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 reactivates memory of previous encounters which can trigger a prediction of what comes next to facilitate responsiveness. However, a prediction error can lead to pruning of the offending memory, a process that weakens its representation in the brain and leads to forgetting. Our goal in this study was to evaluate whether memories are spared from pruning in situations that allow for more abstract yet reliable predictions. We hypothesized that when the category, but not the identity, of a new stimulus can be anticipated, this will reduce pruning of existing memories and also reduce encoding of the specifics of new memories. Participants viewed a sequence of objects, some of which reappeared multiple times ("cues"), followed always by novel items. Half of the cues were followed by new items from different (unpredictable) categories, while others were followed by new 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. These findings demonstrate that how episodic memories are updated is influenced by the reliability of abstract-level predictions in familiar 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 & Pacton, 2006; Turk-Browne, Jungé, & Scholl, 2005; Turk-Browne & Scholl, 2009) – that allow us to acquire long-term knowledge about the statistical structure of the world (Kóbor, Janacsek, Takács, & Nemeth, 2017; Romano, Howard Jr, & Howard, 2010). Studies of visual statistical learning have shown that observers can implicitly learn subtle statistical relationships between visual stimuli in both time (Fiser & Aslin, 2002; Fiser, Scholl, & Aslin, 2007; Schapiro, Kustner, & Turk-Browne, 2012) and space (Chun & Jiang, 1998; Fiser & Aslin, 2001; Preston & Gabrieli, 2008; Turk-Browne & Scholl, 2009). Knowledge of these statistics can build up expectations that trigger predictions about upcoming perceptual events (Turk-Browne, Scholl, Johnson, & Chun, 2010). In some situations, these expectations may be stimulus-specific (Conway & Christiansen, 2006), and in others they may be more abstract, for example, operating at a categorical level that relies on existing conceptual knowledge (Brady & 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, as in many real-world encounters, 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 & 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 & Gilboa, 2014; Preston, Molitor, Pudhiyidath, & Schlichting, 2017; Tse et al., 2007; van Kesteren, Rijpkema, Ruiter, & Fernández, 2010; van Kesteren, Ruiter, Fernández, & Henson, 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 (i.e., a *misprediction occurs*), the resulting error signal can weaken the long-term memory representation of the mispredicted item, leading it to be pruned from memory (Kim, Lewis-Peacock, Norman, & Turk-Browne, 2014; Kim, Norman, & Turk-Browne,

2017). Memory pruning is an adaptive, and error-driven learning process (Pagnoni, Zink, Montague, & Berns, 2002; Schultz & 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, Sherrill, & Preston, 2017; Preston & Eichenbaum, 2013; Schlichting & Frankland, 2017; Schlichting, Mumford, & Preston, 2015; Schlichting & Preston, 2015; Zeithamova, Dominick, & Preston, 2012; Zeithamova, Schlichting, & Preston, 2012). 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 is the *reliability of predictions* in familiar temporal contexts. A context where it is difficult to predict which items will appear next should consistently produce unreliable predictions that are violated by novel items. This should lead to worse memory for the mispredicted items due to pruning, but strong encoding of the novel items because their predictive relationship with the environment is still being learned (Dayan, Kakade, & Montague,

2000). On the other hand, a context that allows for more reliable predictions should yield fewer opportunities for mispredicted items to get pruned, but it could also lead to weaker encoding of the novel items because they were more predictable.

What makes a prediction *reliable*? We define this as a *prediction that is accurate enough to be useful*. Perfect prediction of the future is difficult, but reliable prediction is feasible because our experiences often contain hidden statistical structure that can be learned and generalized to new events. We know that the brain can use existing semantic knowledge of the world to 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 & Oliva, 2008). Over time the brain should adapt to the predictability of these category exemplars and as a result build expectations about a *category* rather than any specific *item* (Brown & Braver, 2005; den Ouden, Friston, Daw, McIntosh, & Stephan, 2009). We will leverage this idea to manipulate the predictability of images in different temporal contexts and evaluate the impact on episodic memory performance.

To date, memory pruning has been studied only in unpredictable situations that consistently generate prediction errors, without any possibility of extracting predictable statistics across repeated exposures. However, it is important to understand how memory is updated when higher-order predictability is embedded into our experiences, as it often is in the real world. 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 1B). 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 (Figure 1A). 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 & Rees, 2006; Lewis-Peacock & 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 1A). 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 greater evidence of memory pruning in the incongruent condition, when the new items in a familiar context were completely unpredictable, as

compared to the congruent condition, when the category of new items could be predicted from abstract statistical learning across repetitions. In addition, we hypothesized that the predictability of category membership in the congruent condition would bias new encoding toward category information (to verify the prediction), rather than item-specific details, thus leading to worse subsequent memory for these items.

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, Zippi, & Preston, in prep.)

including Bank of standardized stimuli (Brodeur, Guérard, & Bouras, 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 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.

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 1A). 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 & Norman, 2014a), and we calculated memory sensitivity using A-prime (*A*’; Stanislaw & 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 (VTC; Grill-Spector & Weiner, 2014; Haxby et al., 2001) 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) with the regressor modeled with a mini-block boxcar and shifted forward 4 s to account for hemodynamic lag. 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 1A). 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*, Figure 2A). 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*, Figure 2A). 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.

Before the major analyses, which linked the classifier evidence strength to memory, we compared both prediction and perception evidence distributions from

category classifier across conditions and positions. Classifier evidence was averaged for each process x condition x position within subject. Across all positions (i.e., 2nd – 4th for prediction, 1st - 4th for perception), repeated-measure two-way ANOVAs were conducted for condition x position within each process, and then three-way ANOVAs were applied for process x condition x position for the final three positions (Figure 2A). Finally, the classifier evidence scores for the target category (e.g., the perceived or predicted category) were compared to the mean of the three non-target categories with a three-way ANOVA for condition x position x target (target/non-target).

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, Turner, Ashby, & Poldrack, 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 $1^{st}/2^{nd}/3^{rd}/4^{th}$ positions), and six regressions conducted for prediction strength and subsequent memory (incongruent/congruent x $2^{nd}/3^{rd}/4^{th}$ 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 to remove subject-specific effect (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) across positions for each condition and process, and bootstrapped for the statistical significance test. Then, a repeated-measure two-way ANOVA for condition x process was conducted on the coefficient of linear regressions (β_{lin}) across positions, and the statistically reliability was assessed via bootstrapping.

