Predictability changes what we remember in familiar temporal contexts

Hyojeong Kim[1*], Margaret L. Schlichting[2], Alison R. Preston[1,3,4],
and Jarrod A. Lewis-Peacock[1,3]

[1]Department of Psychology, University of Texas at Austin, Austin, TX
[2]Department of Psychology, University of Toronto, Toronto, ON, Canada
[3]Center for Learning & Memory, University of Texas at Austin, Austin, TX
[4]Department of Neuroscience, University of Texas at Austin, Austin, TX

***Corresponding Author:**
Hyojeong Kim
Department of Psychology
University of Texas at Austin
Austin, TX 78712
Tel: +1 512 232-1805
hyojeongkim@utexas.edu

Abstract

The human brain constantly anticipates the future based on memories of the past. Encountering a familiar situation can trigger a prediction of what comes next, with a prediction error leading to pruning of the offending memory. Situations with more predictable events should trigger more reliable predictions, and this could have an impact on memory pruning. Our goal was to evaluate whether memories are spared from pruning in situations that allow for more reliable predictions. Participants viewed a sequence of objects, some of which (“cues”) reappeared multiple times, followed always by novel items. Half of the cues were followed by items from different (unpredictable) categories, while others were followed by items from a single (predictable) category. Pattern classification of fMRI data was used to identify category-specific predictions after each cue. Pruning was observed only in unpredictable contexts, while encoding of new items suffered more in predictable contexts.

Introduction

What is past is prologue: similar to the function of autocomplete software on a smartphone, the brain learns from statistical patterns across time to generate expectations that guide future behavior. This process is essential for most of our fundamental abilities including language, perception, action, and memory. It is accomplished in part by domain-general implicit learning mechanisms (alternatively referred to as ‘statistical learning’ (Perruchet and Pacton, 2006; Turk-Browne and Scholl, 2009; Turk-Browne et al., 2005)) that allow us to acquire long-term knowledge about the statistical structure of the world (Kóbor et al., 2017; Romano et al., 2010). Studies of visual statistical learning have shown that observers can implicitly learn subtle statistical relationships between visual stimuli in both time (Fiser et al., 2007; Fiser and Aslin, 2002; Schapiro et al., 2012) and space (Chun and Jiang, 1998; Fiser and Aslin, 2001; Turk-Browne and Scholl, 2009). Knowledge of these statistics can build up expectations that trigger predictions about upcoming perceptual events (Turk-Browne et al., 2010). In some situations, these expectations may be stimulus-specific (Conway and Christiansen, 2006), and in others they may be more abstract, for example, operating at a categorical level that relies on existing conceptual knowledge (Brady and Oliva, 2008). Many real-world situations have relatively stable abstract statistics but highly variable specific details. For example, when entering a coffee shop, one should expect to find a barista behind the counter, but not necessarily the same barista that served coffee on your prior visit. Here, a more abstract prediction (“some barista”) would be more reliable than a specific prediction (“that particular barista”).

The acquisition of new episodic memories is mediated by expectations of what will happen in the near future. Recent neural evidence shows that the brain processes (or encodes) predictable events less strongly than unpredictable events, as evidenced by diminished repetition priming effects for repeated words appearing in predictable temporal contexts (Rommers and Federmeier, 2018). An advantage of predictability is that it can help reduce processing of redundant information during encoding. However, this may come at the expense of detailed stimulus processing, leading to weaker encoding of new memories. Further support for this idea comes from research on “schemas” (Ghosh and Gilboa, 2014; Tse et al., 2007; Marlieke T R van Kesteren et al., 2010; van Kesteren et al., 2012) which have been shown to influence memory encoding. Schemas are associative network structures that develop across multiple episodes and provide an abstract, conceptual framework to facilitate adaptive behavior. Schemas allow for reliable predictions at a fairly abstract level, but this comes at the cost of decreased attention to specific schema-consistent information, which may lead to reduced encoding of the new experience details.

Likewise, the updating of existing memories is also influenced by temporal expectations. When an automatic prediction of an upcoming perceptual item is violated (*misprediction*), the resulting error signal can weaken the long-term memory representation of the mispredicted item, leading it to be pruned from memory (Kim et al., 2017, 2014). Memory pruning is an adaptive, and error-driven learning process (Pagnoni et al., 2002; Schultz and Dickinson, 2000) in which irrelevant item representations are selectively removed from the memory trace to improve access to relevant memories. Prediction error does not always lead to pruning, however. For example, the fidelity of reactivated predictions is often graded, and pruning is most likely for items that are moderately, but not strongly reactivated. When reactivated predictions don't get pruned, they may instead become integrated with new learning episodes (Morton et al., 2017; Preston and Eichenbaum, 2013; Schlichting et al., 2015; Schlichting and Frankland, 2017; Schlichting and Preston, 2015; Zeithamova et al., 2012a, 2012b). Memory integration is a process by which related experiences are stored as overlapping representations in the brain, forming memory networks that span events and support the flexible extraction of novel information. Importantly, it is unclear why certain memories are pruned and others are integrated. Therefore, a central goal of this study is to evaluate how the nature of statistically learned expectations influences how memories are both acquired and updated.

A key factor that may impact whether memories are pruned or integrated is the *reliability of predictions* that are generated in a familiar temporal context. For example, encountering familiar items in a sequence of stimuli can trigger automatic predictions for what stimuli will appear next based on which items appeared in the past. These predictions will be incorrect when novel items appear, and this prediction error can lead to poor memory for the familiar context items (e.g., 'latent inhibition' (Lubow, 1989)) but strong memory for the novel stimuli because their predictive relationships with the environment are not yet known (Dayan et al., 2000). Stronger encoding of these novel items and their temporal context following prediction errors may facilitate their anticipatory prediction when the same familiar context is reencountered. The anticipatory reactivation of their memory trace may contribute to their being pruned from memory following a prediction error – that is, stronger predictions may lead to more pruning in unpredictable contexts.

On the other hand, encountering familiar items should generate more reliable predictions if those items are always followed by more predictable events. Reliable predictions should be formed even if the event details are novel and unpredictable, so long as some predictive information about them can be learned. Using existing semantic knowledge about the world (e.g., categorical information of visual objects), the brain can link conceptually related episodic events across time. For example, if familiar items are always followed by the same *category* of item, this should allow for more abstract, categorical-based statistical learning (Brady and Oliva, 2008). That is, over time the brain should adapt to the predictability (Brown and Braver, 2005; den Ouden et al., 2009) of these category exemplars and as a result build expectations about categories rather than items. When the familiar items are reencountered, the reactivation fidelity (i.e., the item-specificity) of the automatic predictions should be lower (because the predictions are category-based), and this may reduce the likelihood of pruning those items from memory.

Our episodic experiences are always changing, and details of precisely what will happen when we reencounter a familiar situation can be hard to predict. Yet, our experiences often contain hidden statistical structure that can be learned and generalized to new events. To date, memory pruning has been studied only in situations that consistently generate context-based prediction errors, without any predictable statistics across repeated exposures. However, it is also important to understand how memory is updated when higher-order predictability is embedded into our experiences.

Here, we tested whether the predictability of a familiar temporal context impacts the forgetting of previous episodic memories and the acquisition of new memories. We modified the paradigm of Kim et al. (Kim et al., 2014) to include multiple repetitions of particular items in a continuous sequence of visual object presentations (Figure 1A). FMRI data was collected during this incidental encoding task when observers made subcategory judgments about each object in the sequence, and memory for these items was tested later with a surprise item-recognition memory test. As a hidden rule, certain items ("cues") appeared four times across the experiment, and all other items appeared only once. Cue repetitions were always followed by novel items. To manipulate the reliability of predictions generated by the cues, half of the cues were followed by items from *different* categories across repetitions (e.g., *cue*-airplane, *cue*-taco, *cue*-horse, *cue*-pliers; "incongruent" condition), and the other half of cues were followed by items from the *same* category (e.g., all animals: *cue*-badger, *cue*-tiger, *cue*-cow, *cue*-peacock; "congruent" condition).

