Research Article

Psychological Science
2024, Vol. 35(10) 1178–1199

DOI: 10.1177/09567976241256617
www.psychologicalscience.org/PS

Consolidation Enhances Sequential Multistep Anticipation but Diminishes Access to Perceptual Features

Hannah Tarder-Stoll[1,2], Christopher Baldassano[1], and Mariam Aly[1]
[1]Department of Psychology, Columbia University, and [2]Baycrest Health Sciences, Rotman Research Institute, Toronto, Canada

Abstract

Many experiences unfold predictably over time. Memory for these temporal regularities enables anticipation of events multiple steps into the future. Because temporally predictable events repeat over days, weeks, and years, we must maintain—and potentially transform—memories of temporal structure to support adaptive behavior. We explored how individuals build durable models of temporal regularities to guide multistep anticipation. Healthy young adults (Experiment 1: $N = 99$, age range = 18–40 years; Experiment 2: $N = 204$, age range = 19–40 years) learned sequences of scene images that were predictable at the category level and contained incidental perceptual details. Individuals then anticipated upcoming scene categories multiple steps into the future, immediately and at a delay. Consolidation increased the efficiency of anticipation, particularly for events further in the future, but diminished access to perceptual features. Further, maintaining a link-based model of the sequence after consolidation improved anticipation accuracy. Consolidation may therefore promote efficient and durable models of temporal structure, thus facilitating anticipation of future events.

Keywords

episodic memory, consolidation, temporal structure, prediction, model-based learning, open data

Received 4/27/23; Revision accepted 4/19/24

Our daily lives are filled with temporal structure, from our commute to work to the steps in cooking a favorite dinner. Memories for temporal regularities allow us to build internal models that are useful for planning and anticipating future events (Behrens et al., 2018; Momennejad, 2020; O'Keefe & Nadel, 1979; Tolman, 1948). Accordingly, for these models to continue to be useful over a lifetime, they must persist over delays of days, weeks, or years. How do we build durable internal models of temporal structure, and how do these models transform over time?

Past work has investigated how individuals use temporal structure to anticipate events at a relatively short timescale, such as one step into the future (Davachi & DuBrow, 2015; de Lange et al., 2018; Hindy et al., 2016; Kok et al., 2012, 2014; Kok & Turk-Browne, 2018; Schapiro et al., 2012; Summerfield & Egner, 2009). Recent work has further shown that sequence memories guide anticipation of events multiple steps in the future (Brunec & Momennejad, 2022; Caucheteux et al., 2023; Lee et al., 2021; Elliott Wimmer & Büchel, 2019) and judgments about which items are likely to appear imminently (Tiganj et al., 2022), which may help individuals efficiently plan trajectories through learned environments (Bonasia et al., 2016; Brown et al., 2016).

Because we often repeatedly visit an environment whose temporal structure is stable over time, our internal models must be maintained —and potentially transformed—over periods of memory consolidation (Rauss & Born, 2017). Indeed, consolidation improves memory for temporal order (Drosopoulos et al., 2007), including temporal structure acquired through statistical learning (Arciuli & Simpson, 2012; Durrant et al., 2011, 2013; Kim et al., 2009; H. Liu et al., 2023). Consolidation

Corresponding Author:
Hannah Tarder-Stoll, Rotman Research Institute, Baycrest Health Sciences
Email: htarder-stoll@research.baycrest.org

can also enhance integration of sequences with overlapping items (Tompary & Davachi, 2022) and anticipation of upcoming items (Lutz et al., 2018). However, these studies overwhelmingly investigate how motor sequences are consolidated (Galea et al., 2010; Janacsek & Nemeth, 2012; Kóbor et al., 2017; Lutz et al., 2018; Romano et al., 2010; Sanchez et al., 2010; Walker et al., 2002, 2003), and the processes underlying learning and consolidation of procedural versus episodic memories likely differ in important ways (Stickgold, 2005). Studies that have investigated episodic learning of temporal structure have investigated only consolidation of shorter sequences (e.g., three items) and did not test anticipatory judgments (Drosopoulos et al., 2007; Tompary & Davachi, 2022). Thus, we have a limited understanding of how episodic memories of extended sequences are learned, maintained, or transformed over time to enable anticipation multiple steps in the future.

How might memory for temporal structure change with consolidation? Consolidation promotes schematization (Dudai et al., 2015; McClelland et al., 1995; Tompary & Davachi, 2017) so that events are remembered at a high level, but often at the expense of perceptual details (Audrain & McAndrews, 2022; Krenz et al., 2023; Lifanov et al., 2021; Robin & Moscovitch, 2017; Sekeres et al., 2016, 2018; Winocur et al., 2010), and with concomitant changes in neural representations that promote structured or schematic knowledge (Audrain & McAndrews, 2022; Krenz et al., 2023; Lifanov et al., 2021; Robin & Moscovitch, 2017; Sekeres et al., 2016, 2018; Winocur et al., 2010). If memories for temporal structure are likewise schematized, anticipatory judgments may become more efficient after a period of consolidation—particularly for events in the distant future. In tandem, such memories may at first contain visual detail, but they may lose this detail over time in favor of representing schematized sequential structure.

Thus, our first aim was to test whether consolidation of sequence memories promotes efficient far-reaching anticipation at the expense of maintaining incidental perceptual details. Our second aim was to determine what types of internal models are most useful for multistep anticipatory judgments (Momennejad, 2020), both soon after learning and after a consolidation delay. One possibility, inspired by model-based frameworks of decision-making (Daw & Dayan, 2014), is that individuals may form an internal model of a sequence that contains representations of each link between items (Kalm & Norris, 2014). An alternative possibility, inspired by successor representation frameworks (Dayan, 1993; Momennejad et al., 2017; Momennejad & Howard, 2018; Stachenfeld et al., 2017), is that information about future items in a sequence becomes cached into the current item representation.

Statement of Relevance

Experiences in our daily lives tend to unfold predictably over time. Whether traveling on our commute or making dinner, we can use the predictability of our experiences to anticipate upcoming steps beyond just our immediate future, allowing for efficient and adaptive behavior. Although temporal structure tends to be stable over days, months, or years (e.g., the order of subway stops is unlikely to change), it remains poorly understood how memory for temporal structure transforms over time. Here, we asked how memory for temporal structure changes after a period of consolidation to guide anticipation of events multiple steps in the future. We found that anticipating future events in a sequence became more efficient after consolidation, particularly for events further in the future. In contrast, consolidation reduced access to perceptual details. These results provide a crucial link between our understanding of memory consolidation and anticipation of temporal structure, and they demonstrate how memory for temporal structure is adaptively prioritized to anticipate future events.

Across two experiments, we therefore asked how internal models of temporal structure are (a) used to make anticipatory judgments at multiple timescales, (b) transformed with consolidation, and (c) represented in memory to guide accurate anticipation. Participants learned two sequences of 10 images that were predictable at the scene-category level. Participants were then cued with a scene category and a sequence and made anticipatory judgments about upcoming scene categories up to four steps into the future. We applied sequence-specific perceptual filters to each image; this allowed us to investigate whether incidental perceptual details were represented in participants' internal models by occasionally swapping the perceptual filter from the cued sequence to the one associated with the uncued sequence.

In Experiment 1, participants learned the sequences and made anticipatory judgments immediately after. We hypothesized (a) that anticipatory judgments would be less efficient for events further (vs. closer) in the future, indicating a behavioral cost to memories that are multiple steps away from the cued event (Tiganj et al., 2022); (b) that anticipatory judgments would be more efficient when perceptual filters matched (vs. did not match) expectations, reflecting incorporation of perceptual details into anticipated information; and (c) that the effect of perceptual filter on anticipatory judgments would be stronger for closer (vs. further) events. This latter hypothesis reflects our prediction that anticipated information will decline in vividness or detail the further in the future it is.

In Experiment 2, participants learned the sequences and made anticipatory judgments both immediately and after a period of consolidation. We hypothesized (a) that anticipatory judgments would become more efficient after (vs. before) a delay—particularly for events in the distant future, indicating a consolidation benefit for sequence memory; and (b) that perceptual filters would exert less of an influence on anticipatory judgments after a delay (vs. before), reflecting a loss of perceptual details with consolidation. Finally, we modeled response times from Experiments 1 and 2 to determine the strategies that promoted better retention after consolidation. We created two models: a *link-based* model, in which links between successive items in the sequence predict response times, and a *cue-based* model, in which response times varied depending on which category was used as the cue. Because participants were encouraged to create stories that linked adjacent images in the sequences, we hypothesized that the link-based (vs. cue-based) model would better predict response times and influence anticipatory judgments.

Open Practices Statement

All data, stimuli, and analysis scripts are publicly available and can be accessed at https://osf.io/8tuc4/. The experiments reported in this article were not preregistered.

Experiment 1

Method

Participants. Our target sample size was determined from a priori power analyses. We had two primary measures of interest: the effect of steps into the future and the effect of trial validity. A pilot study (n = 15) indicated that the effect of steps into the future was substantially more robust than that of trial validity. We therefore powered our studies to detect the smaller effect of trial validity (dz = 0.59). A power analysis conducted with the *pwr* package in R (Champely et al., 2020) determined that 100 participants would achieve 98% power (at an alpha of .05). We opted to use a higher power threshold than the traditional 80% because pilot samples can overestimate effect sizes (Gelman & Carlin, 2014).

We recruited participants through the Columbia University Psychology Department Participant Pool and through Prolific (www.prolific.co; an online participant-management tool) until our target sample size was met. We recruited a total of 162 participants (116 through Prolific and 46 through Columbia University). Two participants were excluded from data analysis because they failed to respond on more than 50% of trials, and 60 participants were excluded from data analysis because they did not perform statistically above chance on the anticipation task (56.875%, as determined by a binomial test; see the Procedure section). The data from the remaining 100 participants are analyzed here (M_{age} = 26.74 years, SD = 7.33; $M_{\text{education}}$ = 14.88 years, SD = 1.91; see Table 1 for demographic information). To be eligible for the experiment, participants had to report that they were between the ages of 18 and 40, fluent English speakers, and resided in the United States. Participants were compensated $7 per hour (Prolific) or received course credit (Columbia Psychology) for participating. All participants provided informed consent, and all procedures were in accordance with the policies of the Institutional Review Board at Columbia University.

Stimuli. Stimuli consisted of images of 10 different scene categories. The categories were airplane cabins, beaches, bedrooms, castles, city skylines, forests, kitchens, lecture halls, restaurants, and ski slopes. Thirty exemplars of each scene category were used in the experiment for a total of 300 unique images. Images were obtained through the SUN database (Xiao et al., 2010) and through Google image searches.

The 10 scene categories were used to form both Sequence A (i.e., the first learned sequence) and Sequence B (i.e., the second learned sequence), which had the same 10 scene categories in a different order. We used overlapping sequences to encourage rich episodic encoding of context-specific (here, sequence-specific) memories (Chanales et al., 2017) and to make the task difficult enough to avoid ceiling performance. The sequences were circular, meaning that the final category in the sequence connected back to the first category. The order of the scene categories across Sequence A and Sequence B were designed to be as distinct as possible: For any given category, the two preceding and two succeeding categories were different across sequences (Fig. 1a). The order of the scene categories in the sequences was randomized across participants with the constraint that Sequence A was shuffled in the same way across participants to create Sequence B, as described above.

