. In order to measure the amount of reactivation of the third C items, we used multivariate pattern analysis to decode its category during the repeated context, and related this prediction strength to subsequent recognition memory of the item. We found a negative relationship: strongly predicted items were recognized less well than weakly predicted stimuli and control stimuli not part of a context triplet. This observation is consistent with the predictions of NMPH: A large prediction error is associated with moderate activation of the mispredicted item; this moderate activation, in turn, triggers weakening of the item’s

representation in memory. Our findings provide insights into how the brain prunes irrelevant or unstable memories based on automatic context-based prediction.

Chapter 2 describes an fMRI study that investigated how a previously weakened item's representation (based on context-based prediction) is reconstituted when it is re-experienced in a new context. Based on the NMPH (Norman et al., 2005, 2006, 2007), we hypothesized that the brain differentiates the neural representation of the previously mispredicted item from its initially associated context when it is encountered. Specifically, the initial misprediction weakens connections between the mispredicted item and initial context and — upon restudy — the mispredicted item becomes linked to new features that were not shared with the initial context (Hulbert & Norman, 2015). Participants were exposed to a continuous sequence of stimulus pairs (AB). These pairs repeated three times, but upon the fourth repetition, the B item was replaced by a new stimulus X to create a prediction error. Crucially, the B item was subsequently restudied without its preceding context A item, and was followed by one more cycle of violation and restudy trials (AB-AB-AB-AX-B-AY-B). We hypothesized that the weakened representation of B from violation trials would be incrementally differentiated from A after subsequent restudy opportunities, and found a neural differentiation effect consistent with this prediction in the left CA2/3/DG subregion of the hippocampus: That is, representational overlap between A and B items decreased more after the cycles of violation and restudy compared to a control where there was no prediction error and differentiation was not expected. Supporting our hypothesis more directly, this differentiation was mediated by the degree to which the (missing) B item was activated on violation trials (i.e., prediction strength).

This series of studies sheds light on how the memory system *self-organizes* overlapping memories based on competition signals. That is, existing associations between a context and item

are weakened when their contextual relationship is discovered to be incorrect or unstable, which consequently leads to impaired memory of the mispredicted item. However, a previously irrelevant memory can be re-experienced in a novel situation. How is the weakened item's memory representation reconstituted after a subsequent restudy? In the second study, we found that the weakened associations between the context and item cause differentiation of the two memory representations when the item is restudied later. The brain might prevent potential interference between the context and item thorough a cycle of “weakening and differentiation”.

We have so far investigated how *existing* competing memory representations are modified by context-based prediction. Chapter 3 investigates another mechanism whereby the brain avoids interference by forming a distinct representation of a new (but overlapping) experience. Using multivariate pattern analysis on fMRI data, we report the discovery that less representational overlap across repeated experiences leads to better subsequent source memory of an initial experience. Specifically, when a previously encountered item reappears in a novel context, the brain automatically reinstates the representation of the item that was formed during the initial encounter, and this reinstated item representation becomes associated with the novel context, which interferes with subsequent retrieval of the initial context. Conversely, lower reinstatement keeps the initial item from being contaminated by the novel context, which leads to better subsequent source memory for the initial context.

Chapter 1: Pruning of memories due to context-based prediction error

A version of this chapter was published as an article in PNAS (Kim, Lewis-Peacock, Norman, & Turk-Browne, 2014)

Introduction

Our experience of the world is recorded in long-term memory every moment of every day. Such memory formation occurs continuously and incidentally, resulting in a potentially astronomical number of memory traces. This cluttering can be problematic for the efficient functioning of memory systems in the brain. At retrieval, irrelevant memories will compete with the sought-after memory and can prevent it from coming to mind. To avoid these costs, we propose that memory systems regulate themselves by adaptively “pruning” item representations.

How does the brain decide which items to prune from memory? We hypothesize that the brain makes this determination based on how accurately an item is predicted by its context. Specifically, the brain may automatically predict which items are likely to appear in a given context based on prior experience, and then compare this prediction against the actual contents of experience. When the prediction is wrong, the representations of the expected items in long-term memory are weakened. This weakening is manifested in a graded reduction of the accessibility of these items during later retrieval. In this way, the brain can use context-based prediction error to determine when an item is not a stable aspect of the world, and this determination may in turn mark the item for pruning.

The notion that the brain automatically predicts which items should appear in a context is supported by previous research. Such predictions result from learning of relationships between items (Cohen & Eichenbaum, 1993), which in turn allow the appearance of one item to cue the reactivation other associated items (Hirsh, 1974; Howard & Kahana, 2002; Miyashita, 1993).

Behavioral studies have obtained evidence for automatic prediction by showing that task performance is facilitated for items that are predictable in the current context (Nissen & Bullemer, 1987; Olson & Chun, 2001). More recently, neuroimaging studies have obtained evidence for automatic prediction by showing that the medial temporal lobe and sensory cortex reinstate representations of predicted items (Kok, Rahnev, Jehee, Lau, & Lange, 2012; Schapiro, Kustner, & Turk-Browne, 2012; Turk-Browne, Simon, & Sederberg, 2012).

Although context-based prediction has previously been found, its consequences for item learning are unknown. As noted above, such prediction provides an opportunity to discover whether an item is a stable aspect of the world. When an item fails to appear in a context with which it has been previously associated, the resulting error signal may cause the item to be pruned from memory. To test this pruning hypothesis, we set out to relate prediction strength to subsequent memory. Specifically, our hypothesis posits that in situations where a predicted item does not appear as expected, prediction strength should be negatively related to subsequent memory for the item: Stronger predictions lead to larger prediction errors, which in turn lead to more weakening of the predicted item's memory trace, and ultimately to reduced confidence in having seen the item before and a greater likelihood of forgetting the item altogether.

Numerous studies have investigated how prediction error shapes learning in the brain (O'Doherty et al., 2004; Pagnoni, Zink, Montague, & Berns, 2002; Schultz & Dickinson, 2000). Our study differs from these studies in two important ways: First, existing studies have primarily focused on learning to predict future rewards. That is, the specific identities of the predicting stimuli were irrelevant other than in terms of how much reward they predicted (Niv & Schoenbaum, 2008). In contrast, we examine an unsupervised form of stimulus-stimulus learning, in which relationships are formed between, and predictions made about, stimulus

identities with no inherent motivational significance. Second, prediction error is typically viewed as a way of updating associative strength between cues and outcomes (Rescorla & Wagner, 1972). Here, that would correspond to learning how strongly an item (the “outcome”) should be predicted by its context (the “cue”). Instead, we test the novel idea that context-based prediction error weakens the long-term memory representation of the predicted item itself.

To relate the context-based prediction of an item to its later accessibility in memory, we developed a trial-by-trial measure of prediction strength. There is no clear behavioral signature of automatic prediction that can be measured for a single trial. Thus, we used the output of a multivariate pattern classifier applied to fMRI data (Norman, Polyn, Detre, & Haxby, 2006). With this technique, we quantified how strongly an item was predicted when its previously associated temporal context was repeated, and then related this prediction strength to subsequent recognition memory for that item. We also tested our hypothesis with a second multivariate technique, pattern similarity analysis (Kriegeskorte, Mur, & Bandettini, 2008). This allowed us to replicate and extend the main findings, confirming that the relationship between classifier output and subsequent memory reflected prediction of the previously associated item *per se* (see Supplementary).

Twenty-four participants completed an fMRI session in two phases: an incidental encoding phase and a subsequent memory test phase. In the incidental encoding phase, participants were exposed to a continuous stream of face and scene photographs while performing a categorization cover task (discriminating male/female or indoor/outdoor, respectively). Unbeknownst to participants, the stream was generated from triplets (Fig. 1A). The first two “context” items in each triplet were from one category and the final item was from the other category (e.g., $A_{\text{face}} \rightarrow B_{\text{face}} \rightarrow C_{\text{scene}}$). These two context items were repeated later in the

stream, but were followed this time by a novel final item from their same category (e.g., $A_{\text{face}} \rightarrow B_{\text{face}} \rightarrow D_{\text{face}}$). The triplets were used to construct the sequence, but the triplet structure was not overtly signaled to participants — items appeared continuously every 4.5s. Additional single items from each category (i.e., items whose preceding context items never repeated) were inserted into the stream between triplets (e.g., X_{scene}).

We expected that the repeated context items would automatically trigger a prediction that the original final item (C) would appear next — a prediction that would then be violated by the appearance of the novel final item (D). During these “repeated” triplets, we assessed the strength of the prediction of C by measuring how much information about the C item’s category was available in the brain. Note that C always came from a different category than the A, B, and D items in the corresponding repeated triplet, making it possible to resolve how much participants were expecting C using a category-based classifier.

In the subsequent memory test phase, participants performed a recognition task for items from the encoding phase (Fig. 1B). In addition to new lures, there were three types of old items: final items from initial (C) and repeated triplets (D), and control items that were not part of a triplet (X). Importantly, all old items appeared only once during encoding, and thus differences in memory performance between conditions must result from the repetition of context items and any associated prediction errors.

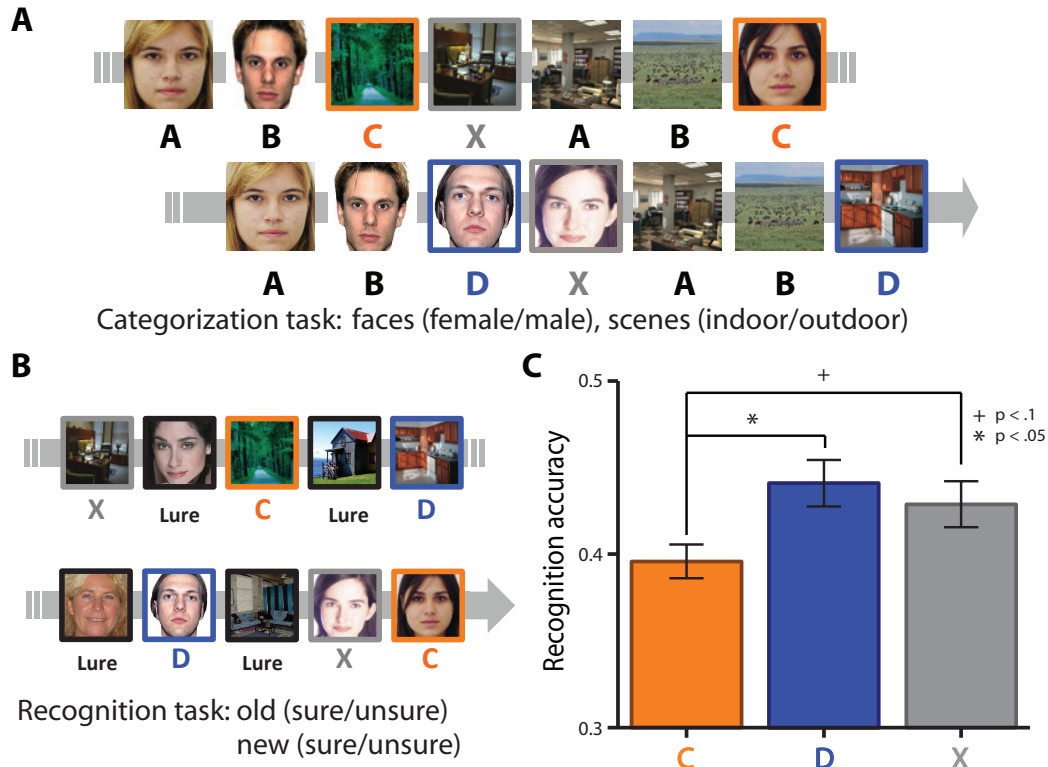


Figure 1.1. Experimental design and behavioral results. (A) During incidental encoding, the trial sequence was constructed from triplets (A→B→C) that repeated once with a novel final item (A→B→D), and unrepeatd single items (X). The categorization task was orthogonal to the triplet structure. (B) In the subsequent memory test, old/new judgments were collected for final items from initial (C) and repeated (D) triplets, single control items (X), and novel lure items. All old items appeared once during incidental encoding. (C) The high-confidence hit rate for C items was significantly lower than for D items, and marginally lower than for X items. Error bars reflect ± 1 SEM.

Results

Subsequent Memory Behavior

Participants displayed reliable memory performance overall, successfully discriminating old items in all conditions from lures ($ps < .001$; Supplementary Fig. 1). Although the essential test of our hypothesis relies on relating memory to prediction strength on an item-by-item basis, we first considered whether there was overall memory suppression for C items relative to both D and X items irrespective of prediction strength (Fig. 1.1C). Indeed, the high-confidence hit rate

for C items was significantly lower than for D items ($t_{23} = 2.33, p = .029$) and marginally lower than for X items ($t_{23} = 1.73, p = .096$). These results did not depend on the greater novelty of items preceding C items during encoding, nor on the earlier average serial position of C items in the trial sequence (see Supplementary).

For the fMRI analyses below, we treat high-confidence hits as “remembered” and other items as “forgotten.” Requiring high confidence for an item to be considered remembered is consistent with prior studies (e.g., Turk-Browne, Yi, & Chun, 2006; Wagner et al., 1998), and is supported by the fact that participants exhibited greater sensitivity for high-confidence responses (mean $A' = 0.81$) than low-confidence responses (mean $A' = 0.58$; $t_{23} = 6.07, p < .001$).

Relating Prediction Strength to Subsequent Memory

Our hypothesis does not state that memory will be impaired for *all* mispredicted items. Rather, context-based prediction error triggers pruning, and thus only C items that were strongly predicted in the repeated triplet should be more likely to be forgotten. That is, we expected that the amount of prediction for a given C item should be *negatively* related to its likelihood of being remembered: weak predictions should generate small errors with little impact, while strong predictions should generate large errors and lead to weakening of the memory. To test this, we measured prediction strength with an fMRI pattern classifier.

For classification, we used regularized logistic regression (penalty=1) to identify face and scene information in brain activity patterns from anatomical regions of interest (ROIs) in bilateral ventral temporal cortex. A separate classifier was trained for each participant using an independent functional localizer (Supplementary Fig. 2A). The classifier was then applied to continuous brain patterns from the incidental encoding phase. For each C item, we calculated the relative amount of classifier evidence for its category during the repeated triplet. We interpreted

information about C's category as evidence of prediction for two reasons: (1) the C item (e.g., C_{scene}) was not shown in the repeated triplet, and (2) the items that did appear in the repeated triplet were all from the other category (e.g., A_{face} → B_{face} → D_{face}). For the main analysis, we averaged the relative evidence for C's category over all time points in the repeated triplet. To test the hypothesis that greater misprediction of a C item increases the likelihood of it being forgotten, we related this average category evidence to the memory outcome for the same C item using logistic regression (Fig. 1.2A).

Consistent with our hypothesis, there was a significant negative relationship between prediction strength and memory (Fig. 1.2B): Greater category evidence was linked to an increased likelihood of forgetting ($\beta_1 = -.85, p < .001$). To examine how this negative trend developed, we performed this same analysis over different time windows in the repeated triplet (Fig. 1.2C). We binned the three time points around each of the three items in the repeated triplet and around the triplet's onset (as a baseline). If the overall negative trend reflects context-based prediction, the trend should be most negative later in the triplet. Namely, there should be no trend before the context items appear, and it should grow to maximum negativity at the time of D, after both predictive context items have appeared and at the anticipated time of C. Indeed, the beta coefficient during the D time period was reliably negative ($p = .002$) and significantly lower than baseline ($p = .014$). All of these findings were validated with a second multivariate analysis approach based on pattern similarity (Supplementary Fig. 3), which confirmed that prediction of the C item *per se* was negatively related to subsequent memory. Note that since only high-confidence hits were treated as remembered, increased forgetting may reflect reduced recognition confidence (e.g., low-confidence hits that were treated as forgotten). Regardless, this

would still be consistent with the hypothesis that strong prediction of the C item weakened its memory.

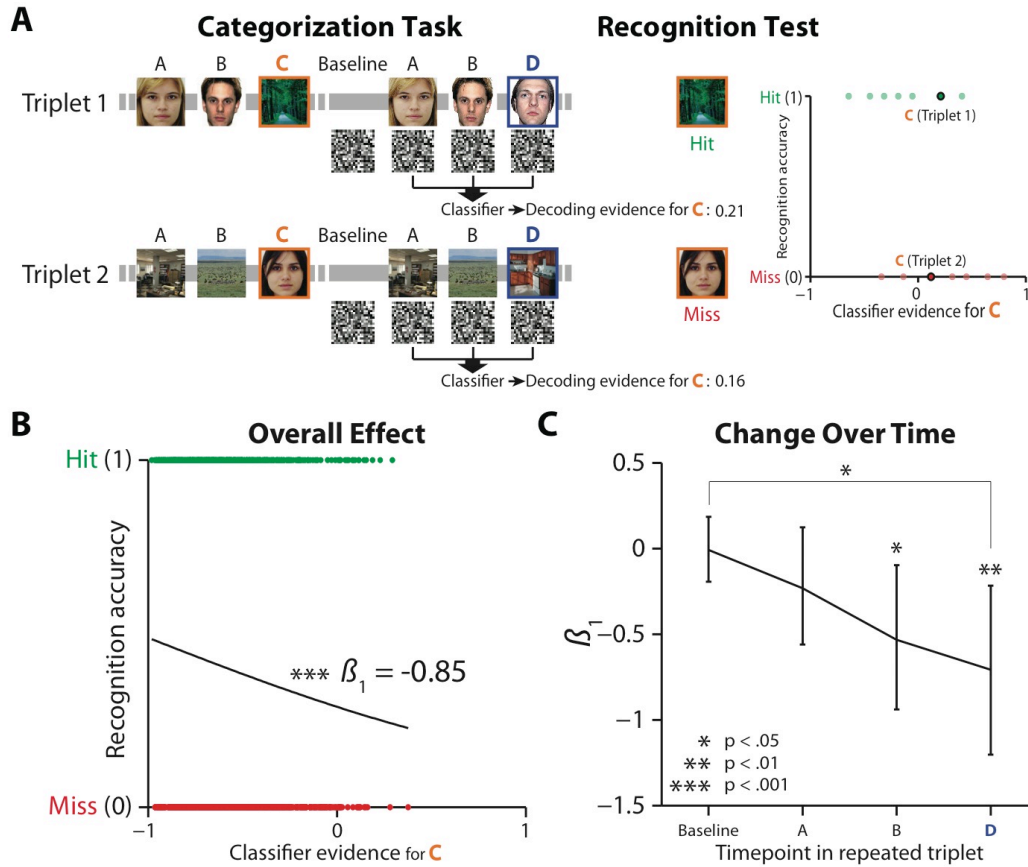


Figure 1.2. Pattern classification analysis results. (A) A classifier trained to decode visual categories was applied to brain patterns from the repeated triplets. Classifier evidence for the category of a C item was related to subsequent memory for that item. This category was not presented in the repeated triplet, and thus classifier evidence for it was interpreted as prediction. (B) Category evidence for C was first averaged over all time points in the repeated triplet. Dots indicate the distribution of classifier evidence for remembered (green) and forgotten (red) items. Logistic regression analyses revealed a reliable negative relationship (i.e., negative β_1), where greater classifier evidence was associated with more forgetting. (C) The same analysis was performed separately during the Baseline, A, B, and D time periods. The negative relationship was maximal during the anticipated time of C, but was evident earlier in the repeated triplet (during B). Error bars reflect 95% bootstrap CIs.