We applied category-based multivoxel pattern analysis (MVPA (Haxby et al., 2014; Haynes and Rees, 2006; Lewis-Peacock and Norman, 2014a; Norman et al., 2006)) to the fMRI data in high-order visual brain areas during this encoding task to quantify the perception of each stimulus, and to covertly measure the automatic prediction of items following the reappearance of familiar cues. These trial-by-trial neural measures were then linked to item-recognition performance on the subsequent memory test (Figure 1B). Memory pruning was evaluated at each cue repetition by assessing the relationship between neural evidence of automatic prediction of the previous item and its subsequent memory strength (stronger predictions leading to worse memory would be consistent with previous results on memory pruning (Kim et al., 2017, 2014)). We hypothesized there would be more evidence of memory pruning in the incongruent condition, when the expectation-violating items were unpredictable and didn't share features with the predicted items, as compared to the congruent condition, when the violating items were predictable at the category level. In addition, we hypothesize that there will be weaker encoding of new items in the congruent condition, relative to the incongruent condition, due to more reliable predictions (and reduced stimulus processing (Brady and Oliva, 2008)) afforded by conceptual-level statistical learning in these familiar temporal contexts.

RESULTS

Note that for all behavioral results, we report combined results (*N* = 46) from a group of fMRI participants (*n* = 22) and behavior-only participants (*n* = 24) who performed the same task outside the scanner (see Methods).

Encoding task performance. Participants were shown a continuous stream of images, one at a time, for the purpose of incidental encoding for a surprise subsequent memory task at the end of the experiment (Figure 1). As a cover task, participants were asked to make a subcategory judgment for each image. The category of the stimulus changed every trial, and therefore, participants were required to maintain their attention and constantly update their response mappings. Subcategory judgments were fast (*M* = 0.67 s, *SEM* = 0.01) and accurate (*M* = 0.87, *SEM* = 0.01; Figure 1A). Performance differed across conditions (incongruent, congruent), position (1st, 2nd, 3rd, 4th), and non-paired items for both trial types (cue, item) in both accuracy and RT (one-way ANOVA, omnibus *Fs* > 35, *Ps* < .001). Responses for cues were faster (0.62 s) and more accurate (0.94) than for non-cued items (0.70 s, 0.83) and non-paired items (0.70 s, 0.82; pairwise *t* test, all *Ps* < .0003, significant after Bonferroni correction), suggesting that repeated encoding enhanced subcategorization performance. Across repetitions, a simple linear regression (coefficient estimate: β_{lin}) shows that the accuracy to cues increased (all β_{lin} > 0.01, *Ps* < .01) and the RTs decreased (all β_{lin} < -0.01, *Ps* < .001), with no difference between conditions (condition x position ANOVA for cues, *Fs* < 3.84, *Ps* > .05). Categorization performance on the non-cued items did not change significantly across repetitions and conditions (*Ps* > .05).

There was a significant interaction of condition x trial type on both RT and accuracy (both *Ps* < .001), with no difference for cues, but with both faster and more accurate responses for non-cue items in the incongruent condition (0.71 s vs. 0.70 s, 0.81 vs. 0.85; both *Ps* < .001). This reflects greater alertness, and perhaps stronger encoding, following repeated cues in the incongruent condition. There was no three-way interaction of condition x position x trial type on either RT or accuracy. Overall the behavioral metrics on the encoding task indicate that participants were properly engaged in the task, and performance differences between the incongruent/congruent conditions demonstrate that they were sensitive to this manipulation.

Subsequent recognition memory. Memory for all items was tested in a surprise recognition test at the end of the experiment (Figure 1B). There was a statistical trend for an interaction of condition (incongruent/congruent) x position ($1^{st}/2^{nd}/3^{rd}/4^{th}$) on recognition accuracy (two-way ANOVA, *F*(3, 135) = 2.32, *P* = .078). In follow-up analyses, this interaction was significant across the first two item positions alone (two-way ANOVA, $F(1, 45) = 5.25$, $P = .027$), with worse memory for the 1st items ($M = 0.81$, $SEM = 0.01$) compared to the 2nd items ($M = 0.83$, $SEM = 0.01$) in the incongruent condition (pairwise t test, $t_{45} = -2.84$, $P = .007$, significant after Bonferroni correction). This interaction was not significant for the final two item positions ($F(1, 45) = 2.14$, $P = .15$), and the results trended in the opposite direction, with significantly worse memory for the 4th items ($M = 0.81$, $SEM = 0.01$) compared to the 3rd items ($M = 0.83$, $SEM = 0.01$) in the congruent condition ($t_{45} = 2.35$, $P = .023$). These results suggest that the 1st items in the incongruent condition may have been pruned, while this was not the case for the 1st items in the congruent condition. The key difference between these conditions is that, when the cue item repeated, automatic predictions for the 1st item had a greater degree of mismatch with the 2nd item (a novel item from a different category), but these predictions only mismatched at the exemplar level in the congruent condition (a novel item from the same category). Furthermore, we find evidence for decreased encoding for the 4th (and final) items in the congruent condition, consistent with the idea that in more predictable temporal contexts (e.g., when the category of the next item can be anticipated) the processing of new item-specific details is reduced.

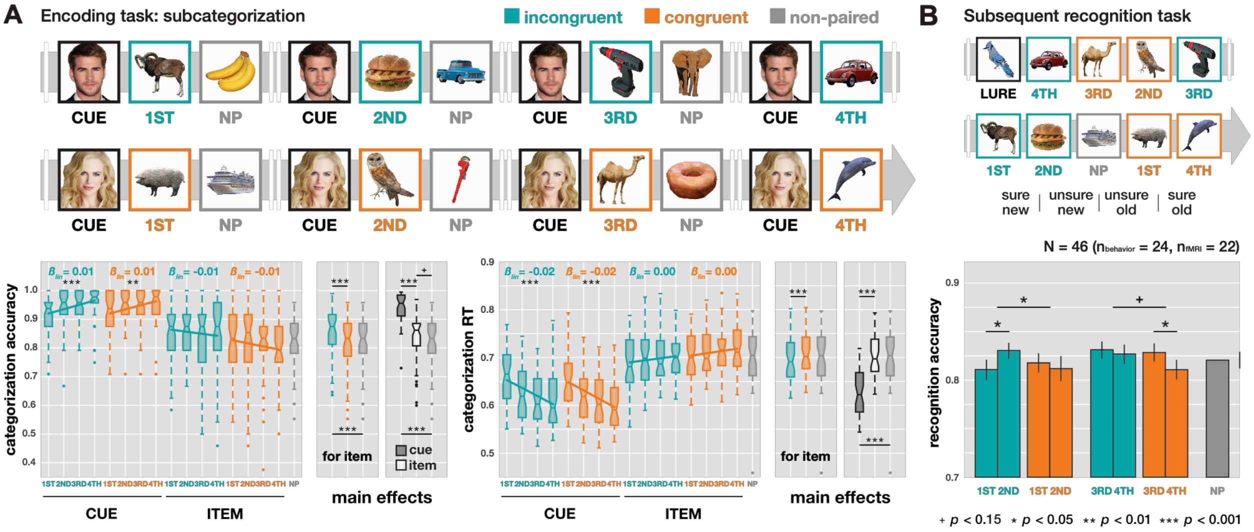

Figure 1. Experimental design and behavioral performance. (A) Incidental encoding task with categorization accuracy and RT. Subcategory judgments for each picture in a sequence (CUE: items that were presented four total times; 1st/2nd/3rd/4th ITEM: unique items that followed a cue in the specified order; NP: non-paired items that did not follow cues). The category for the cues was either face or scene, and there were four categories (animal, food, tool, and vehicle) for items and NP items. For main effects, Bonferroni adjusted alpha levels (α=.05/3) were applied per test. (B) Subsequent recognition task and recognition accuracy (A') for all non-cue items studied previously. Four options with old/new and sure/unsure were given for the response. Error bars represent SEM with a Bonferroni adjusted alpha level (α=.05/2).

