Downloaded from learnmem.cshlp.org on August 26, 2022 - Published by Cold Spring Harbor Laboratory Press

Research

Mnemonic prediction errors promote detailed memories

Oded Bein,[1] Natalie A. Plotkin,[2] and Lila Davachi[2,3]

[1]*Princeton Neuroscience Institute, Princeton University, Princeton, New Jersey 08540, USA;* [2]*Department of Psychology, Columbia University, New York, New York 10027, USA;* [3]*Center for Clinical Research, The Nathan S. Kline Institute for Psychiatric Research, Orangeburg, New York 10962, USA*

When our experience violates our predictions, it is adaptive to update our knowledge to promote a more accurate representation of the world and facilitate future predictions. Theoretical models propose that these mnemonic prediction errors should be encoded into a distinct memory trace to prevent interference with previous, conflicting memories. We investigated this proposal by repeatedly exposing participants to pairs of sequentially presented objects (A → B), thus evoking expectations. Then, we violated participants' expectations by replacing the second object in the pairs with a novel object (A → C). The following item memory test required participants to discriminate between identical old items and similar lures, thus testing detailed and distinctive item memory representations. In two experiments, mnemonic prediction errors enhanced item memory: Participants correctly identified more old items as old when those items violated expectations during learning, compared with items that did not violate expectations. This memory enhancement for C items was only observed when participants later showed intact memory for the related A → B pairs, suggesting that strong predictions are required to facilitate memory for violations. Following up on this, a third experiment reduced prediction strength prior to violation and subsequently eliminated the memory advantage of violations. Interestingly, mnemonic prediction errors did not increase gist-based mistakes of identifying old items as similar lures or identifying similar lures as old. Enhanced item memory in the absence of gist-based mistakes suggests that violations enhanced memory for items' details, which could be mediated via distinct memory traces. Together, these results advance our knowledge of how mnemonic prediction errors promote memory formation.

[Supplemental material is available for this article.]

Most of our daily experiences are highly repetitive and predictable. We typically take the same route to work every day, or we enjoy our favorite soup at the neighborhood restaurant over and over again. Through repetition, we develop predictions and expectations of what will happen within a specific context. Once in a while, however, we encounter surprising events that violate these expectations. For example, we might enter our neighborhood restaurant expecting to have our favorite soup, but we found out on that evening that the restaurant is offering brussels sprouts as an appetizer, instead. We term such surprising events "mnemonic prediction errors"—situations in which we expect one thing based on our memory, but the reality is different.

When we encounter a mnemonic prediction error, it is adaptive to update our memories in order to make better and more accurate predictions in the future. How does memory updating happen? Interestingly, despite the proposed beneficial role that novelty and prediction errors play in learning and memory (Rescorla and Wagner 1972; Schultz et al. 1997; Niv and Schoenbaum 2008; Henson and Gagnepain 2010; Schomaker and Meeter 2015; Friston 2018; Ergo et al. 2020; Reichardt et al. 2020; Frank and Kafkas 2021; Quent et al. 2021), very little is known about how mnemonic prediction errors modulate memory encoding. Theoretical models propose that mnemonic prediction errors should be encoded as distinct memory traces (McClelland et al. 1995; Love et al. 2004; Gershman et al. 2014; Frank et al. 2020). That is, events that violate our expectations should be allocated a unique memory representation distinct from prior memories. This may facilitate memory for the unexpected event, while also mitigating interference with existing memories that may still be relevant. Indeed, the complementary learning systems framework shows computationally that the absence of a separated memory trace for mnemonic prediction errors results in catastrophic interference—incorrectly erasing previous memories (McClelland et al. 1995; Kumaran et al. 2016). In the restaurant example, this would mean that immediate integration of the evening that the restaurant served the brussels sprouts instead of your favorite soup could lead to updating your memory to hold that the restaurant no longer serves the soup. This might be maladaptive, as the restaurant may still serve the soup on other nights. Thus, theoretically, memory enhancement of mnemonic prediction errors via a distinct and separated memory trace enables remembering the event that violated our expectations, while also protecting previous, potentially relevant memories.

However, empirical evidence that mnemonic prediction errors facilitate encoding of distinct memory traces is scarce. Previous studies provide some evidence that novel and potentially unexpected events enhance memory (von Restorff 1933; Tulving and Kroll 1995). Most notable is the "von Restorff effect," or, more broadly, oddball manipulations, whereby a rarely occurring item that is clearly distinct from the rest of the items in a list (e.g., a dog in a list of fruits) is better remembered later on (von

Corresponding author: oded.bein@princeton.edu
Article is online at http://www.learnmem.org/cgi/doi/10.1101/lm.053410.121.

© 2021 Bein et al. This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first 12 months after the full-issue publication date (see http://learnmem.cshlp.org/site/misc/terms.xhtml). After 12 months, it is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at http://creativecommons.org/licenses/by-nc/4.0/.

Downloaded from learnmem.cshlp.org on August 26, 2022 - Published by Cold Spring Harbor Laboratory Press

Restorff 1933; Hunt 1995; Ranganath and Rainer 2003; Schomaker and Meeter 2015). One potential account for this effect is that such isolated items are remembered better because they are unexpected and thus salient (e.g., Green 1956; Axmacher et al. 2010; Murty et al. 2016). However, alternative accounts that do not involve violation of expectations posit that isolation per se can facilitate memory, perhaps because isolated items elicit less interference during retrieval (Waddill and McDaniel 1998; Hunt 2006). These accounts rely on repeated observations that enhanced memory for isolated items is also obtained when the isolated item is the second item in the list, before any expectations can be formed (e.g., Hunt 1995; Schmidt and Schmidt 2017). Hence, it is likely that enhanced memory for oddballs need not result from violation of prior expectations.

Recent studies have directly manipulated violation of expectations and found enhanced memory for such violations (Greve et al. 2017; Brod et al. 2018; Kafkas and Montaldi 2018a; Antony et al. 2020). In one study, Greve et al. (2017) taught participants through repeated exposure that different scene categories predict either positive or negative valence words, and then violated this expectation by altering the valence of the words. Memory of the word-scene association was higher for words that violated prior contingencies (Greve et al. 2017). Similarly, Kafkas and Montaldi (2018a) taught participants that a specific symbol-cue is followed by either a man-made object or natural object, and later switched the contingency in some of the trials to violate participants' predictions. Recollection rates were higher for objects that violated participants' prior expectations, suggesting that mnemonic prediction errors did indeed enhance memory for unexpected items (Kafkas and Montaldi 2018a).

Nonetheless, this prior work cannot speak to whether encountering a mnemonic prediction error results in a detailed memory representation that is potentially distinct from other memories. This is because enhanced memory can result from either having a distinct representation of a specific event, or from integrating across memories (e.g., LaRocque et al. 2013; Schlichting and Preston 2015; Favila et al. 2016; DuBrow and Davachi 2017; Clewett et al. 2019). For example, it has been shown that items that were later remembered were less similar to each other (more distinct) in their hippocampal multivoxel activity patterns, compared with items that were later forgotten (LaRocque et al. 2013), suggesting that distinct item representations promote memory (see also Favila et al. 2016; Jenkins and Ranganath 2016). Other studies, however, have shown that integration across experiences can benefit memory (e.g., Richter et al. 2016) and that similarity between multivoxel representations of items mediated accurate temporal memory (DuBrow and Davachi 2014; see also Schlichting et al. 2015). Thus, it is currently unknown whether mnemonic prediction errors promote detailed and distinct memories.

In the current behavioral study, we addressed this question by violating participants' predictions and then conducting a memory test that gauges memory distinctiveness (Bakker et al. 2008; Stark et al. 2019). To evoke expectations, we used a statistical learning paradigm in which participants were repeatedly presented with a stream of objects. Unknown to the participants, we embedded neighboring pairs of objects within the stream that always appeared in the same order (Fig. 1A). Past work has established that through repetition and learning, participants come to predict the second object in the pair upon seeing the first object (Schapiro et al. 2012; Kim et al. 2014, 2017; Kok et al. 2017). Following the prediction learning phase, we violated expectations in half of the pairs by replacing the second object in the pair with a novel object (Fig. 1B). The other half of the pairs were presented intact; thus, no violations occurred. We scattered novel objects after these intact pairs to serve as a no-violation baseline condition where no sequential predictions were violated.

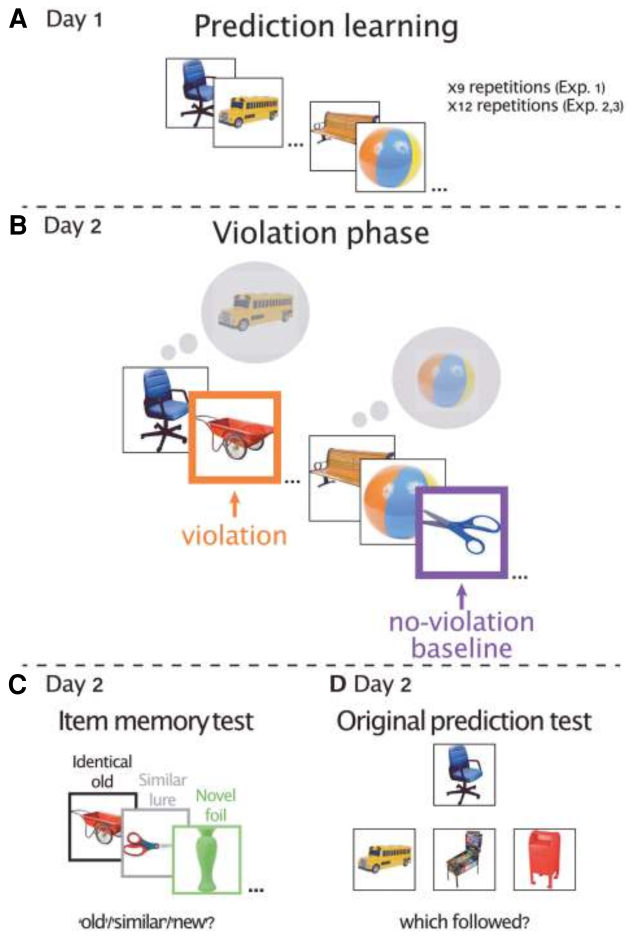

Figure 1. Experimental design, all experiments. (*A*) During prediction learning (day 1), participants repeatedly viewed pairs of sequentially presented objects embedded within a stream of objects. Participants indicated whether each object was bigger or smaller than the previous object (experiments 1 and 2) or than a shoe box (experiment 3). (*B*) In the violation phase (day 2, preceded by a reminder of the predictions, not shown) (see the text for details), novel items were inserted to the sequence of objects, either instead of the second object in the pair, thus violating learned predictions (violation, in orange), or after the second object in a pair (no-violation, purple). The colors appear here for illustration; no color frames appeared on the screen. The task was identical to the prediction learning phase. (*C*) During the item memory test (day 2, immediately following the violation phase), the participants were presented with either identical copies of the violation and no violation items presented during the violation phase (identical old), or with another exemplar of the same item (similar lure) or novel items that did not appear in the experiment before (novel foil). Participants indicated whether an item was "old," "similar," or "new." (*D*) We tested memory for the original predictive pair (day 2), by presenting participants with the first object in a pair and asking which of three bottom objects followed the top object during the study. Distractors were intralist within condition.