To test the specificity of these results, we examined whether classifier evidence for categories other than the C category was related to subsequent memory. We first considered whether the observed negative relationship would persist after controlling for evidence of the D category with partial correlation, and found that it remained robust ($p = .008$). When this analysis

was reversed, testing how D category evidence was related to memory for C after controlling for C category evidence, there was no relationship ($p = .99$). Moreover, classifier evidence for neither the object category nor “rest” (from the localizer) predicted C memory ($ps > .36$). Thus, worse memory of C items was uniquely explained by evidence for the C category in repeated triplets.

Ruling Out an Alternative Interpretation

The observed negative relationship between prediction strength and subsequent memory is consistent with our hypothesis that context-based prediction error leads to pruning. However, a potential alternative explanation is that forgetting occurred because of poor encoding of C in the initial triplet rather than strong prediction of C in the repeated triplet. In this scenario, repeated context items triggered the attempted retrieval of C, which was slower and more difficult for items that were poorly encoded. During this memory search, other items from the same category were examined and rejected, giving rise to greater category information. Under this interpretation, the negative relationship between classifier evidence and memory is spurious: The quality of encoding of the C item in the initial triplet may determine both the amount of category information in the repeated triplet (poor quality means more search and thus more category information) and the subsequent memory outcome (poor quality means more forgetting), without any direct effect of the former on the latter.

According to this alternative, the observed negative relationship should be eliminated if the quality of initial encoding is controlled. We operationally defined encoding quality as the amount of classifier evidence for C’s category when it was perceived in the initial triplet. This is based on the finding of enhanced activation during initial encoding for subsequently remembered vs. forgotten items (Brewer, Zhao, Desmond, Glover, & Gabrieli, 1998; Wagner et al., 1998).

We then repeated all analyses after controlling for the effects of this “perception strength” measure on both prediction strength and subsequent memory (Fig. 1.3). Contrary to the alternative explanation, the negative relationship between prediction strength and subsequent memory remained robust ($p = .004$).

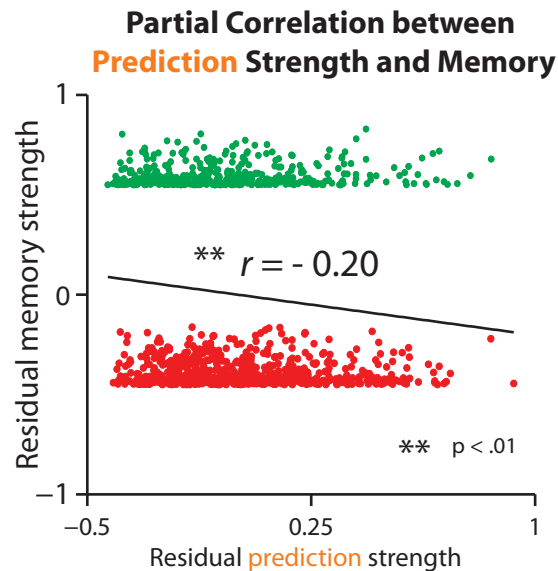


Figure 1.3. Partial logistic regression. Scatterplot and correlation of residual prediction strength and residual memory outcome after partialing out item-by-item perception strength. Green and red dots indicate the distributions of prediction strength values for subsequently remembered and forgotten items, respectively.

Evidence of Prediction

Having ruled out a contribution of initial encoding to the negative relationship between prediction strength and memory, we are left with the interpretation that prediction of C items during the repeated triplet *per se* was responsible for their worse memory. To support this interpretation, we performed two additional tests of whether C items were predicted overall.

First, as would be expected, there was more classifier evidence for the C category during A and B in repeated triplets (when C could be predicted) than during A and B in initial triplets ($p=.038$). There is a potential confound with this analysis, however: A and B were novel in initial triplets but not in repeated triplets. If repetition suppression reduced the amount of activity for A

and B (Grill-Spector, Henson, & Martin, 2006), then classifier evidence for their category may also decrease, spuriously increasing evidence for other categories, including C's. This confound is improbable because repetition suppression has been linked to increased, not decreased, classifier evidence for repeated items (Kok, Rahnev, Jehee, Lau, & Lange, 2012). Indeed, more repetition suppression for A and B (initial minus repeated univariate activity; see Supplementary) was associated with greater evidence for their category ($p = .040$) and lower evidence for C's category ($p < .001$). Thus, repetition suppression worked against the observed increase in C category evidence during context items in repeated vs. initial triplets, so the increase can be interpreted more soundly as prediction of C.

Second, insofar as C was predicted in repeated triplets, then the amount of classifier evidence for C's category in initial and repeated triplets should be correlated across items. That is, idiosyncratic variance in the prototypicality of the faces and scenes should lead to systematic variance across items in the amount of evidence for their category; since we posited that the same C item was processed in matching initial and repeated triplets, prediction strength should thus be correlated with perception strength. Indeed, this correlation was reliably positive at the time of C in the repeated triplet ($p = .022$). There is also a potential confound in this analysis: A and B were present in both initial and repeated triplets and carryover activity from these items may have driven the correlation. However, the relationship remained marginally significant after controlling for the context items with partial correlation ($p = .060$). These results are also consistent with our interpretation that C items were predicted.

Non-Monotonic Plasticity Mechanism

Our discovery of a negative relationship between memory activation and subsequent memory is striking, given that positive relationships have routinely been reported (Brewer, Zhao, Desmond,

Glover, & Gabrieli, 1998; Wagner et al., 1998). How does context-based prediction lead to memory weakening? We interpret these findings in terms of the non-monotonic plasticity hypothesis (Detre, Natarajan, Gershman, & Norman, 2013; Newman & Norman, 2010), which posits a U-shaped relationship between memory activation and learning: low activation does not lead to learning, moderate activation leads to memory weakening, and high activation leads to strengthening. This is based on neurophysiological findings (Artola, Bröcher, & Singer, 1990; Hansel, Artola, & Singer, 1996): moderate post-synaptic depolarization causes long-term depression (i.e., synaptic weakening) whereas greater post-synaptic depolarization causes long-term potentiation (i.e., synaptic strengthening).

Neural network modeling suggests that this learning principle should scale up from synapses to neural ensembles (Norman, Newman, Detre, & Polyn, 2006; Norman, Newman, & Detre, 2007). In these models, memories are composed of distributed populations of neurons, where the strength of the memory is proportional to the degree of interconnectivity between neurons in the ensemble. After incorporating non-monotonic plasticity, moderate activation of these distributed ensembles weakens synapses within the ensemble (thereby reducing behavioral expression of the memory), and stronger activation strengthens synapses within the ensemble (thereby increasing behavioral expression of the memory). Importantly, these models predict that non-monotonic plasticity should apply to the initial presentation of an item as well as to subsequent presentations: When an item is first presented, its neural representation will have some (nonzero) level of interconnectivity, due to prior experience with related stimuli — this initial level of interconnectivity can be reduced (in the case of moderate activation) or increased (in the case of strong activation). In keeping with the predictions of these models, prior studies have found a U-shaped relationship between activation (in EEG or fMRI) and subsequent

accessibility (Detre et al., 2013; Newman & Norman, 2010). For example, moderate levels of activation elicited by the first presentation of a stimulus lead to negative priming (i.e., slower subsequent responding) whereas higher levels of activation lead to positive priming (i.e., faster subsequent responding; Newman & Norman, 2010).

This model can explain the observed relationship between prediction and memory if strong but unconfirmed predictions trigger moderate item activation, and lesser predictions trigger weaker activation. Indeed, even the strongest predictions in the current task may only result in moderate activation: there is only a single opportunity to associate C items with their context, predictions were incidental with respect to the categorization task during encoding, and prediction is internally generated activation — which is generally weaker than perceptual activation (Johnson, Mitchell, Raye, D’Esposito, & Johnson, 2007). When activation levels fall in this low-to-moderate range, the non-monotonic plasticity hypothesis posits that the relationship of item activation to memory will be *negative*: weak predictions will be neutral, whereas stronger predictions (leading to moderate activation) will induce forgetting. By this account, learning does not depend on the explicit computation of an error signal. Rather, prediction error is implicitly realized in the brain as moderate activation of the unconfirmed prediction.

It was hard to know *a priori* that context-based predictions would elicit low-to-moderate activation values. Nevertheless, based on this claim, we can evaluate the non-monotonic plasticity hypothesis by exploring the consequences of higher levels of activation: According to non-monotonic plasticity, the relationship of item activation to memory should be *positive* when activation values range from moderate (which causes weakening) to high (which causes strengthening). To test this idea, we related perception strength — i.e., classifier evidence for the

C category in initial triplets — to subsequent memory. We expected that the perception of an item would elicit substantially higher activation levels than its later context-based prediction (Johnson, Mitchell, Raye, D’Esposito, & Johnson, 2007), leading to a positive relationship with subsequent memory.

As expected, perception strength values were robustly greater on average than prediction strength values ($t_{23} = 38.57, p < .001$; Fig. 1.4B, bottom). Furthermore, the relationship between perception strength and subsequent memory was reliably positive ($\beta_1 = .81, p = .002$; Fig. 1.4A). As shown in Fig. 1.4B, moderate activation (derived from strong prediction *or* weak perception) led to worse memory relative to both low activation (weak prediction) and high activation (strong perception).

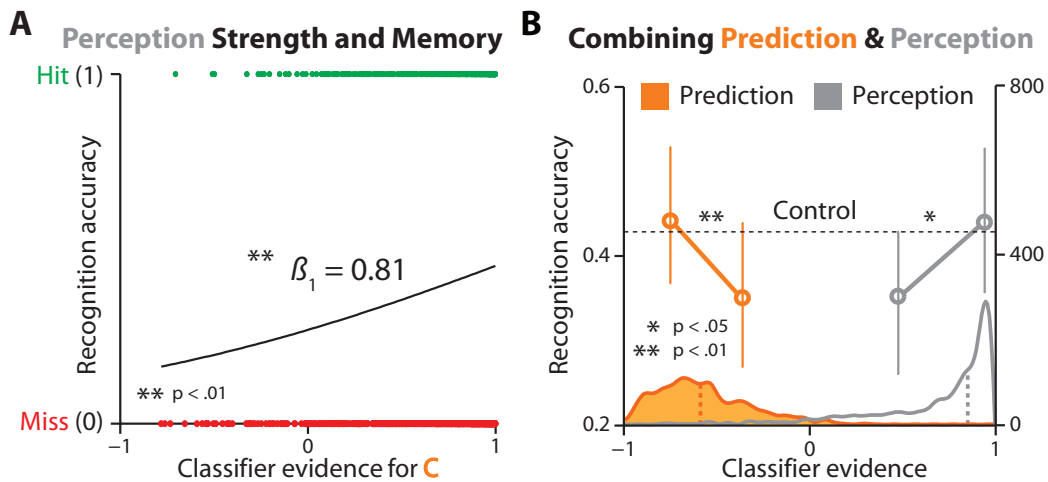


Figure 1.4. Non-monotonic relationship. (A) Logistic regression fit of perception strength to subsequent memory. (B) Relationship between classifier evidence and subsequent memory. Bottom: histograms of classifier evidence for prediction (orange) and perception (gray) over C items. Vertical dotted lines indicate the median. Top: memory for C in each half of median splits. Error bars reflect 95% bootstrap CIs.

To formally model the combined influence of perception and prediction, we used the P-CIT Bayesian curve-fitting algorithm (Detre, Natarajan, Gershman, & Norman, 2013). This algorithm estimates the posterior distribution over “plasticity curves” relating activation

(classifier evidence during incidental encoding) to learning (subsequent memory behavior). For each C item, perception in the initial triplet and prediction in the repeated triplet were treated as separate learning events whose effects were summed to model recognition. The curves recovered by P-CIT were reliably U-shaped, consistent with our interpretation (Supplementary Fig. 4).

Discussion

We obtained evidence that context-based prediction error can lead to forgetting using multivariate pattern analyses of fMRI data. This supports our “pruning” hypothesis that item memories are weakened when they are mispredicted by their context. The non-monotonic plasticity hypothesis provides a neurobiologically plausible mechanism for these findings: A large prediction error is associated with moderate activation of the mispredicted item; this moderate activation, in turn, triggers weakening of the synapses that support the item’s representation in memory.

There are some boundary conditions for this account. As we showed, weak predictions induce low activation and leave memory intact. At the same time, *very* strong predictions (e.g., after extensive experience) may induce high enough activation to shield against pruning, or even enhance memory (Smith, Hasinski, & Sederberg, 2013). Moreover, if a prediction is correct, the sum of bottom-up perception and top-down prediction will likely yield high enough activation to shield the memory from pruning.

Note that our use of the term “pruning” is not meant to imply that traces are being deleted completely from memory. Rather, the pruning we describe refers to a mnemonic regulation process in which the accessibility of less-reliable memories is reduced in a graded manner. This

graded decrease in accessibility might also show up as a decrease in recognition confidence. Indeed, we obtained the same negative relationship between prediction strength and memory when we used a continuous measure of memory defined linearly over the range of high confidence new, low confidence new, low confidence old, and high confidence old ($p = .004$). We interpret reduced confidence for mispredicted items as reflecting weakening of their representation rather than interference during retrieval from a new memory trace created during the repeated triplet (see Supplementary).

Forgetting seems disadvantageous but plays an essential role in maintaining the efficiency of memory operations (Anderson, 2003). Our study sheds new light on the adaptive role of forgetting. Previous studies examined the impact of controlled retrieval on forgetting, whereby executive control processes inhibit or suppress undesirable memories competing for retrieval (Kuhl, Dudukovic, Kahn, & Wagner, 2007). Here we demonstrate for the first time that such forgetting can occur without deploying control processes, simply as a result of *automatic* retrieval during context-based prediction. Participants reported being completely unaware that contexts were repeating, suggesting that automatic retrieval occurs constantly in the background, pruning invalid memories without burdening our conscious mind.

Materials and Methods

Participants

Twenty-four adults (14 women, 19 right-handed, mean age 22.8 years) participated for monetary compensation. All participants had normal or corrected-to-normal vision and provided informed consent. The Princeton University IRB approved the study protocol.

Stimuli

Participants were shown color photographs of male and female faces, indoor and outdoor scenes, and natural and manmade objects. Stimuli were displayed on a projection screen behind the scanner, viewed with a mirror on the head coil (subtending $8.8 \times 8.8^\circ$). Participants fixated a central dot that remained onscreen.

Procedure

Participants completed one scanning session, including the incidental encoding phase, the subsequent memory test, and a functional localizer. During incidental encoding, participants viewed a sequence of faces and scenes and made male/female and indoor/outdoor judgments. Unbeknownst to them, this sequence was generated from triplets. The first two “context” items in a triplet were either both faces or both scenes, and the final item was from the other category. Each triplet used novel exemplars. The context items were shown again after other intervening items (average lag=12.5 items), with the final item replaced by a new item from the context category. Other “single” items — whose preceding context items never repeated — were inserted between triplets. Three runs of incidental encoding were collected, each lasting 8 mins 33 s and containing: 16 initial triplets ($8 A_{\text{face}} \rightarrow B_{\text{face}} \rightarrow C_{\text{scene}}$, $8 A_{\text{scene}} \rightarrow B_{\text{scene}} \rightarrow C_{\text{face}}$), their repetitions ($8 A_{\text{face}} \rightarrow B_{\text{face}} \rightarrow D_{\text{face}}$, $8 A_{\text{scene}} \rightarrow B_{\text{scene}} \rightarrow D_{\text{scene}}$, respectively), and 16 control items ($8 X_{\text{face}}$, $8 X_{\text{scene}}$). The resulting 112 trials started with a blink of fixation to signal an upcoming trial, followed by the stimulus for 1 s and a blank interval of 3.5 s.

The subsequent memory test phase began ~10 mins after encoding. The memory test was a surprise to participants. It consisted of a recognition task for 144 old items and 48 novel lure items. There were three types of old items: 48 final items each from initial (C) and repeated triplets (D), and 48 control items (X). Old items were interleaved with lures randomly.

Participants judged familiarity on a 4-point scale: 1=sure old, 2=unsure old, 3=unsure new, and 4=sure new. As in previous studies (e.g., Turk-Browne, Yi, & Chun, 2006; Wagner et al., 1998), items receiving high-confidence “sure old” responses were treated as remembered and items receiving other responses were treated as forgotten.

After the test, participants completed two runs of a functional localizer. Each run contained 15 blocks, with 5 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 (15 s duration). Each block was followed by 10.5 s of fixation, which was treated as a baseline “rest” category. Total run duration was 6 mins 40.5 s.

Data acquisition

Experiments were run with the Psychophysics Toolbox (<http://psycho toolbox.org>). Neuroimaging data were acquired using a 3T MRI scanner (Siemens Skyra) with a 16-channel head coil. We collected a scout anatomical to align axial functional slices. Whole-brain functional images for the encoding phase and functional localizer were acquired with a gradient-echo EPI sequence (TR=1.5 s; TE=28 ms; flip=64°; iPAT=2; matrix=64x64; slices=27; thickness=4 mm, resolution=3x3 mm). High-resolution (MPRAGE) and co-planar (FLASH) T1 anatomical scans were acquired for registration, along with field maps to correct B0 inhomogeneity.

Preprocessing

The 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 cut-off), and aligned to the first volume. All multivariate pattern analyses were conducted on voxels within anatomically delineated ventral temporal cortex (Kuhl, Rissman, Chun, & Wagner, 2011).

We generated these ROIs in standard space by summing left and right masks of temporal fusiform cortex and parahippocampal gyrus from the Anatomical Automatic Labeling atlas. We converted the ROIs to subject space by inverting the transformations obtained from registering functional scans to standard space.

Classification Analyses

Classification was conducted with the Princeton Multi-Voxel Pattern Analysis Toolbox (<http://www.pni.princeton.edu/mvpa>), using penalized logistic regression with L2-norm regularization (penalty=1). To validate our classifier, we first performed a cross-validation analysis within the localizer data. We trained a separate model for each of four categories — face, scene, object, and rest — using one of the localizer runs, and tested it on the other run (and then swapped training and test runs). All regressors were shifted forward in time by 4.5 s to adjust for the hemodynamic lag. For each fMRI volume in the test set, the classifier estimated the extent to which the brain activity pattern matched the patterns of activity for the four categories on which it was trained (from 0 to 1). We refer to these category level pattern-match values as *classifier evidence*.

To classify the incidental encoding data, we trained a model on both localizer runs. We operationalized prediction strength as the activation of a C item's category during the repeated triplet. Note that all items in the repeated triplet were from the other category: for example, if C was a scene, all of the items in the repeated triplet were faces. We averaged classifier evidence over the repeated triplet (3–16.5 s after trial onset, adjusted for hemodynamic lag), and calculated the difference between the evidence for C's category and A/B/D's category (e.g., scene minus face evidence for $A_{\text{face}} \rightarrow B_{\text{face}} \rightarrow D_{\text{face}}$).