Neural decoding results. All neural decoding was performed separately for each individual. Data from a functional localizer task was collected independently from the encoding task. Note that the stimuli used in the localizer task were separate from those used in the encoding task. The localizer consisted of a one-back working memory task with six categories of images (face, scene, animal, food, tool, and vehicle) and these data were used to train category-specific fMRI pattern classifiers (see Methods). Within the localizer data, we verified that brain activity patterns associated with processing each stimulus category were reliably differentiated in ventral temporal cortex (M = 0.58, SEM = 0.02, chance = 0.14 for 6 stimulus categories + rest; one-sample t test, Ps < .001, Figure 2-figure supplement 1B), using independent training and testing sets with cross-validation analysis. Decoding accuracy for the cue categories (face and scene) was reliably higher than for the other four categories (animal, food, tool, and vehicle; paired *t* test, t_{21} = -15.34, $P < .001$). Pattern classifiers were then re-trained on all data (2 runs) from the localizer task and applied to the encoding task to decode every timepoint in the experiment. The category of each object was reliably decoded during its presentation (*M* = 0.32, *SEM* = 0.01, chance = 0.14, *Ps* < .001, Figure 2-figure supplement 1C).

For every item that was viewed in the encoding task, we defined its “perception strength” as the amount of classifier evidence for that item during its presentation. There was no difference in perception strength across the four repetitions and two conditions (one-way ANOVA, *F*(7, 147) = 0.73, *P* = .672). For repeated cues, we also calculated the “prediction strength” corresponding to the item that followed the cue on the previous iteration (e.g., the prediction for “ram” during “Chris$_2$” in the sequence: Chris$_1$-ram, …, Chris$_2$-sandwich, … in Figure 1A) as the amount of classifier evidence for the category of that item during the repeated presentation of the cue (see Methods for further details). There was also no difference in prediction strength across the three repetitions (2nd– 4th) and two conditions (*F*(5, 105) = 0.40, *P* = .849). Evidence for such cue-based predictions comes from the observation that decoding accuracy for the perception of cues (which appear four times each) was *lower* compared to the decoding accuracy for the perception of non-cue items which appeared only once (paired *t* test, t_{21} = 3.24, $P <$ .01). This was true in both the incongruent (t_{21} = 2.75*,* $P < .05$) and congruent conditions (t_{21}= 2.95, $P < .01$). This relationship is a reversal from the results in the localizer data alone where the cue categories (faces, scenes) were decoded with greater accuracy than the other four categories (t_{21} = 3.24, $P < .01$). This reduction in decoding accuracy for the cues likely reflects the co-mingling of cue processing and automatic predictions of items from other categories triggered by the reappearance of the cue. This possibility will be addressed further in the Discussion. Next, both of these neural measures (perception strength and prediction strength) were linked to subsequent memory behavioral outcomes for each item that appeared in the sequence. The distributions of classifier evidence scores for both of these measures are shown in Figure 2C, showing that perception strength of new items was reliably higher than the prediction strength of the expected items (M = 0.69 vs. M = 0.52, P < .001).

Prediction and subsequent memory. To evaluate the maximum impact of the congruency of the context manipulation, we focused our attention first on results for the final (4th) repetition of the cues in each condition when participants would have had the most opportunity to stabilize their learning of the statistical relationships between cues and items in each condition. This allows us to directly evaluate our main hypothesis that the reliability of context-based predictions (which should be maximally divergent across the two conditions during the final repetition) impacts episodic memory. In the incongruent condition, there was a negative relationship between the prediction strength for the previous (3rd) items and their subsequent memory (logistic regression, $\beta = -0.74$, P = .019, bootstrap one-tailed, Figure 2A). Stronger predictions for these mispredicted items (i.e., 3rd items) led to worse subsequent memory for them. In the congruent condition, however, this relationship was not reliably different than zero ($\beta = -0.03$, P = .531, Figure 2B). There was a statistical trend that this relationship was more negative in the incongruent vs. congruent condition (P = .082).

Control analyses confirmed that these relationships were specific to classifier evidence for the target category, by subtracting the mean classifier evidence of the three non-target categories from the target classifier evidence scores (see (Detre et al., 2013; Kim et al., 2014)), and also by controlling for non-target category evidence using partial correlation (Figure 2-figure supplement 2A). To rule out the possibility that memory for these 3rd items was worse due to poor initial encoding, we also controlled for the perception strength of each item during encoding (Figure 2-figure supplement 2C), and the relationship between prediction strength and subsequent memory remained negative in the incongruent condition (β = −0.18, *P* = .028), and non-existent in the congruent condition (β = −0.01, *P* = .521), suggesting an item-specific pruning effect existed only in the incongruent condition.

Perception and subsequent memory. Focusing again on the final repetition of the cues, we evaluated the relationship between perception strength (measured neurally) of the final (4th) item in each set and the subsequent memory for those items measured with the surprise item-recognition test at the end of the experiment. If abstract memory structures were stabilized in the congruent contexts, the new items would have been also encoded in the abstract level at the expense of item details to facilitate learning efficiency. In the incongruent condition, there was a positive relationship between the perception strength of the 4th items and their subsequent memory (β = 1.24, *P* = .004, Figure 2A): stronger encoding of final items in this condition led to better subsequent memory for them. This relationship did not exist for the congruent condition (β = 0.39, *P* = .215, Figure 2B), and it was more positive in the incongruent vs. congruent condition (*P* = .046). The control analysis controlling non-target category evidence using partial correlation confirmed that these relationships were specific to the evidence for the target category (Figure 2-figure supplement 2B).

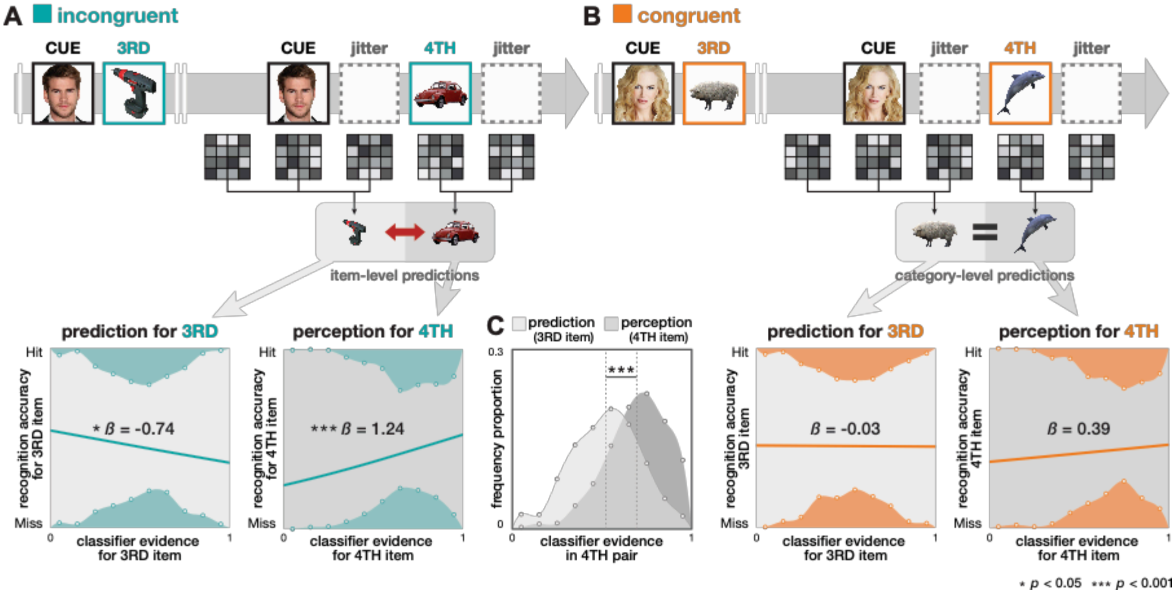

Figure 2. Predicting subsequent memory from brain activity during the final (4th) repetition of cues in the (A) incongruent and (B) congruent condition. Logistic regression results (coefficient estimate: β) linking classifier evidence and recognition accuracy are shown separately for 3rd items (prediction strength) and for 4th items (perception strength). Statistics are based on bootstrap analyses with 1,000 iterations.