Critically, the violation phase was followed by a memory test that targets memory distinctiveness. We presented the participants with either old items from the violation phase, similar lures (a different exemplar of an object presented during the violation phase; e.g., a different pair of scissors than the scissors presented in the violation phase) (Fig. 1C), or novel foils that were only presented during the memory test (Fig. 1C; Bakker et al. 2008; Lacy et al. 2011; Stark et al. 2019). The participants indicated whether an item was identical to an item they had seen before ("old"), "similar," or "new." It is thought that such a fine-grained memory

Downloaded from learnmem.cshlp.org on August 26, 2022 - Published by Cold Spring Harbor Laboratory Press

discrimination requires retrieval of perceptual details of the learned items, to know whether the item presented during the memory test is identical—or rather, only similar—to the item seen during learning. Successful discrimination therefore potentially indicates a distinct memory of the learned items, and can manifest in two ways. The first way is specifically endorsing an old item as "old," without making gist-like mistakes of endorsing an old item as "similar." Endorsing an item as "old" suggests that participants remembered that this particular item appeared during the violation phase. In contrast, endorsing an identical old item with a "similar" response might indicate a more gist-like and less detailed memory representation, because participants remembered that they have seen the item, but did not remember the specific exemplar. Thus, if mnemonic prediction errors enhance memory via a detailed and distinct memory trace, we would expect higher rates of correctly identifying identical old violation items as "old" compared with no-violation items, and no difference between "similar" responses to identical old violation items compared with no-violation items. The second, not mutually exclusive, possibility is that violations will lead to higher rates of correctly identifying similar lures as "similar," potentially indicating that participants were able to distinguish a similar lure from the original old item, indicating memory distinctiveness (Stark et al. 2019; Frank et al. 2020). The memory test could allow us to investigate both possibilities (see the Discussion).

We also considered that memory enhancement for violations might depend on prediction strength (Kim et al. 2014, 2020; Chen et al. 2015; Greve et al. 2017; Kafkas and Montaldi 2018a). Interestingly, previous studies that found a consistent advantage for violations used extensively trained strong predictions (Greve et al. 2017; Kafkas and Montaldi 2018a). Other studies that only presented a sequence of items once or a few times prior to the violation of that sequence did not consistently report a memory advantage for violations (Kim et al. 2014, 2017, 2020; Chen et al. 2015; See also Ortiz-Tudela et al. 2018). Thus, it may be that the strength of the prediction prior to the violation is a critical factor in eliciting a memory advantage for mnemonic prediction errors (Reichardt et al. 2020). To address this possibility, we explicitly tested associative memory for the AB predictive pairs at the end of the experiment (Fig. 1D). This allowed us to compare memory for violations (e.g., C) for which participants remembered the original AB predictive pairs and when they did not remember the AB pairs. We hypothesized that the memory advantage for violations, if observed, should be most pronounced when participants remembered the prediction, indicating that a mnemonic prediction was formed, and then violated.

To sum up, we set out to behaviorally test the idea that mnemonic prediction errors enhance encoding of fine details, promoting high fidelity memory. Briefly, we found a specific advantage for violations in identifying an identical old item as "old," but not as "similar" (Fig. 2). Interestingly, we further found that this specific memory enhancement was modulated by the strength of the memory predictions: The specific memory advantage for violations was only observed when participants remembered the original prediction (experiments 1 and 2), and was diminished when we experimentally reduced prediction strength (experiment 3). We did not find a difference in "similar" responses to similar lures. Together, our results suggest that the memory advantage for mnemonic prediction errors is dependent on prediction strength and might be supported by the creation of a new, distinct memory trace.

Results

In all three experiments, the participants ($N = 28$ in each experiment) were highly accurate during all the phases of learning as well as the violation phase, indicating that our participants performed the task adequately (accuracy was >90%) (see Supplemental Table S1). We found slower reaction times for violation compared with no-violation items during the violation phase, potentially reflecting a slowdown due to a violation of expectations. The data from the learning and violation phases were not the primary interest of the current study and are reported in detail in the Supplemental Material, Supplemental Table S2, and Supplemental Figures S2 and S3.

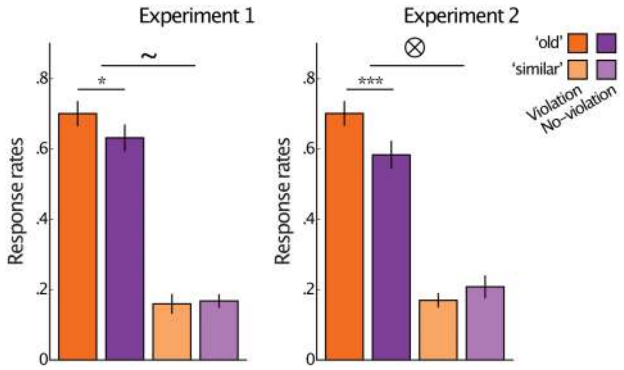

Figure 2. Memory for violation (orange) and no-violation (purple) identical old items in experiments 1 (*left*) and 2 (*right*). "Old" responses are presented in darker colors, "similar" responses are presented in lighter colors. Results are for items for which participants remembered the corresponding original pair. (*) $P < 0.05$, (***) $P < 0.005$, (~) marginally significant interaction ($P < 0.1$), (⊗) significant interaction ($P < 0.05$). Error bars reflect ±SEM.

Experiments 1 and 2

Experiments 1 and 2 were similar in their design and the pattern of results and are therefore reported together. Generally, experiment 2 sought to replicate and further emphasize experiment 1's results in a design that allowed stronger predictions, and thus stronger violations. To that end, in experiment 2 we added more repetition cycles during the prediction learning phase prior to the violation and reduced the ISI from 2.5 to 0.5 sec (see the Materials and Methods). These modifications were used to improve associative binding between original item pairs, leading to stronger predictions and thus to stronger violations.

Turning to the memory results, we first focused our item memory analysis (memory for C items) on items for which participants remembered the original predictive pairs (AB pairs), and thus likely had a prediction for the B item to appear when we violated this prediction. To test memory for the original predictive pairs at the end of the experiment, on each trial, the first object from a neighboring item pair was presented at the top of the screen. Participants then had to choose which one of three objects appearing at the bottom of the screen had followed the top object during learning (distractors were intralist; i.e., second items in other pairs that were violated or remained intact, corresponding to the pair of the target object) (see the Materials and Methods). In both experiments 1 and 2, accuracy rates in the memory test of the original pairs were approximately 0.60, and did not differ between pairs that were later violated and those that remained intact during the violation phase (experiment 1: violated: $M = 0.59$, $SD = 0.16$; intact: $M = 0.58$, $SD = 0.17$; $t_{(27)} = 0.23$, $P = 0.82$; experiment 2: violated: $M = 0.59$, $SD = 0.17$; intact: $M = 0.60$, $SD = 0.17$; $t_{(27)} = 0.51$, $P = 0.62$).

Item memory: specific memory traces for items that violated predictions

Prior to testing our main hypothesis, we established that our participants differentiated between identical old items, similar lures.

Downloaded from learnmem.cshlp.org on August 26, 2022 - Published by Cold Spring Harbor Laboratory Press

Indeed, participants gave more "old" than "similar" responses for identical old items, and more "similar" than "old" responses for similar lures (see the Supplemental Material for full results).

Turning to our main results, we asked whether item memory was enhanced for violations compared with no-violation baseline. Indeed, in both experiments, we found higher rates of responding "old" to identical old violation items compared with no-violation items (experiment 1: violation: $M = 0.70$, SD = 0.19, no-violation: $M = 0.63$, SD = 0.20; $t_{(27)} = 2.28$, $P = 0.03$, Cohen's $d = 0.43$; experiment 2: violation: $M = 0.70$, SD = 0.19, no-violation: $M = 0.58$, SD = 0.21; $t_{(27)} = 3.31$, $P = 0.003$, Cohen's $d = 0.63$) (Fig. 2). Note that these rates are for items for which participants remembered the corresponding predictive pair (for results when the predictive pair is forgotten, see below; Supplemental Fig. S1). Notably, rates of responding "similar" to identical old items did not increase for violations compared with no-violations (experiment 1: violation: $M = 0.16$, SD = 0.15, no-violation $M = 0.17$, SD = 0.10; $t_{(27)} = 0.33$, $P = 0.75$; experiment 2: violation: $M = 0.17$, SD = 0.11, no-violation: $M = 0.21$, SD = 0.17; $t_{(27)} = 0.1$, $P = 0.31$) (Fig. 2). A repeated-measures ANOVA with violation (violation or no-violation) and response ("old" or "similar") revealed an interaction between violation and response (marginally significant in experiment 1: $F_{(1,27)} = 3.05$, $P = 0.09$; $\eta_p^2 = 0.1$, and significant in experiment 2 that involved stronger predictions and violations: $F_{(1,27)} = 5.59$, $P = 0.024$; $\eta_p^2 = 0.17$). These results suggest that violations enhanced item memory but did not enhance gist-based mistakes. Thus, violations of prior expectations enhance accurate item memory, potentially supported by memory for items' details.

Another type of gist-based mistake would be to respond "old" to a similar lure. In these responses as well, we found no difference between violation and no-violation items (experiment 1: violation: $M = 0.22$, SD = 0.19, no-violation $M = 0.22$, SD = 0.16; $t_{(27)} = 0.07$, $P = 0.94$; experiment 2: violation: $M = 0.20$, SD = 0.17, no-violation: $M = 0.20$, SD = 0.18; $t_{(27)} = 0.05$, $P = 0.95$). The interaction of violation (violation or no-violation) by item type (identical old/similar lure) did not reach significance in experiment 1 ($F_{(1,27)} = 1.45$, $P = 0.24$; $\eta_p^2 = 0.05$) and was marginally significant for experiment 2 ($F_{(1,27)} = 4.18$, $P = 0.051$; $\eta_p^2 = 0.13$). This might be due to slightly noisier responses, as can be seen in overall higher SDs in the "old" responses to similar lures compared with "similar" responses to identical old items. Finally, in both experiments, participants numerically identified more similar lures as "similar" for violation compared with no-violation items (experiment 1: violation: $M = 0.54$, SD = 0.23, no-violation $M = 0.51$, SD = 0.22; $t_{(27)} = 0.48$, $P = 0.63$; experiment 2: violation: $M = 0.59$, SD = 0.24, no-violation: $M = 0.58$, SD = 0.22; $t_{(27)} = 0.46$, $P = 0.65$). Together, these results show that violations of prior predictions selectively enhance correct memory for identical old items.