We created two versions of each image of each scene category, using Photoshop's sponge and mosaic filters (Fig. 1b). This resulted in 600 total images that varied

Table 1. Demographic Information for Participants Across Experiments

	Experiment 1	Experiment 2 (1-day delay)	Experiment 2 (1-week delay)
Sample size	100	99	105
Age (*M* ± *SD*)	26.74 ± 7.33	31.65 ± 5.03	30.65 ± 6.73
Years of education (*M* ± *SD*)	14.88 ± 1.91	15.18 ± 1.85	14.99 ± 2.12
Recruitment method	116 Prolific 46 CU	99 Prolific	105 Prolific
Gender	61 F 38 M 1 NB	51 F 47 M 1 NB	55 F 45 M 4 NB
Race	63 W 19 A 10 BR 6 B/AA 1 AI/AN 1 O	77 W 7 A 7 B/AA 7 BR 1 O	80 W 8 B/AA 8 BR 6 A 1 AI/AN 1 O
Ethnicity	97 NH/L 3 H/L	91 NH/L 8 H/L	88 NH/L 16 H/L

Note: CU = Columbia University Psychology Department Participant Pool; gender: F = female, M = male, and NB = nonbinary; race: W = white, A = Asian, BR = biracial, B/AA = Black or African American, AI/AN = American Indian or Alaskan Native, and O = other; ethnicity, NH/L = not Hispanic or Latino, and H/L = Hispanic or Latino. Note that one participant failed to report demographic information for Experiment 2 (1-week delay).

by scene category (10 categories), unique exemplar (30 of each category), and perceptual filter (two filters). For each participant, each filter was assigned to either Sequence A or Sequence B: For half of the participants, all images in Sequence A had the sponge filter applied to them, and all images in Sequence B had the mosaic filter applied to them; the opposite was true for the remaining participants (Fig. 1c). Participants were not informed about this sequence to filter mapping.

Procedure. The experiment was conducted on the Gorilla platform (www.gorilla.sc; Anwyl-Irvine et al., 2020) and was composed of two tasks: sequence learning and anticipation (Fig. 2a).

Sequence learning. During the sequence-learning task, participants were instructed to learn the order of two sequences (A and B; see the Stimuli section). First, participants were shown all 10 scene categories in the Sequence A order on the screen and told to generate a story to link the 10 categories in order. By pressing a button, participants indicated that they were finished generating a story. Then they were shown the sequence as pairs of adjacent scene categories with a text box displayed underneath (e.g., images 1 and 2, then images 2 and 3). Participants were told to write down the story that they had generated (Fig. 2b; see the Supplemental Material available online for story examples). During the story generation and writing phases, perceptual filters were not applied to the scene images (see the Stimuli section) so that participants could not incorporate information about the perceptual filters into their stories; this ensured that these perceptual features were incidental to the main anticipation task. Participants had unlimited time to generate and write down their stories.

After writing down their Sequence A stories, participants then viewed exemplars of the scene categories in the same sequence order, with the perceptual filter for Sequence A (either mosaic or sponge, counterbalanced across participants) applied to each image (Fig. 2b). Participants were told to rehearse their stories. Each image was displayed for 5 s with a 0.5-s intertrial interval between images. After 20% of trials (i.e., test trials), instead of seeing the next image in the sequence, participants were shown two images of upcoming scene categories and were told to indicate which of those two categories was coming up sooner in the sequence, relative to the preceding category. Participants had 8 s to respond and were given feedback about whether their answer was correct or incorrect. The entire Sequence A order was shown three times, with the exemplars of each scene category changing on each repetition.

Following Sequence A learning, the story generation, writing, and rehearsal phases were repeated for Sequence B. In the rehearsal phase, each exemplar had the Sequence B filter (either mosaic or sponge, counterbalanced across participants) applied to it.

After Sequence A and Sequence B learning, the rehearsal phase (including test trials) was repeated for each sequence in an interleaved fashion. Participants

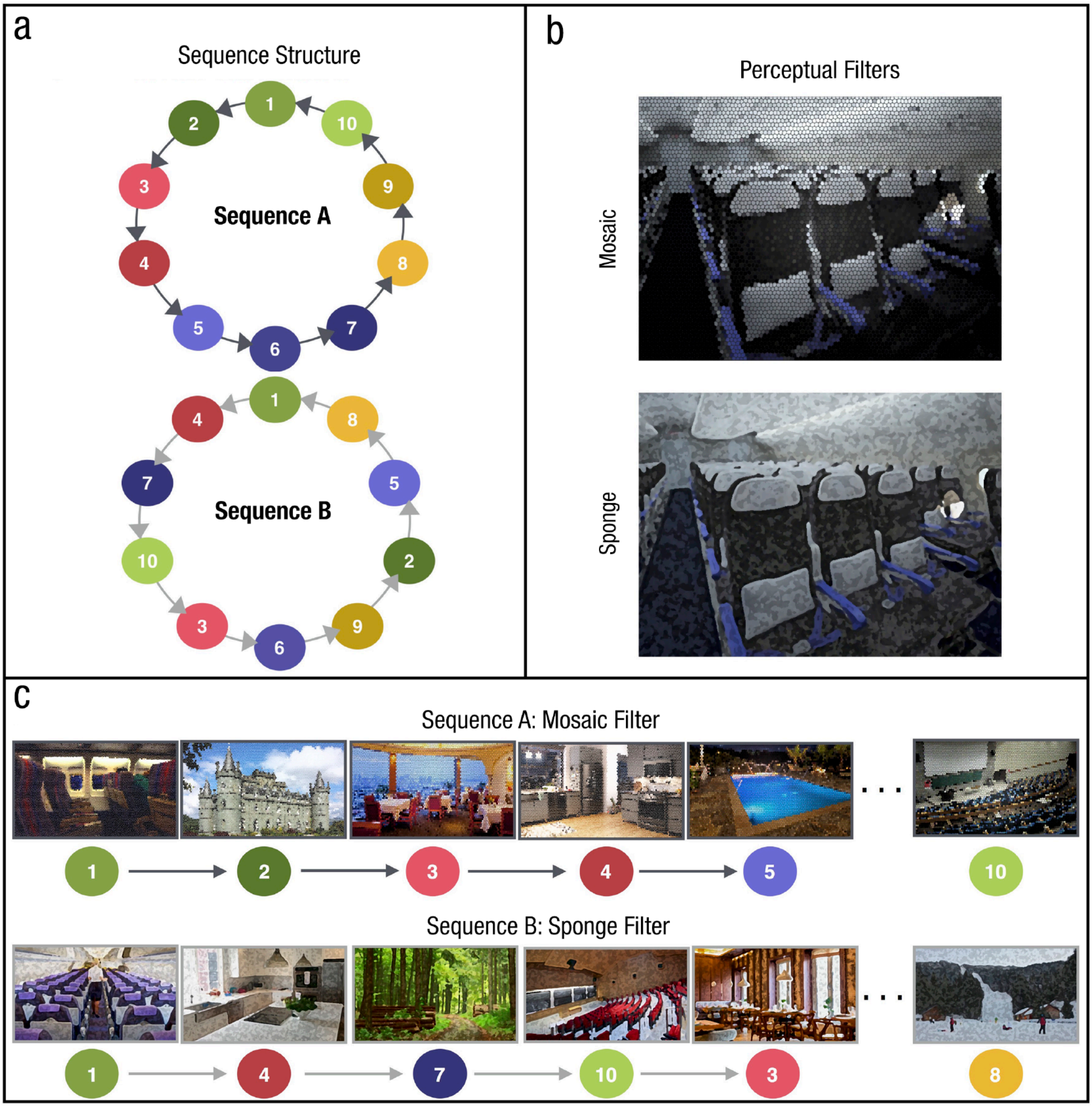

Fig. 1. Sequence structure and stimuli. Sequence structure is illustrated in (a). The sequences consisted of 10 scene categories (e.g., airplane cabins, swimming pools), indicated by the colored nodes. Sequence A (dark arrows, top) and Sequence B (light arrows, bottom) consisted of the same scene categories in a different order. The sequences were constructed to be as distinct as possible; for a given category, the two preceding and two succeeding categories were different across the sequences. Perceptual filters are shown in (b). Each image had a perceptual filter applied to it in either the mosaic style (top) or the sponge style (bottom). Each filter was assigned to one of the two sequences (counterbalanced across participants), but participants were not informed of this sequence-filter mapping. Example sequences are shown in (c), with perceptual filters applied to each scene exemplar for sample Sequence A (top) and sample Sequence B (bottom). In this example, Sequence A has the mosaic filter and Sequence B has the sponge filter. Only six of the ten categories are shown here for illustrative purposes. Sequence A and Sequence B were defined by a fixed order of scene categories, but different exemplars of each category were shown on each trial with minimal exemplar repetition. The sequence order of the categories was randomized across participants.

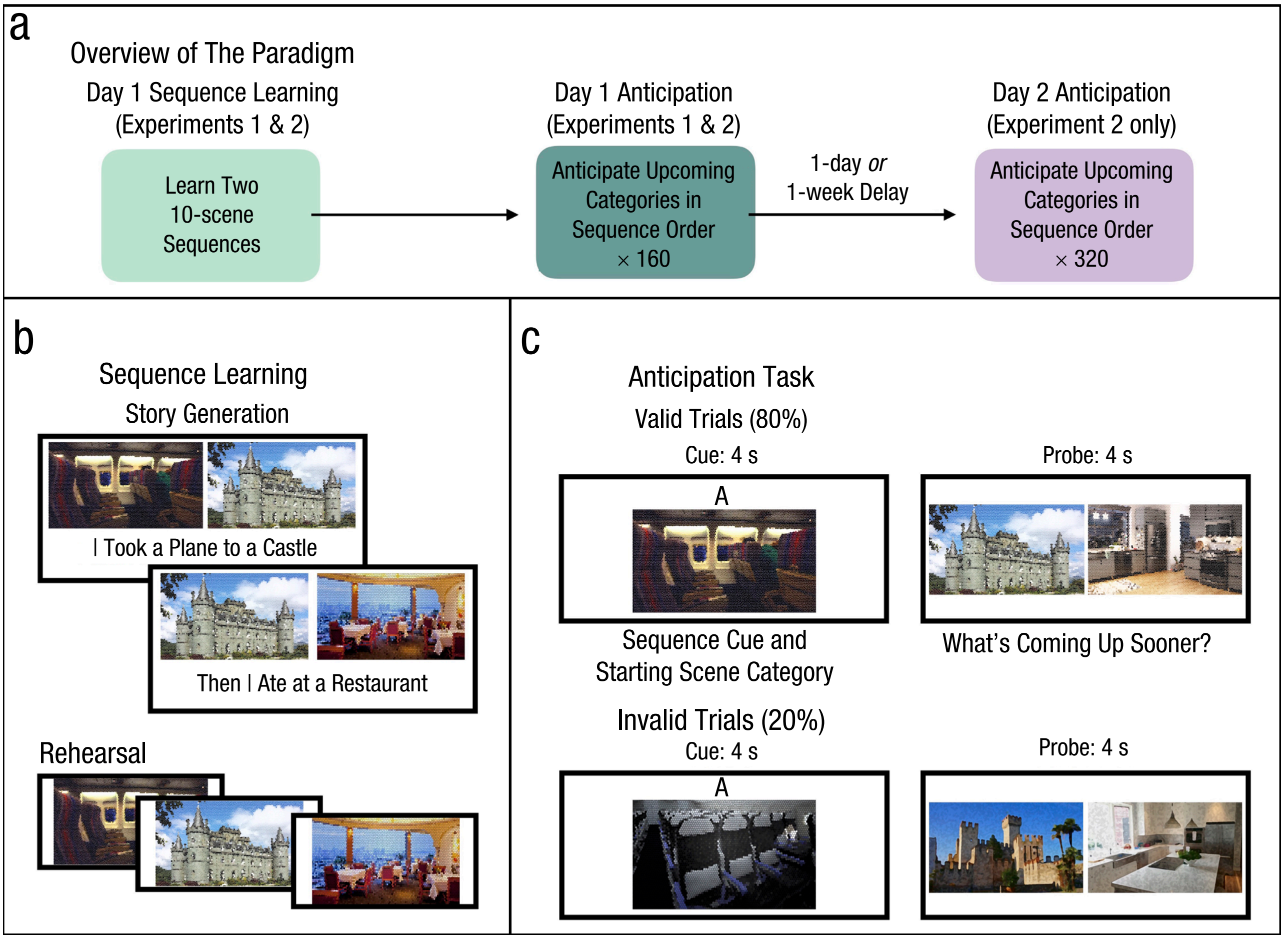
a
Overview of The Paradigm
Day 1 Sequence Learning (Experiments 1 & 2)
Learn Two 10-scene Sequences
Day 1 Anticipation (Experiments 1 & 2)
Anticipate Upcoming Categories in Sequence Order × 160
1-day or 1-week Delay
Day 2 Anticipation (Experiment 2 only)
Anticipate Upcoming Categories in Sequence Order × 320
b
Sequence Learning
Story Generation
I Took a Plane to a Castle
Then I Ate at a Restaurant
Rehearsal
c
Anticipation Task
Valid Trials (80%)
Cue: 4 s
A
Probe: 4 s
Sequence Cue and Starting Scene Category
What's Coming Up Sooner?
Invalid Trials (20%)
Cue: 4 s
A
Probe: 4 s