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 1B). Performance differed across conditions (incongruent, congruent), positions (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 $Ps < .001$, significant after Bonferroni correction for 3 comparisons), 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 $\beta_{lin} > 0.01$, $Ps < .01$) and the RTs decreased (all $\beta_{lin} < -0.01$, $Ps < .001$), with no difference between conditions (condition x position two-way 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 1C). 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, $F(3, 135) = 2.32$, $P = .078$). This interaction was significant only for the subset of participants who were tested outside the scanner at the end of the experiment ($n = 14$, see Methods; $F(3, 39) = 3.13$, $P = .036$). In follow-up analyses across all participants, 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 for 2 comparisons). This interaction was not significant for the final two item positions ($F(1, 45) = 2.14$, $P = .15$) trended in the opposite direction, yet 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.

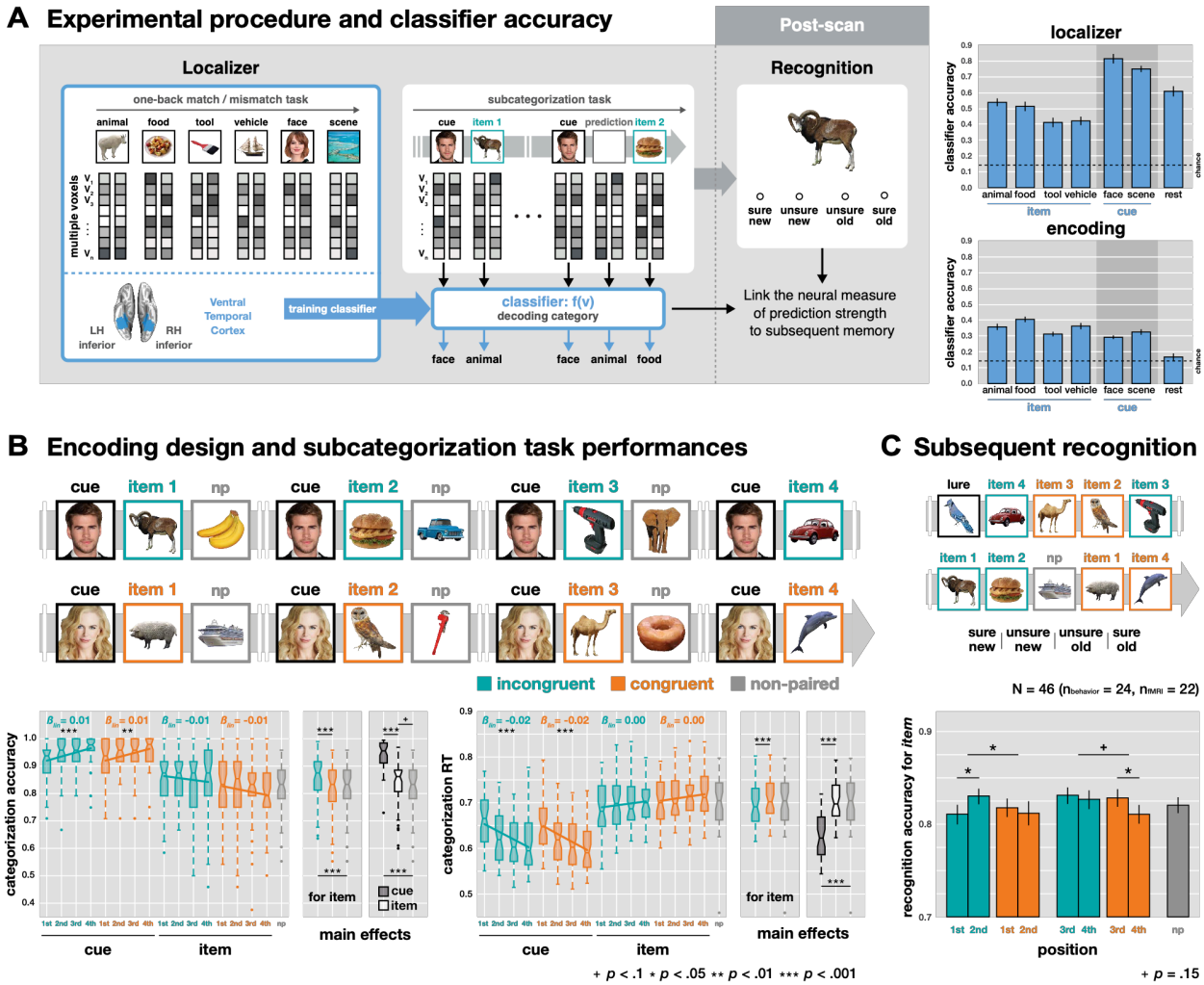

Figure 1. Experimental procedure and behavioral performance. (A) Experimental procedures and fMRI pattern classifier accuracy. (left) Pattern classifiers were trained on fMRI data from the localizer task to identify category-specific patterns of brain activity in ventral temporal cortex mask. These classifiers were then applied to the encoding task, and the neural results were linked to subsequent memory scores in the surprise recognition test. (right/top) Pattern classification was successful for all categories in the localizer task (chance = 0.14). (right/bottom) These classifiers, trained on all localizer data, successfully identified the category of each item presented in the encoding task. (B) Incidental encoding task with categorization accuracy and RT. Subcategory judgments for each picture in a sequence (CUE: items that were presented four total times; $1^{st}/2^{nd}/3^{rd}/4^{th}$ 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. (C) 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).

fMRI pattern classification. The encoding task data were analyzed by fMRI pattern classifiers trained, separately for each individual, on category-specific data from an independent localizer task (Figure 1A, see Methods). The stimuli used in the

localizer task were representatives of the same categories but were separate exemplars 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). 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 1A right/top), 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), which is unsurprising given the ubiquitous use of faces and scenes in the literature (see Grill-Spector & Weiner, 2014). 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.

As a sanity check, we verified that the category of all items presented in the encoding task was being accurately classified (M = 0.32, SEM = 0.01, chance = 0.14, Ps < .001, Figure 1A, right/bottom). However, for cues (which appeared four times each), the classification accuracy was *lower* than for non-cue items (which appeared only once; paired t test, t_{21} = 3.24, P = .004). This was true in both the incongruent condition (t_{21} = 2.75, P = .012) and congruent condition (t_{21}= 2.95, P = .008). This relationship is a reversal from the results in the localizer data described above where the cue categories (faces and scenes) were classified with greater accuracy than the other four categories. This reduction in decoding accuracy for the cues in the encoding task 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.