Next, we sought to further support the notion that the prediction strength modulates memory for violation of these predictions. To that end, we directly compared memory of violation items for which participants remembered the original pair to violation items for which participants forgot the original pair (namely, selected the wrong object in the memory test for the original pair; note that one participant in experiment 1 and one participant in experiment 2 were removed from the analyses based on forgotten pairs due to having no forgotten violation pairs). Memory rates for identical old items were entered to a repeated-measures ANOVA with original-pair memory (remembered or forgotten) and response ("old" or "similar"), which revealed a significant interaction (experiment 1: $F_{(1,26)} = 5.58$, $P = 0.026$; $\eta_p^2 = 0.18$; experiment 2: $F_{(1,26)} = 6.12$, $P = 0.020$; $\eta_p^2 = 0.19$). The interaction stemmed from participants responding "old" proportionally more often to identical old violation items when the original pair was remembered compared with forgotten (experiment 1: original-pair remembered, as above: $M = 0.70$ SD = 0.19; original-pair forgotten: $M = 0.59$, SD = 0.20; $t_{(26)} = 2.88$, $P = 0.008$, Cohen's $d = 0.56$; experiment 2: original-pair remembered, as above: $M = 0.70$, SD = 0.19; original-pair forgotten: $M = 0.63$, SD = 0.20; $t_{(26)} = 2.15$, $P = 0.04$, Cohen's $d = 0.41$). In experiment 1, there was no difference in "similar" responses between violation items for which the original pair was remembered or forgotten ($t_{(26)} = 1.11$, $P = 0.28$, Cohen's $d = 0.21$). In experiment 2, participants significantly responded "similar" to identical old items when they forgot the original pair more than when they remembered the original pair, consistent with the notion that memory for original associations reduced generalization mistakes ($t_{(26)} = 2.24$, $P = 0.03$, Cohen's $d = 0.43$). Together, these results demonstrate that memory for violations of mnemonic predictions is modulated by the strength of the original predictions, as measured by participants' memory for these predictions.

Additionally, when looking at items for which participants forgot the original predictive pair, there was no memory enhancement for violations compared with no-violations. Specifically, there was no significant difference between violation and no-violation items in responding "old" to identical old items in experiment 1 (violation: $M = 0.59$ SD = 0.20, no-violation $M = 0.56$, SD = 0.29; $t_{(26)} = 0.46$, $P = 0.64$). In experiment 2, participants made more "old" responses to identical old no-violation items compared with violation items (violation: $M = 0.63$, SD = 0.20, no-violation $M = 0.69$, SD = 0.19; $t_{(26)} = 2.16$, $P = 0.04$; Cohen's $d = 0.42$). This result was not predicted, and we note that in experiment 1 this difference was not significant and numerically in the opposite direction of experiment 2. Thus, we do not further interpret this result. There was no difference between violation and no-violation items in responding "similar" to identical old items (experiment 1: violation: $M = 0.19$, SD = 0.14, no-violation $M = 0.22$, SD = 0.18; $t_{(26)} = 0.87$, $P = 0.39$; experiment 2: violation: $M = 0.23$, SD = 0.19, no-violation $M = 0.18$, SD = 0.15; $t_{(26)} = 1.33$, $P = 0.20$). The response rates for items for which participants forgot the corresponding original pair are presented in Supplemental Figure S1.

While not the main focus of the current study, we also tested associative memory for the novel associations, namely, the association between the violation or no-violation item, and the preceding item (AC) (see the Materials and Methods). As expected due to the incidental nature of the task and the fact that these pairs were only presented once, memory rates were low. However, we note that we did not find any differences in associative memory between violation and no-violation items (see the Supplemental Material for detailed results). Thus, in the current study, the item memory advantage for mnemonic prediction error was not accompanied by associative memory advantage.

Experiment 3

In experiments 1 and 2 we found that items that violated participants' predictions were remembered better, and that this memory advantage could be attributed to a more detailed memory representation. We have also found that the memory advantage for mnemonic prediction errors is dependent on prediction strength, defined as the participants' memory of the original pair. In experiment 3 we directly tested the possibility that prediction strength modulates the memory advantage of prediction errors by experimentally manipulating prediction strength. To do this, during all learning phases and the violation phase, we used an item-focused task ("bigger or smaller than a shoebox") rather than a task that promoted associative learning ("bigger or smaller than previous object" in experiments 1 and 2). This alteration of the task allowed us to reduce later associative memory for the sequentially presented items (AB), while maintaining the general structure of the sequences, the familiarity of the items in the prior pairs, as well as the total duration of the task. We further reasoned that an item-

Downloaded from learnmem.cshlp.org on August 26, 2022 - Published by Cold Spring Harbor Laboratory Press

focused task should not impair overall item memory rates for the violation and no-violation items themselves. We hypothesized that if we reduced the prediction strength of the original pairs, the violation should be weaker as well, and, consequently, we should observe no memory advantage for such violations.

Prior to reporting the item test results, we wished to verify that memory for the original pairs was reduced. Indeed, in comparison with experiments 1 and 2, in which memory rates were ~0.60, memory for the original pairs was significantly reduced in experiment 3 to ~0.40. Memory rates were entered to two mixed ANOVAs, one comparing experiment 1 with experiment 3, and one comparing experiment 2 with experiment 3. Each ANOVA included experiment (experiment 1/experiment 3; experiment 2/experiment 3) as between-participants variable and Violation (violated/intact pairs) as within-participant variable. Both ANOVAs revealed a significant effect of experiment ($F_{(1,54)}$s > 25.91; Ps < 0.0001), with no effects of violation, or interactions of violation by experiment ($F_{(1,54)}$s < 2.14; Ps > 0.15). As in experiments 1 and 2, in experiment 3, no difference was observed between violated and intact pairs (violated: $M = 41$, SD = 0.13; intact: $M = 0.38$, SD = 0.12; $t_{(27)} = 1.49$, $P = 0.15$). Memory for both the violation and the intact pairs was significantly above chance (violated vs. chance: $t_{(27)} = 3.36$, $P = 0.002$; intact vs. chance: $t_{(27)} = 2.37$, $P = 0.02$).

Item memory: no memory advantage for violations

As in the previous experiments, we first established that our participants differentiated between identical old, similar lures, and foils. We also obtained similar levels of memory to those observed in the previous experiments, showing that modifying the task did not impair overall item memory (see the Supplemental Material).

Turning to our main results, as predicted, here we did not find any advantage for violations, regardless of participants' responses during the original-pair memory test (AB-correct: responding "old" to identical old: violation: $M = 0.64$, SD = 0.31, no-violation $M = 0.61$, SD = 0.31; $t_{(27)} = 0.67$, $P = 0.51$; responding "similar" to identical old items: violation: $M = 0.20$, SD = 0.18, no-violation $M = 0.22$, SD = 0.23; $t_{(27)} = 0.48$, $P = 0.63$; AB-incorrect: responding "old" to identical old: violation: $M = 0.59$, SD = 0.29, no-violation $M = 0.60$, SD = 0.26; $t_{(27)} = 0.43$, $P = 0.67$; responding "similar" to identical old items: violation: $M = 24$, SD = 0.21, no-violation $M = 0.21$, SD = 0.12; $t_{(27)} = 1.09$, $P = 0.29$; note that in the current experiment, AB-correct responses likely reflect weak memory or guesses). There was also no significant difference between violation items for which participants correctly or incorrectly identified the original pair in the associative memory test (responding "old" to identical old: items: $t_{(27)} = 1.1$, $P = 0.28$; responding "similar" to identical old items: $t_{(27)} = 0.98$, $P = 0.34$). These results demonstrate that when memory for the predictive pairs was low, and presumably predictions are weak, item memory for violations of these weak predictions was not enhanced. Thus, strong predictions might be important for boosting memory of mnemonic prediction errors.

One concern is that experiment 3 did not have sufficient power to detect a difference between violation and no-violation items. This is due to lower rates of remembered predictive pairs, which result in less items included in analysis of the item memory test. To address this possibility, we performed a post-hoc power-analysis (note that post-hoc power analysis is useful to determine the potential power of a study because the power is the same regardless of when the test is done) (O'Keefe 2007). We used Pangea, a software that computes power while taking into account the number of observations per participant (the number of participants in experiment 3 was identical to experiments 1 and 2) (J Westfall, unpubl. [http://jakewestfall. org/publications/pangea]); G*Power, another commonly used software, does not take the number of observations into account, and thus would not allow us to consider the difference between the experiments. We used the effect size of experiment 2, as this study had identical parameters to experiment 3, aside from the change in the task. Importantly, we performed the analysis with the number of samples we had on average per participant in experiment 3 (seven observations, computed by taking 18 items presented as identical old items in the item memory test per violation/no-violation condition times 0.40 memory of the corresponding predictive pairs). The power to obtain an effect of higher "old" responses for violation compared with no-violation items was 0.94. In addition, we reduced the effect size to examine what is the smallest effect size we could detect with probability of 0.80 (a common threshold in power analysis) and saw we could detect a potential reduced effect size of 0.495. We thus conclude that we had sufficient power to detect an effect in experiment 3.

Discussion

When we encounter a surprising event that violates our expectations, it is adaptive to update our memory in order to facilitate more accurate predictions in the future (Rescorla and Wagner 1972; Niv and Schoenbaum 2008; Henson and Gagnepain 2010; Friston 2018; Sinclair and Barense 2019; Ergo et al. 2020). While theoretical accounts suggest that violations of memory predictions should lead to the creation of a distinct memory trace (McClelland et al. 1995; Love et al. 2004; Gershman et al. 2014; Frank et al. 2020), empirical evidence for this notion is scarce. We addressed this issue by violating learned predictions and then testing memory for these violations. Critically, we used a memory test that asked participants to discriminate between identical old objects and similar lures, allowing us to gauge the specificity of the memory trace (Bakker et al. 2008; Stark et al. 2019). In two experiments, we found that objects that violated predictions were remembered better than objects that did not violate predictions. Importantly, this advantage was specific to correctly identifying identical old items as old. Participants did not mistakenly identify more identical old items as similar lures, which could have suggested gist-like representations of these items. Thus, the specific advantage we observed suggests that the memory enhancement for violations was supported by detailed and high-fidelity memory representations. Moreover, we found that this enhancement was dependent on memory of the prediction. In experiments 1 and 2, we only found memory enhancement for violations when participants had strong predictions, defined as correct memory for the predictive pair in a later memory test. In experiment 3 we reduced prediction strength by lowering associative binding during encoding and found that while item memory remained intact overall, the memory advantage for violations was diminished. Together, our findings suggest that strong mnemonic prediction errors facilitate item memory. Moreover, they may do so through the creation of a distinct memory representation.

Our results are consistent with previous findings showing enhanced memory for mnemonic prediction errors (Greve et al. 2017; Brod et al. 2018; Kafkas and Montaldi 2018a; Wahlheim et al. 2019; Wahlheim and Zacks 2019). However, the current findings advance our knowledge in an important way: We provide evidence that mnemonic prediction errors enhance memory via detailed and distinct memory traces, in line with theoretical suggestions (McClelland et al. 1995; Love et al. 2004; Gershman et al. 2014; Frank et al. 2020). A previous study by Frank et al. (2020) drew similar conclusions to our study. In that study, predictions were violated by teaching participants that symbol cues predict either a man-made or a natural object, and then switching the contingency

Downloaded from learnmem.cshlp.org on August 26, 2022 - Published by Cold Spring Harbor Laboratory Press

in some of the trials. At test, participants were presented with identical old: items, similar lures, and novel foils. In contrast to our study, however, participants were only given two response options: "old" or "new," and similar lures were to be responded with "new." In a single behavioral experiment, the investigators did not observe higher rates of responding "old" to old violation items compared with prediction-consistent items. They did report higher rates of responding "new" to similar lures of items that violated predictions during learning compared with items that were consistent with participants' predictions. The investigators concluded that these "new" responses reflected accurate memory of the original item, and therefore correct rejection of similar lures. This is ambiguous, however, because "new" responses could reflect that these items were merely forgotten—a possibility that is further strengthened by the lack of item memory advantage (i.e., correctly identifying old items as "old") for violations, suggesting that memory for violations was not superior in the study by Frank et al. (2020). Since in our study participants were given the option to respond "similar" in addition to "old" or "new," they could respond "similar" if they indeed recognized an item as a similar lure, and "new" if they thought that the item was new, indicating forgetting. Thus, there is less ambiguity regarding the interpretation of participants' responses during retrieval. Critically, here we report better item memory for violations. This memory advantage was specifically observed in correctly identifying old items as "old," but without higher rates of gist-like mistakes (identifying old items as "similar"). Therefore, we provide strong empirical evidence that mnemonic prediction errors facilitate memory, potentially through the creation of distinct memory traces.