Fig. 2. Task schematic. An overview of the paradigm is shown in (a). All participants learned the two category sequences (see Fig. 1) and then immediately completed 160 trials of the anticipation task. In Experiment 2 only, participants returned either 1 day or 1 week later and completed another 320 trials of the anticipation task. In sequence learning (b), participants learned the two sequences by generating stories that linked adjacent scene categories. After story generation, participants rehearsed their stories (see text for details). Finally, we tested sequence learning with a recall test for each sequence (not shown). In the anticipation task (c), participants were cued with an image from one of the learned categories along with a sequence cue (A or B) for 4 s. They were then probed with two images and had 4 s to indicate which of the two scene categories was coming up sooner in the cued sequence, relative to the cue image. The correct answer could be one to four steps away from the cue image. On valid trials (80% of trials), images in both the cue and probe screen had the filter from the cued sequence. In this example, based on the sample sequence in Figure 1, both the cue and probe filters are mosaic, the filter assigned to Sequence A. On invalid trials (20% of trials), images in the cue screen had the filter from the cued sequence, but images in the probe screen had the filter from the uncued sequence. In this example, the cue filter is mosaic (assigned to Sequence A in our example), but the probe filter is sponge (assigned to Sequence B in our example). Participants performed this anticipation task for both Sequence A and Sequence B in interleaved blocks, with the order of the cue images randomized within each block.

saw all categories from Sequence A in order, with the sequence-specific filter applied, and then all categories from Sequence B in order, with the sequence-specific filter applied. This procedure occurred three times for a total of six presentations for each sequence. The exemplar of a given scene category was different on each presentation of a given sequence. Finally, participants were prompted to recall the order of Sequence A and Sequence B by writing the order of the categories in provided text boxes. If participants could not remember a category, they were instructed to write "don't know" in the text box. In total, the sequence-learning task took approximately 30 min to complete.

Anticipation task. During the anticipation task, participants were cued with an exemplar of a scene category on the screen, along with a sequence cue (A or B) for 4 s (Fig. 2c). The cue image always had the assigned sequence-specific filter applied to it. Participants were then probed with two exemplars of upcoming categories and were told to indicate which of the two categories was coming up sooner in the cued sequence, relative to the cue image (Fig. 2c). The correct probe could be one to four steps away from the cue image ("steps into the future" variable), and the correct and incorrect probes could be one to four steps away from each other in the sequence (granularity variable). Because the sequences were circular, participants could be cued with any of the scene categories and probed with successors up to eight steps away. Participants had 4 s to respond. On 80% of trials, the probe images had the correct sequence-specific filter applied to them (e.g., a mosaic filter on a mosaic sequence trial; valid trials). However, on 20% of trials, the probe images had the filter from the other sequence applied to them (e.g., a sponge filter on a mosaic sequence trial; invalid trials; Fig. 2c). Together, the cue and the probe screens comprised a single trial, with a 1-s intertrial interval between trials.

Participants performed 160 trials of the anticipation task, with 128 valid trials and 32 invalid trials. The correct answer was equally distributed across steps into the future (one to four) and granularity (one to four), and valid and invalid trials were equally distributed across each step into the future and granularity conditions. Participants performed the anticipation task for both Sequence A and Sequence B in alternating blocks (with the starting sequence counterbalanced across participants). There were eight blocks (10 trials each) of Sequence A anticipation and eight blocks (10 trials each) of Sequence B anticipation. In each block, participants were cued with unique exemplars of each of the 10 scene categories in the sequences in a randomized order. Participants were never cued with the same image (same scene category, exemplar, and filter) more than once. Images were not reused from the sequence-learning phase and were, at most, shown once as a cue and once as a probe during the anticipation task. Participants were given three 60-s breaks spaced evenly throughout the task. The anticipation task took 27 min to complete.

Following the anticipation task, participants completed a short post-task questionnaire in which we asked them about their strategies during the task and whether or not they noticed the sequence-to-perceptual-feature mapping.

Analyses. All analyses were conducted in the R programming language using generalized linear and linear mixed-effects models (GLMMs and LMMs)—the *glmer* and *lmer* functions in the *lme4* package (Bates et al., 2014). For analyses that modeled multiple observations per participant, such as accuracy or response time on a given trial, models included random intercepts and slopes for all within-participant effects. All response time models examined responses on correct trials only. For analyses that modeled summary statistics, such as inverse efficiency, models included random intercepts and slopes for all within-participant main effects, but not interactions. Significance of mixed-effects models was assessed using the *summary()* function from the *lmerTest* package (Kuznetsova et al., 2017), which estimates degrees of freedom using Satterthwaite's method (Satterthwaite, 1941) and obtains t statistics and p values for beta coefficients.

We first checked whether accuracy and response time differed across the sequences (A and B) and the perceptual filters (mosaic and sponge). To examine sequence effects, we fitted separate models for accuracy (a GLMM) and response time (an LMM) as a function of sequence (A = −0.5, B = 0.5). We used the following R-based formulas ("participant" indicates participant number):

$$\text{glmer}\begin{pmatrix}\text{correct} \sim \text{sequence} + \left(1 + \text{sequence} \mid \text{participant}\right), \\ \text{family} = \text{“binomial,” data}\end{pmatrix}$$

$$\text{lmer}\begin{pmatrix}\text{RT} \sim \text{sequence} \;+\; \left(1 \;+\; \text{sequence} \mid \text{participant}\right), \\ \text{data, subset} \;=\; \left(\text{correct} \;==\; 1\right)\end{pmatrix}. \tag{1}$$

To examine effects of the perceptual filters, we fitted separate models for accuracy (a GLMM) and response time (an LMM) as a function of perceptual filter (mosaic = −0.5, sponge = 0.5). We used the following R-based formulas:

$$\text{glmer}\begin{pmatrix}\text{correct} \sim \text{filter} + \left(1 + \text{filter} \mid \text{participant}\right), \\ \text{family} = \text{“binomial,” data}\end{pmatrix}, \text{ and}$$

$$\text{lmer}\begin{pmatrix}\text{RT} \sim \text{filter} + \left(1 + \text{filter} \mid \text{participant}\right), \\ \text{data, subset} = \left(\text{correct} = 1\right)\end{pmatrix}. \tag{2}$$

For our primary analyses, we used inverse efficiency as our dependent variable (average response time divided by accuracy; Townsend & Ashby, 1978). We opted to use inverse efficiency because performance generally got worse, as measured by both response times and accuracy, with further steps into the future (e.g., slower response times and less accurate responses with further steps in the future). Thus, inverse efficiency effectively captured both aspects of performance. Importantly, we did not observe any evidence for speed/accuracy trade-offs. Separate accuracy and response time models are reported in the Supplemental Material.

Because we had a priori hypotheses about the interaction between steps into the future and trial type (valid vs. invalid), we calculated this inverse-efficiency score for each participant in each trial type and steps-into-the-future bin. Finally, we conducted an LMM predicting inverse efficiency as a function of steps into the future (−1.5 = 1 step, −0.5 = 2 steps, 0.5 = 3 steps, 1.5 = 4 steps), trial type (−0.5 = valid trials, 0.5 = invalid trials), and their interaction. We used the following R-based formula:

$$\text{lmer}\left(\begin{matrix}\text{inverseEfficiency} \sim \text{trialType} * \text{steps} + \\ \left(1 + \text{trialType} + \text{steps} \mid \text{participant}\right), \text{data}\end{matrix}\right). \quad (3)$$

On the basis of our predictions in the Introduction, we hypothesized that we would find (a) a main effect of trial type showing that responses would be more efficient for trials with a perceptual-filter match than for trials with a mismatch; (b) a main effect of steps showing that responses would be more efficient for closer (vs. further) future events; and (c) an interaction between trial type and steps, showing that the effect of trial type would be stronger for closer (vs. further) future events.

Finally, we used an LMM to model inverse efficiency as a function of granularity (−1.5 = 1 step, −0.5 = 2 steps, 0.5 = 3 steps, 1.5 = 4 steps). We used the following R-based formula:

$$\text{lmer}\left(\begin{matrix}\text{inverseEfficiency} \sim \text{granularity} \\ + \left(1 + \text{granularity} \mid \text{participant}\right), \text{data}\end{matrix}\right). \quad (4)$$

We did not include granularity or sequence (A vs. B) in our main inverse-efficiency model because we did not have enough trials to separately estimate the effects of steps into the future, trial type, granularity, and sequence in the same model; further, there were no significant interactions between granularity and any other variable or between sequence and any other variable. However, because steps into the future, granularity, and sequence were orthogonalized, our design effectively controls for these variables at each step in the future. We return to the effects of granularity in the Results section and the General Discussion.

Results

Sequence learning. Participants successfully learned the category sequences, achieving an accuracy of 84.33% (95% confidence interval, or CI = [81.61%, 87.05%]) on test trials from the rehearsal phase of the sequence-learning task—greater than chance level of 50%, $t(99)$ = 25.042, $p < .001$. Accuracy was high across both sequences (Sequence A: 85.3%, Sequence B: 81.5%), but was significantly higher for Sequence A (i.e., the first learned sequence) than Sequence B (i.e., the second learned sequence), β = −0.313, 95% CI = [−0.537, −0.089], z = −2.746, p = .006. When asked to explicitly recall the category orders, participants' average accuracy was 97%, and recall accuracy was not significantly different between sequences (Sequence A: 96.9%; Sequence B: 97%; V = 78.5, p = .488; because most participants scored 9/10 or 10/10 on recall, we used a Wilcoxon signed-rank test to account for nonnormality). Overall, these results verify that participants learned the order of both sequences.

Anticipation task.

Overall task performance. We first conducted analyses to ensure that participants performed effectively on the anticipation Task. Participants successfully used their memory to anticipate upcoming events (mean accuracy = 75.22%; 95% CI = [73.2%, 77.3%]), $t(99)$ = 72.19, p < .00001 (compared to chance performance). Participants were overall more accurate for Sequence B than for Sequence A, β = 0.149, 95% CI = [0.044, 0.255], z = 2.778, p = .005, although accuracy was high and significantly above chance for both sequences—Sequence A: 74%, $t(99)$ = 21.117, p < .00001; Sequence B: 76.4%, $t(99)$ = 23.22, p < .00001. Response times did not significantly differ between the two sequences, β = −23.43, 95% CI = [−56.718, 9.806], $t(97.63)$ = −1.501, p = .137. Furthermore, neither accuracy, β = −0.063, 95% CI = [−0.161, 0.035], z = −1.266, p = .206, nor response times, β = 26.12, 95% CI = [−3.777, 56.027], $t(91.72)$ = 1.712, p = .09, differed between trials with a mosaic filter and trials with a sponge filter.