Changes across repeated contexts. We replicated the analyses reported above, now also for each of the three prior appearances of the cues. The relationship between *prediction strength* and subsequent memory in the incongruent condition was consistently negative for all cue repetitions (2nd/3rd/4th, *Ps* < .05), with no differences between the repetitions (*Ps* > .05; Figure 3A*)*. There was no significant relationship for any repetition in the congruent condition. The positive relationship between *perception strength* and subsequent memory in the incongruent condition increased across repetitions (linear regression, β_{lin} = 0.30, *P* = .073) but was significant only in the 4th repetition (*P* = .004) and a statistical trend in the 3rd repetition (*P* = .051; Figure 3A). This is likely due to the fact that the 1st, 2nd, and 3rd items all had potential contributions to their subsequent memory strength from both their initial perception and their subsequent misprediction (and possible pruning) during the next appearance of the cue. This could obscure the relationship between perception and memory for these earlier items. However, the 4th items only had contributions from their perception: there was no further appearance of the cue and thus no opportunity to mispredict the 4th item. Stronger perception strength for these items was, intuitively, associated with better subsequent memory for these items.

In the congruent condition, this positive relationship decreased across repetitions (β_{lin} = −0.51, P = .037) and was statistically significant only in the first two repetitions (1st and 2nd, *Ps* < .01). There was a significant interaction of condition and repetition on this relationship (P = .002). The link between perception strength and subsequent memory for these first two repetitions were significantly different between the conditions. We suggest that the influence of perception remained strong in the congruent condition because, unlike in the incongruent condition, these items were not specifically predicted and pruned upon subsequent appearances of the cues. Rather, predictions in the congruent condition may have become categorical in nature as contexts were repeated. This is supported by the prediction-to-memory results in Figure 2B, showing that prediction strength was unrelated to item-specific memory recognition performance in the congruent condition.

Changes within a repeated context. To examine when prediction information emerged during a cue presentation, classifier evidence was divided into three periods on each trial: *baseline* (3 s before the cue), *cue* (3-5 s during the cue), and *item* (3-5 s during the item that followed the cue). All regressors were shifted 4 s to account for hemodynamic lag, and data were combined across repetitions. Results are shown separately for each repetition in Figure 3-figure supplement 1. As expected, there was no relationship between prediction strength and subsequent memory in the baseline period of either condition (Figure 3B). In the incongruent condition, this relationship was consistently negative, and more negative than in the baseline period, for both the cue and item periods (*P*s < .025). In contrast, there was no relationship, and no difference from baseline, in the congruent condition (*P*s > .119, Figure 3B).

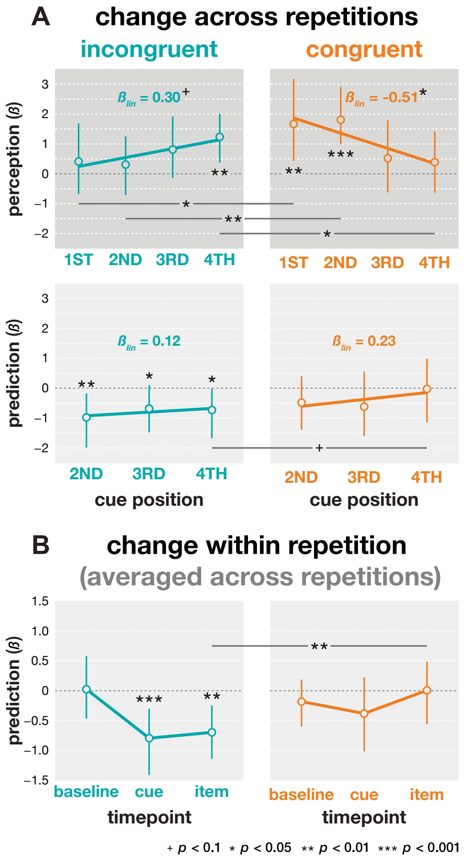

Figure 3. Changes across time in the links between perception, prediction, and subsequent memory. (A) (Top) The relationship between perception strength and subsequent memory across all four cue repetitions in both the incongruent and congruent conditions. (Bottom) The relationship between prediction strength and subsequent memory across the 2^{nd}, 3^{rd}, and 4^{th} repetitions of each cue (when prediction was possible) in each condition. (B) The relationship between prediction strength and subsequent memory is shown for different time windows *within* a cue presentation (baseline, cue, item), and averaged across the final three repetitions of each cue. Statistics are based on bootstrap analysis with 1,000 iterations. Error bars represent 95% CI.

DISCUSSION

This study demonstrates how encountering familiar situations in the visual world influences how we remember past events and how we process new experiences. Specifically, we show that when the temporal statistics of our experience allow for the learning of abstract, conceptual relationships between events, this changes our expectations of future events to be more abstract and less focused on item-specific details. Abstract predictions appear to have two main consequences: (1) memory for past events is better preserved (existing memories of specific items do not get "pruned" during prediction errors), and (2) memory for new events is worse because encoding is less focused on item-specific details.

We hypothesized that incongruent contexts would result in the pruning of memories for past events experienced in that context, whereas congruent contexts would not. Consistent with this hypothesis, we found behavioral evidence of pruning in the incongruent condition, but not in the congruent condition, for the 1st items that followed a cue. Moreover, neural measures of prediction strength for *each* item in the incongruent condition were associated with worse subsequent memory for those items. This replicates the memory pruning effect observed previously (Kim et al., 2014) that describes a form of error-driven statistical learning in which the memory trace for a mispredicted event is weakened, which leads to subsequent forgetting of that event. For cues in the congruent condition, i.e., cues that were always followed by new exemplars from the same semantic category, no memory pruning was observed in the brain-behavior relationships of those items. However, encoding of the details of new items at the end of the cue sequences was impaired compared to the incongruent condition, leading to worse overall memory for these items and no reliable relationship between brain measures of perception strength and subsequent memory. Note that there were no differences on category evidence for both prediction and perception across condition and position, suggesting that the effects we found derived from differences in predictions to which our neural measures were insensitive (e.g. a prediction of a specific item from category A vs. an abstract expectation about category A). Together these results show that the predictability of events impacts how episodic memories are updated and how new memories are formed.

Evidence for automatic predictions of items and categories. In our incidental encoding task, repeated cues in the *incongruent* condition were always followed by new items from new categories. Participants could not anticipate which item would follow a cue when it repeated, nonetheless, there was evidence for an automatic prediction for the previous item that followed the cue each time it repeated. Moreover, stronger predictions for previous items (none of which would ever reappear) were associated with weaker subsequent memory for those items (Figure 2A). This suggests that, in the incongruent condition, participants were generating *item-specific* predictions upon each cue reappearance, which in turn led to pruning of the item-specific episodic memories.

Additional evidence for context-based predictions comes from the observation of *reduced* classifier decoding accuracy for the perception of the repeated cues compared to the single-exposure items (Figure 2-figure supplement 1C). Notably, this is the opposite relationship of classifier performance from within the localizer data alone, where the cue categories (face, scene) were decoded *better* than the other categories (animal, food, tool, and vehicle; Figure 2-figure supplement 2B). The elimination of the decoding advantage for the repeated cues may have arisen from two sources: first, from reduced processing of the now-familiar cues across repeated presentations, and second, from the co-mingling of cue processing with an automatically triggered prediction for an item from another category, which would dilute the measurement of cue-specific neural activation.

The argument for item-specific predictions is further supported by the observation of a strong relationship between perception strength and memory for the final items in the incongruent conditions. These items were remembered well, and their perception strength was directly related to their memory strength. This direct link between perception strength and memory suggests that individual item details were being encoded, which should in turn facilitate predictions of these items upon the next appearance of the cue. Note that it is likely that the encoding and subsequent prediction for these items would decrease if and when it was learned that these predictions were always violated, but this did not seem to occur in this study after only four repetitions.

On the other hand, in the *congruent* condition, cues were always followed by new items from a single category (e.g., a *cue* was always followed by an animal: *cue*-badger, *cue*-tiger, *cue*-cow, *cue*-peacock). It is possible that participants could learn this relationship for each cue and make explicit predictions at the category level. However, post-experiment questionnaires confirmed that, aside from noticing repeated presentation of the cues, participants did not detect any structure in the order of any of the stimulus presentations. Any differences in context-based predictions between the experimental conditions would therefore be due to implicit learning of the transition probabilities associated with the different cues.