The specificity of the memory advantage we observed might point toward the mechanism underlying encoding of mnemonic prediction errors. While we found higher rates of identifying old items as "old," we did not observe significantly higher rates of "similar" responses to similar lures of violations compared with no-violation items. The ability to respond "similar" to similar lures was previously attributed to pattern separation, a process by which similar experiences are allocated with distinct neural representations, allowing their discrimination in memory (Bakker et al. 2008; Stark et al. 2019). Pattern separation is thought to be mediated by the hippocampus (Norman and O'Reilly 2003; Leutgeb et al. 2007; Bakker et al. 2008; Lacy et al. 2011; Baker et al. 2016; Berron et al. 2016; Knierim and Neunuebel 2016). Recently, a hippocampal network model showed more distinct hippocampal activity patterns for mnemonic prediction errors compared with prediction-consistent events, presumably reflecting pattern separation for violations (Frank et al. 2020). Empirical neuroimaging studies demonstrate that average BOLD signal in the hippocampus increases in response to novelty, and more specifically to mnemonic prediction errors (Kumaran and Maguire 2006, 2007; Axmacher et al. 2010; Chen et al. 2011, 2015; Duncan et al. 2012; Allen et al. 2016; Long et al. 2016). This hippocampal involvement might support the creation of distinct memory traces (Davis et al. 2012). However, the magnitude of BOLD signal cannot indicate the type of hippocampal representations, thus whether this involvement reflects pattern separation remains unknown. Alternatively, distinct memory for violations might be mediated by increased processing of perceptual information input from the entorhinal cortex to the hippocampus (Hasselmo et al. 1996; Hasselmo and Stern 2014; Colgin 2016). Consistent, we have previously shown that functional connectivity between the hippocampus and entorhinal cortex increases during mnemonic prediction errors (Bein et al. 2020a). Thus, perceptual input from the entorhinal cortex potentially facilitated detailed memory representations that later enabled correct identification of an old item as "old"; however, this perceptual input might not suffice to clearly distinguish a similar lure from a previously seen item, which might require pattern separation. This might be why we did not observe higher rates of correctly identifying similar lures as "similar" for violation items compared with no-violation items. Additionally, the perirhinal cortex, an adjacent brain region, supports recognition memory for items (Brown and Aggleton 2001; Davachi 2006; Eichenbaum et al. 2007; Staresina and Davachi 2008; Staresina et al. 2011) and is preferentially engaged during mnemonic prediction errors (Chen et al. 2015). Thus, detailed memories may be supported by multiple mechanisms that may lead to potentially different memory phenomena (Kafkas and Montaldi 2018b; Frank and Kafkas 2021).

Mnemonic prediction errors may enhance item memory or associative memory, depending on the violation event. In a previous study, Greve et al. (2017) found that when items (words or faces) violated previously learned scene-item associations, it resulted in higher associative memory between the items and their background scene images. Another study found that mnemonic prediction errors enhanced recollection judgments for the violation item—perhaps reflecting additional details that were remembered from encoding—but did not enhance familiarity judgements that would reflect only item recognition (Kafkas and Montaldi 2018a). Together, these studies suggest that violations enhance memory not only for the violating item, but also for contextual associations. In the current study, however, we found a memory enhancement for items that violated expectations, but without a concomitant enhancement of associative memory (see the Supplemental Material). Note that overall associative memory rates in our experiments were low, and thus this result should be interpreted with caution. Nevertheless, one important difference between our study and the previous study by Greve et al. (2017) is that in their study, the background-scene was presented together on the screen with the violating item during the violation. In contrast, in the current study, only the violating item was presented on the screen, as objects were presented sequentially (the aforementioned study by Kafkas and Montaldi [2018a] used a remember-know paradigm in which participants only reported whether their memory is supported by recollection or not; thus, it is unclear what additional details were recollected). Possibly, violations promote a distinct representation of the event that is happening in the moment of the violation. If only an item is presented during the violation, memory would be enhanced for that item. If, however, more components are included in the violation event (Greve et al. 2017), memory would be facilitated for all components. Thus, mnemonic prediction errors may promote associative versus item memory results depending on the task (Quent et al. 2021).

Consistent with previous studies (Kafkas and Montaldi 2018a), our study revealed that the task, or more broadly, goals, might influence memory for violations. In experiment 3, instead of a task orienting participants toward the associations between items (as was done in experiments 1 and 2), we guided participants toward individual items. We aimed at reducing associative binding, and indeed, associative memory rates of the predictive pairs were lower in experiment 3 compared with experiments 1 and 2. We do not argue, however, that no associations were formed. Some associations might have been created between items in the predictive pairs, for example, considering temporal context models (Kahana 1996), or statistical learning processes (Schapiro and Turk-Browne 2015). It might be, however, that violations of predictions born through statistical learning might not suffice to produce a memory advantage for violations. This can be related to reduced strength of such predictions, or because goal-oriented associative processing during prediction-learning or during the violation is required to boost memory for violations (DuBrow and Davachi 2013). For example, associative processing during the moment of the violation might draw attention to the change from prior

Downloaded from learnmem.cshlp.org on August 26, 2022 - Published by Cold Spring Harbor Laboratory Press

experience, which has been shown to promote memory for potential violations (Wahlheim and Zacks 2019; Garlitch and Wahlheim 2020). Future research could further explore how attention and goals interact with memory in the processing of prediction errors (Ortiz-Tudela et al. 2018; Kafkas and Montaldi 2018b; Garlitch and Wahlheim 2020).

One limitation of the current study is that we estimated prediction strength by testing memory at the end of the experiment. This was done to avoid contamination of the learning or the item memory test. We thus cannot argue we perfectly evaluated the strength of prediction in the moment of the violation. We used pairs that were remembered later as an approximation of stronger predictions during the violation, compared with pairs that were forgotten. A couple of recent studies, however, show that strong predictions, as measured by classification of fMRI multivoxel activity patterns, correlated with subsequent forgetting of predicted items that were violated, as if these predicted items were pruned from memory (Kim et al. 2014, 2020). It is possible that the subsequently remembered predictive pairs in the current study might reflect weaker predictions that were not pruned during the violation, while forgotten predictive pairs reflect strong predictions that were violated but then pruned. Additionally, a recent study showed that violations of predictions led to a differentiation in the neural representation between the cue and the predicted item, which could also result in impaired associative memory of the predictions (Kim et al. 2017); note, however, that such impairment was not reported, and alternative explanations for differentiation exist (Greve et al. 2018). Since we did not measure predictions during the violation, it is difficult to estimate whether eventual forgetting reflects poor initial learning or good initial learning followed by pruning or differentiation. However, a few data points suggest that pruning did not occur in the current study. First, if violation leads to pruning of prior predictions from memory, associative memory for pairs that were violated should be lower compared with pairs that were not violated. Second, the predictive pairs in experiment 2 (in which predictions were stronger) should have been remembered worse than in experiment 1 (see the Materials and Methods). However, these two predictions did not materialize in our data. Furthermore, reaction times for the predicted items in subsequently remembered predictive pairs became quicker during initial learning, compared with predicted items in subsequently forgotten pairs. This suggests that prior to any violations, the predicted items in remembered pairs were more accessible to our participants, potentially reflecting stronger predictions (Supplemental Material). Thus, it could be that in our study neither pruning nor differentiation occurred, and we indeed estimated strong predictions using the final memory test. It is likely that the associative predictions in our study were stronger than those set up in past studies because participants were exposed to specific item pairings that were learned over multiple repetitions. In contrast, in these prior studies category predictions were learned via only a single or few repetitions (Kim et al. 2014, 2017, 2020; more below on category- vs. item-level predictions). In our view, it makes sense that if we only experience an event once, and then it changes, we might not need to remember that unstable event, and pruning might be adaptive. If, however, we have experienced an event multiple times, and then it only changes once (as in our study), we might not want to forget the more typical occurrence of this event. Indeed, according to the nonmonotonic plasticity hypothesis, items that are strongly reactivated are not altered, and only items that are moderately activated are modified and potentially pruned (Newman and Norman 2010; Detre et al. 2013; Ritvo et al. 2019). Thus, suggesting that our predictions were strong and were not pruned is consistent with this perspective, though other possibilities exist (for example, items that are weakly activated are not modified as well according to this theoretical perspective). Future research, potentially using fMRI, could evaluate predictions in the moment to better elucidate the conditions by which pruning of violated predictions occurs.

Relatedly, the specific relationship between old and new memories is an interesting topic that we did not address in the current study. We were motivated by the notion of distinct memory traces for violations and tested the specific hypothesis that distinct traces should lead to detailed memories. However, whether the memory of the predictive pair was updated or whether the violation created a separate memory trace and the memory of the predictive pair remained unchanged are open questions. We can offer some clues: First, a memory advantage for violations was found when participants remembered the previous predictive pairs, and not when the previous pairs were forgotten (see also Wahlheim et al. 2019; Wahlheim and Zacks 2019). Thus, to the extent that the old memory trace was updated, it was not at the expense of the old memory. Moreover, we did not find a difference in memory for predictive pairs that were violated compared with pairs that were not violated. While inconclusive, this as well is consistent with the notion that the prior memory trace was not impaired, potentially because the new violation was encoded separately, which reduced interference with previous memories (McClelland et al. 1995; Kuhl et al. 2010; Davis et al. 2012; Gershman et al. 2014, 2017). Future research, potentially with additional measures (e.g., memory dependency [Horner and Burgess 2013, 2014]), could better elucidate the specific relationship between memory for new events and prior memories (see also Wahlheim and Zacks 2019; Bein et al. 2020b).

An intriguing question is whether different types of prediction errors facilitate memory for violations via shared or distinct mechanisms. In the current study participants learned relatively specific item predictions; e.g., that a chair predicts a bus (Fig. 1). These predictions were later violated by another object (Kumaran and Maguire 2006, 2007; Chen et al. 2015). Other studies, however, taught participants "category-level" predictions, namely, that a cue or a category of items (e.g., faces), predict another category of items (e.g., objects) (Kim et al. 2014, 2020; Greve et al. 2017; Kafkas and Montaldi 2018a; Frank et al. 2020). In these studies, violations were defined as the presentation of items from a different category than expected. These papers show consistent results, namely, that items that violate category-level predictions are remembered better (Greve et al. 2017; Kafkas and Montaldi 2018a). An interesting result by Kim et al. (2020) suggests that the memory advantage for violations might be attributed (at least in part) to reduced memory of prediction-consistent items. Kim et al. (2020) show that strong category predictions, as measured by classification of fMRI BOLD activity patterns, correlate with poorer memory for items that meet these category predictions. An additional study also shows that predictions impair memory for the items that cue these predictions (Sherman and Turk-Browne 2020). These studies suggest that generating a memory prediction may reduce processing of external details as long as these predictions are met, and thus impair memory of external details like the specific item presented (see also Bein et al. 2020a). How such effects might generalize to item-specific predictions is currently unknown, as category and item-level predictions can differ. For example, a study by Long et al. (2016) directly compared category versus item violations and found that only for category violations, but not for item-violations, hippocampal univariate BOLD response correlated with the strength of decoding the category of the previous paired image (i.e., the prediction) in the default-mode network. Future research could explore the intricate relationship between different types of predictions, prediction errors, and memory (Kafkas and Montaldi 2018b; Frank and Kafkas 2021).