Primary analyses. Turning to our hypothesized effects of interest, we calculated each participant's inverse efficiency (i.e., average response time divided by accuracy; Townsend & Ashby, 1978) for each step into the future and conducted an LMM (see Fig. 3a for model coefficients; see the Supplemental Material for separate accuracy and response time models). Steps into the future robustly affected inverse efficiency, with less efficient responses for progressively further steps into the future, β = 310.41, 95% CI = [243.846, 376.974], $t(102.47)$ = 9.140, p < .000001 (Fig. 3b). Granularity also influenced inverse efficiency, with less efficient responses when the number of steps between the probes was smaller, β = −64.60, 95% CI = [−99.822, −29.37], $t(99)$ = −3.594, p = .0005.

Participants therefore accurately anticipated scene categories multiple steps into the future. How visually detailed were their predictions? To determine whether participants were only anticipating the upcoming category or were also representing sequence-specific perceptual features, we investigated whether trial type affected performance and whether this interacted with steps into the future. If participants anticipate visually

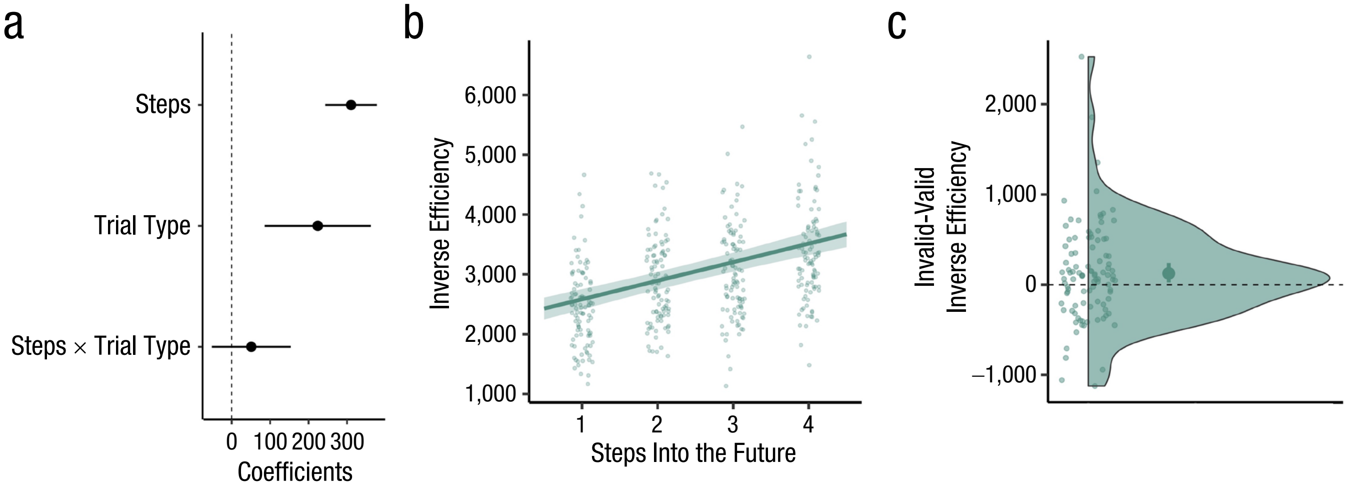

Fig. 3. Anticipation task performance in Experiment 1. Coefficient estimates with 95% confidence intervals for the inverse-efficiency model are shown in (a); predictors in the model were effect coded. In (b), we show inverse efficiency as a function of the number of steps between the cue and the correct probe (i.e., steps into the future). Higher inverse-efficiency values indicate less efficient (slower and/or less accurate) responses. Responses were less efficient when the correct probe was more (vs. fewer) steps away from the cue. Green lines and error ribbons indicate model predictions with 95% confidence intervals; green points indicate each participant's average inverse efficiency for each step into the future. Inverse-efficiency differences between valid and invalid trials (trial type) are shown in (c). The dashed line at 0 indicates equally efficient responses on valid and invalid trials. Responses on invalid trials were overall less efficient compared to valid trials, as indicated by a positive difference score. Small green points indicate the average inverse-efficiency difference for each participant. The large green points indicates the average inverse-efficiency difference across participants with 95% confidence intervals.

specific information, then performance should decrease on invalid trials (in which participants are probed with an incorrect sequence filter) compared to valid trials (in which the perceptual filter matches learned expectations). Consistent with our hypothesis, participants responded less efficiently to invalid versus valid trials, β = 224.01, 95% CI = [87.069, 360.950], $t(145.35) = 3.206$, $p = .002$ (Fig. 3c). This cost to performance is striking because the sequence-specific filters were not relevant to the correct answer, and when asked in the post-task questionnaire, most participants reported not noticing the sequence-to-filter mapping (68 did not notice; 32 noticed). Importantly, the effect of trial type on inverse efficiency was not different between participants who noticed the manipulation compared to those who did not, β = −203.44, 95% CI = [−495.686, 88.798], $t(144.59) = -1.364$, $p = .175$.

Finally, we tested whether the effect of valid versus invalid trials was larger for closer versus further predictions. If anticipated information declines in vividness or detail the further in the future it is, then trial validity should have a larger effect on nearby versus farther-away predictions. Contrary to our hypothesis, however, there was no interaction between steps into the future and trial type on inverse efficiency, β = 51.12, 95% CI = [−51.542, 153.786], $t(597.25) = 0.976$, $p = .330$.

Hence, the validity manipulation demonstrates that responses were less efficient on trials in which participants were probed with the perceptual filter from the incorrect (vs. correct) sequence. We interpreted this efficiency cost as suggesting that individuals were incorporating perceptual features into their sequence representations and experienced a decrease in performance when those perceptual features were not presented. However, an alternative explanation of the invalid versus valid behavioral cost is a *perceptual oddball effect*. Because valid trials were more common, and because perceptual filters were consistent for a given sequence during learning, the sudden shift from the cue's perceptual filter to the probes' filters on the infrequent invalid trials may be unexpected. This may lead to a performance cost because of the perceptual surprise of suddenly changing filters—even if no perceptual features were incorporated into anticipated information. We conducted a control experiment with 101 participants to determine whether such a perceptual oddball effect could explain the validity effect we observed (see the Supplemental Material for details). In this experiment, individuals were first exposed to images in the same category order and with the same perceptual filters as in our sequence-learning task, but were not explicitly taught the category sequence order and performed only an indoor versus outdoor cover task. Individuals then performed a task analogous to our anticipation task, in which they were first cued with a scene and then saw two upcoming scenes on the screen. Critically, individuals were simply told to indicate which of the two scenes was from the same indoor versus outdoor category as the cued scene. Thus, in this experiment, individuals were not asked to make predictions about upcoming scene categories. If the difference in inverse efficiency for valid versus invalid trials reflects a perceptual oddball effect rather than incorporation of perceptual details into predictions of upcoming scene categories, then we should observe the same cost for invalid versus valid trials when participants are making indoor versus outdoor judgments.

In this control experiment, there was no effect of distance into the future on efficiency, $\beta = -8.513$, 95% CI = [−25.186, 8.160], $p = .318$ (see Figs. S1a and S1b in the Supplemental Material), showing that we eliminated our behavioral hallmark of multistep anticipation when individuals were performing an unrelated task. Critically, we found that there was no effect of trial type (invalid vs. valid trials) on the efficiency of indoor versus outdoor judgments, $\beta = 12.294$, $p = .503$, 95% CI = [−23.563, 48.152] (see Fig. S1c in the Supplemental Material). Thus, we eliminated the valid-invalid gap when individuals were making scene judgments but not anticipating upcoming scene categories. This suggests that our original findings were unlikely to be due to the perceptual surprise of a filter swap and instead are consistent with our interpretation that participants were incorporating perceptual features into their prediction of sequence structure.

Together, our results show that individuals accurately anticipate events multiple steps in the future, but do so less efficiently for further steps, and that anticipated information at multiple timescales contains task-irrelevant perceptual features regardless of whether participants explicitly noticed these features and regardless of how far into the future participants anticipated.

Experiment 2

Having shown that individuals anticipate events multiple steps into the future with perceptual detail in Experiment 1, we next investigated how these effects change with consolidation. We ran a second experiment in which participants completed the anticipation task both immediately after sequence learning (to replicate Experiment 1), and either 1 day or 1 week later, to determine how varying lengths of consolidation influence multistep anticipation.

Method

Participants. Our target was to double the sample size of Experiment 1, with participants split between two delay conditions. We recruited 355 participants through Prolific (www.prolific.co) to meet our target sample size. Sixty-two participants did not complete the second session of the experiment (21 participants in the 1-day condition and 41 participants in the 1-week condition), leaving 293 participants who completed the full experiment. An additional 6 participants were excluded from data analysis because they failed to respond on more than 50% of trials, and 83 participants were excluded from data analysis because they did not perform statistically above chance on the anticipation task during the first session (56.875%, as determined by a binomial test; see the Procedure section). Applying these exclusions resulted in 204 participants (1-day condition: $n = 99$, $M_{\text{age}} = 31.65$ years, $SD = 5.03$, $M_{\text{education}} = 15.18$ years, $SD = 1.85$; 1-week condition: $n = 105$, $M_{\text{age}} = 30.65$ years, $SD = 6.73$, $M_{\text{education}} = 14.99$ years, $SD = 2.12$; see Table 1 for demographic information).

To be eligible for the experiment, participants had to report that they were between the ages of 18 and 40, fluent English speakers, and resided in the United States. Participants were compensated $8 per hour for participating in session 1 and $8.50 per hour for participating in session 2. All participants provided informed consent and all procedures were in accordance with the policies of the Institutional Review Board at Columbia University.

Stimuli. Stimuli were identical to Experiment 1.

Procedure. The experiment was conducted on the Gorilla platform (www.gorilla.sc; Anwyl-Irvine et al., 2020) and was composed of three tasks: sequence learning, session 1 anticipation, and session 2 anticipation (Fig. 2a).

Sequence learning. The sequence-learning task was identical to that of Experiment 1.

Session 1 anticipation task. The Session 1 anticipation task was identical to that of Experiment 1, except that participants did not complete the post-task questionnaire following the task.

Session 2 anticipation task. Half of the participants were invited to return 1 day later and half were invited to return 1 week later to take part in the Session 2 anticipation task. Participants were informed about the opportunity to return for a follow-up experiment the evening before they could complete the Session 2 test. The task was identical to the Session 1 anticipation task except that the total number of trials was increased to 320, with 256 valid trials and 64 invalid trials. Consequently, there were 16 blocks of Sequence A anticipation (10 trials each) and 16 blocks of Sequence B anticipation (10 trials each). Sequence A and Sequence B blocks alternated, with the starting sequence counterbalanced across participants.

Participants were given six 60-s breaks spaced evenly throughout the task. The Session 2 anticipation task took 60 min to complete. Following the task, participants completed the posttask questionnaire.

Analyses. All analyses were identical to Experiment 1's, with the following exceptions. In our primary inverse-efficiency model, we included (a) delay (immediate test = –0.5, delayed test = 0.5) as a main effect and a random effect and (b) the interaction between delay, steps into the future, and trial type. Additionally, to control for the varying delay lengths (1 day or 1 week) and to determine whether they differentially influenced any consolidation-dependent effects, we included delay length (–0.5 = 1-day delay, 0.5 = 1-week delay) as a main effect and the interaction between delay length, delay, steps into the future, and trial type. We report significant interactions with delay length in the Results section. However, because there were no major differences between the 1-day and 1-week delay, we report the main results collapsed across delay length. As in Experiment 1, we did not observe any evidence for speed and accuracy tradeoffs. Separate accuracy and response time models are reported in the Supplemental Material.