To quantify both the perception of viewed items and the prediction of subsequent items in the encoding task, we defined two neural metrics using the fMRI pattern classifiers. The "*perception strength*" for each non-cue item was defined as the amount of classifier evidence for that item during its presentation (Figure 2A). There were no differences in mean perception strength for non-cue items across the four sequence positions (1st, 2nd, 3rd, 4th) and two conditions (congruent, incongruent; repeated measure two-way ANOVA, condition (main effect): $F(1, 63) = 0.04$, $P = .835$, position: $F(3, 63) = 1.73$, $P = .171$, interaction: $F(3, 63) = 0.32$, $P = .811$). The "*prediction strength*" for an item was defined as the amount of classifier evidence for that item when its cue reappeared later in the sequence. For example, in the example sequence in Figure 2A ("*Liam Hemsworth*, *ram, …, Liam Hemsworth*, *sandwich*"), the prediction strength for "*ram*" would be the amount of animal-category classifier evidence detected when "*Liam Hemsworth*" appeared the next time (prior to the appearance of "*sandwich*"). There were no differences in mean prediction strength across the three repetitions of the cue (2nd, 3rd, 4th positions) and two conditions (congruent, incongruent; condition: $F(1, 42) = 0.01$, $P = .942$, position: $F(2, 42) = 0.05$, $P = 0.955$, condition x position: $F(2, 42) = 1.05$, $P = .360$). The distributions of scores, which were concatenated across 2nd to 4th positions, for both perception strength and prediction strength are shown separately for each condition in Figure 2A. Perception strength for new items was reliably higher than the prediction strength for the expected items in both conditions and all positions ($M = 0.70$ vs. $M = 0.52$; three-way ANOVA, process: F(1,

42) = 244.41, $P < .001$, with no other significant effects). To assess the specificity of these neural measures, the classifier evidence values for the category of target items were compared with those of non-target categories. The perception evidence was significantly higher for target categories than non-targets categories across conditions and positions (three-way ANOVA, target: $F(1, 63) = 272.88$, $P < .001$), while no such differences were found for prediction evidence (all *Ps* > .164).

Our next step was to link these neural measurements of prediction and perception with subsequent memory outcomes. Note that for all items that appeared in the 1st, 2nd, or 3rd positions of a cue sequence, we will be assessing *two* contributions to each item's subsequent memory outcome: its perception strength in position N, and its prediction strength in position N+1. Memory for items in the 4th position had a contribution only from perception strength, because the cue did not repeat a 5th time. We will first address these two contributions separately, and then together.

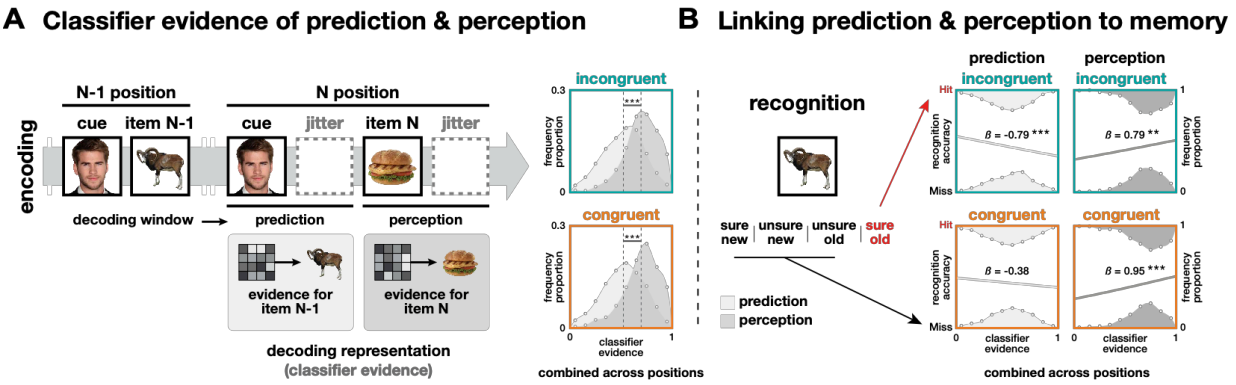

Figure 2. Predicting subsequent memory from brain activity during encoding. (A) Classifier evidence of prediction (item N-1) and perception (item N) in the N position (B) linking to memory performance of those items. (A) The distributions of classifier evidence for prediction (light gray) and perception (dark gray) combined across positions are shown in both conditions. (B) Logistic regression results (coefficient estimate: β) linking classifier evidence and recognition accuracy are shown separately predicted items and for perceived items combined across positions. Statistics are based on bootstrap analyses with 1,000 iterations.

Prediction and subsequent memory. To evaluate our main hypothesis that the reliability of context-based predictions impacts pruning of episodic memories, we used logistic regression to link our neural measures of prediction with subsequent recognition memory from the end of the experiment. Evidence of pruning exists when stronger predictions are associated with worse memory for the mispredicted items (see Kim et al., 2014). Figure 2B shows evidence consistent with pruning in the incongruent condition, such that stronger predictions were indeed associated with worse memory for the mispredicted items (β = -0.79, P < .001). This relationship was found in each repetition of the cue ($2^{nd}/3^{rd}/4^{th}$ positions, Ps < .045), with no change across the positions (β_{lin} = 0.12, P = .323, Figure 3A). Consistent with our hypothesis, however, no evidence of pruning was found in the congruent condition (β = -0.38, P = .119, Figure 2B), where prediction strength was unrelated to subsequent memory for the mispredicted items. This relationship was absent for each repetition of the congruent cues (Figure 3B). Across all positions, there was no significant difference in the prediction-memory relationship between conditions (β_{inc} = -0.79 vs. β_{con} = −0.38, P = .161, Figure 2B). However, there was a statistical trend that it was stronger in the last position for the incongruent condition (β_{inc} = -0.74 vs. β_{con} = 0.23, P = .082, Figure 3). Although prediction strength in the congruent condition was unrelated to subsequent memory for the mispredicted item, it was, however, related to subsequent memory for the *new item*. Stronger predictions from a congruent cue were associated with worse subsequent memory for the next item that appeared (β = -0.64, P = .008). This was not true for incongruent cues (β = 0.01, P = .470; P = .063 vs. congruent cues). This result suggests that congruent cues, which afforded categorical predictions, biased attention

toward category-general features, and away from item-specific features during encoding of the next items in the sequence. This idea will be developed further in later sections.