For each repetition of the congruent cues, there was no relationship between the neural evidence for prediction and subsequent memory for the previous (3rd) item (Figure 3A). Also, subsequent memory for the final (4th) item in congruent sets was poor and unrelated to neural perception strength for those items. Unlike in the incongruent condition, these data do not support an inference that participants were making item-specific predictions for these cues. Rather, they suggest that implicit predictions in the congruent condition were made at the *category* level rather than at the *item* level (e.g., “expect some animal” instead of “expect that cow”). These results reflect the consequences of statistical learning focused not on item-specific details, but rather on abstract conceptual information (Brady and Oliva, 2008). Across multiple experiences, overlapping features (i.e., the semantic category) of individual events were extracted to form more generalized knowledge about these specific situations, similar to the formation of memory schemas (O’Reilly and Norman, 2002; Tse et al., 2007; van Kesteren et al., 2016, 2012). Once this general inference was developed, only category-level information was encoded and the specific episodic details of new items were forgotten (van Kesteren et al., 2016, 2012).

Our main analyses relied on category-specific fMRI pattern classifiers to covertly measure implicit predictions of previous items from a repeated context. Category classifiers can produce more robust decoding than sub-category classifiers or item-level classifiers, but of course they lack item specificity. We decided against proceeding with an item-level decoding approach due to insufficient decoding accuracy in early pilot data. Instead we used category information as a proxy for item information, and therefore we could not distinguish predictions for individual exemplars of a category from generic category predictions. We must therefore rely on the relationship (or lack thereof) between these category-specific neural estimates and the item-specific behavioral measures for each stimulus to speculate on the nature of these predictions. Future work should use neural analyses sensitive to item-level representations (e.g., representational similarity analysis (Kriegeskorte et al., 2008)) to more directly test this idea.

Memory pruning. Results for the incongruent condition in the present study, with cues followed by novel items from novel categories, were consistent with previous findings on memory pruning [19] which found a negative relationship between neural prediction strength for mispredicted items in a temporal sequence and subsequent memory for those items. Memory pruning is a form of error-driven learning that is consistent with predictions of the non-monotonic plasticity hypothesis (NMPH) (Detre et al., 2013; Lewis-Peacock and Norman, 2014b; Newman and Norman, 2010) which claims that moderately activated memories can lead to weakening and subsequent forgetting of those memories. Here, moderately active memories were created by the automatic context-based predictions that occurred during the incidental encoding task. In Figure 2A, the *prediction strength* for the 3rd item following a repeated cue is contrasted with the *perception strength* for the 4th item that actually appeared. The distributions of classifier evidence values show that perception strength (M = 0.69) was reliably higher than prediction strength (M = 0.52, P < .001, Figure 2C). Taking classifier evidence as an index of the strength of “memory activation”, we see that prediction leads to more moderately active representations (compared to perception), and the

NMPH predicts that these memory representations would be more vulnerable to weakening and long-term forgetting.

According to this framework, predictions of a specific event are realized by anticipatory activation of its memory representation, resulting in relatively weak activation of its memory trace (compared to activation during the initial perception of the event). If this prediction is confirmed, its neural activation will increase, as will the strength of its representation in long-term memory, due to additional processing of the event. If the prediction is violated, its reactivated memory representation may persist in this moderately activated state which, according to the hypothesis, can trigger its weakening and forgetting. In temporal contexts that allow for more abstract statistical learning, the predictions may not contain representations of specific events from the past (e.g., “some animal” might be expected but not “that brown cow”). The lack of specificity in these predictions may have the effect of shielding the memories of those specific events from modification.

These results are consistent with recent work in psycholinguistics. For example, the memory pruning of prior cue associates observed in our incongruent condition is consistent with work by Oppenheim and colleagues (2010) (Oppenheim et al., 2010), who describe a *cumulative semantic interference* effect in which retrieving a word can impair the retrieval of other words from the same semantic category. They argue that the ‘negative’ impact on related words and the ‘positive’ effect of repetition priming for the target words are two sides of the same coin, both resulting from error-driven implicit learning processes. Similar to the NMPH framework described above, the mechanism by which this is accomplished is modeled as the inhibition of weakly activated competing words during the gradual learning of the associations between concepts and words.

Memory pruning vs. memory integration. It could be argued that cues in the incongruent condition should trigger memory *integration* (Greve et al., 2018; Morton et al., 2017; Preston and Eichenbaum, 2013; Schlichting et al., 2015; Schlichting and Frankland, 2017; Schlichting and Preston, 2015; Zeithamova et al., 2012a, 2012b) rather than memory pruning, such that all items maintain their associations to the repeated context, similar to inferential learning (Morton et al., 2017; Preston and Eichenbaum, 2013; Schlichting et al., 2015; Schlichting and Frankland, 2017; Schlichting and Preston, 2015; Zeithamova et al., 2012a, 2012b). Recent evidence suggested that prediction error weakens overlapping representation between the mispredicted item and its context, leading to differentiation of their neural patterns in the hippocampus (Kim et al., 2017). However, the same neural consequences have also be observed for memory integration (Zeithamova et al., 2012a). The hippocampus has a critical role, not only for memory integration (Eichenbaum, 2000; van Kesteren et al., 2016) but also in mismatch detection (Kumaran and Maguire, 2007; Long et al., 2016) for prediction errors. In a recent study by Long and colleagues (Long et al., 2016), the activation of the hippocampus was found to be positively correlated with prediction errors, and even more so if the mispredicted item was semantically related to the actual item. (This is similar to the congruent condition in the present study.) The hippocampus was not recruited when predictions were correct or unrelated semantically to the novel events. This suggests that mismatch signals are key for triggering updating of existing memories (Kumaran and Maguire, 2007; Long et al., 2016; Schlichting and Preston, 2015). In our data, however, we were unable to find any relationship between prediction strength and hippocampal activation, or any evidence of hippocampal involvement when mismatched predictions were semantically related to the new events (i.e., in the congruent condition). Unlike Long and colleagues' study or inferential learning studies (Schlichting et al., 2015; Zeithamova et al., 2012b) in which the participants explicitly learned word-picture or picture-picture associations in the pre-training phase, the cue-item associations were implicitly learned in our study. Explicit predictions based on over-learned associations might be too strong to trigger pruning of existing memories, but rather may promote integration of the new semantically related items (Mortan et al., 2018; Schlichting et al., 2015).

Reduced encoding in predictable contexts. In our study, the congruent condition involved repeated visual cues that were consistently followed by items from a single category. Results suggest that participants implicitly learned these relationships, as both behavioral evidence and neural evidence in this condition diverged from the incongruent condition in which the cues were always followed by a new item from an unpredictable category. Specifically, memory for the final item in the congruent sets was worse than previous items in the set (Figure 1C), and there was no relationship between neural evidence of perception for these items and their subsequent memory strength (Figure 2B). Together these results suggest that in temporal contexts with greater predictability (e.g., when the category of the next item can be anticipated), encoding of new items is reduced (but see (Friedman, 1979; Gronau and Shachar, 2015; Marlieke

T. R. van Kesteren et al., 2010; Zwaan and Radvansky, 1998)). Consistent with this idea, Rommers and Federmeier (2018) (Rommers and Federmeier, 2018) recently demonstrated evidence that predictable information is processed more weakly than unpredictable information. When words reappeared in predictable contexts, the neural responses measured using electroencephalography (EEG) indicated that the words were processed less than repeated words in unpredictable contexts. Specifically, the repetition priming effects were diminished in the N400 and LPC components of the EEG signal. The authors suggest that predictability allowed the brain to operate in top-down "verification mode" at the expense of detailed stimulus processing.

Conclusions. The learning processes observed in this study are examples of adaptive forgetting that allow for the efficient use of the brain's memory systems (Kim et al., 2017, 2014; Lewis-Peacock and Norman, 2014b; Wylie et al., 2008). Being able to anticipate the demands required of us in familiar situations can help us to respond more effectively and proactively. Pruning unreliable memories, via statistical learning, supports this behavior by reducing interference during context-based retrieval of relevant memories (Kumaran and Maguire, 2007). Here, we demonstrated that the stability of episodic memories is evaluated over multiple exposures to the context in which those memories were acquired. When a context afforded no accurate predictions, previous experiences were nonetheless anticipated, perhaps reflecting a persistent, but futile, effort to learn the statistics of the environment. Memory for these experiences was pruned when their predictions were violated. When a context afforded general information (but not specific details) about what to expect, the previously encountered events were no longer predicted, and their memories were shielded from pruning.