We used this relative measure of category information because it has proven more sensitive in prior studies (Detre, Natarajan, Gershman, & Norman, 2013). The pattern of results was identical when we used evidence for C's category alone (without subtracting evidence for A/B/D's category). To quantify how category information changed over time, we locked trial regressors for the incidental encoding phase to trial onset time (Supplementary Fig. 2B). We then binned three volumes around the onset of the repeated triplet (Baseline) and around each item in the repeated triplet shifted forward by 3 s to account for the hemodynamic lag: Baseline=-1.5–1.5 s, A=3–6 s, B=7.5–10.5 s, D=12–15 s.

Logistic Regression Analyses

The main goal of our study was to examine the relationship between classifier evidence for C in the repeated triplet and subsequent memory for that item. Characterizing this relationship in a quantitatively precise manner required a substantial amount of data, and thus we pooled trials across participants before performing the logistic regression analyses.

There are two potential concerns with this kind of “super-subject” analysis: First, effects could be driven by a subset of participants and not generalize to the population. We therefore assessed the reliability of our results across participants (random-effects) using a bootstrap test in which we resampled entire participants with replacement and performed the same logistic regression analyses on the resampled data (Efron, 1979). This provided a population-level confidence interval for each effect in terms of the proportion of bootstrap samples (out of 1,000) in which the effect was present.

The second potential concern is that effects may not be reliable within individual participants. For example, the super-subject relationship between prediction strength and subsequent memory might reflect variance across rather than within participants. To address this

concern, we standardized prediction strength values within each participant (by z-scoring classifier evidence across C items), and re-ran all super-subject analyses. Since this eliminates differences in mean prediction strength across participants, any remaining effects are attributable to within-participant variance in prediction strength. We obtained the same pattern of results as when not standardizing: There was a negative relationship between average classifier evidence and subsequent memory ($\beta_1 = -.26, p = .006$), and the relationship in the D time period was reliably negative ($p = .004$) and significantly more negative than the relationship in the Baseline time period ($p = .026$).

We also examined the relationship between the perception of C and subsequent memory for that item. Similar to prediction strength, perception strength was defined as the relative classifier evidence for C's category. The only difference is that we measured this evidence during C's presentation in the initial triplet, 4.5 s after its onset. We then performed the same logistic regression analyses to relate perception to memory.

Chapter 2: Differentiation of Incorrectly Predicted Memories after Restudy

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 given a familiar context weakens episodic associations between the mispredicted item and the context, which leads to worse subsequent memory of 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 might need to regain its mnemonic strength. How does the brain re-strengthen such memories without simply recreating bonds to the now-irrelevant context? Based on our previous network modeling work (Norman et al., 2006, 2007), and empirical findings (Hulbert & Norman, 2014), we propose that the brain resolves this by *adaptively differentiating* the memory from its previous context.

The model's predictions are illustrated in Fig. 2.1. 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, but 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). Thus, in this example, 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 with A. This reduced mass in the 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, the unique features of B will be activated, but not features previously shared with A. Instead, because of set-point inhibition, activation will spread to other new features (not previously shared with A), and connections to these features are strengthened. This process of swapping out shared for unshared features decreases overlap between the A and B memories.

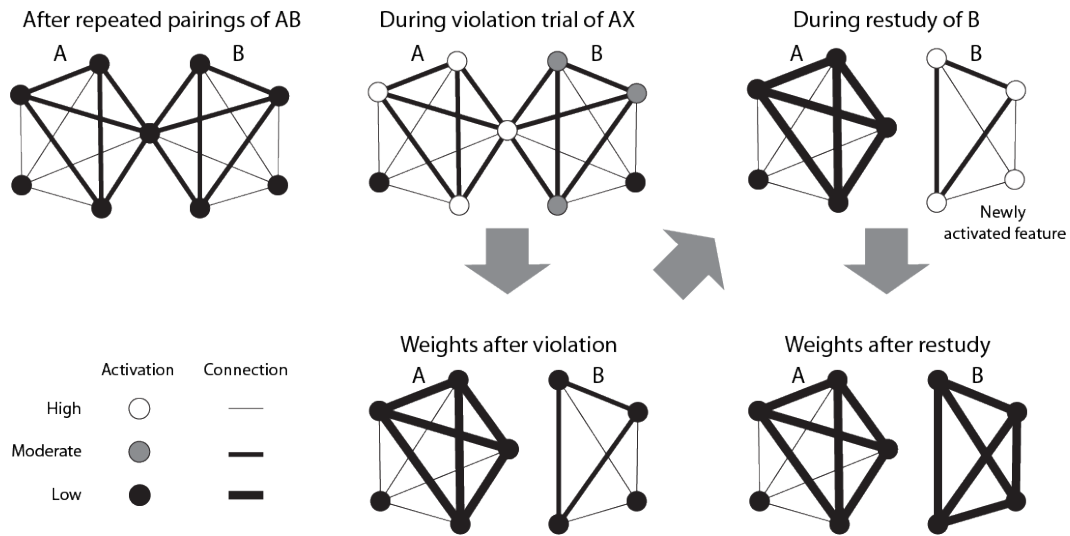


Figure 2.1. How interleaved misprediction and restudy lead to differentiation. Consider two items, A and B, that have been paired previously (AB), but in this particular instance of A, B does not appear (AX). This unconfirmed prediction leads to moderate activation of the features of B. According to our theory (Norman et al., 2006, 2007; Hulbert & Norman, 2014), 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.

Other studies have demonstrated neural differentiation from learning of inter-related materials (Favila et al., 2016; Schapiro et al., 2012; Schlichting et al., 2015). 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 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 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 to a control condition consisting of pairs that were not violated. To test this, we used fMRI pattern similarity analysis, both to track changes in neural overlap between A and B, and also to track how strongly pattern B was predicted on violation trials. Our results show, for the first time, a direct relationship between competitive dynamics (i.e., misprediction) during learning and representational change, thereby providing support for our mechanistic model of learning.

Materials and Methods

Participants

Thirty-two adults (19 women, 27 right-handed, mean age 20.06 years) participated for monetary compensation. All had normal or corrected-to-normal vision and provided informed consent. The Princeton University IRB approved the study protocol.

Stimuli

Participants were shown color photographs of indoor and outdoor scenes (including from <http://cvcl.mit.edu/MM/sceneCategories.html>), male and female faces (including from

www.macbrain.org/resources.htm), and natural and manmade objects. Stimuli were projected on a screen behind the scanner, viewed with a mirror on the head coil (subtending $8.8 \times 8.8^\circ$).

Participants fixated a black central dot that changed to white when a response was recorded.

Procedure

The experimental procedure unfolded over two days (Fig. 2.2A). All phases of the experiment were scanned with fMRI. The first session consisted of six runs of an incidental encoding task. Participants viewed face and scene stimuli one at a time and made male/female judgments for faces and indoor/outdoor judgments for scenes. Unbeknownst to participants, the stimulus sequence contained image pairs (e.g., scene A-scene B). There were eight pairs for each condition (violation and non-violation) within each run. 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 (“pre-learning snapshot”), uncontaminated by the pairmate (Schapiro et al., 2012).

There was no explicit distinction between the pre-learning snapshots and the presentation of pairs. After the items in a given pair were shown separately (to acquire pre-learning 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 (e.g., scene A-face X); this event was omitted for non-violation 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-restudy was repeated again 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 the non-violation pairs, leading to a sequence matched for the number of exposures to B just 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 pre-learning, 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 hours) and consisted of three tasks: post-learning snapshot (2 runs), surprise memory test (2 runs), and functional localizer (3 runs). The post-learning snapshots were collected in the same manner as the pre-learning 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 non-violation), randomly intermixed with 48 novel scenes. A two-step memory response was made for each scene: old or new, 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 post-learning and memory runs were interleaved and their order was counterbalanced across participants (odd participants: memory run 1, post-learning run 1, post-learning run 2, memory run 2; even participants: post-learning run 1, memory run 1, memory run 2, post-learning run 2). Half of the studied pairs were assigned to memory run 1 and post-learning run 1, and the other half to memory run 2 and post-learning run 2. In other words, for

half of the B items from the first session, memory was tested before the post-learning snapshot was taken, and for the other half of B items, the post-learning snapshot was taken before memory was tested. We designed the procedure this way because we were concerned that first testing memory for an item could contaminate a subsequent post-learning snapshot for that item, and vice-versa. We wanted some items to get an uncontaminated post-learning snapshot (before memory was tested). Although behavioral recognition memory performance was not the focus of this paper, we also wanted some items to get an uncontaminated memory judgment (before the snapshot was taken). For the present focus on neural differentiation (and related follow-up analyses), we limited analysis to the pairs for which the post-learning snapshot preceded the recognition memory test, thereby ensuring that the post-learning snapshot would not be affected by any learning that occurred as a result of recognition memory test.

After the post-learning 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 total duration of each run was 465 s. For the analyses described below, we only used the scene blocks. Specifically, we calculated a “template” activity pattern for the scene category based on the scene blocks; this template was later used to evaluate whether our results are item-specific (see below).

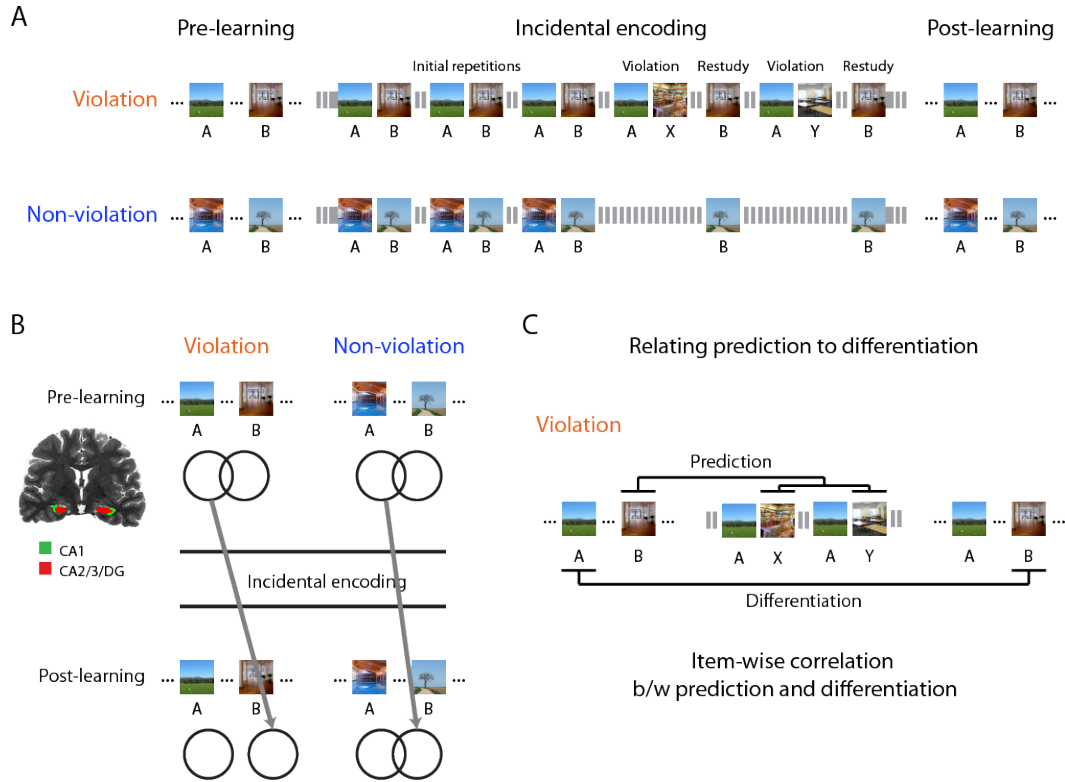


Figure 2.2. Experimental design and main analyses (A) During incidental encoding, participants performed a categorization task for scenes (indoor/outdoor) and faces (male/female). Before the encoding, all scene images were presented in a random order so that we could take pre-learning neural snapshots of the images. During the encoding, a trial sequence of 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 non-violation condition (AB-AB-AB-B-B). After the incidental encoding phase, the same scene images were presented again in a random order, where we measured post-learning neural snapshots. (B) Based on the left CA2/3/DG region (see *Region of interest*), we measured pattern similarity between A and B items across the pre-, and post-learning snapshot periods. (C) We measured prediction strength by calculating pattern similarity between pre-learning B and second members of the violation trials (X/Y in AX/AY). We then measured a correlation coefficient between prediction strength and neural differentiation in a trial-by-trial fashion.

Data acquisition

The experiment was programmed using the Psychophysics Toolbox (<http://psycho toolbox.org>).

MRI data were acquired with a 3T MRI scanner (Siemens Prisma) with a 64-channel head coil.

Functional scans came from a multi-band EPI sequence (TR = 1.5 s, TE = 39 ms, flip angle =

50°, voxel size = 1.5 mm iso), with 52 oblique axial slices (transverse to the long axis of the

hippocampus) in an interleaved order. A whole-brain T1 MPRAGE image and a co-planar T1

FLASH image were acquired for registration with 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 x 0.44 mm in-plane, 1.8 mm thick). A field map was acquired to correct for B0 inhomogeneity.

Regions of interest

This study involves rapidly learning new, arbitrary associations between stimuli. As the hippocampus supports the encoding and retrieval of such memories (Davachi, 2006; Norman & O'Reilly, 2003; Schapiro et al., in press), we expected that representational changes would occur in the hippocampus. Based on a previous study of neural differentiation (Hulbert & Norman, 2015; Schapiro et al., 2012), we focused specifically on the left hippocampus. Within the hippocampus, we were primarily interested in CA3 and dentate gyrus (DG) subfields, as they are the core storage sites for episodic memories and generate predictions via pattern completion (e.g., Hindy, Ng, & Turk-Browne, 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 & Turk-Browne, 2015), and then registered to an MNI template. Voxels in the MNI template were assigned subfield labels if the probability was greater than 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 cut-off), 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 MPAGE scans across sessions.

Measuring differentiation

For our first analysis of overall differentiation, we measured how much B's neural representation after learning had moved away from the original representation of A, and whether this differed for violation and non-violation pairs. Specifically, we measured the Pearson correlation between the pre-learning snapshot of A and the post-learning snapshot of B (Fig. 2.2B). These snapshots were defined by the spatial pattern of activity over voxels in an ROI elicited by each item, 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 (Favila et al., 2016; Schapiro et al., 2012). For example, Favila et al. (2016) measured neural overlap between competing items *within* a post-learning phase, relative to non-competing 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 non-violation condition (because of AX and AY violation trials), so any comparison between violation and non-violation conditions that includes the post-learning snapshot of A is confounded by item frequency. Consequently, we used the

pre-learning snapshot of A (prior to any difference between conditions) as the baseline for representational change.

Our hypothesis makes an item-specific claim: 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, and not just generically more distinct from other items. The basic measure of differentiation above is consistent with both possibilities — i.e., is A special or just another item? — and thus 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 re-calculated 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 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 (i.e., the matching A is not special), 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, in each participant we computed the difference between the violation and non-violation conditions in pattern similarity of pre-learning A to post-learning B, in both the original (unpermuted) data and the permuted data; then we computed the z score of this original difference relative to the 1,000 permuted differences. We then examined the reliability of these z scores compared to zero with a one-sample t-test across participants.

Relating prediction to differentiation

Beyond showing item-specific neural differentiation, a key contribution of our study is to provide an explanation for how it arises — as a result of misprediction. The main effect of differentiation for violation vs. non-violation conditions is consistent with this, as they only differed in terms of the presence of violation trials that allowed for 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 non-violation condition. To measure prediction on violation trials (i.e., AX and AY), we calculated the amount of evidence for B during the unexpected items X and Y that appeared when B should have been presented. Specifically, we measured pattern similarity between the pre-learning snapshot of B (not used to calculate differentiation) and the pattern of activity evoked by 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 score 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 particular 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 pre-learning 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 pre-learning snapshots of different B items should be interchangeable when calculating pattern similarity with X and Y. Thus, we scrambled the original pairings of pre-snapshot B and X/Y 1,000 times, re-calculated their pattern similarity, and then re-calculated the prediction-differentiation relationship. For example, if the original prediction scores were derived by correlating pre-B₁ with X₁/Y₁ (where X₁ and Y₁ followed A₁), etc., and the original differentiation scores were derived by correlating pre-A₁ with post-B₁, etc., a permutation might involve re-computing the prediction scores (comparing pre-B₇ to X₁/Y₁, etc.) but keeping the differentiation scores the same; the re-computed prediction scores are 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 from the permuted distribution for this relationship within participant and 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 prior to 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 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.

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: $t(31) = 20.99, p < .001$), and did not differ across the pre-learning snapshots, pair sequences, or post-learning snapshots ($F < 1$). For the recognition memory test, we restricted analysis to the B items that were tested before the post-learning snapshot was taken. Performance was good in general, with B items more likely to be endorsed as “old” than new items (mean $A' = 0.83, t(31) = 64.95, p < .001$). We did not have a specific expectation for a difference in memory between violation and non-violation conditions because neural differentiation can have opposite effects on recognition memory (see *Discussion*). Indeed, there was no difference ($t(31) = -.24, p > .81$) between the two conditions. Neural analyses were restricted to the other B items, whose post-learning snapshots were collected prior to 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 post-learning snapshot of B and the pre-learning snapshot of A. We focused initially on left CA2/3/DG ROI, given prior findings of differentiation (Hulbert & Norman, 2015; Schapiro et al., 2012). We hypothesized that the misprediction of B items in the violation condition would reduce subsequent neural overlap with A after restudy. Indeed, as shown in Fig. 2.3A, pattern similarity was lower for pairs from the violation vs. non-violation conditions ($t(31) = -2.82, p = .008$). We conducted additional

exploratory analyses outside of the main ROI, in other potentially relevant hippocampal subfields: right CA2/3/DG, and left and right CA1. None of these regions showed a significant difference between two conditions (left CA1: $t(31) = .36$, $p = .070$, right CA1: $t(31) = 1.88$, $p = .070$, right CA2/3/DG: $t(31) = .68$, $p = .496$).