To examine when prediction information emerged during a cue presentation, we linked subsequent memory with prediction strength measured separately from three time periods of each cue: *baseline* (3 s before the cue), *cue* (3-5 s during the cue), and *item* (3-5 s after the cue when the new item appeared, Figure 4). As expected, there was no relationship between prediction strength and subsequent memory in either condition during the baseline period, a time at which there would be no basis for a prediction (Figure 4A). As expected, no relationship emerged during the cue period or item period in the congruent condition (*P*s > .270). However, in the incongruent condition, the negative relationship between prediction strength and subsequent memory emerged during the cue period and persisted through the item period (*P*s < .024).

Perception and subsequent memory. To evaluate our hypothesis that increased predictability of items in a familiar context will decrease perceptual encoding of novel stimuli, we again used logistic regression to link our neural measures of perception strength with subsequent recognition memory. In the incongruent condition, when new items were unpredictable, we found that on average stronger perception strength for an item was associated with better memory for that item ($\beta = 0.79$, $P = .005$, Figure 2B). This relationship strengthened across repetitions ($\beta_{lin} = 0.46$, $P = .002$) and was strongest by the final (4th) repetition ($\beta = 1.24$, $P = .004$, Figure 3A). This indicates that new items in this condition were being actively encoded because their neural index

of perception strength was directed related to their subsequent memory. The influence of perception strength on memory for earlier items could have been diminished by the contribution from the (mis)prediction of those items during the next cue repetition (we will address this more in the next section). This could not be the case for the final items.

For the congruent condition, we found a similar result that, on average, stronger perception was associated with better subsequent memory ($\beta = 0.95$, $P = .001$, Figure 2B). However, this relationship weakened significantly across repetitions ($\beta_{lin} = -0.71$, $P = .006$) and disappeared altogether by the final (4th) position ($\beta = 0.39$, $P = .215$, Figure 3B). There was a significant interaction of condition (incongruent, congruent) on the perception-memory relationship. For earlier items, the influence of perception on memory was stronger for congruent items compared to incongruent items (1st and 2nd, $Ps < .027$), but this relationship was reversed for the final items (4th, $P = .046$). Across the three repetitions, the shift in perception-memory relationship differed significantly between the two conditions (linear: $\beta_{inc} = 0.46$ vs. $\beta_{con} = -0.71$, $P = .001$). Our primary analyses focus on items in the final three repetitions, but these interaction results for perception and memory hold when including 1st position items as well (linear: $\beta_{inc} = 0.30$ vs. $\beta_{con} = -0.51$, $P = .002$). Finally, there was a significant interaction of condition x process (prediction/perception) on the shift in brain-behavior relationships across repetitions ($F(1, 1999) = 9053.83$, $P < .001$) indicating that the anticipation and processing of new items impacted memory differently across time in predictable vs. unpredictable contexts.

Together, these results indicate that new items in the congruent condition, whose category could be predicted from prior statistics, were not being actively encoded after

multiple repetitions of their context. This is consistent with the idea that statistical learning at the categorical level was gradually stabilized across multiple repetitions of the context, which led to new items being also encoded at the categorical level at the expense of item details. We will now discuss various control analyses that rule out possible confounds and alternative explanations of our results.

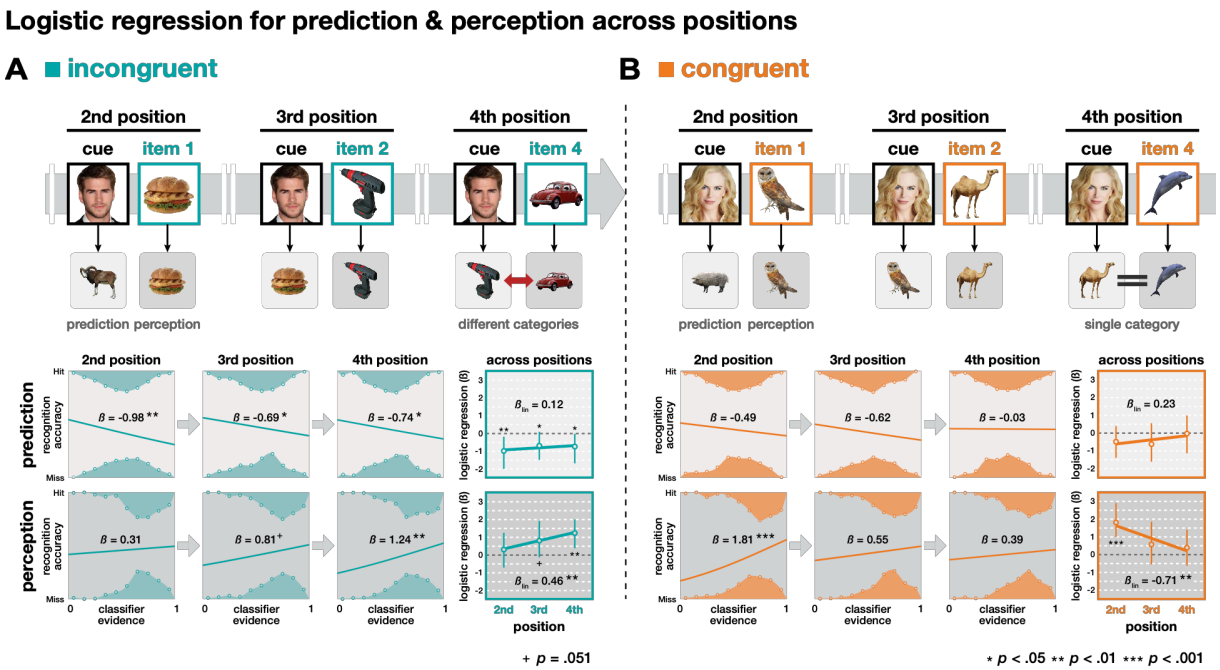

Figure 3. Changes across time in the links between perception, prediction, and subsequent memory in the (A) incongruent and (B) congruent condition. Hypothesized mental representations for prediction and perception across given trials (top). Logistic regression results (coefficient estimate: β) linking classifier evidence and recognition accuracy are shown separately for prediction (middle) and perception (bottom) for each position and across positions in both the incongruent and congruent conditions. Statistics are based on bootstrap analyses with 1,000 iterations. Error bars represent 95% CI.