Another well-known type of prediction error is reward prediction error in which participants obtain a different reward than

Downloaded from learnmem.cshlp.org on August 26, 2022 - Published by Cold Spring Harbor Laboratory Press

what is expected (Rescorla and Wagner 1972; Schultz et al. 1997; O'Doherty et al. 2004; Niv and Schoenbaum 2008; Gläscher et al. 2010). Reward prediction errors may modulate memory via similar or different mechanisms than mnemonic prediction errors. While reward prediction errors have long been established as promoting learning and decision making, their role in long-term memory has only been appreciated more recently (e.g., Wimmer et al. 2014; Rouhani et al. 2018, 2020; Jang et al. 2019; Ergo et al. 2020; Rouhani and Niv 2021). These studies have generally shown memory benefits for stimuli appearing during a reward prediction error (Wimmer et al. 2014; Davidow et al. 2016; De Loof et al. 2018; Rouhani et al. 2018; Jang et al. 2019; Kalbe and Schwabe 2019; Ergo et al. 2020; Rouhani and Niv 2021). Interestingly, a recent study (Rouhani et al. 2020) evinced that reward prediction errors reduced linking items together in memory; associative priming as well as temporal memory was reduced between items studied before and after high prediction errors. These findings suggest that reward prediction errors resulted in memories that are distinct from each other, similarly to the current findings. Thus, the creation of distinct memory representations might be a general mechanism for encoding prediction errors.

To conclude, we found that mnemonic prediction errors enhance memory by facilitating detailed, high-fidelity memory, potentially reflecting distinct memory traces. Interestingly, the notion that prediction errors should lead to separation of memory traces has been suggested across domains in cognition and neuroscience. For example, prediction errors were postulated to cause separation in category learning (McClelland et al. 1995; Love et al. 2004; Davis et al. 2012) and reinforcement learning and state representations (Gershman and Niv 2010; Gershman et al. 2010, 2014, 2017). Likewise, prediction errors have also been posited to determine event segmentation processes, namely, identifying boundaries in ongoing experiences (Zacks and Tversky 2001; Zacks et al. 2011; Franklin et al. 2020; but see Clewett and Davachi 2017 for complications in this view). That said, empirical evidence directly supporting these notions is still nascent and many questions remain open. Together with recent literature, the current study opens exciting avenues for future research investigating the cognitive and neural mechanisms underlying mnemonic prediction errors, which will hopefully lead toward a deeper understanding of how prediction errors modulate learning and memory across different domains.

Materials and Methods

Experiment 1

Participants

Twenty-eight participants were included in this study (19 females, aged 18–31 yr, mean age: 22.46). One additional participant was excluded due to a technical error. Eleven additional participants were excluded due to poor compliance with the task, namely, <40% memory for both types of old associations (that were violated or remained intact during the violation phase; see below for more details, we address the potential concern about high exclusion rate in experiment 2). The participants were members of the New York University community, with normal or corrected to normal vision. They provided written informed consent to participate in the study and received a payment at a rate of $10/h for their participation. The study was approved by the New York University Institutional Review Board.

The sample size was determined based on prior literature. We aimed for a final sample size of 20–28 participants, given prior similar studies ($N = 20$ [experiment 1 in Greve et al. 2017]; $N = 28$ [Kafkas et al. 2018a]). Note that statistical learning paradigms tend to have low and variable memory (e.g., Siegelman et al. 2017). Thus, we ran 40 participants knowing that we might need to exclude participants based on poor performance. Indeed, our final sample was within the expected range ($N = 28$).

Materials

The stimuli consistent of 180 images of everyday nameable objects from sets used in previous studies (Polyn et al. 2005; Kuhl et al. 2011; DuBrow and Davachi 2013; Tompary and Davachi 2017). We complemented some of the objects with analogous images from the internet to achieve similar lures that would also comply with our size-judgment task during encoding. The objects were presented in the center of a white square background, sized 350 × 350 pixels. The objects were all matched in size of appearance on the screen by scaling the objects so that the larger dimension of the object (horizontal or vertical) would fully occupy the 350-pixel length of the white square (the ratio between the horizontal and vertical dimensions of the object was kept, to avoid distortion). The stimuli were presented on a gray background. Of the total number of images, 180 images were allocated as images to compose the original predictive pairs, later to be violated or not violated (of these, 90 were classified as big objects and 90 were classified as small objects for the learning task) (see below). Items that would serve as novel foils in the memory test were also taken from that pool of 180 images. The allocation of stimuli to predictive pairs that would be violated or not, or to novel foils, as well as to location in the predictive pair (first or second item) was randomized per participant. In addition, 144 images were composed of 72 pairs of objects and similar exemplars. Thirty-six pairs were classified as big items, and 36 as small items. The allocation of objects to either violation or no-violation items was randomized per participant. For each participant, we also randomized which of the two exemplars per object would appear during the critical violation-phase, as well as which items within each type of item will be presented as identical old or similar lures in the item memory test.

Procedure

The experiment was conducted over 2 d and included a prediction learning phase, a reminder phase, a violation phase, an item memory test and associative memory tests. The prediction learning phase was conducted on day 1, and the rest of the experiment on day 2, and was scheduled ∼24 h apart. All phases of the task were controlled by Matlab (R2018b), using Psychtoolbox3 extensions (Brainard 1997; Pelli 1997; Kleiner et al. 2007). Generally, during the prediction learning, reminder, and violation phases, we used a statistical learning paradigm in which participants were presented with a stream of objects that included neighboring pairs of objects that followed each other back to back. Previous studies have established that after some learning, the prediction of the second object in the pair arises upon seeing the first object (Schapiro et al. 2012; Turk-Browne et al. 2012; Kim et al. 2014, 2017). Then, we violated this prediction in half of these pairs during the critical violation phase (Kim et al. 2014, 2017) and tested memory for these violations.

Prediction learning. During the prediction learning phase (Fig. 1A), participants were presented with a stream of objects. Each object was presented alone on the center of the screen for 1.5 sec and was followed by a 2.5-sec fixation cross located at the center of the screen as an interstimulus interval (ISI). The participants had to indicate, for each object, whether it is bigger or smaller than the previous object. The participants indicated "bigger" or "smaller" using a key press, and the keys were counterbalanced across participants. The initial learning phase was preceded by detailed instructions and a practice round. The participants were informed that during the experiment all objects have relatively the same size on the screen, but that they should make their judgments based on real life. They were additionally asked to respond as quickly as possible while still being accurate. Unknown to the participants, the stream of objects included 72 pairs that always followed each other (referred to here as "original pairs"). Half of these pairs of objects would later on be violated and half would remain intact during the violation

Downloaded from learnmem.cshlp.org on August 26, 2022 - Published by Cold Spring Harbor Laboratory Press

phase. Each pair was composed of a big object and a small object. In half of the pairs (within each pair type: to be violated or remain intact), the first object was the big object, and in the other half, it was the small object. The pairing of objects was fully randomized for each participant (maintaining the limitation that each pair included a small object and a big object), as well as the allocation of objects to either to-be violated/remain intact pairs. Pairs repeated nine times during the prediction learning phase, in nine cycles. Within each cycle, all pairs appeared, and the order of the pairs was randomized. We limited the randomization such that there would be a gap of at least two pairs between repetitions of the same pair (which could have happened across cycles). Prior to every odd-number cycle (cycles 1, 3, 5, and 7), a gray screen appeared with a short reminder of the task's instructions, and participants pressed a button to continue. To allow participants some additional breaks, before the even-number cycles (cycles 2, 4, 6, and 8) we introduced a 1-min break in which a gray screen with the sentence, "We'll continue in a bit," appeared. Participants were instructed that they do not need to do anything to continue the experiment, just stay concentrated for when the task starts again, which happened automatically. After the sixth cycle, we introduced a longer break in which we asked the participants to get the experimenter from the other room. When the experimenter and the participants returned to the room, the experimenter instructed the participants that they would now continue the task, as before, and the task proceeded. Following the prediction learning phase, participants were thanked for their participation and were reminded to come back for the second day of the experiment.

Reminder. Day 2 took place ∼24 h after day 1, and started with a reminder session, which was identical to the initial learning phase, but it only included one cycle in which all 72 pairs were presented again once. This phase was preceded by detailed instructions and a practice, identical to the initial learning phase. Prior to the beginning of the reminder cycle, a gray screen with a short reminder of the instructions (identical to prior to odd-number cycles in the initial learning phase) appeared, and participants pressed a button shortly when they were ready to start the task.

Violation phase. The critical violation phase (Fig. 1B) immediately followed the reminder phase, with no explicit instructions to the participants, aside from a short reminder of the task instructions on the computer screen (identical to the reminder task and odd-number cycles in the prediction learning phase). The violation phase was identical to the initial learning phase and the reminder in terms of the timing and the task that participants performed, but with modifications to the sequences of the objects to induce violations of prior expectations. The violation phase was divided into four blocks. Each block included nine pairs that were violated, and nine that remained intact (18 pairs per block, pairs did not repeat across blocks). Within each block, each pair appeared twice: The first appearance of all 18 pairs was intact, to allow an additional reminder of the pairs. In the second presentation of the pairs, half of the pairs were violated, and the other half remained intact. Thus, in total across the 4 block of the violation phase, 36 pairs were violated and 36 pairs remained intact. To induce violations, we introduced novel items that did not appear in the experiment before. Half of the novel items (36 in total, nine in each block) violated previous expectations (referred to here as violation items); these items were inserted in the sequence instead of the second item in the second presentation of the pairs (nine violated pairs per block). We made sure to violate the item's identity, but not the response, by inserting small novel items instead of a previously presented small item, and likewise for big items. The other half of the novel items did not violate any prediction (referred to here as no-violation items): These items were placed in the sequence after the second object in the second presentation of pairs that remained intact (36 in total, nine in each block). Since in the violation items the size of the object was switched from the previous object, such that a big violation object was presented after a small object, and vice-versa, we maintained the same switch for no-violation items, presenting big violation object after small objects, and vice versa (note that for both violation and no-violation items, half of the items were big, eliciting "bigger" response, while the other half was small items eliciting a "smaller" response). The no-violation items served as a control baseline during the item-memory test. Importantly, these objects were novel like the violation items; but, as they were scattered after the pairs, they did not violate any prediction. Placing a novel item after each second presentation of an intact pair would allow us to further test associative memory for these novel items (see associative memory test below). Since each block consisted of both first and second presentations of pairs, the violation rate was 25%. In 75% of the pairs, the learned predictions were valid. This might be important to ensure that the predictions still occur during the violation phase (Smith et al. 2013; Chen et al. 2015; Greve et al. 2017). Likewise, note that no-violation items only followed 25% of the pairs in each block, making it unlikely that participants started to expect these items after the second item in the pairs, even if participants came to recognize the pairs during the stream of objects in this statistical learning task. We further ensured that there will be at least six pairs between each presentation (first/second) of the same repeating pair in the violation phase. To allow that gap, each block started with four to-be violated pairs, and four to-remain intact pairs, presented in a random order, and proceeded with a mix of first-presentation pairs (five in each pair type) and second presentations of the pairs (violation/intact). We additionally controlled the sequence such that there would be no more than three consecutive violation trials, to prevent participants from expecting or habituating to violations. We further equated the average gap between the first and second presentation of pairs that were violated or remained intact. To allow breaks, the four blocks were separated by a gray screen with the sentence, "We'll continue in a bit." After a minute, the next block continued automatically (identical to the prediction learning phase). Thus, to our participants, the violation phase was similar to the reminder phase or any of the cycles in the prediction learning phase, and there was no explicit transition to the violation phase.