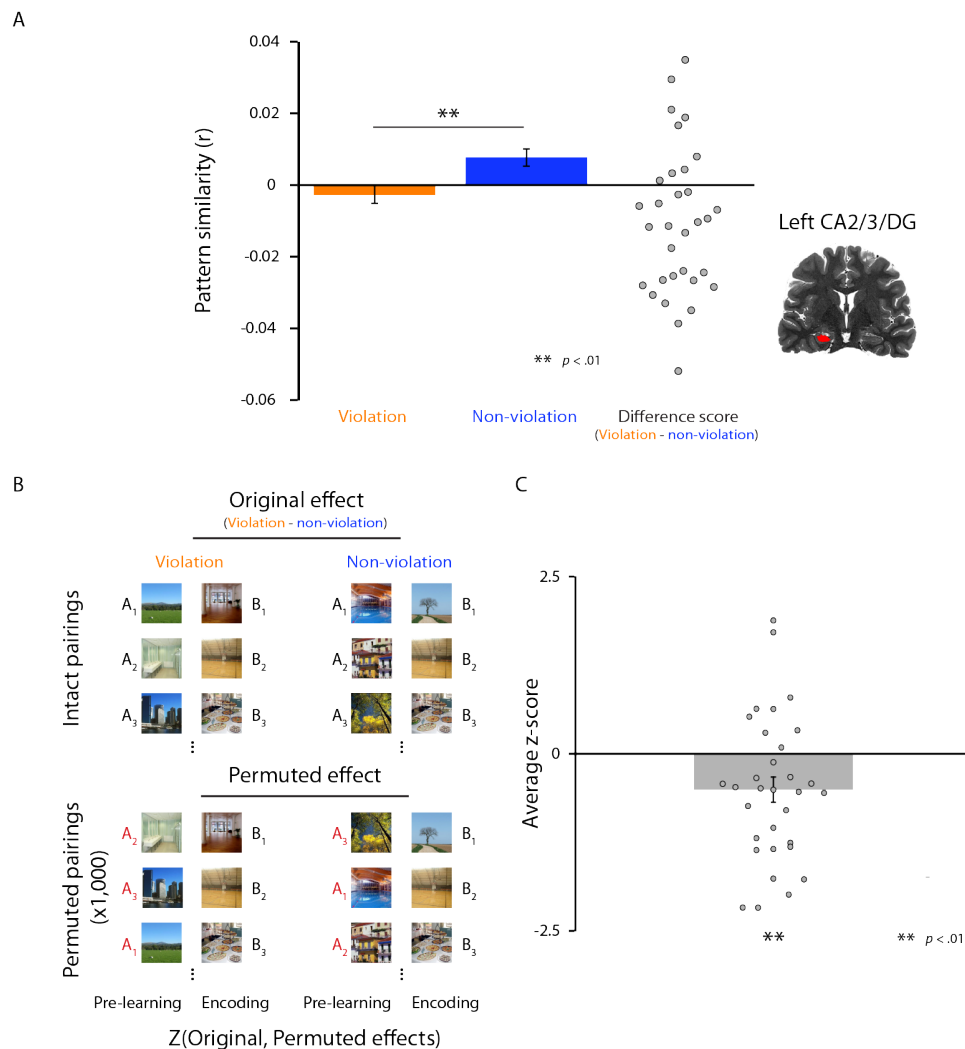


Figure 2.3. Neural differentiation effect and item-specificity test in left CA2/3/DG. (A) Pattern similarity between the pre-learning snapshot of A and the post-learning snapshot of B item was significantly lower for the violation vs. non-violation conditions. (B) Schematic of procedure for randomization analysis. For each participant, we shuffled the AB pairings 1000 times within each condition. For each iteration, we calculated the same pre-learning A to post-learning 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) Pattern similarity was lower for the original vs. permuted AB pairings, consistent with the neural differentiation effect being item-specific.

Our hypothesis explains this neural differentiation in terms of swapping out shared features with A for new features of B. Thus, 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 pre-learning snapshots of A and the post-learning snapshots of B (Fig. 2.3B). If neural differentiation is item-specific, then this scrambling should destroy the difference in pattern similarity between violation and non-violation conditions. 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 ($t(31) = 2.81, p = .009$; Fig. 2.3C).

Relationship between prediction and differentiation

The analysis above relies on the categorical manipulation of condition 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 with pre-learning B snapshot) and the strength of neural overlap for the corresponding pair (pattern similarity between pre-learning A and post-learning B snapshots). Indeed, as shown in Fig. 2.4A, there was a reliable negative correlation between these measures ($t(31) = -2.60, p = .014$).

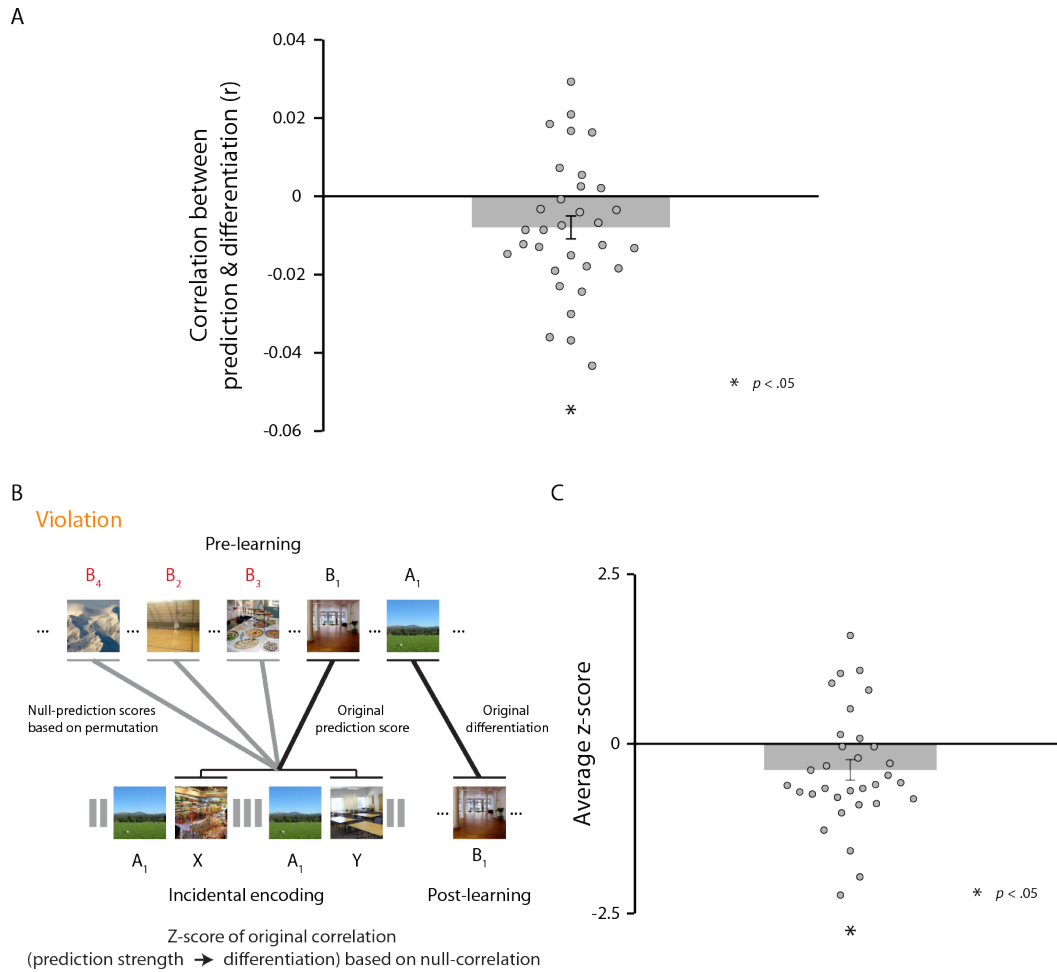


Figure 2.4. Relationship between prediction strength and differentiation, and item-specificity test for this relationship. (A) The relationship between prediction strength and differentiation was significantly negative. (B) Schematic of procedure for permutation analysis. For each participant, we permuted the pairings of the pre-learning B snapshots from the violation condition and the violation trials for the associated A items 1,000 times. We then re-calculated 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) The original relationship between prediction strength and differentiation was significantly more negative than the relationships based on permuted pairs.

As with the differentiation effect, our model also posits that prediction of B *per se* is critical, as 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 pre-learning 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. 2.4B). Consistent with item-specific prediction being the critical ingredient, the relationship between prediction strength and differentiation was stronger for the original pairings B items and violation trials, relative to the null distribution from permuted pairings ($t(31) = -2.55, p = .016$; Fig. 2.4C). Second, we regressed generic category-level information out of the pre-learning B snapshots and violation trials prior to calculating prediction scores, thus attenuating the contribution of non-item-specific features of B to pattern similarity. As expected, the negative relationship between prediction and differentiation persisted ($t(31) = -2.45, p = .020$).

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, 2016), 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 post-learning 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 univariate activation for B items in the post-learning phase is responsible for lower pattern similarity, 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 non-violation

conditions during the post-learning phase ($t(31) = -.55, p = .585$). Second, there was no trial-by-trial relationship between univariate activation in the post-learning phase and differentiation ($t(31) = .37, p = .711$). Third, the negative relationship between prediction strength and differentiation persisted after controlling for the univariate activation level with partial correlation ($t(31) = -2.63, p = .013$). These observations are consistent with our interpretation of learning reflecting differentiation in 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 activation will spread to new features that were not activated by previous presentation of B, and

that these features will be 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 approach used here).

A few other recent studies have used a similar pre-post “snapshot” approach to study differentiation (Schapiro et al., 2012; Hulbert & Norman, 2014; Favila & Kuhl, 2016; Schlichting et al., 2015). 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 (transitional 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 pre- to post-learning 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 (2014) used a retrieval practice 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 and Kuhl (2016) showed that linking two scene stimuli to a shared face associate leads to hippocampal differentiation of the scenes. 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 prior to 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 you get

differentiation vs. integration. We hypothesize that the key mediating variable is competitor activation — moderate activation followed by restudy leads to differentiation, strong activation leads to integration. For example, in Schlichting et al. (2015), blocked AB study may increase competitor activation during BC learning, tilting the balance towards integration. Likewise, regional differences in differentiation vs. integration may relate to how tightly activity is controlled: In regions like the hippocampus that have sparse activation, it is harder for related memories to activate strongly, biasing learning toward differentiation; other regions, including in cortex, with less sparse activation would be biased toward integration. 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 (see Hulbert & Norman, 2014). Pattern separation refers to the fact that the hippocampus automatically assigns distinct representations to stimuli due to sparse coding in DG and CA3 (Yassa & Stark, 2011). While 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. The brain might resolve potential interference between related memories through this active differentiation process, above and beyond pattern separation.

This study was not designed to identify the behavioral consequences of neural differentiation, although that remains an important goal of future work. Specifically, 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 (e.g., Norman & O'Reilly, 2003; Norman, 2010) and empirical studies (e.g., 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. Relatedly, our model predicts that B becomes differentiated specifically from A, but does not provide comprehensive predictions about B's relationship to other items (which would govern both interference and global match). The aforementioned study by Favila and Kuhl (2016) suggests a better way of behaviorally measuring differentiation: They found that neural differentiation between A and B facilitated subsequent learning of distinct associations with each (A-X, B-Y), presumably by reducing interference.

Summary

We found that interleaved misprediction and restudy leads to neural differentiation. These findings are in line with predictions from our neural network modeling work (Norman et al., 2006, 2007) and other recent studies in the field (Schapiro et al., 2012; Favila and Kuhl, 2016; Hulbert and Norman, 2014; Schlichting 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 in memories. This complements prior findings that activation of non-target memories can weaken them (Detre et al., 2013; Kim et al., 2014;

Newman & Norman, 2010; Poppenk and Norman, 2014), suggesting 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 avoid subsequent interference, and to minimize prediction errors.

Chapter 3: Neural overlap across episodes allows for retroactive interference

Introduction

Our experience is highly repetitive, with the same relatively small set of objects appearing repeatedly over time and often in different contexts. For example, we might move a piece of furniture to many different apartments, see somebody from work at the grocery store, or look for our car in various parking areas. How does experiencing a familiar item in a novel context affect pre-existing memories of the item and its prior contexts?

It has long been known that memory for the initial context in which an item was experienced can be impaired by a later encounter of the item in a new context. Such *retroactive interference* has been widely investigated using the AB/AC paradigm (McGovern, 1964; Postman & Underwood, 1973; Richter et al., 2016). In this paradigm, participants learn an episode with components A and B, then another episode with components A and C. Because of the shared component A, learning AC can trigger retrieval of the previously-learned AB memory. How does this memory reinstatement relate to retroactive interference (i.e., forgetting of B and/or AB)? One prominent account is that reactivation of a prior context B during later AC learning builds resistance to interference, leading to better subsequent retrieval of the initial context B when cued with A (Koen & Rugg, 2016; Kuhl, Shah, DuBrow, & Wagner, 2010).

Here we investigate a different — though not mutually exclusive — account of how memory reinstatement relates to retroactive interference. We focus on the fact that mental representations of an item can differ over time even when we putatively experience the “same” item. In the AB/AC paradigm, for example, this would correspond to variance in the extent to

which the representation of A during AC learning is the same as the representation of A during the prior AB learning. Although representational overlap across repeated presentations of an item can be associated with better memory for that item (Ward et al., 2013; Xue et al., 2010), it is unknown how such representational overlap of an item affects source memory for prior contexts with which the item was associated.

We hypothesize that retroactive interference occurs when the same item representation is reinstated across episodes with different contexts. Specifically, reinstatement of the item representation engaged by the initial processing of A (from AB learning) during AC learning allows this prior item to become associated with the novel context (C), which interferes with later retrieval of the initial context B. In contrast, if the representation of A during AC learning differs from that of the prior AB episode, memory of the initial context B might be less affected by retroactive interference. In short, we predict a negative relationship between item-specific representational overlap and subsequent source memory of an initial context. Note that this is the opposite of the positive relationship found for context (rather than item) reinstatement mentioned above, where more context information leads to greater initial source memory.

To test this hypothesis, we presented object items (A) sequentially during two different orienting tasks (B and C). These tasks served as the contexts to which the items could be bound (Johnson, Kounios, & Nolde, 1997). Using fMRI, we measured pattern similarity for a given item across the two task contexts in the lateral occipital cortex (LOC), which is thought to represent the visual features of objects (Grill-Spector et al., 2001). We then related these item-wise pattern similarity scores to subsequent source memory for the initial task. In addition to testing for the hypothesized negative effect of item reactivation, we also tested for the positive effect of context reactivation observed in previous studies (Koen & Rugg, 2016; Kuhl et al.,

2010). Such a dissociation would provide strong evidence that item and context reactivation have differential effects on source memory. Consistent with our hypothesis, we found a negative relationship between item reactivation and subsequent source memory: greater item-wise pattern similarity was associated with worse source memory of initial tasks.

Materials and methods

Overview

This study consisted of three phases. In an initial encoding phase, participants were exposed to a sequence of object images and performed one of two orienting tasks (artist or function task). These tasks served as the initial context to which the object items could be bound during the encoding phase. In an item repetition phase, half of these objects were presented again while participants performed a new orienting task (the organic task). We measured how reliably the initial representations of the items were reinstated in this phase by calculating item-specific pattern similarity between the initial and repeated presentations. In a memory test phase, source memory for the initial task was measured and related to the item-specific pattern similarity scores calculated in the second phase. A final task localizer phase was used to generate template neural patterns for each task, which were used to ensure that pattern similarity in the second phase reflected item rather than task reactivation.

Participants

Thirty-two adults (14 women, all right-handed, mean age 21.88 years) participated for monetary compensation. All participants had normal or corrected-to-normal vision and provided informed consent. The Princeton University IRB approved the study protocol.

Stimuli

Participants were shown color photographs of natural and manmade real-world objects. Stimuli were displayed on a projection screen behind the scanner bore, viewed with a mirror on the head coil (subtending $8.8 \times 8.8^\circ$). Participants fixated a central dot that remained onscreen throughout.

Procedure

Participants completed one scanning session with four phases: initial encoding, item repetition, source memory test, and task localizer. During the initial encoding phase (Fig. 3.1A), participants viewed a series of objects that were randomly assigned to one of two orienting tasks: How easy would it be to draw the object? (artist task); or, How useful is the object? (function task). Participants provided responses on a 4-point scale (artist/function): 1 = very easy/very useless, 2 = easy/useless, 3 = hard/useful, 4 = very hard/very useful. We used these tasks because previous studies have shown that they can be decoded with fMRI (Johnson, McDuff, Rugg, & Norman, 2009; Koen & Rugg, 2016; McDuff, Frankel, & Norman, 2009). Four runs of encoding were collected, and each run contained 2 blocks of objects from each of the two tasks (the order of the 4 blocks was randomized). The task was instructed with a cue at the beginning of the block (e.g., “Artist task”). Each object stimulus was presented for 1 s, followed by a blank interval of 2 s. There were 12 trials per block (48 s duration), followed by 15 s of rest. The total duration of each run was 3 min 42 s.

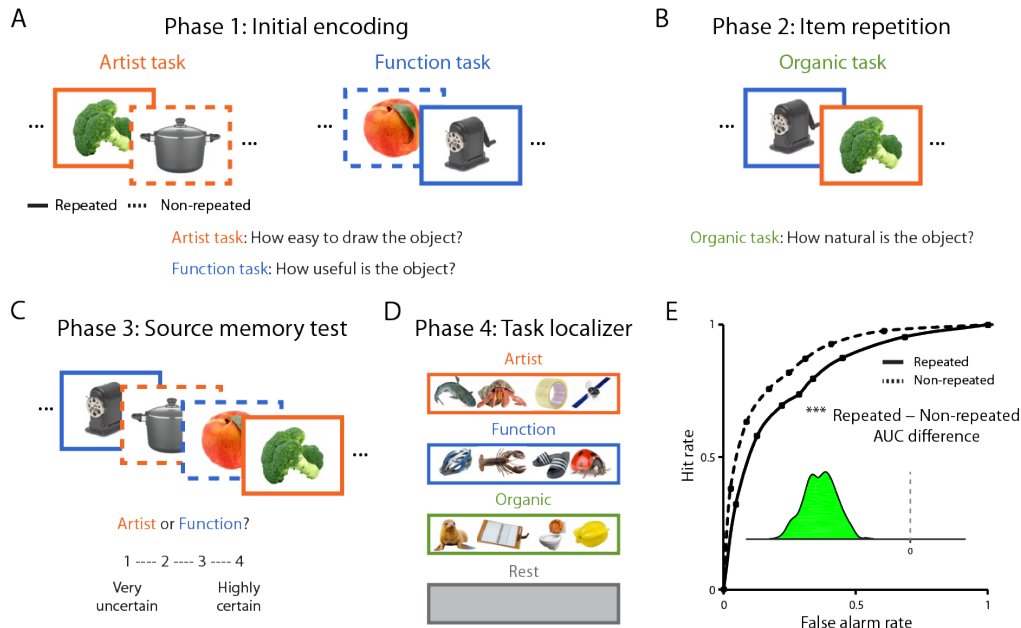


Figure 3.1 Experimental design and behavioral results. (A) During initial encoding, object images were randomly assigned to one of the two orienting tasks (artist and function tasks). (B) In the item repetition phase, half of objects from the first phase were presented again while the participants were doing a third (organic) task. (C) In the subsequent source memory test, judgments were collected about which task was first performed on each object, both for objects presented twice (repeated condition) and objects presented once (non-repeated condition). (D) In the task localizer, a new set of objects was presented in each of the three tasks to define task-specific neural activity patterns. (E) The area under the curve (AUC) of memory judgments was calculated by plotting hit rate against false alarm rate for each of eight possible memory responses: Very unsure (1) to very sure (4) “artist” or “function”. Lower AUC indicates worse memory, and so memory for the first task was worse in the repeated vs. non-repeated condition. The inset plot depicts the sampling distribution of the repeated minus non-repeated AUC difference from random-effects bootstrap resampling of participants. Almost all resampled AUC differences were below zero (green area), indicating a reliable retroactive interference effect. $*** p < .001$.

In the item repetition phase (Fig. 3.1B), half of the objects from each task in the initial encoding phase (i.e., 48 objects for each of artist and function) were presented again, and participants determined how organic the object was on a 4-point scale: 1 = very artificial, 2 = artificial, 3 = natural, 4 = very natural. Each of the 96 objects was presented for 1 s, followed by a blank interval of 3 s. All stimuli were presented in a single run without a rest period, lasting 7 min 30 s.