Controlling for potential confounding factors. Control analyses confirmed that the negative relationships of prediction strength and memory in the incongruent condition (i.e., memory pruning) were specific to classifier evidence for the category of the predicted "target" item. We used partial correlation to control for the mean evidence

of the three non-target categories for each prediction (for similar approaches see Detre et al., 2013; Kim et al., 2014), and the pruning relationship between target evidence and subsequent memory held for the predicted items (Figure 4B). To rule out the possibility that memory for the mispredicted items was worse due to poor initial encoding of those items, we controlled for the perception strength of each item during encoding and found that the relationship between prediction strength and subsequent memory remained negative in the incongruent condition and absent in the congruent condition (Figure 4C). Finally, we ran an analysis to evaluate the specificity of the relationship between perception strength and subsequent memory. We used partial correlation to control for the amount of non-target category evidence during perception and confirmed that these relationships were specific to the evidence for the target category (Figure 4D).

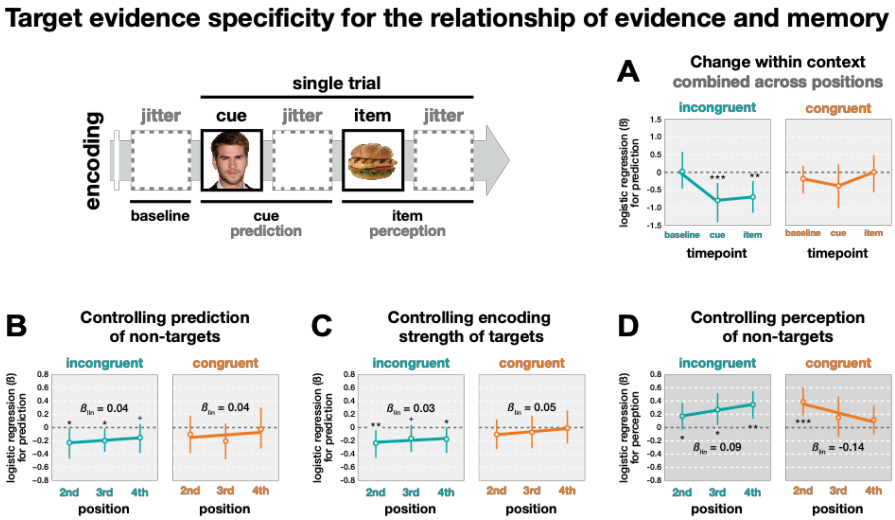

Figure 4. Control analyses for the specificity of the target category in the relationship between prediction/perception strength and subsequent memory across positions. (A) The relationship between prediction strength and subsequent memory is shown for different time windows *within* a cue presentation (baseline, cue, item), and averaged across the final three repetitions of each cue. (B) Controlling for the prediction evidence for non-target categories. (C) Controlling for the perception strength of the predicted item when it was encoded on its previous position following cue. (D) Controlling for the perception evidence for non-target categories. Statistics are based on bootstrap analysis with 1,000 iterations. Error bars represent 95% CI.

Mismatch signals in hippocampus. Linear regression analyses were applied to link mismatch signals in the hippocampus (Long, Lee, & Kuhl, 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).

DISCUSSION

When we reencounter a familiar situation, our brain anticipates what might come next based on previous encounters. This study demonstrates that the predictability of our experiences influences whether we prune memories of past events and also how we form memories of new ones. Our results show that when the temporal statistics across repeated experiences allow for the learning of abstract, conceptual relationships between events, this has two main consequences: (1) memory for past events is no longer impacted by prediction errors for new events, and (2) memory for new events declines because encoding becomes less focused on item-specific details.

We used fMRI pattern classifiers to track the explicit encoding and the covert prediction of visual stimuli during a continuous sequence of image presentations. The sequence was configured such that some of the stimuli ("cues") repeated multiple times across the experiment. Cue stimuli were followed either by new items that belonged to different categories (incongruent condition) or new items that belonged to the same category (congruent condition). For cues in the incongruent condition, which afforded

participants no reliable predictive information across repetitions, stronger neural evidence for a (incorrect) prediction that the previous item to follow the cue will reappear were associated with worse subsequent memory for that mispredicted item. 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. A key result of this study is that for cues in the congruent condition, which afforded participants the ability to learn across repetitions the category of item to expect, no memory pruning was observed. Furthermore, encoding of new items after the final appearance of congruent cues was impaired relative to incongruent cues: subsequent memory on average was worse for these congruent items, and the memory strength of individual items was unrelated to neural measures of their perception strength.

Context-based predictions of items. Our neural analyses of automatic predictions relied on pattern classifier estimates for a "target" category following the repetition of a cue stimulus (e.g., classifier evidence for the target category of "animal" following the repetition of *cue* in the sequence "*cue,* badger*, …, cue, …*"). On average, the amount of classifier evidence following a cue repetition for the target category was not greater than for non-target categories (in this example: food, tools, vehicles). This was true for both incongruent cues and congruent cues, and by itself is inconsistent with the idea that cues triggered predictions of items from the previous repetition. However, the results found from linking these neural measures of "prediction" on a trial-by-trial basis to subsequent memory outcomes suggests otherwise. For incongruent cues,

stronger classifier evidence for the (predicted) target category was associated with weaker subsequent memory for those items (Figure 3A). This relationship held even after controlling for the amount of classifier evidence for the non-target categories at the time of prediction and for the initial perception strength of those target items from the previous repetition (Figure 4B and C). This robust link to item-specific behavioral outcomes is consistent with the interpretation that incongruent cues triggered specific predictions of previous items that were consequently pruned from memory after the prediction error.

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 1A). 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 1A). The elimination of the decoding advantage for the repeated cues could 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 (or items) 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 condition. 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.

Context-based predictions of categories. 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 participants noticing repeated presentation of the cues, they did not detect any structure in the order of stimulus presentations. Any differences in prediction between the experimental conditions would therefore be due to implicit learning of the transition probabilities associated with incongruent cues vs. congruent cues.