However, new learning was also diminished in these more stable contexts. These findings deepen our understanding of how episodic memories are formed and updated by demonstrating how our ability to predict the future influences how we remember the past.

METHODS

Participants. Thirty healthy young adults (13 male; age, *M* = 22 yr, *SD* = 3.48, all right-handed) were recruited from the student body and campus community of the University of Texas at Austin to participate in the neuroimaging experiment. Five participants were excluded due to low classifier accuracy in the localizer task (5 *SEM* below the mean), and three participants were excluded due to low recognition accuracy (10 SEM below the mean), resulting in final sample size of *n*=22. Twenty-four additional participants (13 male; age, *M* = 23.29y, *SD* = 4.81, left-handed = 1) were recruited for a behavior-only version of the experiment. All participants had normal or corrected-to-normal vision. The study was approved by the University of Texas at Austin Institutional Review Board and informed consent was obtained from all participants.

Stimuli. Colored pictures of common objects were used for this experiment. They were selected from six categories (with two subcategories each): famous faces (female/male), famous scenes (manmade/natural), animals (land/non-land), food (cooked/uncooked), tools (power/non-power), and vehicles (land/non-land). Object images were obtained from various resources (Morton et al., 2017) including Bank of standardized stimuli (Brodeur et al., 2014), and Google Images.

Procedure. The experiment proceeded in three tasks: incidental encoding, functional localizer, and subsequent recognition memory test. fMRI brain data was acquired for the encoding task (6 runs, 335s/run) and the localizer task (2 runs, 513s/run) (*N* = 22, Figure 2-figure supplement 1A). Participants performed the subsequent recognition memory test either after the localizer (outside the scanner, *N* = 14) or before the localizer (in the scanner, *N* = 8). For behavior-only participants (*N* = 24), encoding and recognition phases were conducted sequentially. There was a significant condition (incongruent/congruent) x position ($1^{st}/2^{nd}/3^{rd}/4^{th}$) interaction on recognition accuracy for the post-scan participants ($F(3, 39) = 3.13$, $P = .036$) but not for the collapsed ($F(3, 135) = 2.32$, $P = .078$), during-scan ($F(3, 21) = 0.40$, $P = .751$), and behavioral-only participants ($F(3, 69) = 2.16$, $P = .101$).

Incidental encoding task: Participants were shown a steady stream of images, one at a time, for the purpose of incidental encoding for a surprise subsequent memory task at the end of the experiment. In the stream, there were hidden sequences consisting of cue-item pairs. Each cue was associated with four different items, which made four cue-item pairs as one set. For half of the cues, all items were selected from a single category (*congruent condition*). For the other half of the cues, the items were selected from a new category each time (*incongruent condition*). There were 24 sets (96 total cue-item pairs) for each condition, and 96 *non-paired* items that were not part of a set and never directly followed a cue. The pairs from a given set were not adjacent but appeared intermingled with other sets and non-paired items (mean lag = 8 trials). All cues appeared four times, and all other items appeared only once. In each run (80 trials), there were four sets for both conditions and 16 non-paired items. Across all six runs (480 trials total), the categories (e.g., animal, food, etc.) and subcategories (e.g., land/non-land, cooked/uncooked, etc.) of items and non-paired items were counter-balanced.

As a cover task, participants were asked to make a subcategory judgment for each image using one of two buttons on a 4-button box (in the scanner) or on a keyboard (outside the scanner). The category of the stimulus changed every trial, and therefore participants were required to constantly update their response mappings. To facilitate performance, we provided the two subcategory options for each stimulus (e.g., female/male for faces). On a trial, the stimulus displayed for 1 s on a white background box (visual angle: 21.8° x 21.8°), with empty feedback circles and text underneath the image displaying the subcategory choices, during which participants had to make a response. When the stimulus disappeared, a blank white box remained with feedback circles underneath, in which one of the circles was colored for 1 s based on performance (green: correct, red: incorrect, yellow: missed). The inter-trial interval was pseudo-randomly jittered at 2, 3, or 4 s.

Either faces or scenes, but not both, were used as cue stimuli for each participant (*N* = 13/22 fMRI, and *N* = 12/24 behavioral participants had face cues). The non-selected category was not used for the encoding task for that participant. These two categories were chosen as cues based on their superior classification accuracy in ventral temporal cortex (face, scene; *M* = 0.78, *SE* = 0.03), relative to the other four categories (*M* = 0.47, *SEM* = 0.03), from a separate pilot sample (*N* = 3) on the localizer task (see Figure 2-figure supplement 1B). We chose famous people and famous places to facilitate recognition of the cues, which in turn should facilitate the generation of context-based predictions when the cues repeated. The other four categories (animals, food, tools, vehicles) were used for the stimuli that appeared (only once) as items following a cue or as non-paired items. Participants practiced the task before scanning with a separate set of images until they reached a criterion of 80% accuracy for the subcategory judgment task. Categorization performance was calculated with accuracy and RT of the responses, and a simple linear regression was applied to track the performance changes across repetitions for trial type and condition.

Subsequent recognition memory test: In this phase, the participants were given a surprise memory test for the objects that they saw in the encoding task. All objects used for non-cue items (288 old; 96 items for each incongruent, congruent, and non-paired condition) and 96 novel lures were tested in a random order. Participants made a recognition judgment using a 4-point scale: 1 = sure new, 2 = unsure new, 3 = unsure old, and 4 = sure old. Only “sure old” responses were treated as hits (Kim et al., 2014; Lewis-Peacock and Norman, 2014b), and we calculated memory sensitivity using A-prime (*A*’) (Stanislaw and Todorov, 1999). A subset of participants (*N* = 8/22) took the memory test right after the encoding task in the scanner prior to the localizer phase to minimize any possible memory interference from stimuli in the localizer. However, there was no observed impact of task order on memory performance (*F*(1, 20) = 1.137, *P* = .299). For assessing statistical reliability of the subsequent memory results, we combined data from the behavioral and fMRI groups (*N* = 46 total).

Functional localizer: Participants performed a one-back task with six categories of images: face, scene, animal, food, tool, vehicle. These stimuli were unique to the localizer and were never shown again. Each image was presented for 1.5 s on a white background box followed by an inter-trial interval for 0.5 s in which only the white background box remained on the screen. Stimulus display parameters were similar to the encoding task. However, rather than making a subcategory judgment, participants responded “same” if the object matched the previous object or “different” otherwise (on average, there was 1 repeat every 5 trials). Responses were to be made within 1 s, and visual feedback was given using the color of the frame of the background box (green: correct, red: incorrect) immediately after the response, or after stimulus offset if no response was made. Stimuli were blocked by category with 10 trials per mini-block, lasting 20 s, and 6 mini-blocks (6 categories x alternate subcategory across blocks) per block, followed by 6 s of blank inter-block interval. There were two fMRI runs of the localizer task, each with 4 blocks (24 mini-blocks) presented in randomized order. Fifteen seconds were added to the end of each run to account for hemodynamic delay on the last trial. To verify the accuracy of the classifier, the one-sample *t* test was conducted for each category.

Data acquisition. The Psychophysics Toolbox (http://psychtoolbox.org) was used to run experiments. Neuroimaging data were acquired on a 3.0-T Siemems Skyra MRI (Siemens, Erlangen, Germany) with a 64-channel head coil. High-resolution anatomical images were collected for registration from a T1-weighted 3-D MPRAGE volume (256 × 256 × 192 matrix, 1 mm^3 voxels). A gradient-echo echo planar imaging sequence was applied for functional images with following parameters: TR = 1 s, multiband factor = 4, TE = 30 ms, 63° flip, 96 × 96 x 64 matrix, 2.4 mm^3 voxels, 56 slices, no gap.

Preprocessing. FSL (http://fsl.fmrib.ox.ac.uk) was used to preprocess the fMRI data. Functional volumes were corrected for motion, aligned to the mean volume of the middle run, detrended and temporal high-pass filtered (128s). Timepoints with excessive motion were removed (framewise displacement, threshold = 0.9 mm (Power et al., 2014); *M* = 6.7 TRs removed, *SD* = 14.5).