The source memory test (Fig. 3.1C) came as a surprise to participants. It contained the 96 objects presented in both the encoding and reactivation phases (repeated condition) and the 96

objects shown only in encoding phase (non-repeated condition). Participants were instructed to discriminate which task had been performed on each object during the initial encoding phase (i.e. artist or function), and then to report their confidence level on a 4-point scale: 1 = very unsure, 2 = unsure, 3 = sure, and 4 = very sure. Each object was presented for 6 s, though participants were encouraged to respond within 5 s. If they failed to respond on a given trial, the object was omitted from later analyses (around 3% of total trials).

After the memory test, participants completed three runs of a functional localizer (Fig. 3.1D), in which new object images were presented in one of the three tasks: artist, function, and organic. Each run contained 6 blocks, with 2 blocks from each of the three tasks in a random order. Each object was presented for 1 s, followed by a blank interval of 2 s. There were 12 trials per block (36 s duration). Each block was followed by 15 s of fixation, which was treated as a baseline “rest” category. Total run duration was 3 mins 36 s.

Behavioral analysis

We measured memory performance by dividing responses from the source memory test into eight levels of confidence: 4 = very sure “artist” to -4: very sure “function”. These judgments were quantified using receiver operating characteristic (ROC) analyses (Macmillan & Creelman, 2005; Green & Swets, 1966). For each of the repeated and non-repeated conditions, we created an ROC curve across the eight confidence levels and calculated the area under the curve (AUC). Calculating these curves precisely requires a substantial amount of data, and thus we pooled trials across participants beforehand. We assessed the reliability of the AUC difference between conditions across participants (random-effects) using a bootstrapping approach in which entire participants were resampled with replacement 1000 times (Efron, 1979). This provided a

population-level confidence interval (CI) for each effect, and also allowed for null hypothesis testing based on the proportion of bootstrapped samples in which the effect was reversed.

Data acquisition

Experiments were run with the Psychophysics Toolbox (<http://psychtoolbox.org>). Neuroimaging data were acquired using a 3T MRI scanner (Siemens Skyra) with a 16-channel head coil. A scout anatomical scan was used to align axial functional slices. Functional images covering the whole brain were acquired with a T2* gradient-echo EPI sequence (TR = 1.5 s; TE = 28 ms; flip = 64°; iPAT = 2; matrix = 64x64; slices = 26; thickness = 4 mm, resolution = 3x3 mm). High-resolution (MPRAGE) and co-planar (FLASH) T1 anatomical scans were acquired for registration, along with field maps to correct B0 inhomogeneities.

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 cut-off), spatially normalized (5mm FWHM), and aligned to the middle volume.

Selection of ROIs

We defined ROIs for object processing (LOC) and for task processing. LOC was defined anatomically from the Harvard-Oxford cortical atlas in FSL and transformed into each participant's space. The task ROI was defined in two steps: (1) We picked voxels selective to each of the three tasks (artist, function, and organic) by performing a general linear model (GLM) analysis of the localizer, with regressors for the each of four tasks (artist, function, organic, and rest). We ran three contrasts (artist vs. others, function vs. others, and organic vs. others) and selected voxels whose absolute z -values were above 2.3 ($p < .01$). (2) We then took the union of the surviving voxels of each contrast.

Measuring item reactivation

We measured how reliably the initial representation of each item was reinstated in the second phase by calculating the Pearson correlation of the patterns of activity elicited in the LOC on the initial and repeated presentations 4.5 s after onset.

Measuring generic task information

We measured the amount of task information in the first phase (to index task encoding) and in the second phase (to index task reactivation). Based on the localizer, we defined an activation template for each of the initial tasks (i.e., artist and function) over the voxels of the task mask by averaging patterns from the corresponding task (e.g., patterns from the artist task). Patterns of activity from the first and second phases were then correlated with the templates. We subtracted the correlation with the irrelevant task template from that of relevant template to get a single index of task reactivation (e.g., for an item whose initial task was artist, the artist template is relevant and the function template is irrelevant).

Relating neural measures to source memory

We measured the relationship between each of the neural measures (item reactivation, task encoding, and task reactivation) and subsequent source memory. For each neural measure, we divided trials into high vs. low pattern similarity based on a median split, and computed an ROC curve within each split. We then measured the area under each of the ROC curves (AUC), which reflects sensitivity of source memory (i.e., larger AUC is associated with better source memory). To calculate the effect of pattern similarity, we computed the difference between the high-pattern-similarity AUC and the low-pattern-similarity AUC. The sign of this AUC difference represents the direction of the relationship between the neural measure and memory: positive

sign = positive relationship (i.e., better memory when pattern similarity is higher) and negative sign = negative relationship (i.e., worse memory when pattern similarity is higher).

Simulations to identify the best way to establish item specificity

We were mainly interested in the relationship between item-specific pattern similarity and source memory. However, our basic measure of item reinstatement (pattern similarity between initial and repeated presentations of an item) could be confounded by generic reactivation of the task from the first presentation during the second presentation. Such task-based pattern similarity could also be related to source memory: The more engaged participants were in the task during the initial presentation, the more that the task would get associated with the object; this could lead to both greater automatic reactivation of the task during the second presentation and to better source memory for the task when cued with the object in the test.

We set out to find a measure of item-specific pattern similarity that avoids this confound. To evaluate the susceptibility of different measures to this confound, we performed two simulations: (1) where pattern similarity was driven by specific information about the item, and (2) where pattern similarity was driven by generic information about the task. We then considered how analysis approaches used previously (Kim et al., 2014; Koen & Rugg, 2016) picked up on these sources of information.

Approach 1: same minus different analysis

Similar to our study, Koen and Rugg (2016) investigated the effect of item-specific task reactivation on source memory. To measure item-specific reactivation (while excluding task reactivation), they averaged pattern similarity across pairs of different items (e.g., A to B, C, D, ...), and subtracted it from pattern similarity for the same item pair (e.g., A to A); the same minus different item pattern similarity was then related to memory. In our simulations, this

measure reliably predicted memory *for both cases* ($ps < .001$), that is, even when there was no item-specific information. This result can be explained by the fact that the same minus different pattern similarity can still contain task information when there is item-wise variance in task reactivation. Thus, this analysis does not decisively discriminate between item and task reactivation effects on memory.

Approach 2: permutation analysis

We previously used permutation analysis to establish item specificity (Kim et al., 2014). This involves scrambling the pairings of items 1000 times and, for each scramble, re-calculating pattern similarity and its relationship to memory. A z-score of the original effect (based on the intact pairings) can then be calculated with respect to this null distribution. This analysis provided different results across the two simulations. When item reactivation drives memory, the original effect was reliably greater than the permuted effects ($p < .001$). In contrast, when task reactivation drives memory, the original effect was not different from the permuted effects. This pattern of results confirms that the permutation analysis can identify item-specific relationships and is not fooled by item-wise variance in task reactivation.

Approach 3: regression + permutation analysis

Although the permutation analysis handles the case where task reactivation in the second phase varies across items, it fails if task activation at encoding (for a given item) is correlated with task reactivation for that item at retrieval. That is, if a task is really active during encoding, it might be reactivated more during the item repetition, and permuting the item pairings will eliminate this, giving the appearance of item-specific information. Indeed, in a further simulation of this situation, the original effect is stronger than the permuted distribution ($p = .008$). To address this issue, we adopted a different approach of regressing out template activity patterns for each task

from every item's representation prior to calculating pattern similarity and its relationship to source memory. After removing task information in this manner on a trial-by-trial basis, the item-specific case survives the permutation test, but both forms of the task simulation — (1) item-wise variance in task reactivation alone, and (2) correlated item-wise variance in task encoding and task reactivation — fail for the first time ($ps > .64$).

Having confirmed the validity of regression + permutation analysis, we applied it to our data in three steps: First, we regressed out the localizer template for each task from the corresponding patterns in the first and second phase (e.g., for an item whose initial task was artist, we regressed out the artist template from both patterns for that item). We then measured item-wise pattern similarity in the residual patterns, and related it to source memory by measuring the AUC difference for items with high vs. low pattern similarity. Second, we performed a permutation analysis by scrambling the original pairings of first and second phase trials within each task and participant (e.g., permuting items whose initial task was artist). For each of 1,000 permutations, we re-calculated pattern similarity and related it to memory. Finally, a z-score of the original relationship (AUC difference score) based on the intact pairings was calculated with respect to the null distribution of AUC differences.

Results

Subsequent source memory behavior

Overall, participants successfully discriminated the correct from incorrect initial task (mean $A' = 0.56$, bootstrap $p < .001$). Consistent with the idea that item repetition increases retroactive interference, source memory for the initial task was lower for repeated compared to non-repeated items (AUC difference = -0.07, CI = -0.09– -0.04, bootstrap $p < .001$; Fig. 3.1E).

Initial pattern similarity results

We first confirmed that LOC activation patterns contained information about items, by showing that pattern similarity between the first and second phases was greater when the item was the same vs. when two different items from the same task were compared (bootstrap $p < .001$; Fig. 3.2A). We next confirmed that activation patterns from the task masks (which were defined from the localizer) contained task information in the first phase. Indeed, across trials, pattern similarity was higher for relevant than the irrelevant task template (bootstrap $p < .001$; Fig. 3.2B).

Task encoding results

Several previous studies (Gordon, Rissman, Kiani, & Wagner, 2014; Kim, Lewis-Peacock, Norman, & Turk-Browne, 2014; Koen & Rugg, 2016; Kuhl, Rissman, & Wagner, 2012) have reported positive relationships between multivariate measures of encoding strength and subsequent memory. Consistent with this, greater neural evidence for the corresponding task template during encoding was associated with better task memory on the final test (AUC difference = 0.05, CI = 0.02 – 0.09, bootstrap $p < .001$; Fig. 3.2C).

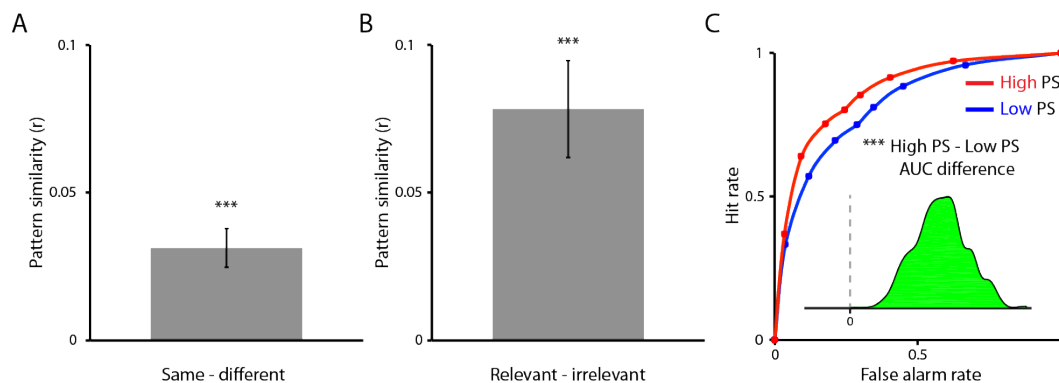


Figure 3.2. Manipulation checks. (A) Pattern similarity in LOC was higher on average between the same items from the first and second phases than between different items. (B) During the first phase, pattern similarity was higher for the relevant (currently-performed) task than the irrelevant task. (C) Greater evidence for the corresponding task template during encoding was associated with better task memory on the final test, as reflected in a greater AUC for high vs. low pattern similarity. *** $p < .001$.

Relationship between item reactivation and subsequent source memory

We hypothesized that reactivation of the item representation from the first phase during the second phase would lead to retroactive interference with the source memory for the initial task. Consistent with our hypothesis, higher pattern similarity between the first and second phases was associated with worse source memory (AUC difference = -0.04, CI = -0.08 – 0.002, bootstrap $p = .008$; Fig. 3.3C).

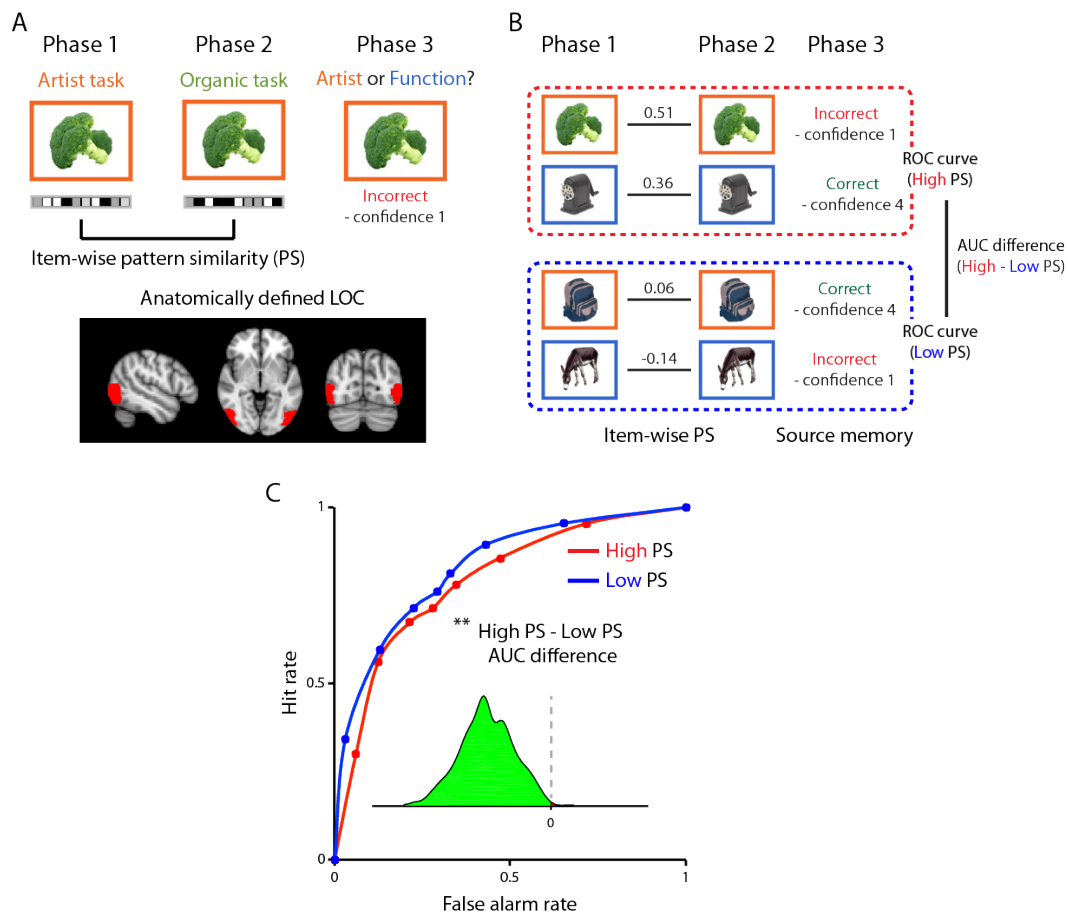


Figure 3.3. Relating pattern similarity to source memory across items. (A) Pattern similarity was measured for each item between the first and second phases. (B) After performing a median split of the pattern similarity scores, an ROC curve was constructed for high and low pattern similarity items from the source memory test and AUC was calculated. A difference in these AUCs indicates a relationship between pattern similarity and source memory, and the sign indicates the direction of the relationship. (C) There was a reliable negative relationship between pattern similarity and source memory across items. $**p < .01$

This observed negative relationship might be driven by item-wise variance in task reactivation rather than item reactivation. We controlled for this confound using a two-step permutation analysis (see Methods): First, we regressed out generic task information from patterns in the first and second phases (Fig. 3.4A), and re-calculated pattern similarity and its relationship to source memory. Second, we scrambled the original pairings of items across the first and second phases and re-calculated the relationship for each permutation (Fig. 3.4C). If item reactivation modulates subsequent source memory, as hypothesized, then the relationship based on residual intact pairings (excluding task information) should be more negative than the null distribution of relationships from permuted pairings (excluding both task and item information). After regressing out task information, the relationship based on intact pairings remained negative (AUC difference = -0.03, CI = -0.06 – 0.003, bootstrap p = .020; Fig. 3.4B). In addition, this relationship was significantly more negative than the null distribution of permuted relationships (average z -score = -1.28, CI = -2.97 – 0.44, one-tailed bootstrap p = .029; Fig. 3.4D).

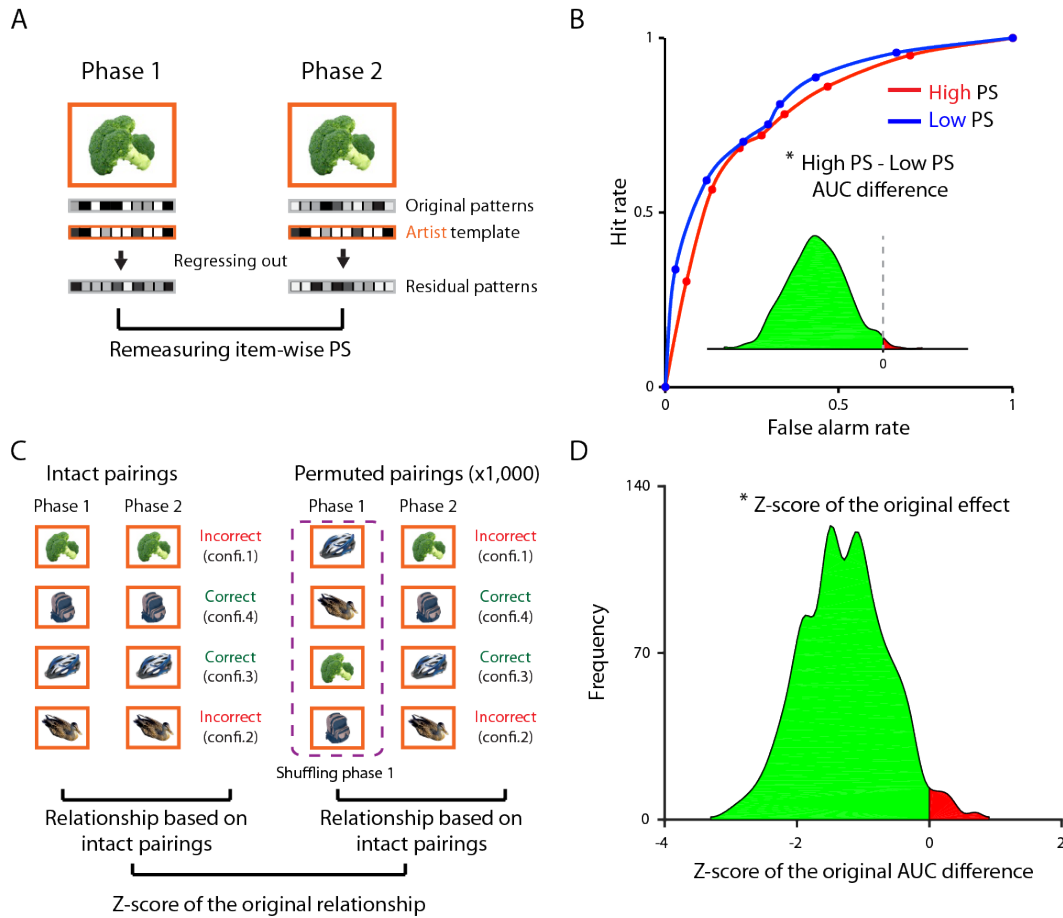


Figure 3.4. Testing specificity of item reinstatement effects. (A) Generic task information was regressed out of the item patterns from the first and second phase using the task templates from the localizer. (B) Using the residual patterns, the relationship between item-wise pattern similarity and source memory was re-calculated and remained reliably negative. (C) We further narrowed in on item-specific variance by submitting the residual item patterns to a permutation test. * $p < .05$. (D) The z-score of the original negative relationship was more negative than the null relationships calculated after permuting the item pairings between the first and second phases. * $p < .05$ (one-tailed).