Similar to the results for incongruent cues, there was no direct neural evidence of prediction (target classifier evidence was not greater than non-target evidence) for congruent cues. Unlike incongruent cues, however, there was no relationship between the neural evidence for the "predicted" item (i.e., the item that followed the cue on its prior appearance) and that item's subsequent memory (Figure 3B). These data do not support an inference that participants were making item-specific predictions for congruent cues. Rather, they suggest that congruent cues either (1) triggered no predictions at all, or (2) triggered predictions at the *category* level rather than at the *item*

level (e.g., “expect some animal” instead of “expect that badger”). Two additional results provide converging evidence for the latter interpretation. First, if predictions were categorical, this should focus attention on categorical information, rather than item-specific details, during the encoding of new items, which in turn should lead to worse subsequent memory for these items. Indeed, we found that stronger prediction evidence after a congruent cue was associated with worse subsequent memory for the next item, which by design was a new exemplar of that category ($\beta = -0.64$, $P = .008$). This relationship between prediction strength and memory for new items was not true for incongruent cues ($\beta = 0.01$, $P = .47$). Second, as reported in Figure 3B, the perception strength for new items following congruent cues became decoupled from their subsequent memory after the second appearance of the cues.

Together, we believe these results are most consistent with the idea that congruent cues triggered categorical predictions rather than no predictions at all. The congruent condition results reflect the consequences of statistical learning focused not on item-specific details, but rather on abstract conceptual information (Brady & 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 & Norman, 2002; Tse et al., 2007; Marlieke T. R. van Kesteren, Brown, & Wagner, 2016; Marlieke T R van Kesteren et al., 2012). Once this general inference was developed, this focused future encoding on category-level information and the specific episodic details of new items were forgotten (Marlieke T. R. van Kesteren et al., 2016; Marlieke T R van Kesteren et al., 2012).

Our analyses relied on category-specific fMRI pattern classifiers to covertly measure implicit predictions of previous items from repeated contexts. 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 classifier sensitivity in early pilot data. Instead we relied on category information and therefore could not distinguish predictions for individual exemplars of a category from generic category predictions. Instead, we leveraged 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, Mur, & Bandettini, 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 (Kim et al., 2014) 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 & Norman, 2014a; Newman & 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 following cue repetitions in the

incidental encoding task. In Figure 2A, the *prediction strength* for the predicted items following a repeated cue is contrasted with the *perception strength* for the items that actually appeared. As would be expected, the distributions of classifier evidence values show that perception strength (M = 0.70) was reliably higher than prediction strength in both conditions (M = 0.52, Ps < .001, Figure 2A). 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 will not be enhanced beyond this moderately activated state which, according to the hypothesis, can lead to weakening and forgetting of the item. In temporal contexts that allow for more abstract statistical learning (e.g., our congruent condition), the predictions may not contain representations of specific events from the past (e.g., "some animal" might be expected but not "that badger"). The lack of specificity in these predictions may have the effect of shielding the memories of those specific events from modification.

Memory pruning vs. memory integration. It could be argued that cues in the incongruent condition should trigger memory integration (Greve, Abdulrahman, & Henson, 2018; Morton et al., 2017; Preston & Eichenbaum, 2013; Schlichting & Frankland, 2017; Schlichting et al., 2015; Schlichting & Preston, 2015; Zeithamova, Dominick, et al., 2012; Zeithamova, Schlichting, et al., 2012) rather than memory pruning, such that all items maintain their associations to the repeated context, similar to inferential learning (Morton et al., 2017; Preston & Eichenbaum, 2013; Schlichting & Frankland, 2017; Schlichting et al., 2015; Schlichting & Preston, 2015; Zeithamova, Dominick, et al., 2012; Zeithamova, Schlichting, et al., 2012). 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, Dominick, et al., 2012). The hippocampus has a critical role, not only for memory integration (Eichenbaum, 2000; Marlieke T. R. van Kesteren et al., 2016) but also in mismatch detection (Kumaran & 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 & Maguire, 2007; Long et al., 2016; Schlichting & 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, Schlichting, et al., 2012) 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 (Morton et al., 2017; Schlichting & Preston, 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 a new category. Specifically, memory for the final item in the congruent contexts was worse than previous items in those same contexts (Figure 1C), and there was no relationship between neural evidence of perception for these items and their subsequent memory strength (Figure 3B). 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 & Shachar, 2015; van Kesteren et al., 2010; Zwaan & Radvansky, 1998). Consistent with this idea, 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. From these results, the authors suggested that predictability allows the brain to operate in a top-down "verification mode" at the expense of detailed stimulus processing. Our data are consistent with these findings in language processing, and they extend the idea that predictability of perceptual events (here, in sequences of visual objects) leads to reduced encoding of new stimuli.

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., 2014, 2017; Lewis-Peacock & Norman, 2014a; Wylie, Foxe, & Taylor, 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 mechanisms supports this behavior by reducing interference during context-based retrieval of relevant memories (Kumaran & 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 predictable contexts. These findings deepen our understanding of how episodic memories are formed and updated by demonstrating that our ability to predict the future influences how we remember the past.

REFERENCES

Brady, T. F., & Oliva, A. (2008). Statistical learning using real-world scenes: extracting categorical regularities without conscious intent. *Psychological Science*, *19*(7), 678–685. http://doi.org/10.1111/j.1467-9280.2008.02142.x

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

Brown, J. W., & Braver, T. S. (2005). Learned predictions of error likelihood in the anterior cingulate cortex. *Science*, *307*(5712), 1118–1121. http://doi.org/10.1126/science.1105783

Chun, M. M., & Jiang, Y. (1998). Contextual cueing: implicit learning and memory of visual context guides spatial attention. *Cognitive Psychology*, *36*(1), 28–71. http://doi.org/10.1006/cogp.1998.0681

Conway, C. M., & Christiansen, M. H. (2006). Statistical learning within and between modalities: pitting abstract against stimulus-specific representations. *Psychological Science*, *17*(10), 905–912. http://doi.org/10.1111/j.1467-9280.2006.01801.x

Dayan, P., Kakade, S., & Montague, P. R. (2000). Learning and selective attention. *Nature Neuroscience*, *3*(11s), 1218–1223. http://doi.org/10.1038/81504

den Ouden, H. E. M., Friston, K. J., Daw, N. D., McIntosh, A. R., & Stephan, K. E. (2009). A dual role for prediction error in associative learning. *Cerebral Cortex*, *19*(5), 1175–1185. http://doi.org/10.1093/cercor/bhn161