Region-of-interest. Bilateral ventral temporal cortex (Grill-Spector and Weiner, 2014; Haxby et al., 2001) (VTC) was anatomically delineated as a region of interest (ROI). The ROI in standard space was generated by combining bilateral temporal occipital fusiform cortex and posterior parahippocampal gyrus from the Harvard–Oxford cortical atlas. This was converted to native space and resampled to functional resolution for each subject using the transformation matrix derived from registration of that subject's data to standard space. This size of this ROI across subjects was *M* = 4,266 voxels, *SD* = 283.

Classification Analyses. The Princeton Multi-Voxel Pattern Analysis Toolbox (www.pni.princeton.edu/mvpa) was used for multivoxel pattern classification using L2-regularized, non-multinomial (one-vs-others, for each category) logistic regression. Classifiers were trained separately for each participant using localizer data in bilateral ventral temporal cortex. Regressors for all seven categories (face, scene, animal, food, tool, vehicle, rest) were shifted by 4 s to adjust for hemodynamic lag. To validate classifier performance, cross-validation was performed across the two runs of localizer data. This was done 22 times with different penalties (from 0 to 1000) to find the optimal penalty for each participant (*M* = 156, *SD* = 244). Prior to classification, feature selection was performed for each training set using a voxel-wise ANOVA across all categories and timepoints (threshold: *P* = 0.0001). The selected voxels were used to train and test the classifier (19.2% of the original voxels; *M* = 820 voxels, *SD* = 237). Across subjects, classifier performance was reliably above chance for each category (*M* = 0.58, *SEM* = 0.02, chance level = 0.14, Figure 2-figure supplement 1B). Five subjects were excluded from further analyses due to low classifier accuracy (5 *SEM* below the mean). Data from both localizer runs were then used to re-train the classifiers which were then applied to data from the encoding task. This produced classifier evidence scores (from 0 to 1) for each category at every timepoint in the encoding task. These scores reflect the likelihood that a test sample of brain activity contains a representation of a given category. The same individualized penalty derived from the cross-validation of the localizer data was used, and a new feature-selected mask was computed (26.6% of the original voxels; *M* = 1136, *SD* = 264). The *perception strength* of each object during the encoding task was defined as the average classifier evidence for the object's category from its onset until the onset of the next stimulus (i.e., 1 s of display plus 2, 3, or 4 s of inter-trial interval, depending on jitter, shifted forward 4 s to account for hemodynamic lag; *prediction window*). On cue repetitions, the *prediction strength* for the item that previously followed the cue was defined as the average classifier evidence for that item's category during the perception time window for the cue (*perception window*). Note that to minimize influence from the onset of the next item, but to keep the window size equal to the perception window, the prediction window was actually shifted 1 s *backward* prior to then being shifted 4 s forward (net: 3 s forward) to account for hemodynamic lag.

In an attempt to improve decoding sensitivity for predictions in this fast event-related design, we modeled category-level beta estimates for the perception of each stimulus in the encoding task (General Linear Model, GLM, utilizing a hemodynamic response function) to remove the evoked activity from cue presentations. Then we modeled trial-specific beta estimates for the predictions from the residuals of this analysis (GLM including a regressor for that trial + another regressor for other trials) (Mumford et al., 2012). We applied the same classifiers trained on the localizer data with shifted regressors to decode the “prediction betas” from the encoding task. The results obtained from this analysis were qualitatively similar to the results obtained using the unmodeled, shifted regressors, which we chose to report here.

Linking neural data to behavior. Binary logistic regression analysis was used to examine the impact on subsequent memory from prediction strength and perception strength during the encoding task. The result of each of these analyses is a coefficient estimate (*β*) of the relationship between the given neural evidence and the memory outcomes. To increase our ability to detect trial-specific effects, we pooled data from all subjects and then performed bootstrap resampling to evaluate the population reliability of the result (Efron, 1979). On each bootstrap iteration (of 1,000 total), we sampled randomly (with replacement) a collection of participants data to match the size of our experimental sample (*N* = 22). There were eight regressions conducted for perception strength and subsequent memory (incongruent/congruent x 1st/2nd/3rd/4th positions), and six regressions conducted for prediction strength and subsequent memory (incongruent/congruent x 2nd/3rd/4th positions). Statistical significance was calculated with a non-parametric test across bootstrap iterations, evaluating the stability of an effect of interest by calculating the proportion of iterations in which the effect was found. Lastly, to verify that the effects were not arising from variance across participants but from within-subject variance, we repeated the main analyses using standardized (z-scored) classifier evidence for each participant (Kim et al., 2014). Results from the main analyses were qualitatively similar and confirmed. Across repetitions, linear regression analyses were conducted on the binary logistic regression results (β) for each condition and process (prediction/perception), and F-test on the model was calculated for the statistical significance test.

Mismatch signals in hippocampus. Linear regression analyses were applied to link mismatch signals in the hippocampus (Long et al., 2016) and prediction strength decoded from the ventral temporal cortex. The mismatch signal was defined as the average intensity of the signal during the *perception window*, and prediction strength was defined as the average classifier evidence during the *prediction window*. There was no reliable relationship between mismatch signals and prediction strength for either incongruent or congruent conditions (*Ps* > .05).

REFERENCES

Brady TF, Oliva A. 2008. Statistical learning using real-world scenes: extracting categorical regularities without conscious intent. *Psychol Sci* **19**:678–685. doi:10.1111/j.1467-9280.2008.02142.x

Brodeur MB, Guérard K, Bouras M. 2014. Bank of standardized stimuli (BOSS) phase ii: 930 new normative photos. *PLoS One* **9**:e106953. doi:10.1371/journal.pone.0106953

Brown JW, Braver TS. 2005. Learned predictions of error likelihood in the anterior cingulate cortex. *Science (80-)* **307**:1118–1121. doi:10.1126/science.1105783

Chun MM, Jiang Y. 1998. Contextual Cueing: Implicit Learning and Memory of Visual Context Guides Spatial Attention,. *Cogn Psychol* **36**:28–71. doi:10.1006/cogp.1998.0681

Conway CM, Christiansen MH. 2006. Statistical learning within and between modalities: pitting abstract against stimulus-specific representations. *Psychol Sci* **17**:905–912. doi:10.1111/j.1467-9280.2006.01801.x

Dayan P, Kakade S, Montague PR. 2000. Learning and selective attention. *Nat Neurosci* **3**:1218–1223. doi:10.1038/81504

den Ouden HEM, Friston KJ, Daw ND, McIntosh AR, Stephan KE. 2009. A dual role for prediction error in associative learning. *Cereb Cortex* **19**:1175–1185. doi:10.1093/cercor/bhn161

Detre GJ, Natarajan A, Gershman SJ, Norman K a. 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

Efron B. 1979. Bootstrap methods: another look at the jackknife. *Ann Stat* **7**:1–26. doi:10.1214/aos/1176344552

Eichenbaum H. 2000. A cortical-hippocampal system for declarative memory. *Nat Rev Neurosci* **1**:41–50. doi:10.1038/35036213

Fiser J, Aslin RN. 2002. Statistical learning of higher-order temporal structure from visual shape sequences. *J Exp Psychol Learn Mem Cogn* **28**:458–467. doi:10.1037/0278-7393.28.3.458

Fiser J, Aslin RN. 2001. Unsupervised statistical learning of higher order spatial structures from visual scenes. *Psychol Sci* **12**:499–504. doi:10.1037//0278-7393.28.3.458

Fiser J, Scholl BJ, Aslin RN. 2007. Perceived object trajectories during occlusion constrain visual statistical learning. *Psychon Bull Rev* **14**:173–178.