Relationship between task reactivation and subsequent source memory

In addition to examining the effect of item reinstatement on initial source memory, we also considered the effect of task reinstatement. We measured the amount of information about the initial task when each item was repeated in the second phase by calculating pattern similarity with the task templates from the localizer (Fig. 3.5A and 3.5B). In contrast to the negative effect of item reinstatement, there was a numerically positive effect of task reinstatement, although it

failed to reach significance (AUC difference = 0.02, CI = -0.02 – 0.06, bootstrap $p = .23$; Fig 3.5C).

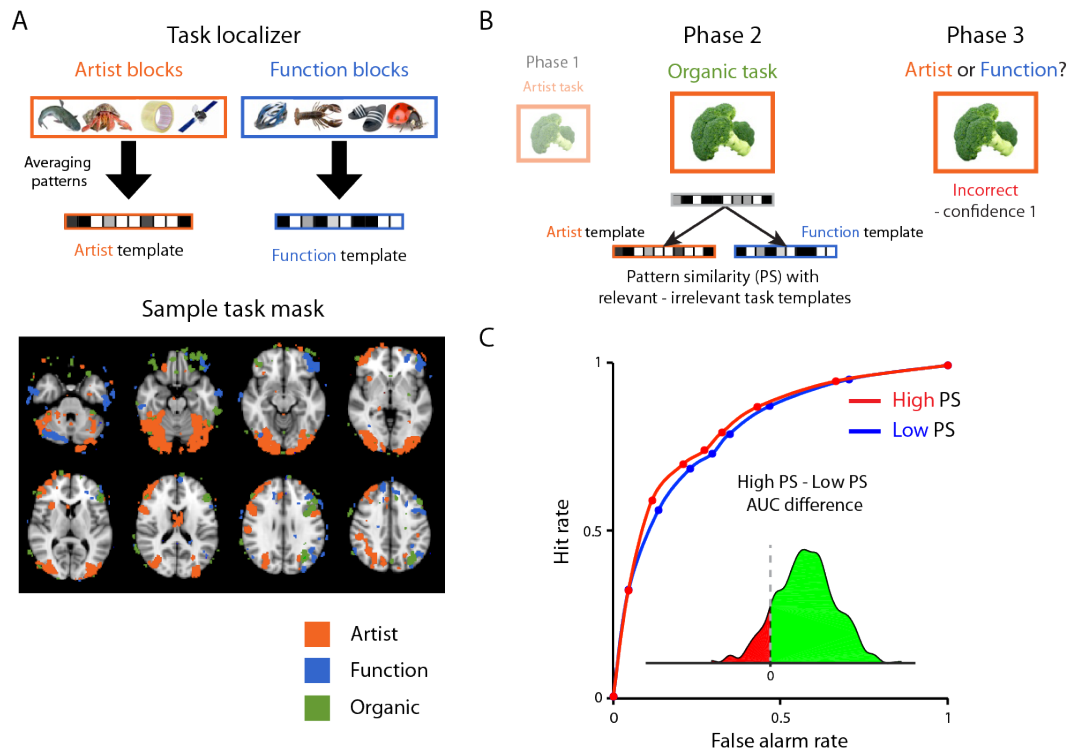


Figure 3.5. Relating task reactivation to memory. (A) We generated template activity patterns for each task from the localizer by selecting task-selective voxels with a GLM and then averaging the activity values in each voxel across all trials of the task. (B) We measured reactivation of the initial tasks (artist or function) in the second phase as the pattern similarity of each item with its corresponding task template minus its non-corresponding task template (both of which were different from the current organic task). (C) The relationship between this index of task reactivation and subsequent source memory was numerically positive, but not significant.

Discussion

We investigated how memory for the context in which an object is encountered is influenced by encountering it again in a novel context. Although the basic behavioral result — that subsequent source memory for the initial context is impaired — is a well-known form of retroactive interference, we tested a novel explanation for this important phenomenon. Specifically, we hypothesized that the extent to which the item is represented the same way across contexts

determines how strongly the new context gets associated with the item, and in turn how much interference the old context suffers when it later needs to be retrieved. Consistent with this hypothesis, we found a negative relationship between neural overlap across item repetitions and subsequent source memory for the initial context.

We used pattern similarity to index how much the initial item representation was reactivated upon repetition. However, taken at face value, this measure does not necessarily reflect item information alone – it can also be influenced by variance in task reactivation. We addressed this issue by showing (a) that the negative relationship persists after regressing out task information, (b) that it is eliminated after permuting item pairings within task, and (c) that task reactivation *per se* leads to an effect in the opposite direction. These results together strongly support our main conclusion that reactivation of an item-specific representation leads to retroactive interference.

Using a paradigm similar to ours, Koen and Rugg (2016) also investigated the consequence of item reactivation on task memory. As in our study, participants performed one task on an item during an initial phase; then performed a different task on that item during a later phase; later, participants were tested on which task they initially performed on the item. Interestingly, Koen & Rugg (2016) found a *positive* relationship between reactivation and memory, which is the opposite of what we found. There are several important differences between their study and ours, and so it is hard to know exactly which is responsible for the discrepancy. One possibility, supported by our simulations, is that the same minus different pattern similarity measure they used to establish item specificity can in fact be influenced by generic task reactivation (i.e., task reactivation on its own can cause a “same minus different” task effect that is above zero). However, we propose this possibility with caution because they

found no correlation between task reactivation, measured using a pattern classifier, and item reactivation, measured using pattern similarity analysis.

Another important difference in the memory test paradigm. In Koen and Rugg (2016), participants were required to recall *both* the first and second tasks during which items were encountered. In contrast, we used a 2AFC paradigm that required choosing between the two possible initial tasks. Compared to their recall paradigm, performance in our test may have been more influenced by associations with the second task. Specifically, when both the first and second tasks are reactivated in the recall test, the best strategy would be to report both, irrespective of any difference in associative strength. In our case, a stronger non-target association with the organic test may have made choosing the correct initial task more difficult. Thus, we conjecture that our memory paradigm might be more sensitive to interference.

A final difference between Koen and Rugg (2016) and our study was the stimulus modalities used. Their materials were words, whereas ours were objects. This allowed us to concentrate our analyses on the LOC, a region known to represent item-level features of objects. In contrast, they measured pattern similarity from a task-selective mask, which may have contained weaker item-level information. They may have thus had less distinctive item representations and less variability in the extent to which they were reactivated across repetitions.

In conclusion, our episodic experiences are highly overlapping, and new learning can reduce the accessibility of earlier related memories. Thus, it is an important challenge to understand how memory systems balance between being sensitive to new experiences while retaining and protecting traces of old experiences (McClelland, McNaughton, & O'Reilly, 1995; Norman & O'Reilly, 2003). Recent fMRI studies (Kuhl et al., 2010; Koen and Rugg, 2016) have found that reactivation of an old memory during new learning makes the old memory more

resistant to retroactive interference. Here we elaborate on this by reporting a new factor affecting retroactive interference: More reliable reinstatement of the initial representation of a *shared* component (here, the item) across old and new experiences can increase interference with a *unique* component of the old experience (here, the initial task).

Conclusions

This work has helped advance our understanding of how the brain resolves competition between overlapping memories. Much of the work has been motivated by a widely held view that rapid relational learning is highly prone to competition when experiences are substantially similar to each other. Based on the nonmonotonic plasticity hypothesis (Norman et al., 2005, 2006, 2007), we provided empirical evidence that the brain re-organizes overlapping memories based on competition signals.

In Chapter 1, we revealed that memories of incorrectly predicted items are weakened by context-based prediction error, and we interpreted the findings in terms of our model in which context-based predictions reactivate item representations just enough for them to be weakened during a misprediction. In Chapter 2, we demonstrated that the weakened item representations are differentiated from the initial context that triggered the misprediction.

Our findings are different from previous studies that examined the impact of controlled retrieval on forgetting and underlying neural representational change (Levy & Anderson, 2002; Wimber, Alink, Charest, Kriegeskorte, & Anderson, 2015). We demonstrated that such behavioral and neural changes could occur simply as a result of automatic memory reactivation during context-based prediction, which does not necessarily rely on control processes. However, it should be noted that our findings do not disprove the potential role of the control processes. Rather we argue that, in addition to the control processes, the local learning processes, which operate on the basis of memory competition, might have different levels of influence on the changes in memory.

What is the adaptive value of neural differentiation observed in our work? We suggest that differentiation might help reduce potential later interference. For example, an item

differentiated from its previous context might be associated more efficiently with a new context due to the reduced competition from the old context. However, we do not have direct behavioral evidence to support that speculation. Future studies should examine the potential benefit of neural differentiation (see Favila & Kuhl, 2016).

Throughout Chapters 1 and 2, we investigated how the brain minimizes potential competition by reorganizing *existing* memory representations. In Chapter 3, we revealed another *online* mechanism whereby initial item reinstatement determines the influence of overlap on new learning.

Though previous studies have found that existing memories can be modulated by their reactivation during new learning, the findings are somewhat mixed. Whereas some evidence supports the claim that reactivation makes old memories susceptible to retroactive interference (Forcato et al., 2007, 2009; Gershman, Schapiro, Hupbach, & Norman, 2013; Hupbach, Gomez, Hardt, & Nadel, 2007; Sederberg, Gershman, Polyn, & Norman, 2011), other studies observed the opposite effect that that reactivation alleviates retroactive interference (Kuhl et al., 2010; Koen and Rugg, 2016). Our findings provide some hints on how the conflicting results can be resolved. We argue that the effect of reactivation on memory depends on the contents of reactivation: If unique features of an initial experience (e.g., B in AB learning) are reactivated during new learning (e.g., AC learning), it becomes more resistant to interference from novel information. However, if shared features (e.g., A in AB/AC learning) across the two overlapping experiences are reactivated, then it is possible for these reactivated features to be associated with novel information that will amplify the retroactive interference effect.

Although our ability of relational learning enables us to predict the future based on past experiences, overlapping memories can cause mental clutter. This dissertation work has made some steps forward in understanding how the brain resolves memory competition.

Reference

- Aggleton, J. P., & Brown, M. W. (1999). Episodic memory, amnesia, and the hippocampal-anterior thalamic axis. *The Behavioral and Brain Sciences*, 22(3), 425-444-489.
- Aly, M., & Turk-Browne, N. B. (2015). Attention Stabilizes Representations in the Human Hippocampus. *Cerebral Cortex*, 26(2), 783–796
- Aly, M., & Turk-Browne, N. B. (2016). Attention promotes episodic encoding by stabilizing hippocampal representations. *Proceedings of the National Academy of Sciences*, 113(4), E420–E429.
- Anderson, M. C. (2003). Rethinking interference theory: Executive control and the mechanisms of forgetting. *Journal of Memory and Language*, 49(4), 415–445.
- Anderson, M. C., & Neely, J. H. (1996). Interference and inhibition in memory retrieval. In E. L. Bjork & R. A. Bjork (Eds.), *Memory, Handbook of perception and cognition* (2 ed., Vol. 10, pp. 237-313). San Diego: Academic Press.
- Anderson, M. C., Bjork, E. L., & Bjork, R. A. (2000). Retrieval-induced forgetting: Evidence for a recall-specific mechanism. *Psychonomic Bulletin & Review*, 7(3), 522–530.
- Anderson, M. C., Bjork, R. A., & Bjork, E. L. (1994). Remembering can cause forgetting: retrieval dynamics in long-term memory. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, 20(5), 1063–1087.
- Anderson, M. C., & Green, C. (2001). Suppressing unwanted memories by executive control. *Nature*, 410(6826), 366–369.
- Andersson JL, Jenkinson M, 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 University, Oxford, UK
- Aron, A. R., Robbins, T. W., & Poldrack, R. A. (2004). Inhibition and the right inferior frontal cortex. *Trends in Cognitive Sciences*, 8(4), 170–177.
- Artola, A., Bröcher, S., & Singer, W. (1990). Different voltage-dependent thresholds for inducing long-term depression and long-term potentiation in slices of rat visual cortex. *Nature*, 347(6288), 69–72.
- Bar, M. (2007). The proactive brain: using analogies and associations to generate predictions. *Trends in Cognitive Sciences*, 11(7), 280–289.
- Bonnici, H. M., Kumaran, D., Chadwick, M. J., Weiskopf, N., Hassabis, D., & Maguire, E. A. (2012). Decoding representations of scenes in the medial temporal lobes. *Hippocampus*, 22(5), 1143–1153.
- Brewer, J. B., Zhao, Z., Desmond, J. E., Glover, G. H., & Gabrieli, J. D. E. (1998). Making Memories: Brain Activity that Predicts How Well Visual Experience Will Be Remembered. *Science*, 281(5380), 1185–1187.
- Brown, A. D., Kouri, N., & Hirst, W. (2012). Memory's Malleability: Its Role in Shaping Collective Memory and Social Identity. *Frontiers in Psychology*, 3.
- Carr, V. A., Rissman, J., & Wagner, A. D. (2010). Imaging the Human Medial Temporal Lobe with High-Resolution fMRI. *Neuron*, 65(3), 298–308.
- Ciranni, M. A., & Shimamura, A. P. (1999). Retrieval-induced forgetting in episodic memory. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, 25(6), 1403–1414.

- Cohen, N. J., & Eichenbaum, H. (1993). Memory, amnesia, and the hippocampal system (Vol. xii). Cambridge, MA, US: The MIT Press.
- Coman, A., Manier, D., & Hirst, W. (2009). Forgetting the Unforgettable Through Conversation Socially Shared Retrieval-Induced Forgetting of September 11 Memories. *Psychological Science*, 20(5), 627–633.
- Coutanche, M. N. (2013). Distinguishing multi-voxel patterns and mean activation: Why, how, and what does it tell us? *Cognitive, Affective, & Behavioral Neuroscience*, 13(3), 667–673.
- Davachi, L. (2006). Item, context and relational episodic encoding in humans. *Current Opinion in Neurobiology*, 16(6), 693–700.
- Davis, T., LaRocque, K. F., Mumford, J. A., Norman, K. A., Wagner, A. D., & Poldrack, R. A. (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.
- Davis, T., & Poldrack, R. A. (2013). Measuring neural representations with fMRI: practices and pitfalls. *Annals of the New York Academy of Sciences*, 1296, 108–134.
- Detre, G. J., Natarajan, A., Gershman, S. J., & Norman, K. A. (2013). Moderate levels of activation lead to forgetting in the think/no-think paradigm. *Neuropsychologia*, 51(12), 2371–2388.
- Duncan, K., Sadanand, A., & Davachi, L. (2012). Memory's Penumbra: Episodic Memory Decisions Induce Lingering Mnemonic Biases. *Science*, 337(6093), 485–487.
- Duvernoy, H. M., Cattin, F., & Risold, P.-Y. (2013). The Human Hippocampus. Berlin, Heidelberg: Springer.
- Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1), 1–26.
- Fan, J. E., & Turk-Browne, N. B. (2013). Internal attention to features in visual short-term memory guides object learning. *Cognition*, 129(2), 292–308.
- Favila, S. E., Chanales, A. J. H., & Kuhl, B. A. (2016). Experience-dependent hippocampal pattern differentiation prevents interference during subsequent learning. *Nature Communications*, 7, 11066.
- Forcato, C., Burgos, V. L., Argibay, P. F., Molina, V. A., Pedreira, M. E., & Maldonado, H. (2007). Reconsolidation of declarative memory in humans. *Learning & Memory*, 14(4), 295–303.
- Forcato, C., Burgos, V. L., Argibay, P. F., Molina, V. A., Pedreira, M. E., & Maldonado, H. (2007). Reconsolidation of declarative memory in humans. *Learning & Memory*, 14(4), 295–303.
- Gershman, S. J., Schapiro, A. C., Hupbach, A., & Norman, K. A. (2013). Neural Context Reinstatement Predicts Memory Misattribution. *The Journal of Neuroscience*, 33(20), 8590–8595.
- Green, D. M., & Swets, J. A. (1966). Signal detection theory and psycho- physics. New York, NY: Wiley.
- Grill-Spector, K., Henson, R., & Martin, A. (2006). Repetition and the brain: neural models of stimulus-specific effects. *Trends in Cognitive Sciences*, 10(1), 14–23.
- Grill-Spector, K., Kourtzi, Z., & Kanwisher, N. (2001). The lateral occipital complex and its role in object recognition. *Vision Research*, 41(10–11), 1409–1422.

- Hansel, C., Artola, A., & Singer, W. (1996). Different threshold levels of postsynaptic $[Ca^{2+}]_i$ have to be reached to induce LTP and LTD in neocortical pyramidal cells. *Journal of Physiology-Paris*, 90(5–6), 317–319.
- Hindy, N. C., Ng, F. Y., & Turk-Browne, N. B. (2016). Linking pattern completion in the hippocampus to predictive coding in visual cortex. *Nature Neuroscience*, 19(5), 665–667.
- Hirsh, R. (1974). The hippocampus and contextual retrieval of information from memory: A theory. *Behavioral Biology*, 12(4), 421–444.
- Howard, M. W., & Kahana, M. J. (2002). A Distributed Representation of Temporal Context. *Journal of Mathematical Psychology*, 46(3), 269–299.
- Hulbert, J. C., & Norman, K. A. (2014). Neural Differentiation Tracks Improved Recall of Competing Memories Following Interleaved Study and Retrieval Practice. *Cerebral Cortex*, 25(10), 3994–4008.
- Hupbach, A., Gomez, R., Hardt, O., & Nadel, L. (2007). Reconsolidation of episodic memories: A subtle reminder triggers integration of new information. *Learning & Memory*, 14(1–2), 47–53.
- Johnson, J. D., McDuff, S. G. R., Rugg, M. D., & Norman, K. A. (2009). Recollection, Familiarity, and Cortical Reinstatement: A Multivoxel Pattern Analysis. *Neuron*, 63(5), 697–708.
- Johnson, M., Kounios, J., & Nolde, S. (1997). Electrophysiological brain activity and memory source monitoring. *Neuroreport*, 8(5), 1317–1320.
- Johnson, M. R., Mitchell, K. J., Raye, C. L., D’Esposito, M., & Johnson, M. K. (2007). A brief thought can modulate activity in extrastriate visual areas: Top-down effects of refreshing just-seen visual stimuli. *NeuroImage*, 37(1), 290–299.
- Johnson, S. K., & Anderson, M. C. (2004). The role of inhibitory control in forgetting semantic knowledge. *Psychological Science*, 15(7), 448–453.
- Kim, G., Lewis-Peacock, J. A., Norman, K. A., & Turk-Browne, N. B. (2014). Pruning of memories by context-based prediction error. *Proceedings of the National Academy of Sciences*, 111(24), 8997–9002.
- Koen, J. D., & Rugg, M. D. (2016). Memory Reactivation Predicts Resistance to Retroactive Interference: Evidence from Multivariate Classification and Pattern Similarity Analyses. *The Journal of Neuroscience*, 36(15), 4389–4399.
- Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational Similarity Analysis - Connecting the Branches of Systems Neuroscience. *Frontiers in Systems Neuroscience*, 2.
- Kuhl, B. A., Dudukovic, N. M., Kahn, I., & Wagner, A. D. (2007). Decreased demands on cognitive control reveal the neural processing benefits of forgetting. *Nature Neuroscience*, 10(7), 908–914.
- Kuhl, B. A., Rissman, J., & Wagner, A. D. (2012). Multi-voxel patterns of visual category representation during episodic encoding are predictive of subsequent memory. *Neuropsychologia*, 50(4), 458–469.
- Kuhl, B. A., Shah, A. T., DuBrow, S., & Wagner, A. D. (2010). Resistance to forgetting associated with hippocampus-mediated reactivation during new learning. *Nature Neuroscience*, 13(4), 501–506.
- LaRocque, K. F., Smith, M. E., Carr, V. A., Witthoft, N., Grill-Spector, K., & Wagner, A. D. (2013). Global Similarity and Pattern Separation in the Human Medial Temporal Lobe Predict Subsequent Memory. *The Journal of Neuroscience*, 33(13), 5466–5474.