Detre, G. J., Natarajan, A., Gershman, S. J., & Norman, K. A. (2013). Moderate levels of

activation lead to forgetting in the think/no-think paradigm. *Neuropsychologia*, *51*(12), 2371–2388. http://doi.org/10.1016/j.neuropsychologia.2013.02.017

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, *7*(1), 1–26. http://doi.org/10.1214/aos/1176344552

Eichenbaum, H. (2000). A cortical-hippocampal system for declarative memory. *Nature Reviews Neuroscience*, *1*(1), 41–50. http://doi.org/10.1038/35036213

Fiser, J., & Aslin, R. N. (2001). Unsupervised statistical learning of higher order spatial structures from visual scenes. *Psychological Science*, *12*(6), 499–504. http://doi.org/10.1037//0278-7393.28.3.458

Fiser, J., & Aslin, R. N. (2002). Statistical learning of higher-order temporal structure from visual shape sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(3), 458–467. http://doi.org/10.1037/0278-7393.28.3.458

Fiser, J., Scholl, B. J., & Aslin, R. N. (2007). Perceived object trajectories during occlusion constrain visual statistical learning. *Psychonomic Bulletin & Review*, *14*(1), 173–178.

Friedman, A. (1979). Framing pictures: the role of knowledge in automatized encoding and memory for gist. *Journal of Experimental Psychology: General*, *108*(3), 316–355. http://doi.org/10.1037/0096-3445.108.3.316

Ghosh, V. E., & Gilboa, A. (2014). What is a memory schema? A historical perspective on current neuroscience literature. *Neuropsychologia*, *53*, 104–114. http://doi.org/10.1016/j.neuropsychologia.2013.11.010

Greve, A., Abdulrahman, H., & Henson, R. N. (2018). Opinion: Neural differentiation of incorrectly predicted memories. *Frontiers in Human Neuroscience*, *12*(278).

http://doi.org/doi: 10.3389/fnhum.2018.00278

Grill-Spector, K., & Weiner, K. S. (2014). The functional architecture of the ventral temporal cortex and its role in categorization. *Nature Reviews Neuroscience*, *15*(8), 536–548. http://doi.org/10.1038/nrn3747

Gronau, N., & Shachar, M. (2015). Contextual consistency facilitates long-term memory of perceptual detail in barely seen images. *Journal of Experimental Psychology: Human Perception and Performance*, *41*(4), 1095–1111. http://doi.org/10.1037/xhp0000071

Haxby, J. V., Connolly, A. C., & Guntupalli, J. S. (2014). Decoding neural representational spaces using multivariate pattern analysis. *Annual Review of Neuroscience*, *37*, 435–456. http://doi.org/10.1146/annurev-neuro-062012-170325

Haxby, J. V, Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distrubuted and overlapping representations of face and objects in ventral temporal cortex. *Science*, *293*(5539), 2425–2430. http://doi.org/10.1126/science.1063736

Haynes, J. D., & Rees, G. (2006). Decoding mental states from brain activity in humans. *Nature Reviews. Neuroscience*, *7*(7), 523–534. http://doi.org/10.1038/nrn1931

Kim, G., Lewis-Peacock, J. A., Norman, K. A., & Turk-Browne, N. B. (2014). Pruning of memories by context-based prediction error. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(24), 8997–9002. http://doi.org/10.1073/pnas.1319438111

Kim, G., Norman, K. A., & Turk-Browne, N. B. (2017). Neural differentiation of incorrectly predicted memories. *The Journal of Neuroscience*, *37*(8), 2022–2031. http://doi.org/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. *Scientific Reports*, *7*(1), 760. http://doi.org/10.1038/s41598-017-00807-3

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*, 4. http://doi.org/10.3389/neuro.06.004.2008

Kumaran, D., & Maguire, E. A. (2007). Match mismatch processes underlie human hippocampal responses to associative novelty. *The Journal of Neuroscience*, *27*(32), 8517–8524. http://doi.org/10.1523/JNEUROSCI.1677-07.2007

Lewis-Peacock, J. A., & Norman, K. A. (2014a). Competition between items in working memory leads to forgetting. *Nature Communications*, *5*, 5768. http://doi.org/10.1038/ncomms6768

Lewis-Peacock, J. A., & Norman, K. A. (2014b). *Multi-voxel pattern analysis of fMRI data*. (M. S. Gazzaniga & G. R. Mangun, Eds.) *The Cognitive Neurosciences* (5th ed.). MIT Press.

Long, N. M., Lee, H., & Kuhl, B. A. (2016). Hippocampal mismatch signals are modulated by the strength of neural predictions and their similarity to outcomes. *The Journal of Neuroscience*, *36*(50), 12677–12687. http://doi.org/10.1523/JNEUROSCI.1850-16.2016

Morton, N. W., Sherrill, K. R., & Preston, A. R. (2017). Memory integration constructs maps of space, time, and concepts. *Current Opinion in Behavioral Sciences*, *17*, 161–168. http://doi.org/10.1016/j.cobeha.2017.08.007

Morton, N. W., Zippi, E. L., & Preston, A. R. (n.d.). Memory reactivation and

suppression modulate integration of the semantic features of related memories in hippocampus.

Mumford, J. A., Turner, B. O., Ashby, F. G., & Poldrack, R. A. (2012). Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. *NeuroImage*, *59*(3), 2636–2643. http://doi.org/10.1016/j.neuroimage.2011.08.076

Newman, E. L., & Norman, K. A. (2010). Moderate excitation leads to weakening of perceptual representations. *Cerebral Cortex*, *20*(11), 2760–2770. http://doi.org/10.1093/cercor/bhq021

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, *10*(9), 424–430. http://doi.org/10.1016/j.tics.2006.07.005

O'Reilly, R. C., & Norman, K. A. (2002). Hippocampal and neocortical contributions to memory: advances in the complementary learning systems framework. *Trends in Cognitive Sciences*, *6*(12), 505–510. http://doi.org/https://doi.org/10.1016/S1364-6613(02)02005-3

Pagnoni, G., Zink, C. F., Montague, P. R., & Berns, G. S. (2002). Activity in human ventral striatum locked to errors of reward prediction. *Nature Neuroscience*, *5*(2), 97–98. http://doi.org/10.1038/nn802

Perruchet, P., & Pacton, S. (2006). Implicit learning and statistical learning: one phenomenon, two approaches. *Trends in Cognitive Sciences*, *10*(5), 233–238. http://doi.org/10.1016/j.tics.2006.03.006

Power, J. D., Mitra, A., Laumann, T. O., Snyder, A. Z., Schlaggar, B. L., & Petersen, S.