On the basis of our predictions in the Introduction, we hypothesized that we would find (a) an interaction between trial type and delay, such that the effect of perceptual filter (match vs. mismatch) would be less strong after (vs. before) a delay, reflecting loss of perceptual details over time, and (b) an interaction between steps and delay, such that the effect of closer (vs. further) future events on response efficiency would be less strong after (vs. before) a delay, reflecting schematization of memory over time.

As in Experiment 1, we report the main effect of granularity on inverse efficiency, but we did not include granularity or sequence in our main inverse-efficiency model after confirming there were no significant interactions between granularity and other variables or between sequence and other variables.

Results

Sequence learning. To verify that participants learned both sequences, we calculated accuracy on test trials from the rehearsal phase of the sequence-learning task (see the Method section). As in Experiment 1, accuracy was significantly higher than chance performance (50%) on the test trials, 82.62%, 95% CI = [80.50%, 84.73%], $t(203) = 30.457$, $p < .000001$. Accuracy was high across both sequences (Sequence A: 79.4%, Sequence B: 83.7%), but was significantly higher for Sequence B (i.e., the second learned sequence) than Sequence A (i.e., the first learned sequence; $\beta = 0.334$, 95% CI = [0.183, 0.485], $z = 4.325$, $p = .00002$).

Next, we calculated recall accuracy for sequences A and B. Participants' average recall accuracy was 96.5%. Recall accuracy was not significantly different between sequences (Sequence A: 96.0%; Sequence B: 97.1%; $V = 324$, $p = .675$); because most participants scored 9/10 or 10/10 on recall, we used a Wilcoxon signed-rank test to account for nonnormality). Overall, these results verify that participants learned the order of both sequences.

Anticipation task.

Overall task performance. We first ensured that participants were performing effectively on the anticipation task for both the immediate and delayed tests in both delay-length conditions. Accuracy was significantly higher than chance performance (50% accuracy) in the anticipation task for Session 1 (i.e., the immediate test) and the anticipation task for Session 2 (i.e., the delayed test) in both the 1-day and 1-week conditions (all *p*s < .00001).

Accuracy was not significantly different between sequences A and B, $\beta = -0.013$, 95% CI [–0.080, 0.054], $z = -0.377$, $p = .706$. Response times, however, were slower for Sequence B compared to Sequence A (i.e., the second vs. first learned sequence), $\beta = 22.264$, 95% CI = [8.387, 35.918], $t(209.533) = 3.251$, $p = .001$, with this effect becoming larger at the delayed test compared to the immediate test, $\beta = 30.203$, 95% CI = [12.764, 47.643], $t(101625.549) = 3.394$, $p = .0007$. Neither accuracy, $\beta = -0.038$, 95% CI = [–0.084, 0.008], $z = -1.596$, $p = .111$, nor response time, $\beta = 0.305$, 95% CI = [–10.74, 11.35], $t(167.105) = 0.054$, $p = .957$, differed between trials with a mosaic filter and trials with a sponge filter.

Primary analyses. Turning to our hypothesized effects of interest, we calculated inverse-efficiency scores to capture both accuracy and response time in a single measure. We then conducted an LMM assessing whether delay, delay length, steps into the future, trial type, and their interactions influenced inverse efficiency on the anticipation task (see Fig. 4a for model coefficients; see the Supplemental Material for separate accuracy and response time models).

Delay robustly influenced inverse efficiency, with participants' responses becoming more efficient in the delayed versus immediate test, $\beta = -318.413$, 95% CI = [–403.126, –233.699], $t(194.788) = -7.367$, $p < .000001$ (see Fig. 4b). There was a significant interaction between delay and delay length such that efficiency in the delayed (vs. immediate) test improved more at the 1-day (vs. 1-week) delay, $\beta = 161.470$, 95% CI = [7.206, 315.735], $t(304.029) = 2.051$, $p = .041$, but there were

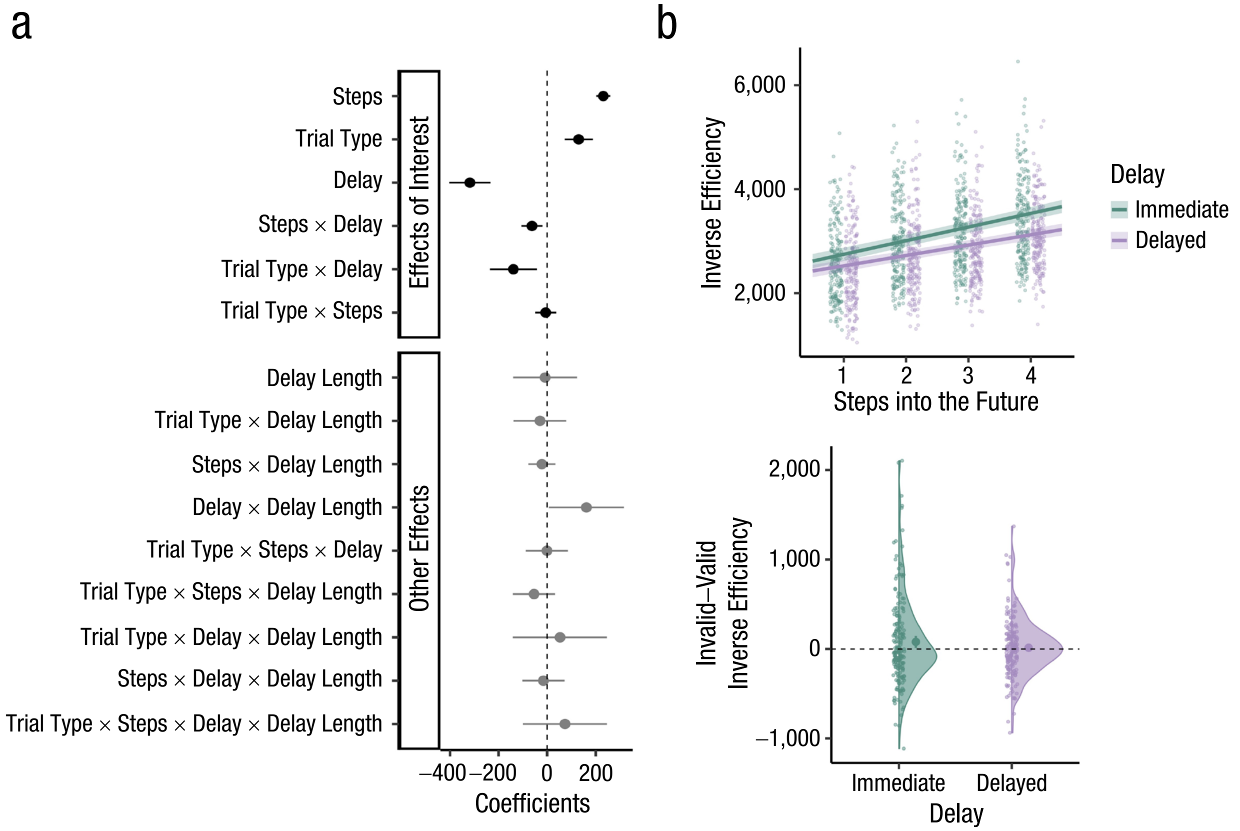

Fig. 4. Anticipation-task performance in Experiment 2. Coefficient estimates with 95% confidence intervals for the inverse-efficiency model are shown in (a). The black points indicate effects of interest, and the gray points indicate all other effects in the model. Predictors in all the models were effect coded. Inverse efficiency as a function of steps into the future and delay is shown in (b). Higher inverse-efficiency values indicate less efficient responses. Responses were more efficient in the delayed (vs. the immediate) test, with the inverse-efficiency benefit for the delayed (vs. immediate) test increasing with steps into the future. Green lines and shaded bands indicate model predictions with 95% confidence intervals for the immediate test; green points indicate each participant's inverse-efficiency score for each step into the future for the immediate test. The same conventions were used for the delayed test, plotted in purple. Inverse-efficiency differences between valid and invalid trials as a function of delay are shown in (c). Participants were less efficient for invalid than valid trials at the immediate test (replicating the results of Experiment 1), but the efficiency difference between valid and invalid trials decreased after a delay. The dashed line at 0 indicates equally efficient responses on valid and invalid trials. Small points indicate the average inverse-efficiency difference for each participant at each delay. The large points indicate the average inverse-efficiency difference across participants with 95% confidence intervals.

no other interactions with delay length. This suggests that delay-dependent effects of other variables (e.g., steps into the future, trial type) on memory were similar for the 1-day and 1-week delays.

Replicating Experiment 1, both granularity and steps into the future affected inverse efficiency: There were less efficient responses when the number of steps between the probes was smaller, β = −64.254, 95% CI = [−76.584, −51.924], *t*(582.052) = −10.21, *p* < .000001, and when the correct answer was further into the future, β = 231.175, 95% CI = [202.401, 259.950], *t*(191.0712) = 15.746, *p* < .000001 (see Fig. 4b).

As hypothesized, there was a significant step into the future by delay interaction: Inverse efficiency improved the most at the delayed test (vs. the immediate test) for further steps into the future, β = −62.635, 95% CI = [−105.992, −19.279], *t*(2660.211) = −2.831, *p* = .005 (Fig. 4b). There was no three-way interaction between steps into the future, delay, and delay length; this suggests that delay-dependent changes in the effect of steps into the future on inverse efficiency were similar at the 1-day and 1-week tests, β = −15.525, 95% CI = [−140.943, 32.483], *t*(2660.211) = −0.351, *p* = .726.

Therefore, participants became more efficient in their responses following consolidation, and this efficiency improvement was larger for further steps into the future. Was this improvement in anticipation for further steps accompanied by forgetting of sequence-specific perceptual details? To answer this question, we again investigated whether trial type (valid vs. invalid) affected performance, and whether this interacted with delay, steps into the future, and delay length (see Fig. 4a for model coefficients). Replicating Experiment 1, participants responded less efficiently to invalid versus valid trials, β = 129.683, 95% CI = [71.837, 187.529], *t*(319.046) = 4.394, *p* = .00002 (see Fig. 4c). Critically, there was also a trial type by delay interaction—the difference in efficiency between invalid and valid trials was smaller in the delayed (vs. the immediate) test, β = −139.373, 95% CI = [−236.321, −42.424], *t*(2660.211) = −2.818, *p* = .005 (Fig. 4c). Contrary to our hypothesis but consistent with Experiment 1, there was no interaction between trial type and steps into the future, β = −6.171, 95% CI = [−49.528, 37.185], *t*(2660.211) = −0.279, *p* = .780.

Together, our results show that anticipation became more efficient following consolidation, with a larger efficiency improvement for further steps into the future compared to closer ones. Consolidation also reduced the extent to which sequence-specific perceptual features were incorporated into anticipated information, as indicated by the more limited impact of the perceptual filter applied to the probe images.

Repeated practice or consolidation? In the current study, individuals repeatedly anticipated upcoming categories in the sequence structure in the immediate session (i.e., within Day 1), prior to our consolidation manipulation. This raises the possibility that some of our observed effects in comparing the first session to subsequent sessions were the result of repeated practice rather than the result of memory processes occurring during the consolidation delay. Alternatively, repeated retrieval and consolidation may have jointly contributed to our observed effects on memory, consistent with theories that retrieval and consolidation are closely intertwined and can cooperatively influence memory retention (Antony et al., 2017). To investigate these possibilities in our data, we examined whether early versus late trials within and across sessions revealed differing contributions of practice and delay to memory performance.