Item memory test

After the violation phase, we conducted a surprise item memory test for violation and no-violation items (Fig. 1C). In each type of item (violation/no-violation), half of the items (18; 36 in total) presented in the test were identical to the item presented during the violation phase (identical old). The other half of the items were similar lures: a different exemplar of an object presented during the violation phase. We additionally included 36 new items (18 big and 18 small) that did not appear in the experiment before and served as novel foils. Each item was presented alone on the screen for 3 sec and was followed by a 3-sec fixation cross. Participants were asked to indicate for each item, whether the item was "old," "similar" or "new." Specifically, the participants were instructed to indicate "old" if the object was identical to an object that appeared in the study. They were instructed to respond "similar" if they have seen the object in the study before, but it is not the exact object that appeared in the study, and they were instructed to respond "new" if they have not seen that object in the study before at all. The participants were also given an example for each type of response, and a short practice before starting the test. The participants indicated their responses by pressing one of three keys, which were counterbalanced across participants. The order of the objects was pseudorandomized for each participant. To maintain a spread of the type of trials (violation: old/similar lures; no-violation: old/similar lures; new) across the test, we made sure that each half of the test included half of the objects in each trial type. The order of the objects was randomized for each participant with the limitation that there would be no more than 4 consecutive identical response type ("old"/"similar"/"new"). The item

Downloaded from learnmem.cshlp.org on August 26, 2022 - Published by Cold Spring Harbor Laboratory Press

memory test preceded by detailed instructions and a practice, and was divided into two blocks to allow participants a short break. Each block was preceded by a gray screen with a short reminder of the instructions.

Associative memory test

Upon completion of the item memory test, participants were given a short distraction task in which they solved simple math equations for 3 min to reduce potential interference between the memory tests. After that distractor task, the associative memory test was administered. The associative memory test was preceded by detailed instructions and a short practice. On each trial, one object appeared on the top of the screen and three objects appeared in a row at the bottom of the screen: a target object and two distractors. Participants had to first indicate which of the three bottom objects followed the top object during the experiment. After they made their choice, the other two objects disappeared and a scale of 1–6 appeared below the chosen object. Participants then rated their confidence that the object they had chosen had followed the top on the 1–6 scale, where one would be guess, and six would be very sure. The participants were encouraged to use the entire range of the 1–6 scale. They were told they could take as long as they want within a 10-sec window for each decision (i.e., choosing the item and confidence rating). Upon rating confidence (after choosing the object), all objects were removed from the screen, participants viewed a fixation cross for 1 sec, and a new trial began. If the participants did not respond within the allocated 10 sec, the objects were removed from the screen and the fixation cross appeared prior to a new trial.

We first tested all the "novel" associations; namely, the associations that included the novel items. Specifically, in the violation pairs, we presented the first object in the pair at the top of the screen, as this is the item that preceded the novel violation item during the violation phase. Since we presented the no-violation items after intact pairs, we could test associative memory for these items as well. For these items, we presented the second object in the pair at the top of the screen, as this was the item that preceded the novel no-violation item. The distractors were always intralist; namely, violation targets appeared with other violation items as distractors, and likewise for no-violation targets. The distractors were also of the same size category (big or small) as the target object. The location of the target (left/middle/right in the row of three bottom objects) was pseudorandomized such that a third of the targets in each trial type (violation/no-violation) appeared in each location. The order of the trials was randomized per participant, maintaining that there was a gap of at least two trials between the appearance of the same object (as target/distractor).

After testing for all of the novel associations, we tested memory of the original pairs, namely the pairs that appeared multiple times during the initial learning and the violation phase and were then violated or remained intact. This test was identical to the memory test for the novel association, only that the cue, at the top of the screen, was the first item of the pair, and the target, at the bottom of the screen with two distractors of the same pair-type (violation or intact) and size category (big or small), was the second item in the pair. There was no explicit notice to the participants between testing the novel and the original associations, aside from having a short break in which the gray screen appeared with a short reminder of the instructions. To our participants, it was merely as if another block of the associative memory test had begun. After this memory test, the participants were debriefed and received compensation.

Analysis

To enable the examination of violation items for which participants remembered the learned prediction, we analyzed associative memory rates for the predictive pairs. We compared between memory rates for these pairs in the violation and no-violation condition using a paired-sample *t*-test (all *t*-tests reported here are two-tailed). Analyses were done using custom-made Matlab code.

Prior to testing our main hypothesis, we established that participants indeed distinguished between similar lures and identical old items in the item memory test. To that end, we collapsed across violation and no-violation items, and examined the rates of responding "old" or "similar" to identical old items versus similar lures. These response rates were entered to a repeated-measures ANOVA with Item Type (identical old or similar lure) and Response ("old" or "similar") as factors. Planned comparisons between "old" and "similar" responses within each item type were conducted using paired-sample two-tailed *t*-tests. We further compared between "old" responses to identical old versus novel foils, as well as "similar" responses to similar lures versus novel foils, using paired-sample *t*-tests.

Our main analysis focused on comparing item memory between violation and no-violation items for which participants remembered the original predictive pairs. To that end, we first identified which items violated (for violation items) or followed (no-violation items) pairs that participants correctly remembered, and which items violated or followed pairs that participants forgot. We then took only items for which participants remembered the preceding pair and calculated the rates of responding "old" or "similar" to identical old violation and no-violation items. These memory rates were tested using a repeated-measures ANOVA with violation (violation or no-violation) and response type ("old" or "similar") as factors. Planned comparisons between violation and no-violation items in each response type were performed using paired-sample *t*-tests. We further compared the rates of "old" responses with violation and no-violation items (where we found a difference in response rate), when participants remembered versus forgot the original predictive pairs. These "old" response rates were examined using a repeated-measures ANOVA with Violation (violation or no-violation) and Memory (remembered or forgotten) as factors. Planned comparisons between remembered and forgotten predictive pairs, within either violation or no-violation items, were conducted using paired-sample *t*-tests.

Finally, associative memory rates for the novel associations (which included the violation or no-violation items) were compared with chance; associative memory rates were also compared between violation and no-violation associations, both analyses used paired-sample *t*-tests. For these associations as well, we examined pairs for which the original predictive pair was remembered or forgotten.

Experiment 2

Participants

Twenty-eight participants were included in this study (24 females, aged 18–30 yr, mean age: 22.57). Four additional participants were excluded due to poor compliance with the task, by the same criterion as experiment 1; namely, <40% memory for both types of old associations (that were violated or remained intact during the violation phase). Indeed, we alleviated the concern of experiment 1, as here only four participants were excluded. The participants were members of the Columbia University community, with normal or corrected to normal vision. They provided written informed consent to participate in the study and received a payment at a rate of $12/h for their participation. The study was approved by the Columbia University Institutional Review Board. The sample size was determined based on experiment 1. Since we aimed for exactly 28 participants, here we collected data until we reached this sample size after exclusion of participants.

Procedure

The procedure was identical to experiment 1, with the only modifications that the initial learning phase included 12 cycles of repetition of the pairs, and the ISI during the initial learning phase, the reminder and the violation phase was 0.5 sec. These changes were made to enhance learning of the pairs prior to the violation phase.

Downloaded from learnmem.cshlp.org on August 26, 2022 - Published by Cold Spring Harbor Laboratory Press

Experiment 3

Participants

Twenty-eight participants were included in this study (20 females, aged 18–34 yr, mean age: 24.64). The sample size was determined based on experiments 1 and 2. Naturally, because this experiment aimed to examine item memory for the violation of weakly encoded predictions, we did not exclude participants based on low memory of the original pairs (also note that, on average, memory rates were ~0.4, which was our exclusion criterion in the previous experiments) (see above).

Procedure

The procedure was identical to experiment 2, with the only modification that during learning, participants were asked to indicate whether an object was bigger or smaller than a shoe box in real-life (rather than whether an object is bigger or smaller than the previous object in the sequence in real life).

Data Availability

Raw data and the stimuli used for this study are available online (https://osf.io/thzub). All analysis code is available at https://github.com/odedbein/simPEL_public.

Acknowledgments

This research was supported by National Institute of Mental Health grant R01MH074692 to L.D. O.B. is further supported by a McCracken fellowship and the Grinker award. We thank Taylor Nicholas for her help in data collection. We further thank David Clewett, Alexandra Cohen, Gail Rosenbaum, and Amir Tal for their thoughtful comments during the preparation of the manuscript.

References

Allen TA, Salz DM, McKenzie S, Fortin NJ. 2016. Nonspatial sequence coding in CA1 neurons. *J Neurosci* **36:** 1547–1563. doi:10.1523/JNEUROSCI.2874-15.2016

Antony J, Hartshorne T, Pomeroy K, Gureckis T, Hasson U, McDougle S, Norman K. 2020. Behavioral, physiological, and neural signatures of surprise during naturalistic sports viewing. *Neuron* **109:** 377–390.e7. doi:10.1101/2020.03.26.008714

Axmacher N, Cohen MX, Fell J, Haupt S, Dümpelmann M, Elger CE, Schlaepfer TE, Lenartz D, Sturm V, Ranganath C. 2010. Intracranial EEG correlates of expectancy and memory formation in the human hippocampus and nucleus accumbens. *Neuron* **65:** 541–549. doi:10.1016/j.neuron.2010.02.006

Baker S, Vieweg P, Gao F, Gilboa A, Wolbers T, Black SE, Rosenbaum RS. 2016. The human dentate gyrus plays a necessary role in discriminating new memories. *Curr Biol* **26:** 2629–2634. doi:10.1016/j.cub.2016.07.081

Bakker A, Kirwan CB, Miller M, Stark CEL. 2008. Pattern separation in the human hippocampal CA3 and dentate gyrus. *Science* **319:** 1640–1643. doi:10.1126/science.1152882

Bein O, Duncan K, Davachi L. 2020a. Mnemonic prediction errors bias hippocampal states. *Nat Commun* **11:** 3451. doi:10.1038/s41467-020-17287-1

Bein O, Reggev N, Maril A. 2020b. Prior knowledge promotes hippocampal separation but cortical assimilation in the left inferior frontal gyrus. *Nat Commun* **11:** 4590. doi:10.1038/s41467-020-18364-1

Berron D, Schutze H, Maass A, Cardenas-Blanco A, Kuijf HJ, Kumaran D, Duezel E. 2016. Strong evidence for pattern separation in human dentate gyrus. *J Neurosci* **36:** 7569–7579. doi:10.1523/JNEUROSCI.0518-16.2016