- Levy, B. J., & Anderson, M. C. (2002). Inhibitory processes and the control of memory retrieval. *Trends in Cognitive Sciences*, 6(7), 299–305.
- Lila, D. (2006). Item, context and relational episodic encoding in humans. *Current Opinion in Neurobiology*, 16(6), 693–700.
- MacDonald, A. W., Cohen, J. D., Stenger, V. A., & Carter, C. S. (2000). Dissociating the role of the dorsolateral prefrontal and anterior cingulate cortex in cognitive control. *Science (New York, N.Y.)*, 288(5472), 1835–1838.
- Macmillan, N. A., & Creelman, C. D. (2005). Detection theory: A user's guide (2nd ed.). Mahwah, New Jersey: Lawrence Erlbaum Associates.
- McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (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. *Psychological Review*, 102(3), 419–457.
- McDuff, S. G. R., Frankel, H. C., & Norman, K. A. (2009). Multivoxel Pattern Analysis Reveals Increased Memory Targeting and Reduced Use of Retrieved Details during Single-Agenda Source Monitoring. *The Journal of Neuroscience*, 29(2), 508–516.
- McGovern, J. B. (1964). Extinction of associations in four transfer paradigms. *Psychological Monographs: General and Applied*, 78(16), 1–21.
- Miyashita, Y. (1993). Inferior Temporal Cortex: Where Visual Perception Meets Memory. *Annual Review of Neuroscience*, 16(1), 245–263.
- Nestor, A., Plaut, D. C., & Behrmann, M. (2011). Unraveling the distributed neural code of facial identity through spatiotemporal pattern analysis. *Proceedings of the National Academy of Sciences*, 108(24), 9998–10003.
- Newman, E. L., & Norman, K. A. (2010). Moderate Excitation Leads to Weakening of Perceptual Representations. *Cerebral Cortex*, 20(11), 2760–2770.
- Nissen, M. J., & Bullemer, P. (1987). Attentional requirements of learning: Evidence from performance measures. *Cognitive Psychology*, 19(1), 1–32.
- Niv, Y., & Schoenbaum, G. (2008). Dialogues on prediction errors. *Trends in Cognitive Sciences*, 12(7), 265–272.
- Norman, K. A. (2010). How hippocampus and cortex contribute to recognition memory: revisiting the complementary learning systems model. *Hippocampus*, 20(11), 1217–1227.
- Norman, K. A., Newman, E., Detre, G., & Polyn, S. (2006). How Inhibitory Oscillations Can Train Neural Networks and Punish Competitors. *Neural Computation*, 18(7), 1577–1610.
- Norman, K. A., Newman, E. L., & Detre, G. (2007). A neural network model of retrieval-induced forgetting. *Psychological Review*, 114(4), 887–953.
- Norman, K. A., Newman, E. L., & Perotte, A. J. (2005). Methods for reducing interference in the Complementary Learning Systems model: Oscillating inhibition and autonomous memory rehearsal. *Neural Networks*, 18(9), 1212–1228.
- Norman, K. A., & O'Reilly, R. C. (2003). Modeling hippocampal and neocortical contributions to recognition memory: A complementary-learning-systems approach. *Psychological Review*, 110(4), 611–646.
- Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: Multivoxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, 10(9), 424–430.
- O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., & Dolan, R. J. (2004). Dissociable Roles of Ventral and Dorsal Striatum in Instrumental Conditioning. *Science*, 304(5669), 452–454.

- O'Keefe, J., & Nadel, L. (1978). *The hippocampus as a cognitive map*. Oxford, UK: Oxford University Press.
- Olson, I. R., & Chun, M. M. (2001). Temporal contextual cuing of visual attention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27(5), 1299–1313.
- Pagnoni, G., Zink, C. F., Montague, P. R., & Berns, G. S. (2002). Activity in human ventral striatum locked to errors of reward prediction. *Nature Neuroscience*, 5(2), 97–98.
- Poppenk, J., & Norman, K. A. (2014). Briefly Cuing Memories Leads to Suppression of Their Neural Representations. *The Journal of Neuroscience*, 34(23), 8010–8020.
- Postman, L., & Underwood, B. J. (1973). Critical issues in interference theory. *Memory & Cognition*, 1(1), 19–40.
- Richter, F. R., Chanale, A. J. H., & Kuhl, B. A. (2016). Predicting the integration of overlapping memories by decoding mnemonic processing states during learning. *NeuroImage*, 124, Part A, 323–335.
- Rescorla RA, Wagner AR (1972) A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical Conditioning II: Current Research and Theory*, eds Black AH, Prokasy WF (Appleton Century Crofts, NewYork), pp 64–99.
- Román, P., Soriano, M. F., Gómez-Ariza, C. J., & Bajo, M. T. (2009). Retrieval-Induced Forgetting and Executive Control. *Psychological Science*, 20(9), 1053–1058.
- Schapiro, A. C., Kustner, L. V., & Turk-Browne, N. B. (2012a). Shaping of object representations in the human medial temporal lobe based on temporal regularities. *Current Biology: CB*, 22(17), 1622–1627.
- Schacter, D. L., Addis, D. R., & Buckner, R. L. (2008). Episodic Simulation of Future Events. *Annals of the New York Academy of Sciences*, 1124(1), 39–60.
- Schapiro, A. C., Kustner, L. V., & Turk-Browne, N. B. (2012b). Shaping of Object Representations in the Human Medial Temporal Lobe Based on Temporal Regularities. *Current Biology*, 22(17), 1622–1627.
- Schapiro, A. C., Turk-Browne, N. B., Botvinick, M. M., & Norman, K. A. (2016). Complementary learning systems within the hippocampus: A neural network modeling approach to reconciling episodic memory with statistical learning. *Philosophical Transactions of the Royal Society*, in press.
- Schlichting, M. L., Mumford, J. A., & Preston, A. R. (2015). Learning-related representational changes reveal dissociable integration and separation signatures in the hippocampus and prefrontal cortex. *Nature Communications*, 6, 8151.
- Schultz W, Dickinson A (2000) Neuronal coding of prediction errors. *Annual Review of Neuroscience* 23: 473–500.
- Sederberg, P., Gershman, S., Polyn, S., & Norman, K. (2011). Human memory reconsolidation can be explained using the temporal context model. *Psychonomic Bulletin & Review*, 18(3), 455–468.
- Shivde, G., & Anderson, M. C. (2001). The role of inhibition in meaning selection: Insights from retrieval-induced forgetting. In *On the consequences of meaning selection: Perspectives on resolving lexical ambiguity* (pp. 175–190).
- Smith, T. A., Hasinski, A. E., & Sederberg, P. B. (2013). The context repetition effect: Predicted events are remembered better, even when they don't happen. *Journal of Experimental Psychology: General*, 142(4), 1298–1308.

- Tomlinson, T. D., Huber, D. E., Rieth, C. A., & Davelaar, E. J. (2009). An interference account of cue-independent forgetting in the no-think paradigm. *Proceedings of the National Academy of Sciences of the United States of America*, 106(37), 15588–15593.
- Turk-Browne, N. B., Yi, D.-J., & Chun, M. M. (2006). Linking Implicit and Explicit Memory: Common Encoding Factors and Shared Representations. *Neuron*, 49(6), 917–927.
- Verosky, S. C., Todorov, A., & Turk-Browne, N. B. (2013). Representations of individuals in ventral temporal cortex defined by faces and biographies. *Neuropsychologia*, 51(11), 2100–2108.
- Watkins, O. C., & Watkins, M. J. (1975). Buildup of proactive inhibition as a cue-overload effect. *Journal of Experimental Psychology: Human Learning and Memory*, 1(4), 442–452.
- Wagner, A. D., Schacter, D. L., Rotte, M., Koutstaal, W., Maril, A., Dale, A. M., ... Buckner, R. L. (1998). Building Memories: Remembering and Forgetting of Verbal Experiences as Predicted by Brain Activity. *Science*, 281(5380), 1188–1191.
- Ward, E. J., Chun, M. M., & Kuhl, B. A. (2013). Repetition Suppression and Multi-Voxel Pattern Similarity Differentially Track Implicit and Explicit Visual Memory. *The Journal of Neuroscience*, 33(37), 14749–14757.
- Wimber, M., Alink, A., Charest, I., Kriegeskorte, N., & Anderson, M. C. (2015). Retrieval induces adaptive forgetting of competing memories via cortical pattern suppression. *Nature Neuroscience*, 18(4), 582–589.
- Xue, G., Dong, Q., Chen, C., Lu, Z., Mumford, J. A., & Poldrack, R. A. (2010). Greater Neural Pattern Similarity Across Repetitions Is Associated with Better Memory. *Science*, 330(6000), 97–101.
- Yassa, M. A., & Stark, C. E. L. (2011). Pattern separation in the hippocampus. *Trends in Neurosciences*, 34(10), 515–525.

Supplementary Information

Overview of Pattern Similarity Analyses

In the pattern classification analyses reported in the main text, we quantified the prediction of C by measuring the amount of classifier evidence for the category of the C item. To provide converging evidence for our claim that context-based prediction error induced forgetting, we repeated these analyses using pattern similarity as a complementary measure of item prediction. Here, we quantified the prediction of C by taking the pattern of activity elicited by the perception of C in the initial triplet and measuring how strongly this pattern was reinstated in the repeated triplet. Pattern similarity was quantified as the spatial correlation in activity over voxels in category-selective ventral temporal cortex during the perception of C and during the repeated triplet (when participants might be predicting C). Using logistic regression, we then related the correlation for a given C item to whether it was remembered or forgotten in the subsequent memory test.

One potential advantage of this approach is that our measure of prediction of C was based on the pattern match to the perception of the same item. This stands in contrast to the main pattern classification analyses, where the classifier was trained on different items from the same category presented in a separate localizer run. Because the pattern similarity approach computes the match to patterns for specific C items, it can — in principle — be used to assess how prediction of the C item *itself* affects subsequent memory. In practice, accomplishing this goal is complicated by the fact that pattern similarity reflects both item-specific information (Schapiro, Kustner, & Turk-Browne, 2012; Xue et al., 2010) and also category information. For example, if a pattern is defined over both face- and scene-selective visual cortex, then different face exemplars will have higher pattern similarity with each other than with any scene.

To verify that the relationship between pattern similarity and subsequent memory reflected item-specific information, we performed a *within-category* permutation test where we repeatedly shuffled the pairing of initial and repeated triplets for all C items from the same category. This allowed us to test whether the relationship between pattern similarity and subsequent memory is stronger when the item matches than when only the category matches.

Relating Pattern Similarity To Subsequent Memory

To perform the logistic regression for pattern similarity, we extracted the patterns of activity evoked by the perception of C in the initial triplet (4.5 s after its onset) and by the prediction of C in the repeated triplet (3–16.5 s after triplet onset). These patterns of activity for C items were obtained from *within* the corresponding category-sensitive ROI. That is, we extracted the patterns from the temporal fusiform cortex when C was a face and from the parahippocampal gyrus when C was a scene. This selection of the category-selective ROIs reduced the amount of variance over voxels in the pattern attributable to category information (thereby isolating item-specific information). These anatomical regions have been used previously as face- and scene-selective ROIs, respectively (Kuhl, Rissman, & Wagner, 2012); they include the peaks of the fusiform face area and parahippocampal place area, respectively, as well as surrounding category-selective voxels that might contribute to representing individual exemplars (Bonnici et al., 2012; Nestor, Plaut, & Behrmann, 2011; Verosky, Todorov, & Turk-Browne, 2013).

We then correlated the patterns from perception and prediction for a given C item to obtain a measure of pattern similarity. We interpreted greater correlation as more prediction of C, given that C was never presented in the repeated triplet and that patterns were obtained from the ROI selective for the category of C (note that A, B, and D were from the other category). As before, we used logistic regression to relate this new measure of prediction strength to

subsequent memory (Supplementary Fig. 1.3A). As hypothesized, greater pattern similarity for C was associated with more forgetting ($p=.034$; Supplementary. 1.3B). Separately correlating the initial C pattern with the same Baseline, A, B, and D time windows used for the classification analyses revealed that this trend grew more negative over time in the repeated triplet (Supplementary Fig. 1.3C). These pattern similarity results provide a clear replication of the pattern classification results.

Ruling Out Carryover Effects

One potential concern with the pattern similarity analysis is that the same A and B context items were shown in both the initial and repeated triplets. If the pattern obtained during the perception of C in the initial triplet was contaminated by lingering traces of A and B, then the observed pattern similarity in the repeated triplet could reflect the repeated perception of A and B rather than the prediction of C. The use of an ROI selective for C but not A and B mitigates this concern. To further rule out this possibility, we performed additional control analyses that seeded the pattern similarity analysis with patterns obtained from the time of A and B in the initial triplet. If the repeated perception of these items drove the negative relationship, then A and B seeds should yield the same results. However, no reliable negative relationship with memory for C was observed when pattern similarity was based on A or B seed patterns ($ps>.27$).

Item-Specific Permutation Analysis

In addition to replicating the main results, the pattern similarity analysis provides additional explanatory leverage. By definition, the pattern classification analysis described in the main text only identified information about the category of C items. In contrast, the pattern similarity analysis might also be sensitive to information about specific exemplars within each category. To examine this, we conducted a permutation analysis by shuffling the pairing of initial and

repeated triplets such that the category of C was preserved. If the observed negative relationship between pattern similarity and subsequent memory were stronger when the exemplar matched than when only the category matched, this would decisively support our interpretation that forgetting reflects misprediction of the C item itself.

To perform the permutation analysis, we first identified pairs of initial and repeated triplets that shared the same context items, and separated them based on whether the C item was a face or a scene. We then shuffled these pairings 1,000 times within each category, each time calculating “pattern similarity” across the scrambled pairs and relating it to subsequent memory for the C item in the initial triplet with logistic regression. Based on the resulting null distribution of 1,000 beta coefficients, a z-score for the true beta coefficient (when the context items were aligned) was calculated. To assess the random-effects reliability of this z-score across participants, we used the same kind of bootstrap test as was performed for the basic logistic regression analysis of classifier evidence and pattern similarity. Specifically, we resampled entire participants with replacement and performed the same permutation test on the resampled data. The distribution of z-scores across bootstrap samples provides us with a population-level confidence interval on these z-scores.

We found a significantly stronger negative trend at the time of D for the true pairing of initial and repeated triplets, relative to the null distribution acquired from shuffled data (bootstrap $p=.036$). Because triplet pairings were shuffled within category, this provides evidence that activation of the specific C exemplar (above and beyond activation of C’s category) contributed to forgetting.

Curve-Fitting Analysis

We used the P-CIT Bayesian curve-fitting algorithm (Detre, Natarajan, Gershman, & Norman, 2013) to estimate the shape of the “plasticity curve” relating item activation during the incidental encoding phase (indexed by category classifier evidence) to recognition during the subsequent memory test. The P-CIT algorithm approximates the posterior distribution over plasticity curves (i.e., which curves are most probable, given the neural and behavioral data). P-CIT generates this approximation via the following three steps: First, the algorithm defines a parameterized family of curves (piecewise-linear curves with three segments) and randomly samples 100,000 curves from this parameterized space. Importantly, this family of curves includes some curves that fit with the non-monotonic plasticity hypothesis and other curves that do not fit. Second, for each randomly generated curve, the algorithm assigns an *importance weight* to the curve that explains how well the curve explains the observed relationship between neural and behavioral data. Finally, these importance weights are used to compute the probability of each curve, given the neural and behavioral data.

Perception and prediction were treated as distinct learning events that could both affect subsequent memory. For each one of the randomly sampled curves, we used that curve, coupled with perception and prediction strength values (measured using the classifier), to generate predictions about which C items would be remembered or forgotten. Specifically, for each item, we separately computed the expected effect of perception (by taking the measured perception strength and evaluating the sampled plasticity curve at that value) and the expected effect of prediction (by taking the measured prediction strength and evaluating the sampled plasticity curve at that value). To estimate the probability that the item would be remembered or forgotten, we summed the expected effects of perception and prediction and fed this sum into a logistic

function (the parameters of which were estimated by the model), which gave us an estimated probability of successful recognition for that item. For each sampled curve, we compared these estimated probabilities of successful recognition (for each item) to the actual recognition outcomes, and assigned an importance weight to the curve reflecting how well the estimated recognition outcomes fit with the actual outcomes. This importance weight value summarizes how well that particular curve explains the observed relationship between neural data (i.e., classifier measurements of perception and prediction) and behavioral data.

After assigning importance weights to each of the 100,000 sampled curves, we generated a new set of samples by taking the best curves from the previous generation (i.e., the curves with the highest importance weights) and distorting them slightly. From this point forward, we alternated between assigning importance weights to sampled curves and generating new sampled curves based on these importance weights, and repeated this process for 100,000 iterations (Detre, Natarajan, Gershman, & Norman, 2013).

The collection of weighted curves generated by this process can be interpreted as an approximate posterior probability distribution over curves — such that the weight of a curve is proportional to its probability. To generate a mean predicted curve, we averaged together the sampled curves in the final generation of samples, weighted by their importance values (Supplementary Fig. 1.4). We also computed credible intervals to indicate the spread of the posterior probability distribution around the mean curve. We did this by evaluating the final set of sampled curves at regular intervals along the x-axis (i.e., item activation). For each x coordinate, we computed the 90% credible interval by finding the range of y values that contained the middle 90% of the curve probability mass.

We also computed $P(\text{theory consistent})$ — the overall posterior probability that the true plasticity curve fits with our theory (i.e., that it is U-shaped). We used the following procedure to compute this probability: First, we labeled each sampled curve as theory-consistent or -inconsistent. Curves were labeled as theory-consistent if they showed a “dip” — i.e., the curve dropped below its starting point and then rose above that starting point moving from left to right. We then calculated the proportion of posterior probability mass taken up by theory-consistent samples. To compute this value, we summed together the importance weights associated with theory-consistent samples. This number provides an efficient summary of how well the data support the non-monotonic plasticity hypothesis. The Matlab code used to perform the analyses can be downloaded from <http://code.google.com/p/p-cit-toolbox>.