We found that efficiency on our multistep anticipation task improved during the immediate session (i.e., within Day 1) from early to late trials (first vs. second half, β = −286.350, *p* < .0001, 95% CI = [−374.428, −198.273]). This is consistent with an effect of repeated practice enhancing memory. Critically, however, late trials in the immediate session (Day 1) were less efficient than early trials in the delayed session (Day 2 and Day 7; β = −302.424, *p* < .0001, 95% CI = [−403.404, −201.444]); this was also true when separately comparing the immediate session to the 1-day delay, β = 399.97, *p* < .0001, 95% CI = [−540.118, −259.828], and to the 1-week delay, β = −201.95, 95% CI = [−341.490, −62.419], *p* = .005. This improvement from the end of the first session to the beginning of the delayed session is consistent with a memory-consolidation effect, in addition to the within-session improvements that may be driven by practice effects. Further, there was a discontinuity in the time course of response times between late trials of the immediate session and early trials of the delayed session, with faster response times even in the first few trials in the delayed session compared to the immediate session (see Fig. S2 in the Supplemental Material). Together, this suggests an improvement in sequence anticipation over the course of consolidation in addition to an effect of repeated practice.

Modeling Anticipation Strategies

Memory for temporal structure therefore changes over time to support multistep anticipation. But what strategies do participants use to anticipate upcoming information? Which strategies are most beneficial for behavior, and do they change with consolidation? One class of models predicts that a participant maintains an internal representation of the entire sequence in memory and explicitly rolls out the sequence, link by link, to anticipate upcoming events (Daw & Dayan, 2014). Another class predicts that a participant will build a representation for each item that incorporates cached information about future items, with stronger cached representations for events that are coming up sooner in the sequence (Dayan, 1993).

Drawing inspiration from these models, we tested whether participants were using a link-based strategy or a cue-based strategy. In a link-based strategy, each link between items in a sequence is represented in memory and participants sequentially "traverse" the sequence of links beginning at the cue item until they reach one of the probe items. A weak (vs. strong) link would be more difficult to traverse (increasing response time), and this slowdown would occur whenever the rollout from the cue to the nearest probe includes this link. In a cue-based strategy, information about future items becomes embedded in the representation of each cue, with closer upcoming items being more strongly represented. Identifying which of two probes is coming up sooner would be accomplished by directly comparing the probe items to this learned cue representation to determine which is more similar. Here, response times should depend on the quality of the cached representation at the given cue item, and trials for neighboring cues in the sequence could yield very different response times. Thus, the critical difference between these two models lies in whether response times are driven by the starting point of the anticipation judgments (cue model) or by the connecting links that need to be traversed (link model). The link model predicts that each link between scene categories is associated with a fixed response time, so that traversing that link yields similar response times across trials regardless of the starting point in the sequence (Fig. 5a and 5b). On the other hand, the cue model predicts that the response

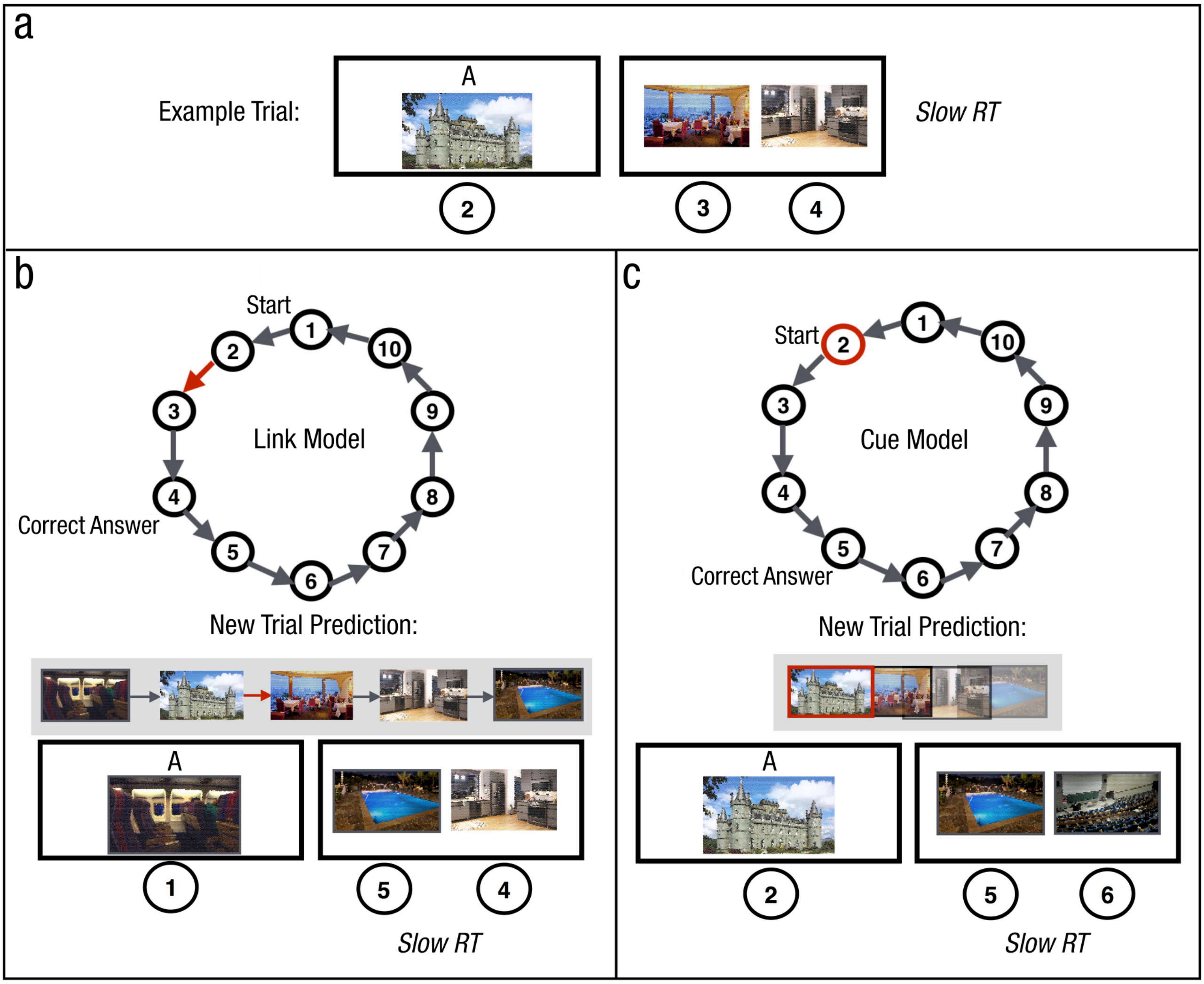

Fig. 5. Schematic of possible anticipation strategies. On this sample trial (a), when a participant was cued with Scene 2 and probed with Scenes 3 and 4, they showed a slower-than-average response time (RT) when identifying that Scene 3 is coming up sooner in the sequence. The link model (b) would explain this poor performance as coming from a weak connection between Scene 2 and Scene 3 (red arrow), and would therefore predict that any trial involving that link (such as being cued with Scene 1 and probed with Scenes 4 and 5) would lead to a slower-than-average response time. The cue model (c), in contrast, assumes that the representation of each scene category contains cached information about upcoming scene categories that is used to make anticipatory judgments. The slow response time in (a) is therefore the result of a poor representation for Scene 2 (red circle), and any trial in which Scene 2 is the cue would lead to a slower-than-average response time.

time penalty for making predictions at progressively farther distances is specific to each cue and is not tied to the specific links being traversed (Fig. 5a, 5c).

To adjudicate between these two strategies, we modeled response times using both a link model and a cue model, and tested which model better explained anticipation behavior (Fig. 5). We hypothesized that the link model would better predict response times in the anticipation task because we asked participants to generate stories about the sequences by creating links between the scene categories. In contrast, a cue-based model may be a more useful and efficient approach when sequences are nonoverlapping (so that transitions between states are not context-dependent) and when there is time pressure for very rapid responses (limiting the opportunity to explicitly roll out future states with a link-based approach; Gershman, 2018). We also hypothesized that using a link model rather than a cue model would benefit anticipation task performance, because the link model is analogous to a model-based strategy (Daw & Dayan, 2014), which is more computationally expensive but generally leads to higher accuracy than cue-based (caching) models (Momennejad et al., 2017).

Method

Analyses. Using the data reported in Experiment 1 and Experiment 2, we created two different kinds of linear

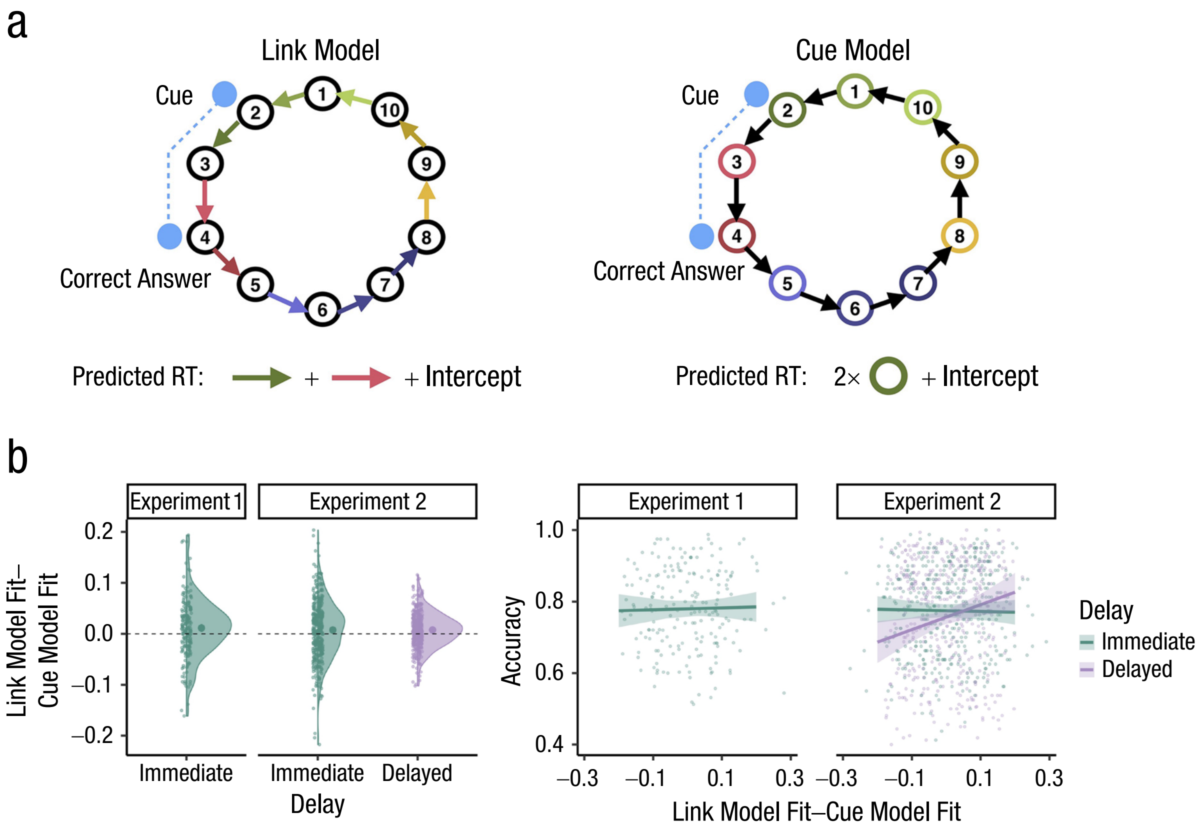

Fig. 6. Link and cue model schematic and results. A model schematic for the link and cue models is shown in (a). Open circles represent scene categories, and lines represent the link between the categories in the sequence. In the link model (left), trial-wise response times were predicted as a function of the links between the cue image and the correct answer. In this example, the predicted response time (RT) depends on link-specific values for the green link (from Category 2 to 3) and the red link (from Category 3 to 4). In the cue model (right), trial-wise response times were instead assumed to vary as a function of the cued category, with linearly increasing response times based on the distance to the correct answer. In this example, the predicted response time depends on a cue-specific value for the green cue (Category 2) modulated by the number of steps to the correct answer (here, two steps). Link model fits were better than cue model fits (b), as indicated by a positive difference score; this difference was significant across both experiments and both test delays in Experiment 2 (immediate and delayed). The dashed line at 0 indicates equal cue and link model fits. Small points indicate each participant's model-fit difference for each experiment and delay. The large points indicate the average model-fit difference across participants with 95% confidence intervals. Higher link (vs. cue) model fits in (c) predict accuracy on the anticipation task, but only in the delayed (vs. the immediate) test. Lines and error bands indicate model predictions with 95% confidence intervals. Small points indicate each participant's model-fit difference at each delay.

models in the R programming language. We modeled response times, as opposed to inverse efficiency, in these models because they required trial-by-trial measures, and inverse efficiency is a summary statistic. We modeled response times only on correct trials to ensure that participants were successfully using learned information.