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

Brod G, Hasselhorn M, Bunge SA. 2018. When generating a prediction boosts learning: the element of surprise. *Learn Instruct* **55:** 22–31. doi:10.1016/j.learninstruc.2018.01.013

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

Chen J, Olsen RK, Preston AR, Glover GH, Wagner AD. 2011. Associative retrieval processes in the human medial temporal lobe: hippocampal retrieval success and CA1 mismatch detection. *Learn Mem* **18:** 523–528. doi:10.1101/lm.2135211

Chen J, Cook PA, Wagner AD. 2015. Prediction strength modulates responses in human area CA1 to sequence violations. *J Neurophysiol* **114:** 1227–1238. doi:10.1152/jn.00149.2015

Clewett D, Davachi L. 2017. The ebb and flow of experience determines the temporal structure of memory. *Curr Opin Behav Sci* **17:** 186–193. doi:10.1016/j.cobeha.2017.08.013

Clewett D, DuBrow S, Davachi L. 2019. Transcending time in the brain: how event memories are constructed from experience. *Hippocampus* **29:** 162–183. doi:10.1002/hipo.23074

Colgin LL. 2016. Rhythms of the hippocampal network. *Nat Rev Neurosci* **17:** 239–249. doi:10.1038/nrn.2016.21

Davachi L. 2006. Item, context and relational episodic encoding in humans. *Curr Opin Neurobiol* **16:** 693–700. doi:10.1016/j.conb.2006.10.012

Davidow JY, Foerde K, Galván A, Shohamy D. 2016. An upside to reward sensitivity: the hippocampus supports enhanced reinforcement learning in adolescence. *Neuron* **92:** 93–99. doi:10.1016/j.neuron.2016.08.031

Davis T, Love BC, Preston AR. 2012. Learning the exception to the rule: model-based fMRI reveals specialized representations for surprising category members. *Cereb Cortex* **22:** 260–273. doi:10.1093/cercor/bhr036

De Loof E, Ergo K, Naert L, Janssens C, Talsma D, Van Opstal F, Verguts T. 2018. Signed reward prediction errors drive declarative learning. *PLoS One* **13:** 1–15. doi:10.1371/journal.pone.0189212

Detre GJ, Natarajan A, Gershman SJ, Norman KA. 2013. Moderate levels of activation lead to forgetting in the think/no-think paradigm. *Neuropsychologia* **51:** 2371–2388. doi:10.1016/j.neuropsychologia.2013.02.017

DuBrow S, Davachi L. 2013. The influence of context boundaries on memory for the sequential order of events. *J Exp Psychol Gen* **142:** 1277–1286. doi:10.1037/a0034024

DuBrow S, Davachi L. 2014. Temporal memory is shaped by encoding stability and intervening item reactivation. *J Neurosci* **34:** 13998–14005. doi:10.1523/jneurosci.2535-14.2014

DuBrow S, Davachi L. 2017. Commentary: distinct neural mechanisms for remembering when an event occurred. *Front Psychol* **8:** 189. doi:10.1002/hipo.22571

Duncan K, Ketz N, Inati SJ, Davachi L. 2012. Evidence for area CA1 as a match/mismatch detector: a high-resolution fMRI study of the human hippocampus. *Hippocampus* **22:** 389–398. doi:10.1002/hipo.20933

Eichenbaum H, Yonelinas AP, Ranganath C. 2007. The medial temporal lobe and recognition memory. *Annu Rev Neurosci* **30:** 123–152. doi:10.1146/annurev.neuro.30.051606.094328

Ergo K, De Loof E, Verguts T. 2020. Reward prediction error and declarative memory. *Trends Cogn Sci* **24:** 388–397. doi:10.1016/j.tics.2020.02.009

Favila SE, Chanales AJH, Kuhl BA. 2016. Experience-dependent hippocampal pattern differentiation prevents interference during subsequent learning. *Nat Commun* **6:** 11066. doi:10.1038/ncomms11066

Frank D, Kafkas A. 2021. Expectation-driven novelty effects in episodic memory. *Neurobiol Learn Mem* **183:** 107466. doi:10.1016/j.nlm.2021.107466

Frank D, Montemurro MA, Montaldi D. 2020. Pattern separation underpins expectation-modulated memory. *J Neurosci* **40:** 3455–3464. doi:10.1523/JNEUROSCI.2047-19.2020

Franklin NT, Norman KA, Ranganath C, Zacks JM, Louis S, Gershman SJ. 2020. Structured event memory: a neuro-symbolic model of event cognition. *Psychol Rev* **127:** 327–361. doi:10.1101/541607

Friston K. 2018. Does predictive coding have a future? *Nat Neurosci* **21:** 1019–1021. doi:10.1038/s41593-018-0200-7

Garlitch SM, Wahlheim CN. 2020. The role of attentional fluctuation during study in recollecting episodic changes at test. *Mem Cognit* **48:** 800–814. doi:10.3758/s13421-020-01018-4

Gershman SJ, Niv Y. 2010. Learning latent structure: carving nature at its joints. *Curr Opin Neurobiol* **20:** 251–256. doi:10.1016/j.conb.2010.02.008

Gershman SJ, Blei DM, Niv Y. 2010. Context, learning, and extinction. *Psychol Rev* **117:** 197–209. doi:10.1037/a0017808

Gershman SJ, Radulescu A, Norman KA, Niv Y. 2014. Statistical computations underlying the dynamics of memory updating. *PLoS Comput Biol* **10:** e1003939. doi:10.1371/journal.pcbi.1003939

Gershman SJ, Monfils M, Norman KA, Niv Y. 2017. The computational nature of memory modification. *Elife* **6:** e23763. doi:10.7554/eLife.23763

Gläscher J, Daw N, Dayan P, O'Doherty JP. 2010. States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron* **66:** 585–595. doi:10.1016/j.neuron.2010.04.016

Green RT. 1956. Surprise as a factor in the Von Restorff effect. *J Exp Psychol* **52:** 340–344.

Downloaded from learnmem.cshlp.org on August 26, 2022 - Published by Cold Spring Harbor Laboratory Press

Greve A, Cooper E, Kaula A, Anderson MC, Henson R. 2017. Does prediction error drive one-shot declarative learning? *J Mem Lang* **94:** 149–165. doi:10.1016/j.jml.2016.11.001

Greve A, Abdulrahman H, Henson RN. 2018. Neural differentiation of incorrectly predicted memories. *Front Hum Neurosci* **12:** 278.

Hasselmo ME, Stern CE. 2014. Theta rhythm and the encoding and retrieval of space and time. *Neuroimage* **85:** 656–666. doi:10.1016/j.neuroimage.2013.06.022

Hasselmo ME, Wyble BP, Wallenstein GV. 1996. Encoding and retrieval of episodic memories: role of cholinergic and GABAergic modulation in the hippocampus. *Hippocampus* **6:** 693–708.

Henson RN, Gagnepain P. 2010. Predictive, interactive multiple memory systems. *Hippocampus* **20:** 1315–1326. doi:10.1002/hipo.20857

Horner AJ, Burgess N. 2013. The associative structure of memory for multi-element events. *J Exp Psychol Gen* **142:** 1370–1383. doi:10.1037/a0033626

Horner AJ, Burgess N. 2014. Pattern completion in multielement event engrams. *Curr Biol* **24:** 988–992. doi:10.1016/j.cub.2014.03.012

Hunt RR. 1995. The subtlety of distinctiveness: what von Restorff really did. *Psychon Bull Rev* **2:** 105–112. doi:10.3758/BF03214414

Hunt RR. 2006. The concept of distinctiveness in memory research. In: *Distinctiveness and memory* (ed. Hunt RR, Worthen JB). Oxford University Press, Oxford, UK.

Jang AI, Nassar MR, Dillon DG, Frank MJ. 2019. Positive reward prediction errors during decision-making strengthen memory encoding. *Nat Hum Behav* **3:** 719–732. doi:10.1038/s41562-019-0597-3

Jenkins LJ, Ranganath C. 2016. Distinct Neural Mechanisms for Remembering When an Event Occurred. *Hippocampus* **26:** 554–559. doi:10.1002/hipo.22571

Kafkas A, Montaldi D. 2018a. Expectation affects learning and modulates memory experience at retrieval. *Cognition* **180:** 123–134. doi:10.1016/j.cognition.2018.07.010

Kafkas A, Montaldi D. 2018b. How do memory systems detect and respond to novelty? *Neurosci Lett* **680:** 60–68. doi:10.1016/j.neulet.2018.01.053

Kahana MJ. 1996. Associative retrieval processes in free recall. *Mem Cognit* **24:** 103–109. doi:10.3758/bf03197276

Kalbe F, Schwabe L. 2019. Beyond arousal: prediction error related to aversive events promotes episodic memory formation. *J Exp Psychol Learn Mem Cogn* **46:** 234–246. doi:10.1037/xlm0000728

Kim G, Lewis-Peacock JA, Norman KA, Turk-Browne NB. 2014. Pruning of memories by context-based prediction error. *Proc Natl Acad Sci* **111:** 8997–9002. doi:10.1073/pnas.1319438111

Kim G, Norman KA, Turk-Browne NB. 2017. Neural differentiation of incorrectly predicted memories. *J Neurosci* **37:** 2022–2031. doi:10.1523/JNEUROSCI.3272-16.2017

Kim H, Schlichting ML, Preston AR, Lewis-Peacock JA. 2020. Predictability changes what we remember in familiar temporal contexts. *J Cogn Neurosci* **32:** 124–140.

Kleiner M, Brainard D, Pelli D, Ingling A, Murray R, Broussard C. 2007. What's new in Psychtoolbox-3? *Perception* **36:** 1–16.

Knierim JJ, Neunuebel JP. 2016. Tracking the flow of hippocampal computation: pattern separation, pattern completion, and attractor dynamics. *Neurobiol Learn Mem* **129:** 38–49. doi:10.1016/j.nlm.2015.10.008

Kok P, Mostert P, De Lange FP. 2017. Prior expectations induce prestimulus sensory templates. *Proc Natl Acad Sci* **114:** 10473–10478. doi:10.1073/pnas.1705652114

Kuhl BA, Shah AT, DuBrow S, Wagner AD. 2010. Resistance to forgetting associated with hippocampus-mediated reactivation during new learning. *Nat Neurosci* **13:** 501–506. doi:10.1038/nn.2498

Kuhl BA, Rissman J, Chun MM, Wagner AD. 2011. Fidelity of neural reactivation reveals competition between memories. *Proc Natl Acad Sci* **108:** 5903–5908. doi:10.1073/pnas.1016939108

Kumaran D, Maguire EA. 2006. An unexpected sequence of events: mismatch detection in the human hippocampus. *PLoS Biol* **4:** 2372–2382. doi:10.1371/journal.pbio.0040424

Kumaran D, Maguire EA. 2007. Match-mismatch processes underlie human hippocampal responses to associative novelty. *J Neurosci* **27:** 8517–8524. doi:10.1523/jneurosci.1677-07.2007

Kumaran D, Hassabis D, McClelland JL. 2016. What learning systems do intelligent agents need? Complementary Learning Systems Theory Updated. *Trends Cogn Sci* **20:** 512–534. doi:10.1016/j.tics.2016.05.004

Lacy JW, Yassa MA, Stark SM, Muftuler LT, Stark CEL. 2011. Distinct pattern separation related transfer functions in human CA3/dentate and CA1 revealed using high-resolution fMRI and variable mnemonic similarity. *Ward* **2001:** 15–18.