Friedman A. 1979. Framing pictures: the role of knowledge in automatized encoding and memory for gist. *J Exp Psychol Gen* **108**:316–355. doi:10.1037/0096-3445.108.3.316

Ghosh VE, Gilboa A. 2014. What is a memory schema? A historical perspective on current neuroscience literature. *Neuropsychologia* **53**:104–114. doi:10.1016/j.neuropsychologia.2013.11.010

Greve A, Abdulrahman H, Henson RN. 2018. Opinion: Neural differentiation of incorrectly predicted memories. *Front Hum Neurosci* **12**. doi:doi: 10.3389/fnhum.2018.00278

Grill-Spector K, Weiner KS. 2014. The functional architecture of the ventral temporal cortex and its role in categorization. *Nat Rev Neurosci* **15**:536–548. doi:10.1038/nrn3747

Gronau N, Shachar M. 2015. Contextual consistency facilitates long-term memory of perceptual detail in barely seen images. *J Exp Psychol Hum Percept Perform* **41**:1095–1111. doi:10.1037/xhp0000071

Haxby J V., Connolly AC, Guntupalli JS. 2014. Decoding neural representational spaces using multivariate pattern analysis. *Annu Rev Neurosci* **37**:435–456. doi:10.1146/annurev-neuro-062012-170325

Haxby J V, Gobbini MI, Furey ML, Ishai A, Schouten JL, Pietrini P. 2001. Distrubuted and overlapping representations of face and objects in ventral temporal cortex. *Science (80-)* **293**:2425–2430. doi:10.1126/science.1063736

Haynes J-D, Rees G. 2006. Decoding mental states from brain activity in humans. *Nat Rev Neurosci* **7**:523–534. doi:10.1038/nrn1931

Kim G, Lewis-Peacock JA, Norman KA, Turk-Browne NB. 2014. Pruning of memories by context-based prediction error. *Proc Natl Acad Sci U S A* **111**:8997–9002. 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:http://dx.doi.org/10.1101/083022

Kóbor A, Janacsek K, Takács Á, Nemeth D. 2017. Statistical learning leads to persistent memory: evidence for one-year consolidation. *Sci Rep* **7**:760. doi:10.1038/s41598-017-00807-3

Kriegeskorte N, Mur M, Bandettini P. 2008. Representational similarity analysis -

connecting the branches of systems neuroscience. *Front Syst Neurosci* **2**:4. doi:10.3389/neuro.06.004.2008

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

Lewis-Peacock JA, Norman KA. 2014a. Multi-voxel pattern analysis of fMRI data, 5th ed, The Cognitive Neurosciences. MIT Press.

Lewis-Peacock JA, Norman KA. 2014b. Competition between items in working memory leads to forgetting. *Nat Commun* **5**:5768. doi:10.1038/ncomms6768

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**:12677–12687. doi:10.1523/JNEUROSCI.1850-16.2016

Lubow RE. 1989. Latent inhibition and conditioned attention theory. New York: Cambridge University Press.

Mortan N, Zippi E, Preston A. 2018. Merging memories: Reactivation of individual event elements during learning supports memory integration. *Prep*.

Morton NW, Sherrill KR, Preston AR. 2017. Memory integration constructs maps of space, time, and concepts. *Curr Opin Behav Sci* **17**:161–168. doi:10.1016/j.cobeha.2017.08.007

Mumford JA, Turner BO, Ashby FG, Poldrack RA. 2012. Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. *Neuroimage* **59**:2636–2643. doi:10.1016/j.neuroimage.2011.08.076

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

Norman K a., Polyn SM, Detre GJ, Haxby J V. 2006. Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends Cogn Sci* **10**:424–430. doi:10.1016/j.tics.2006.07.005

O'Reilly RC, Norman KA. 2002. Hippocampal and neocortical contributions to memory: advances in the complementary learning systems framework. *Trends Cogn Sci* **6**:505–510.

Oppenheim GM, Dell GS, Schwartz MF. 2010. The dark side of incremental learning: A model of cumulative semantic interference during lexical access in speech production. *Cognition* **114**:227–252. doi:10.1016/j.cognition.2009.09.007

Pagnoni G, Zink CF, Montague PR, Berns GS. 2002. Activity in human ventral striatum locked to errors of reward prediction. *Nat Neurosci* **5**:97–98. doi:10.1038/nn802

Perruchet P, Pacton S. 2006. Implicit learning and statistical learning: one phenomenon, two approaches. *Trends Cogn Sci* **10**:233–238. doi:10.1016/j.tics.2006.03.006

Power JD, Mitra A, Laumann TO, Snyder AZ, Schlaggar BL, Petersen SE. 2014. Methods to detect, characterize, and remove motion artifact in resting state fMRI. *Neuroimage* **84**:320–341. doi:10.1016/j.neuroimage.2013.08.048

Preston AR, Eichenbaum H. 2013. Interplay of hippocampus and prefrontal cortex in memory. *Curr Biol* **23**:R764–R773. doi:10.1016/j.cub.2013.05.041

Romano JC, Howard Jr JH, Howard D V. 2010. One-year retention of general and sequence-specific skills in a probabilistic, serial reaction time task. *Memory* **18**:427–441. doi:10.1080/09658211003742680

Rommers J, Federmeier KD. 2018. Predictability's aftermath: downstream consequences of word predictability as revealed by repetition effects. *Cortex* **101**:1016–1030. doi:10.1016/j.cortex.2017.12.018

Schapiro AC, Kustner L V., 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, Frankland PW. 2017. Memory allocation and integration in rodents and humans. *Curr Opin Behav Sci* **17**:90–98. doi:10.1016/j.cobeha.2017.07.013

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

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

Schultz W, Dickinson A. 2000. Neuronal coding of prediction errors. *Annu Rev Neurosci* **23**:473–500.

Stanislaw H, Todorov N. 1999. Calculation of signal detection theory measures. *Behav Res Methods, Instruments, Comput* **31**:137–149. doi:10.3758/BF03207704

Tse D, Langston RF, Kakeyama M, Bethus I, Spooner PA, Wood ER, Witter MP, Morris RGM. 2007. Schemas and memory consolidation. *Science (80-)* **316**:76–82. doi:10.1126/science.1135935

Turk-Browne N, Scholl B. 2009. Neural evidence of statistical learning: Efficient detection of visual regularities without awareness. *J Cogn Neurosci* **21**:1934–1945.

Turk-Browne NB, Jungé JA, Scholl BJ. 2005. The automaticity of visual statistical learning. *J Exp Psychol Gen* **134**:552–564. doi:10.1037/0096-3445.134.4.552

Turk-Browne NB, Scholl BJ, Johnson MK, Chun MM. 2010. Implicit perceptual anticipation triggered by statistical learning. *J Neurosci* **30**:11177–11187. doi:10.1523/JNEUROSCI.0858-10.2010

van Kesteren MTR, Brown TI, Wagner AD. 2016. Interactions between memory and new learning: insights from fMRI multivoxel pattern analysis. *Front Syst Neurosci* **10**:46. doi:10.3389/fnsys.2016.00046

van Kesteren MTR, Fernández G, Norris DG, Hermans EJ. 2010. Persistent schema-dependent hippocampal-neocortical connectivity during memory encoding and postencoding rest in humans. *Proc Natl Acad Sci* **107**:7550–7555. doi:10.1073/pnas.0914892107

van Kesteren MTR, Rijpkema M, Ruiter DJ, Fernández G. 2010. Retrieval of associative information congruent with prior knowledge is related to increased medial prefrontal activity and connectivity. *J Neurosci* **30**:15888–1594. doi:10.1523/JNEUROSCI.2674-10.2010

van Kesteren MTR, Ruiter DJ, Fernández G, Henson RN. 2012. How schema and novelty augment memory formation. *Trends Neurosci* **35**:211–219. doi:10.1016/j.tins.2012.02.001

Wylie GR, Foxe JJ, Taylor TL. 2008. Forgetting as an active process: an FMRI investigation of item-method-directed forgetting. *Cereb Cortex* **18**:670–682. doi:10.1093/cercor/bhm101

Zeithamova D, Dominick AL, Preston AR. 2012a. Hippocampal and ventral medial prefrontal activation during retrieval-mediated learning supports novel inference. *Neuron* **75**:168–179. doi:10.1016/j.neuron.2012.05.010

Zeithamova D, Schlichting ML, Preston AR. 2012b. The hippocampus and inferential reasoning: building memories to navigate future decisions. *Front Hum Neurosci* **6**:70. doi:10.3389/fnhum.2012.00070

Zwaan RA, Radvansky GA. 1998. Situation models in language coomprehension and memory. *Psychol Bull* **123**:162–185. doi:10.1016/0142-9612(94)90271-2