For each participant, we first modeled response times as a function of the links that participants crossed to get from the cue to the correct answer in the probe during each anticipation task trial. In this link model (Fig. 6a), each link between scene categories in a given sequence was modeled separately according to whether or not that link would be traversed from the cue to the correct answer on a given trial (0 = *not used*, 1 = *used*). Because the links between scene categories were different in each sequence, we created separate models for Sequence A and Sequence B. For Experiment 2, we also modeled response times separately for the immediate and delayed sessions. We used the following R-based formula, with separate regressors for each link between adjacent scene categories in the circular sequence:

$$\mathrm{lm}\left(\begin{array}{l}\mathrm{RT} \sim \mathrm{link1}+\mathrm{link2}+\mathrm{link3}+\mathrm{link4}+\mathrm{link5}+\mathrm{link6} \\ +\mathrm{link7}+\mathrm{link8}+\mathrm{link9}+\mathrm{link10},\ \mathrm{data}, \\ \mathrm{subset}=(\mathrm{correct}==1)\end{array}\right). \tag{5}$$

We next modeled each participant's response times as a function of the scene category that they were cued with during each trial in the anticipation task. This model assumed that increasingly distant future states would be increasingly difficult to access (consistent with temporal discounting; Gershman et al., 2012) but that the degree of difficulty could be different for different cues. In this cue model (Fig. 6a), each scene category in a given sequence was modeled separately, coded by whether or not that category was the cue on a given trial and, if it was cued, how many steps into the future the correct answer was (0 = *not cued*, 1 = *cued on a 1-step trial*, 2 = *cued on a 2-step trial*, 3 = *cued on a 3-step trial*, 4 = *cued on a 4-step trial*). We again modeled response times separately for Sequence A and Sequence B (because their future states differed) and, in Experiment 2, separately for the immediate and delayed test. We used the following R-based formula, with separate regressors for each scene category:

$$\mathrm{lm}\left(\begin{array}{l}\mathrm{RT} \sim \mathrm{cue1} + \mathrm{cue2} + \mathrm{cue3} + \mathrm{cue4} + \mathrm{cue5} \\ + \mathrm{cue6} + \mathrm{cue7} + \mathrm{cue8} + \mathrm{cue9} + \mathrm{cue10}, \\ \mathrm{data},\ \mathrm{subset}\ =\ \left(\mathrm{correct}\ ==\ 1\right)\end{array}\right). \tag{6}$$

We assessed goodness of fit for both the link and the cue models by creating null models for each participant in which the data were fitted to a model in which the cue identities were randomly shuffled. We created 100 null models for each participant by repeating this procedure 100 times. Next, we calculated the R^2 from the real model and each of the 100 null models for each participant. We then created a model-fit score for each participant and each sequence by calculating the difference between the R^2 of the real model and the average R^2 of the null models. To test whether the link and cue model fits were significantly better than the permuted model fits, we conducted one-sample *t* tests comparing the model-fit score for each model, averaged across sequences, to 0.

Next, to determine whether the link or the cue model provided a better fit to response times and whether this differed by delay, we created a model difference score by subtracting each participant's cue-model-fit score from the link-model-fit score, separately for each sequence. We then predicted this model difference score as a function of sequence (Experiment 1), or as a function of delay, delay length, their interaction, and sequence (Experiment 2). In these models, the intercept term would provide evidence for better link- or cue-model fits overall across our sample, whereas the delay term would provide evidence for a shift in model fit with consolidation. We used the following R-based formula:

$$\mathrm{lmer}\left(\begin{array}{l}\mathrm{modDiff} \sim 1 + \mathrm{delay} * \mathrm{delayLength} \\ + \mathrm{sequence} + \left(1 \mid \mathrm{participant}\right),\ \mathrm{data}\end{array}\right). \tag{7}$$

We hypothesized that the link model would provide a better fit to response times than the cue model for two reasons: (a) Participants were encouraged to create stories consisting of pairwise links between adjacent images in the sequences; and (b) cue-based models may be a more useful and efficient approach when sequences are nonoverlapping and there is time pressure for very rapid responses (Gershman, 2018). We were agnostic as to whether the link-model versus cue-model fit would change with delay.

Finally, to determine which strategy yielded superior behavioral performance, we investigated whether individual differences in participants' model difference scores predicted accuracy on the anticipation task. We focused on accuracy to have a dependent measure that is independent from response times, which were the models' outcome variable. We predicted participants' average accuracy as a function of their model difference score and sequence (Experiment 1) or model difference score, delay, delay length, their interactions, and sequence (Experiment 2). We used the following R based formula:

$$\mathrm{lmer}\left(\begin{array}{l}\mathrm{accuracy} \sim \mathrm{modDiff} * \mathrm{delay} * \mathrm{delayLength} \\ + \mathrm{sequence} + \left(1 \mid \mathrm{participant}\right),\ \mathrm{data}\end{array}\right). \tag{8}$$

We hypothesized that a better link-model versus cue-model fit would predict more accurate responses on the anticipation task. This is because our link model is meant to approximate model-based strategies which, although computationally costly, enable accurate and flexible responses (Momennejad et al., 2017). We were agnostic as to whether the effect of link-model versus cue-model fits on accuracy would change with delay.

Results

We first tested whether our link model and cue model provided good fits to participants' response times in the anticipation task by comparing their model-fit scores (model R^2 vs. null models' R^2; see the Method section) to 0. In Experiment 1, the link model provided a better fit to response times than the null models—mean R^2 difference: .011; $t(99) = 2.226$, $p = .028$. The cue model, however, was not better than the null models—mean R^2 difference: −.002; $t(99) = -0.53$, $p = .594$. In Experiment 2, the link model provided a better fit to response times than the null models in both Session 1—mean R^2 difference: .0179; $t(203) = 5.106$, $p < .000001$—and Session 2—mean R^2 difference: .0182; $t(203) = 7.865$, $p < .000001$. Contrary to Experiment 1, the cue model also provided a better fit to response times than the null models for both Session 1—mean

R^2 difference: .0104; *t*(203) = 3.936, *p* =.0001138—and Session 2—mean R^2 difference: .0105; *t*(203) = 5.470, *p* = .0000001.

We next investigated whether the link or cue model performed better and whether the superior model changed with delay. The link model provided a better fit to response times than the cue model in both Experiment 1, β = 0.014, 95% CI = [0.006, 0.0215], *t*(198) = 3.48, *p* =.0006 (Fig. 6b), and Experiment 2, β = 0.008, 95% CI = [0.004, 0.012], *t*(225.9) = 4.268, *p* = .00003 (Fig. 6b). This superiority of the link model did not change across sequences—Experiment 1: β = −0.01, 95% CI = [−0.025, 0.006], *t*(198) = −1.24, *p* = .216; Experiment 2: β = −0.0001, 95% CI = [−0.005, 0.005], *t*(628.7) = 0.065, *p* = .948—nor (in Experiment 2) across delay, β = −0.0008, 95% CI = [−0.008, 0.006], *t*(259.8) = −0.259, *p* = .796; delay length, β = 0.0005, 95% CI [−0.007, 0.008], *t*(298.7) = 0.148, *p* = .883; or their interaction, β = −0.009, 95% CI = [−0.022, 0.004], *t*(331.1) = −1.367, *p* = .172.

Finally, we assessed whether individual differences in strategy use, indexed by superior fits for the link model versus the cue model, predicted average accuracy on the anticipation task and whether this effect changed with delay. In Experiment 1, differences in the link versus cue model difference score did not predict accuracy, β = 0.026, 95% CI = [−0.166, 0.219], *t*(123.6) = 0.271, *p* = .787 (Fig. 6c). In contrast, model difference scores did predict accuracy in Experiment 2, β = 0.165, 95% CI = [0.006, 0.323], *t*(666.895) = 2.039, *p* = .042 (Fig. 6c). Importantly, there was a model difference score by delay interaction, β = 0.366, 95% CI = [0.066, 0.666], *t*(653.688) = 2.394, *p* = .017 (Fig. 6c), so that higher link versus cue model difference scores did not predict accuracy in the immediate test session, replicating Experiment 1, β = 0.074, 95% CI = [−0.09, 0.239], *t*(297.291) = 0.883, *p* = .378, but did predict accuracy in the delayed test session, β = 0.299, 95% CI = [0.052, 0.547], *t*(250.673) = 2.375, *p* = .018. Delay length did not interact with model difference score, β = −0.006, 95% CI = [−0.313, 0.300], *t*(681.477) = −0.040, *p* = .968, nor was there a three-way interaction between delay length, model difference score, and delay, β = 0.2333, 95% CI = [−0.368, 0.833], *t*(654.571) = 0.760, *p* = .448.

Thus, the link model more effectively captured participants' response times than the cue model, regardless of delay. Higher link versus cue model fits also predicted accuracy in the anticipation task, but only after consolidation.

General Discussion

Memory for temporal structure is adaptive because it allows us to anticipate what is likely to happen in the future. A burgeoning line of research has explored how memory for temporal structure is represented in the brain (Bellmund et al., 2020; Hsieh & Ranganath, 2015; Kalm et al., 2013; Kalm & Norris, 2014; Lee et al., 2021) and in behavior (Drosopoulos et al., 2007; Tiganj et al., 2022) to guide adaptive future-oriented behavior. Indeed, prediction is thought to be a primary cognitive and neural function (Clark, 2013; Friston, 2005), and offline periods of memory consolidation may play a crucial role in extracting regularities that will be useful for prediction (Hobson & Friston, 2012). Our findings show how memories for such regularities are shaped by consolidation and provide an important link in understanding the adaptive function of memory.

Across two experiments, we found that multistep anticipation became more efficient with consolidation, particularly for further events. Anticipated events contained representations of task-irrelevant perceptual features, but these perceptual features had less influence on behavior after consolidation. Finally, maintaining a link-based, rather than a cue-based, model of the sequence after consolidation benefited multistep anticipation. Overall, these results shed light on how memories adaptively shift to prioritize temporal structure at the cost of perceptual details.

Our findings are consistent with, and build upon, influential theories of memory consolidation, which posit that memories shift from detailed to schematic over time (Robin & Moscovitch, 2017; Sekeres et al., 2018; Winocur et al., 2010). Schematized representations should support memory for the high-level structure of experiences (McClelland et al., 1995) but may lack perceptual detail. Indeed, we found that, with consolidation, perceptual details became more weakly represented in memory. They nevertheless still exerted an influence on behavior, suggesting that perceptual details were not entirely forgotten after consolidation (Gilboa & Moscovitch, 2021; Robin & Moscovitch, 2017). However, it is important to note that, in our experiments, temporal structure was task relevant whereas perceptual features were incidental to the task. Perceptual details may have been more strongly maintained in memory if they were task relevant (Schapiro et al., 2017).

In contrast to perceptual details, information about temporal structure was more efficiently accessed after a delay. Our findings of efficient access of multistep sequences after consolidation extends prior work showing memory improvements after consolidation for category structure (Schapiro et al., 2017) and short-timescale statistical regularities (Durrant et al., 2011). Extending consolidation-related benefits to multistep anticipation, especially for events further in the future, suggests that order judgments—which may involve retrieving a compressed representation of temporal structure (Tiganj et al., 2022)—become even more compressed over time. Thus, our work allows insights into how behavior may become increasingly adaptive with consolidation by allowing fast access to anticipated future events—particularly those that are further in the future.