E. (2014). Methods to detect, characterize, and remove motion artifact in resting state fMRI. *NeuroImage*, *84*, 320–341. http://doi.org/10.1016/j.neuroimage.2013.08.048

Preston, A. R., & Eichenbaum, H. (2013). Interplay of hippocampus and prefrontal cortex in memory. *Current Biology : CB*, *23*(17), R764–R773. http://doi.org/10.1016/j.cub.2013.05.041

Preston, A. R., & Gabrieli, J. D. E. (2008). Dissociation between explicit memory and configural memory in the human medial temporal lobe. *Cerebral Cortex*, *18*(9), 2192–2207. http://doi.org/10.1093/cercor/bhm245

Preston, A. R., Molitor, R. J., Pudhiyidath, A., & Schlichting, M. L. (2017). Schemas. In J.H. BYRNE (Ed.), *Learning and memory: a comprehensive reference* (2nd ed., pp. 125–132). New York: Elsevier.

Romano, J. C., Howard Jr, J. H., & Howard, D. V. (2010). One-year retention of general and sequence-specific skills in a probabilistic, serial reaction time task. *Memory*, *18*(4), 427–441. http://doi.org/10.1080/09658211003742680

Rommers, J., & Federmeier, K. D. (2018). Predictability's aftermath: downstream consequences of word predictability as revealed by repetition effects. *Cortex*, *101*(10), 1016–1030. http://doi.org/10.1016/j.cortex.2017.12.018

Schapiro, A. C., Kustner, L. V., & Turk-Browne, N. B. (2012). Shaping of object representations in the human medial temporal lobe based on temporal regularities. *Current Biology*, *22*(17), 1622–1627. http://doi.org/10.1016/j.cub.2012.06.056

Schlichting, M. L., & Frankland, P. W. (2017). Memory allocation and integration in rodents and humans. *Current Opinion in Behavioral Sciences*, *17*, 90–98.

http://doi.org/10.1016/j.cobeha.2017.07.013

Schlichting, M. L., Mumford, J. A., & Preston, A. R. (2015). Learning-related representational changes reveal dissociable integration and separation signatures in the hippocampus and prefrontal cortex. *Nature Communications*, *6*, 8151. http://doi.org/10.1038/ncomms9151

Schlichting, M. L., & Preston, A. R. (2015). Memory integration: neural mechanisms and implications for behavior. *Current Opinion in Behavioral Sciences*, *1*, 1–8. http://doi.org/10.1016/j.cobeha.2014.07.005

Schultz, W., & Dickinson, A. (2000). Neuronal coding of prediction errors. *Annual Review of Neuroscience*, *23*(1), 473–500.

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, *31*(1), 137–149. http://doi.org/10.3758/BF03207704

Tse, D., Langston, R. F., Kakeyama, M., Bethus, I., Spooner, P. A., Wood, E. R., … Morris, R. G. M. (2007). Schemas and memory consolidation. *Science*, *316*(5821), 76–82. http://doi.org/10.1126/science.1135935

Turk-Browne, N. B., Jungé, J. A., & Scholl, B. J. (2005). The automaticity of visual statistical learning. *Journal of Experimental Psychology: General*, *134*(4), 552–564. http://doi.org/10.1037/0096-3445.134.4.552

Turk-Browne, N. B., & Scholl, B. J. (2009). Neural evidence of statistical learning: Efficient detection of visual regularities without awareness. *Journal of Cognitive Neuroscience*, *21*(10), 1934–1945. http://doi.org/https://doi-org.ezproxy.lib.utexas.edu/10.1162/jocn.2009.21131

Turk-Browne, N. B., Scholl, B. J., Johnson, M. K., & Chun, M. M. (2010). Implicit perceptual anticipation triggered by statistical learning. *The Journal of Neuroscience*, *30*(33), 11177–11187. http://doi.org/10.1523/JNEUROSCI.0858-10.2010

van Kesteren, M. T. R., Brown, T. I., & Wagner, A. D. (2016). Interactions between memory and new learning: insights from fMRI multivoxel pattern analysis. *Frontiers in Systems Neuroscience*, *10*, 46. http://doi.org/10.3389/fnsys.2016.00046

van Kesteren, M. T. R., Fernández, G., Norris, D. G., & Hermans, E. J. (2010). Persistent schema-dependent hippocampal-neocortical connectivity during memory encoding and postencoding rest in humans. *Proceedings of the National Academy of Sciences*, *107*(16), 7550–7555. http://doi.org/10.1073/pnas.0914892107

van Kesteren, M. T. R., Rijpkema, M., Ruiter, D. J., & Fernández, G. (2010). Retrieval of associative information congruent with prior knowledge is related to increased medial prefrontal activity and connectivity. *Journal of Neuroscience*, *30*(47), 15888–1594. http://doi.org/10.1523/JNEUROSCI.2674-10.2010

van Kesteren, M. T. R., Ruiter, D. J., Fernández, G., & Henson, R. N. (2012). How schema and novelty augment memory formation. *Trends in Neurosciences*, *35*(4), 211–219. http://doi.org/10.1016/j.tins.2012.02.001

Wylie, G. R., Foxe, J. J., & Taylor, T. L. (2008). Forgetting as an active process: an FMRI investigation of item-method-directed forgetting. *Cerebral Cortex*, *18*(3), 670–682. http://doi.org/10.1093/cercor/bhm101

Zeithamova, D., Dominick, A. L., & Preston, A. R. (2012). Hippocampal and ventral medial prefrontal activation during retrieval-mediated learning supports novel

inference. *Neuron*, *75*(1), 168–179. http://doi.org/10.1016/j.neuron.2012.05.010

Zeithamova, D., Schlichting, M. L., & Preston, A. R. (2012). The hippocampus and inferential reasoning: building memories to navigate future decisions. *Frontiers in Human Neuroscience*, *6*, 70. http://doi.org/10.3389/fnhum.2012.00070

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language coomprehension and memory. *Psychological Bulletin*, *123*(2), 162–185. http://doi.org/10.1016/0142-9612(94)90271-2