LaRocque KF, Smith ME, Carr VA, Witthoft N, Grill-Spector K, Wagner AD. 2013. Global similarity and pattern separation in the human medial temporal lobe predict subsequent memory. *J Neurosci* **33:** 5466–5474. doi:10.1523/jneurosci.4293-12.2013

Leutgeb JK, Leutgeb S, Moser M, Moser EI. 2007. Pattern separation in the dentate gyrus and CA3 of the hippocampus. *Science* **315:** 961–966. doi:10.1126/science.1135801

Long NM, Lee H, Kuhl BA. 2016. Hippocampal mismatch signals are modulated by the strength of neural predictions and their similarity to outcomes. *J Neurosci* **36:** 1850–1816. doi:10.1523/JNEUROSCI.1850-16.2016

Love BC, Medin DL, Gureckis TM. 2004. SUSTAIN: a network model of category learning. *Psychol Rev* **111:** 309–332. doi:10.1037/0033-295X.111.2.309

McClelland JL, McNaughton BL, Oreilly RC. 1995. Why there are complementary learning-systems in the hippocampus and neocortex: insights from the success and failures of connectionist models of learning and memory. *Psychol Rev* **102:** 419–457. doi:10.1037/0033-295x.102.3.419

Murty VP, LaBar KS, Adcock RA. 2016. Distinct medial temporal networks encode surprise during motivation by reward versus punishment. *Neurobiol Learn Mem* **134:** 55–64. doi:10.1016/j.nlm.2016.01.018

Newman EL, Norman KA. 2010. Moderate excitation leads to weakening of perceptual representations. *Cereb Cortex* **20:** 2760–2770. doi:10.1093/cercor/bhq021

Niv Y, Schoenbaum G. 2008. Dialogues on prediction errors. *Trends Cogn Sci* **12:** 265–272. doi:10.1016/j.tics.2008.03.006

Norman KA, O'Reilly RC. 2003. Modeling hippocampal and neocortical contributions to recognition memory: a complementary-learning-systems approach. *Psychol Rev* **110:** 611–646. doi:10.1037/0033-295X.110.4.611

O'Doherty J, Dayan P, Schultz J, Deichmann R, Friston K, Dolan RJ. 2004. Dissociable Roles of Ventral and Dorsal Striatum in Instrumental Conditioning. *Science* **304:** 452–454. doi:10.1126/science.1094285

O'Keefe DJ. 2007. Brief report: post hoc power, observed power, a priori power, retrospective power, prospective power, achieved power: sorting out appropriate uses of statistical power analyses. *Commun Methods Meas* **1:** 291–299. doi:10.1080/19312450701641375

Ortiz-Tudela J, Milliken B, Jiménez L, Lupiáñez J. 2018. Attentional influences on memory formation: a tale of a not-so-simple story. *Mem Cognit* **46:** 544–557. doi:10.3758/s13421-017-0784-2

Pelli DG. 1997. The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spat Vis* **10:** 437–442.

Polyn SM, Natu VS, Cohen JD, Norman KA. 2005. Category-specific cortical activity precedes retrieval during memory search. *Science* **310:** 1963–1966. doi:10.1126/science.1117645

Quent JA, Henson RN, Greve A. 2021. A predictive account of how novelty influences declarative memory. *Neurobiol Learn Mem* **179:** 107382. doi:10.1016/j.nlm.2021.107382

Ranganath C, Rainer G. 2003. Neural mechanisms for detecting and remembering novel events. *Nat Rev Neurosci* **4:** 193–202. doi:10.1038/nrn1052

Reichardt R, Polner B, Simor P. 2020. Novelty manipulations, memory performance, and predictive coding: the role of unexpectedness. *Front Hum Neurosci* **14:** 152. doi:10.3389/fnhum.2020.00152

Rescorla RA, Wagner AR. 1972. A theory of Pavlovian conditioning. In *Classical conditioning II. Current research and theory* (ed. Black AH, Prokasy WF), pp. 64–99. Appleton-Century-Crofts, New York.

Richter FR, Chanales AJH, Kuhl BA. 2016. Predicting the integration of overlapping memories by decoding mnemonic processing states during learning. *Neuroimage* **124:** 323–335. doi:10.1016/j.neuroimage.2015.08.051

Ritvo VJH, Turk-Browne NB, Norman KA. 2019. Nonmonotonic plasticity: how memory retrieval drives learning. *Trends Cogn Sci* **23:** 726–742. doi:10.1016/j.tics.2019.06.007

Rouhani N, Niv Y. 2021. Signed and unsigned reward prediction errors dynamically enhance learning and memory. *Elife* **10:** e61077.

Rouhani N, Norman KA, Niv Y. 2018. Dissociable effects of surprising rewards on learning and memory. *J Exp Psychol Learn Mem Cogn* **44:** 1430–1443. doi:10.1037/xlm0000518

Rouhani N, Norman KA, Niv Y, Bornstein AM. 2020. Reward prediction errors create event boundaries in memory. *Cognition* **203:** 104269.

Schapiro AC, Turk-Browne N. 2015. Statistical learning. *Brain Mapp* **3:** 501–506. doi:10.1016/B978-0-12-397025-1.00276-1

Schapiro AC, Kustner LV, Turk-Browne NB. 2012. Shaping of object representations in the human medial temporal lobe based on temporal regularities. *Curr Biol* **22:** 1622–1627. doi:10.1016/j.cub.2012.06.056

Schlichting ML, Preston AR. 2015. Memory integration: neural mechanisms and implications for behavior. *Curr Opin Behav Sci* **1:** 1–8. doi:10.1016/j.cobeha.2014.07.005

Schlichting ML, Mumford JA, Preston AR. 2015. Learning-related representational changes reveal dissociable integration and separation signatures in the hippocampus and prefrontal cortex. *Nat Commun* **6:** 8151. doi:10.1038/ncomms9151

Schmidt SR, Schmidt CR. 2017. Revisiting von Restorff's early isolation effect. *Mem Cognit* **45:** 194–207. doi:10.3758/s13421-016-0651-6

Downloaded from learnmem.cshlp.org on August 26, 2022 - Published by Cold Spring Harbor Laboratory Press

Schomaker J, Meeter M. 2015. Short- and long-lasting consequences of novelty, deviance and surprise on brain and cognition. *Neurosci Biobehav Rev* **55:** 268–279. doi:10.1016/j.neubiorev.2015.05.002

Schultz W, Dayan P, Montague PR. 1997. A neural substrate of prediction and reward. *Science* **275:** 1593–1599. doi:10.1126/science.275.5306.1593

Sherman BE, Turk-Browne NB. 2020. Statistical prediction of the future impairs episodic encoding of the present. *Proc Nat Acad Sci* **117:** 22760–22770. doi:10.1073/pnas.2013291117

Siegelman N, Bogaerts L, Christiansen MH, Frost R. 2017. Towards a theory of individual differences in statistical learning. *Philos Trans R Soc B Biol Sci* **372:** 1711. doi:10.1098/rstb.2016.0059

Sinclair AH, Barense MD. 2019. Prediction error and memory reactivation: how incomplete reminders drive reconsolidation. *Trends Neurosci* **42:** 727–739. doi:10.1016/j.tins.2019.08.007

Smith TA, Hasinski AE, Sederberg PB. 2013. The context repetition effect: predicted events are remembered better, even when they don't happen. *J Exp Psychol Gen* **142:** 1298–1308. doi:10.1037/a0034067

Staresina BP, Davachi L. 2008. Selective and shared contributions of the hippocampus and perirhinal cortex to episodic item and associative encoding. *J Cogn Neurosci* **20:** 1478–1489. doi:10.1162/jocn.2008.20104

Staresina BP, Duncan K, Davachi L. 2011. Perirhinal and parahippocampal cortices differentially contribute to later recollection of object- and scene-related event details. *J Neurosci* **31:** 8739–8747. doi:10.1523/JNEUROSCI.4978-10.2011

Stark SM, Kirwan CB, Stark CEL. 2019. Mnemonic similarity task: a tool for assessing hippocampal integrity. *Trends Cogn Sci* **23:** 938–951. doi:10.1016/j.tics.2019.08.003

Tompary A, Davachi L. 2017. Consolidation promotes the emergence of representational overlap in the hippocampus and medial prefrontal cortex. *Neuron* **96:** 228–241. doi:10.1016/j.neuron.2017.09.005

Tulving E, Kroll N. 1995. Novelty assessment in the brain and long-term memory encoding. *Psychon Bull Rev* **2:** 387–390. doi:10.3758/bf03210977

Turk-Browne NB, Simon MG, Sederberg PB. 2012. Scene representations in parahippocampal cortex depend on temporal context. *J Neurosci* **32:** 7202–7207. doi:10.1523/JNEUROSCI.0942-12.2012

von Restorff H. 1933. Über die wirkung yon bereiehsbildungen im spurenfeld (The effects of field formation in the trace field). *Psychol Forschung* **18:** 299–342. doi:10.1007/BF02441202

Waddill PJ, McDaniel MA. 1998. Distinctiveness effects in recall: differential processing or privileged retrieval? *Mem Cognit* **26:** 108–120. doi:10.3758/BF03211374

Wahlheim CN, Zacks JM. 2019. Memory guides the processing of event changes for older and younger adults. *J Exp Psychol Gen* **148:** 30–50.

Wahlheim CN, Smith WG, Delaney PF. 2019. Reminders can enhance or impair episodic memory updating: a memory-for-change perspective. *Memory* **27:** 849–867. doi:10.1080/09658211.2019.1582677

Wimmer GE, Braun EK, Daw ND, Shohamy D. 2014. Episodic memory encoding interferes with reward learning and decreases striatal prediction errors. *J Neurosci* **34:** 14901–14912. doi:10.1523/JNEUROSCI.0204-14.2014

Zacks JM, Tversky B. 2001. Event structure in perception and conceptions. *Psychol Bull* **127:** 3–21.

Zacks JM, Kurby CA, Eisenberg ML, Haroutunian N. 2011. Prediction error associated with the perceptual segmentation of naturalistic events. *J Cogn Neurosci* **23:** 4057–4066. doi:10.1162/jocn_a_00078

Received March 16, 2021; accepted in revised form August 16, 2021.

Downloaded from learnmem.cshlp.org on August 26, 2022 - Published by Cold Spring Harbor Laboratory Press

CSH PRESS LEARNING & MEMORY

Mnemonic prediction errors promote detailed memories

Oded Bein, Natalie A. Plotkin and Lila Davachi

Learn. Mem. 2021, **28:**
Access the most recent version at doi:10.1101/lm.053410.121

Supplemental Material **http://learnmem.cshlp.org/content/suppl/2021/09/29/28.11.422.DC1**

References This article cites 111 articles, 22 of which can be accessed free at:
http://learnmem.cshlp.org/content/28/11/422.full.html#ref-list-1

Creative Commons License This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first 12 months after the full-issue publication date (see http://learnmem.cshlp.org/site/misc/terms.xhtml). After 12 months, it is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at http://creativecommons.org/licenses/by-nc/4.0/.

Email Alerting Service Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or **click here.**

© 2021 Bein et al.; Published by Cold Spring Harbor Laboratory Press