Relating Repetition Suppression to Classifier Evidence

We quantified repetition suppression by first extracting the activity evoked by the presentation of A and B in the initial and repeated triplets (4.5-9 s after triplet onset) for each voxel in category-selective anatomical ROIs (i.e., in temporal fusiform cortex when A and B were faces and in parahippocampal gyrus when they were scenes), and then we averaged over voxels within each ROI and performed the subtraction of initial minus repeated triplets. Classifier evidence for C and D was obtained in the same manner as in the main analysis, but only during the time window when A and B were processed. As before, we used logistic regression to relate the amount of repetition suppression for A and B to the classifier evidence for each of the C and D categories.

Ruling Out Effects of Novelty

The behavioral data nicely fit our memory-pruning hypothesis. However, one concern is that the pattern of results (lower memory for C than D items) might be confounded by the novelty of the preceding items in the trial sequence (i.e., A and B were novel before C and repeated before D).

For example, the difference between C and D may reflect enhanced encoding of D because it stood out as novel against a context of old items and/or reduced encoding of C because the preceding new items captured attention. There are theoretical and empirical reasons to think that contextual novelty cannot explain our results.

First, it was recently shown (Duncan, Sadanand, & Davachi, 2012) that novel items facilitate the formation of new memory representations for a subsequent item (pattern separation), while preceding familiar items engage retrieval of existing memory representations and thus reduce encoding (pattern completion). Since C was preceded by novel items and D by familiar items, this study would predict better memory for C than D — the exact opposite of what we observed.

Second, a behavioral pilot study we ran controlled for novelty but observed the same forgetting effect for mispredicted items. The design was quite similar to the reported fMRI study, but we used pairs of scene images (e.g., $A \rightarrow B$ and $C \rightarrow D$) instead of triplets, and these pairs repeated several times. The prediction of the second item (B) based on the first item (A) was violated by swapping the first and the second items across pairs on the fourth repetition ($A \rightarrow B$, $A \rightarrow B$, $A \rightarrow B$, $A \rightarrow D$ and $C \rightarrow D$, $C \rightarrow D$, $C \rightarrow D$, $C \rightarrow B$). Other pairs (e.g., $E \rightarrow F$) were repeated four times intact as a control condition ($E \rightarrow F$, $E \rightarrow F$, $E \rightarrow F$, $E \rightarrow F$). We measured subsequent memory for the second items in the violation condition (B and D) and the control condition (F). Critically, context items in both conditions (A, C, and E) had equal frequency on the fourth repetition when predictions could be violated. Nevertheless, memory in the violation condition was significantly lower than in the control condition ($p=.017$).

Third, although C and D were by definition preceded by novel and familiar items, the X items had variable contexts. We therefore examined memory for X items as a function of the

number of preceding novel items: 1) preceded by a repeated ABD triplet (1 novel, 2 repeated), 2) preceded by a repeated ABD triplet and another X item (2 novel, 2 repeated), 3) preceded by an initial ABC triplet (3 novel), and 4) preceded by an initial ABC triplet and another X item (4 novel). If contextual novelty impaired encoding (e.g., for C vs. D), then memory for X items should decrease as a function of the number of preceding novel items. However, there was no effect of these conditions ($ps > .17$). Indeed, X memory was numerically highest in the fourth condition, which had the most preceding novelty.

Finally, to further test whether the familiarity of items preceding D boosted its encoding, we measured repetition priming for the A and B items in the initial minus repeated triplets and then related these priming scores to D memory using logistic regression. Not surprisingly, we obtained overall repetition priming in response times ($ps < .001$): participants judged the subcategory of A and B items faster when they were repeated vs. novel (45.46 and 47.13 ms faster, respectively). The familiarity account predicts a positive relationship between repetition priming and D memory, which was not obtained (also no relationship to C memory; $ps > .16$).

All of these findings suggest that our behavioral data result from a detrimental effect of prediction violation, rather than an effect of the novelty/familiarity of preceding context items.

Ruling Out Effects of Serial Position

During incidental exposure, C items appeared earlier in the trial sequence than D items (by definition) and X items (which were distributed uniformly), resulting in a systematic serial position difference across conditions ($ps < .001$). This is a potential confound that may explain why memory was worse for C items than for D and X items.

We were sensitive to this issue when designing the experiment, and attempted to minimize it in two ways. First, the incidental encoding phase was divided into three runs (each

lasting around 9 min), and triplets repeated within run such that the conditions were spread across all runs. Second, there was a 10-min rest period between the encoding and test phases to attenuate recency effects.

Nevertheless, we conducted control analyses to rule out a contribution of serial position to our results empirically. One analysis examined (across trials) whether memory for C or D items could be predicted from their serial position. The logic was that if serial position was solely responsible for the observed behavioral differences, then serial position should be related to subsequent memory within these conditions. However, there was no relationship for either C or D items ($p > .15$).

We also conducted a sub-sampling analysis in which we reversed the serial position bias by selecting C and D items such that the average serial position of D items was earlier than C items. Specifically, we deleted pairs of the earliest remaining C item and the last remaining D item within subject until the reversal occurred. As a manipulation check, this procedure did result in an earlier serial position for D than C ($p < .001$). Nevertheless, the behavioral results from the remaining trials were identical to the original pattern, with worse memory for C than D ($p = .040$).

These findings rule out a serial position explanation of the behavioral data, and remain consistent with our pruning hypothesis.

Arbitrating Between Memory Weakening and Interference During Retrieval

Our preferred explanation of decreased recognition of C items relative to D and X items is that memory for C items was weakened. However, other potential explanations exist. For example, it is possible that when C was predicted (but did not appear) during the repeated triplet, participants encoded that “C is absent”; later, during the recognition test, this “C is absent” memory trace

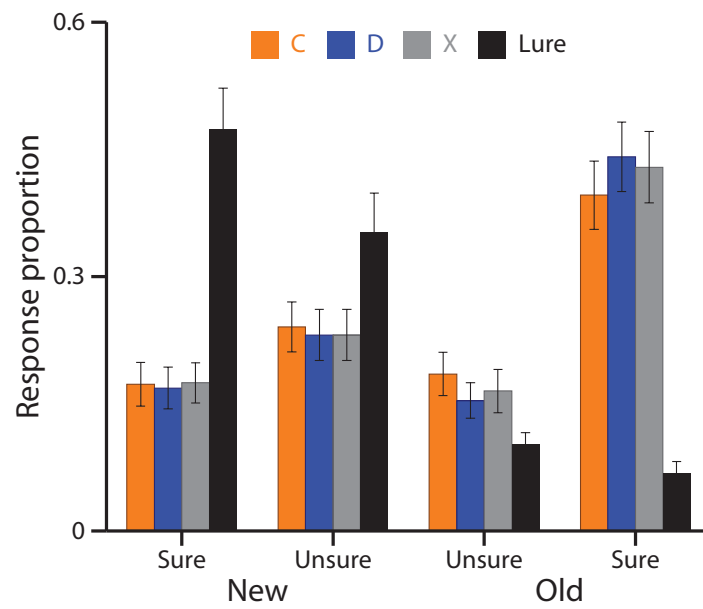
might have been activated, competing with the original C memory and reducing recognition confidence.

When we consider the largely implicit nature of our study, however, we think that this alternative “memory for absence” account cannot fully explain our results. Stimuli in the incidental encoding phase were presented in a continuous stream, and thus it was impossible to know *a priori* whether an item was the A, B, or C item in a triplet. Meanwhile, participants performed an orthogonal categorization task on these stimuli, and they thought that measuring performance on this task was the purpose of the study. Additionally, the context items (A and B) were repeated just once. Thus, it was extremely hard for a participant to detect any structure in the stimulus sequence. After completing the study, we anecdotally asked the participants whether they noticed any regularity during the study phase. Though not recorded systematically, no participant reported explicit awareness of the triplet structure. In other words, participants were likely not aware of the absence of C, and, consequently, it is unlikely that they formed a declarative trace of the thought “C is absent”.

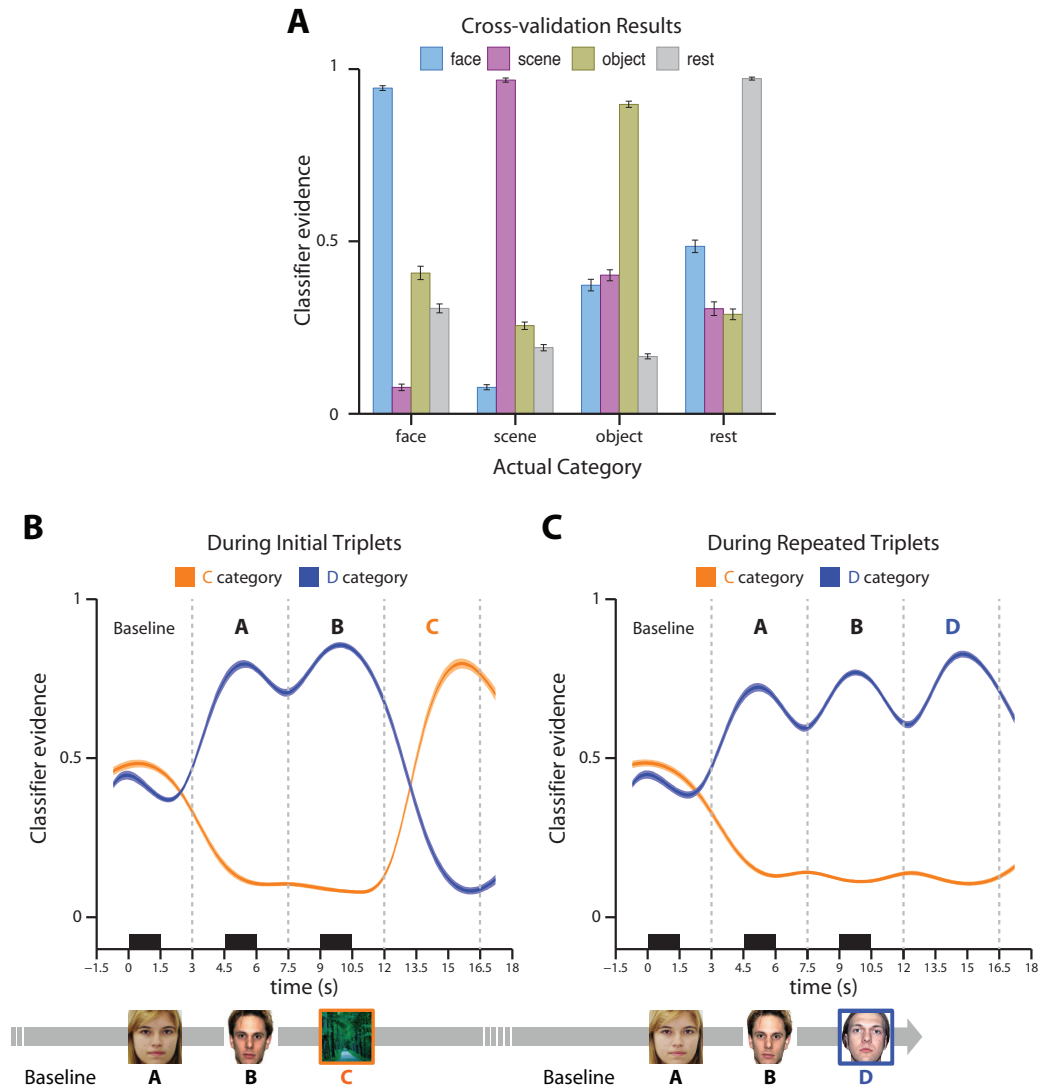
Furthermore, even if participants did form such a trace, knowing that C was predicted-but-absent necessarily implies that C was presented earlier in the experiment, and hence it should increase rather than decrease confidence that C was studied. Imagine that you met a person at a bus stop yesterday, and you notice that he/she is not there today. Later, you meet him/her again accidentally. Intuitively, the additional declarative trace of the absence would help you retrieve that you have met him/her before, rather than impairing the memory.

Another related possibility is that when C was predicted during the repeated triplet, the C representation was bound to the D representation. Later, when C was presented at test, it activated the D representation, thereby causing interference and reducing recognition confidence.

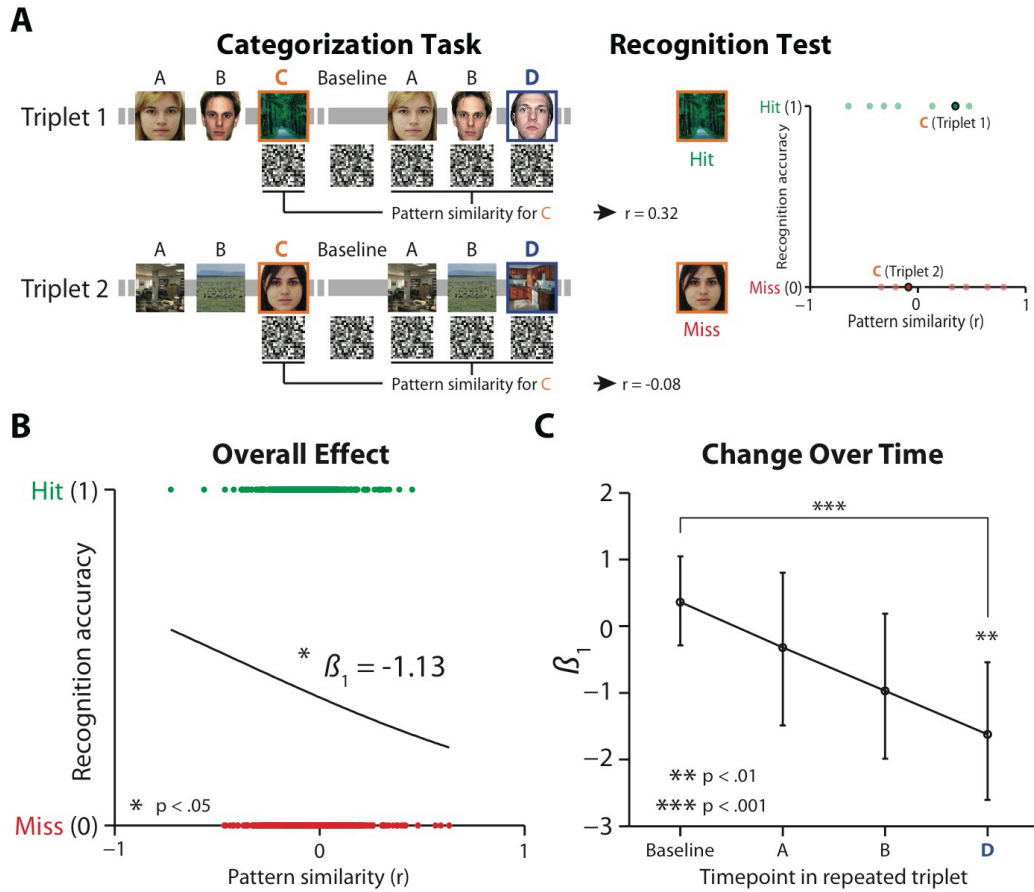
There are two reasons why this account seems unlikely. First, recognition memory tests are thought to provide direct access to stored memory traces and thus to be relatively impervious to these kinds of retrieval interference effects (Tomlinson, Huber, Rieth, & Davelaar, 2009), compared to tests of cued recall. Second, this account also predicts that C should interfere with D (i.e., when D was presented at test, the C representation should have come to mind, causing interference and reducing recognition confidence for D), but this was not the case: Memory for D items did not differ statistically from memory for control X items ($p=.63$), and was in fact numerically higher.



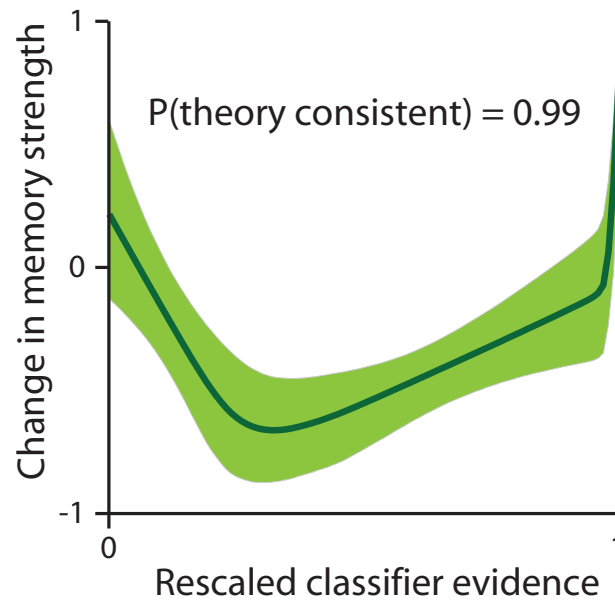
Supplementary Figure 1. Behavioral recognition memory. Response proportions for old items from incidental encoding (C, D, X) and new items (Lure). The four response options are shown on the x-axis. Error bars reflect ± 1 SEM.



Supplementary Figure 2. Details of pattern classification analyses. (A) Cross-validation results by category from the localizer. Error bars reflect ± 1 SEM. (B) Trajectories over time of classifier evidence for stimulus categories in initial triplets. On the x-axis, time=0 indicates the actual time of stimulus onset (i.e., not shifted to account for hemodynamic lag). Classifier evidence peaked around 4.5 s after stimulus onset. Discrete data points were interpolated for visualization. Ribbons reflect ± 1 SEM. (C) Same trajectories in the repeated triplets.



Supplementary Figure 3. Pattern similarity analyses. (A) Pattern similarity was computed as the Pearson correlation between the patterns of voxel activity from the initial triplet when C was perceived and from the repeated triplet when C could have been predicted. The resulting coefficient for each triplet was then related to subsequent memory for C. (B) The pattern similarity for C was first averaged over all time points in the repeated triplet. Dots indicate the distribution of similarity for remembered (green) and forgotten (red) items. There was a reliably negative logistic trend, with greater pattern similarity associated with more forgetting. (C) The same analysis was performed separately during the Baseline, A, B, and D time periods. The negative relationship was maximal during the anticipated time of C. Error bars reflect 95% bootstrap CIs.



Supplementary Figure 4. Curve-fitting analysis. Empirically derived estimate of the “plasticity curve” relating classifier evidence to subsequent memory performance, obtained using the P-CIT curve-fitting algorithm (Detre, Natarajan, Gershman, & Norman, 2013). Behavioral outcomes on the recognition memory test were modeled as the summed effects of perception strength (during the initial triplet) and prediction strength (during the repeated triplet). The x-axis shows rescaled classifier evidence (0=minimum observed classifier evidence, 1=maximum observed classifier evidence) and the y-axis represents the change in subsequent memory strength. The solid green line depicts the mean of the posterior distribution over curves, and the ribbon shows the 90% credible interval (such that 90% of the curve probability mass lies within the ribbon). P-CIT also returns the overall posterior probability that the curve has a U-shape (as predicted by the non-monotonic plasticity hypothesis); in this case $P(\text{theory consistent})=.99$.