We used our trial-type manipulation (valid, matching vs. invalid, mismatching perceptual filters) to probe the extent to which anticipation became less visually detailed over time, as predicted by theories of memory consolidation (Robin & Moscovitch, 2017; Sekeres et al., 2018). A related prediction is that anticipated events should be represented in a more gist-like manner over consolidation (Dudai et al., 2015). One potential way of testing that with our design is by examining whether the effect of granularity changes with delay: When the two probes are close to each other in the sequence, judgments about which is nearer should be more challenging as a memory becomes more gist-like. Although we observed a general effect of granularity, we did not observe any interactions with delay. This does not rule out increasingly gist-like memories over time; instead, in our study, gist-like memories may be reflected primarily in the loss of perceptual details rather than the loss of temporal precision in predicted events.

To further probe how temporal structure is represented in memory, we drew inspiration from reinforcement-learning models to determine what types of internal models were most useful for anticipation, both immediately and at a delay. Individuals tended to use a link-based strategy (akin to a model-based representation) rather than a cue-based strategy (akin to a successor representation), and this preference for the link model over the cue model was present in both the immediate session and after consolidation. Critically, maintaining a link (vs. cue) model after consolidation predicted accuracy on the anticipation task. This finding dovetails with predictions of model-based learning: Such models are thought to lead to highly accurate representations postconsolidation because they store a fully connected internal model of an environment (Y. Liu et al., 2021; Wimmer et al., 2023). However, such models are computationally expensive because they require an individual to traverse individual links between states (Daw & Dayan, 2014). Thus, we propose that consolidation processes may be particularly adaptive because they allow the brain to maintain and stabilize—and perhaps make more efficient—computationally intensive representations.

Although we found that the link model was the optimal internal model in our task, other tasks might be better solved with a cue model. Past studies have shown a preference for successor representation strategies (conceptually similar to our cue model) when short sequences end in monetary reward (Momennejad et al., 2017). A cue-model preference may therefore emerge if sequences end in a goal location or a reward—whereas in our task, the sequences were circular without a salient endpoint. Further, including other features in the sequence, such as event boundaries, could lead to a model in which participants cache future states within an event and skip between event boundaries, rather than traversing each link separately (Michelmann et al., 2023). Finally, other models could contain information about each item's position within the sequence in addition to, or instead of, the links between successive items (Kalm & Norris, 2014).

Despite finding robust differences in memory for temporal structure between the immediate and delayed sessions, we failed to find further improvements from the 1-day to the 1-week delay. This may have occurred because participants repeatedly anticipated the same sequences across the experiment. Repeated retrieval may quicken memory consolidation (Antony et al., 2017). Indeed, repeated retrieval, compared to restudy, increases behavioral markers of semanticization after consolidation (Lifanov et al., 2021), and repeatedly testing regularities in a statistical-learning task reduces forgetting of explicit memories over the course of consolidation (H. Liu et al., 2023). Repeated anticipation may have likewise hastened schematization processes, thus reducing differences between our 1-day and 1-week conditions. In line with these theories, we found that repeated retrieval and a period of consolidation both enhanced efficiency in our sequence-anticipation task. Practice and consolidation may therefore jointly contribute to retention of sequence memory: This has been well documented in the motor-learning literature, which shows that motor sequence memory tends to first improve with training and then further improve with consolidation (Brawn et al., 2010). Our findings provide a novel extension of sequence-consolidation effects to the episodic-memory literature: we show that explicit judgments about temporally extended sequences of images are enhanced both with repeated practice and with consolidation.

Additionally, our findings that consolidation benefited sequence memory is consistent with other recent work showing sleep-related improvements (followed by protracted forgetting) of memory for real-world sequences (Diamond et al., 2024). This pattern of results is similar to ours, and suggests that an initial consolidation-related benefit may stabilize or decrease over time, with no subsequent improvement with longer delays. Thus, the lack of difference between our 1-day and 1-week conditions in the distance and granularity effects is consistent with these findings in that sleep-related memory consolidation benefits do not show further improvements over long delays (Diamond et al., 2024). Despite the convergence of these findings, it is important to note that our findings may be limited to healthy younger adults. Future work could test the generalizability of this work to other samples.

Note that our sample was limited to healthy younger adults who were either undergraduate students or Prolific workers. Future work should test the generalizability of our findings to more heterogeneous samples and to other age groups because the impact of consolidation is known to vary across the lifespan (Gui et al., 2017). Further, future work can investigate whether sleep is critical for driving these consolidation effects and how prediction efficiency changes over much longer delays of months or years.

In summary, we showed that consolidation leads to efficient access of temporal structure in the service of multistep anticipatory judgments, but at the cost of perceptual details. Furthermore, postconsolidation maintenance of internal models that linked experienced events predicted accurate anticipation, raising the intriguing possibility that consolidation may increase the efficiency, accuracy, or stability of computationally intensive strategies. Together, our work shows how memories are maintained and transformed over time to prioritize representations of temporal structure at the expense of incidental perceptual features—allowing us, in turn, to anticipate likely upcoming events and behave adaptively in a complex world.

Transparency

Action Editor: M. Natasha Rajah
Editor: Patricia J. Bauer
Author Contributions

The second and third authors contributed equally to this work.

Hannah Tarder-Stoll: Conceptualization; Formal analysis; Investigation; Methodology; Visualization; Writing – original draft.

Christopher Baldassano: Conceptualization; Formal analysis; Funding acquisition; Methodology; Supervision; Writing – review & editing.

Mariam Aly: Conceptualization; Formal analysis; Funding acquisition; Methodology; Supervision; Writing – review & editing.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

This work was funded by a National Institutes of Health Research Project grant to M. Aly and C. Baldassano (Grant No. R01EY034436-01).

Open Practices

All data, stimuli, and analysis scripts are publicly available on the Open Science Framework and can be accessed at https://osf.io/8tuc4/. The experiments reported in this article were not preregistered. This article has received the badge for Open Data. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges.

ORCID iDs

Hannah Tarder-Stoll https://orcid.org/0009-0007-0957-4499
Christopher Baldassano https://orcid.org/0000-0003-3540-5019
Mariam Aly https://orcid.org/0000-0003-4033-6134

Acknowledgments

We would like to thank Margaret Kuehl and Narjes Al-Zahli for assistance with data collection; the Alyssano Group, Lila Davachi, and Serra Favila for helpful advice on this project; and Andrew Schaumberg for advice on data analysis. Mariam Aly is now at the Department of Psychology, University of California, Berkeley.

Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/suppl/10.1177/09567976241256617

References

Antony, J. W., Ferreira, C. S., Norman, K. A., & Wimber, M. (2017). Retrieval as a fast route to memory consolidation. *Trends in Cognitive Sciences*, *21*(8), 573–576. https://doi.org/10.1016/j.tics.2017.05.001

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, *52*(1), 388–407. https://doi.org/10.3758/s13428-019-01237-x

Arciuli, J., & Simpson, I. C. (2012). Statistical learning is lasting and consistent over time. *Neuroscience Letters*, *517*(2), 133–135. https://doi.org/10.1016/j.neulet.2012.04.045

Audrain, S., & McAndrews, M. P. (2022). Schemas provide a scaffold for neocortical integration of new memories over time. *Nature Communications*, *13*(1), Article 1. https://doi.org/10.1038/s41467-022-33517-0

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). *Fitting linear mixed-effects models using lme4* (arXiv:1406.5823). arXiv. https://doi.org/10.48550/arXiv.1406.5823

Behrens, T. E. J., Muller, T. H., Whittington, J. C. R., Mark, S., Baram, A. B., Stachenfeld, K. L., & Kurth-Nelson, Z. (2018). What is a cognitive map? Organizing knowledge for flexible behavior. *Neuron*, *100*(2), 490–509. https://doi.org/10.1016/j.neuron.2018.10.002

Bellmund, J. L. S., Polti, I., & Doeller, C. F. (2020). Sequence memory in the hippocampal–entorhinal region. *Journal of Cognitive Neuroscience*, *32*(11), 2056–2070. https://doi.org/10.1162/jocn_a_01592

Bonasia, K., Blommesteyn, J., & Moscovitch, M. (2016). Memory and navigation: Compression of space varies with route length and turns. *Hippocampus*, *26*(1), 9–12. https://doi.org/10.1002/hipo.22539

Brawn, T. P., Fenn, K. M., Nusbaum, H. C., & Margoliash, D. (2010). Consolidating the effects of waking and sleep on motor-sequence learning. *Journal of Neuroscience*, *30*(42), 13977–13982. https://doi.org/10.1523/JNEUROSCI.3295-10.2010

Brown, T. I., Carr, V. A., LaRocque, K. F., Favila, S. E., Gordon, A. M., Bowles, B., Bailenson, J. N., & Wagner, A. D. (2016). Prospective representation of navigational goals in the human hippocampus. *Science*, *352*(6291), 1323–1326. https://doi.org/10.1126/science.aaf0784

Brunec, I. K., & Momennejad, I. (2022). Predictive representations in hippocampal and prefrontal hierarchies. *Journal of Neuroscience*, *42*(2), 299–312. https://doi.org/10.1523/JNEUROSCI.1327-21.2021

Caucheteux, C., Gramfort, A., & King, J.-R. (2023). Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nature Human Behaviour*, *7*(3), Article 3. https://doi.org/10.1038/s41562-022-01516-2

Champely, S., Ekstrom, C., Dalgaard, P., Gill, J., Weibelzahl, S., Anandkumar, A., Ford, C., Volcic, R., & Rosario, H. D. (2020). *pwr: Basic functions for power analysis* (1.3-0) [Computer software]. https://CRAN.R-project.org/package=pwr

Chanales, A. J. H., Oza, A., Favila, S. E., & Kuhl, B. A. (2017). Overlap among spatial memories triggers repulsion of hippocampal representations. *Current Biology*, *27*(15), 2307–2317.e5. https://doi.org/10.1016/j.cub.2017.06.057

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(3), 181–204. https://doi.org/10.1017/S0140525X12000477

Davachi, L., & DuBrow, S. (2015). How the hippocampus preserves order: The role of prediction and context. *Trends in Cognitive Sciences*, *19*(2), 92–99. https://doi.org/10.1016/j.tics.2014.12.004

Daw, N. D., & Dayan, P. (2014). The algorithmic anatomy of model-based evaluation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1655), Article 20130478. https://doi.org/10.1098/rstb.2013.0478

Dayan, P. (1993). Improving generalization for temporal difference learning: The successor representation. *Neural Computation*, *5*(4), 613–624. https://doi.org/10.1162/neco.1993.5.4.613

de Lange, F. P., Heilbron, M., & Kok, P. (2018). How do expectations shape perception? *Trends in Cognitive Sciences*, *22*(9), 764–779. https://doi.org/10.1016/j.tics.2018.06.002

Diamond, N. B., Simpson, S., Pérez, D. B., Murray, B., Fogel, S., & Levine, B. (2024). *Sleep selectively and durably enhances real-world sequence memory*. bioRxiv. https://doi.org/10.1101/2024.01.10.575038

Drosopoulos, S., Windau, E., Wagner, U., & Born, J. (2007). Sleep enforces the temporal order in memory. *PLOS ONE*, *2*(4), Article e376. https://doi.org/10.1371/journal.pone.0000376

Dudai, Y., Karni, A., & Born, J. (2015). The consolidation and transformation of memory. *Neuron*, *88*(1), 20–32. https://doi.org/10.1016/j.neuron.2015.09.004

Durrant, S. J., Cairney, S. A., & Lewis, P. A. (2013). Overnight consolidation aids the transfer of statistical knowledge from the medial temporal lobe to the striatum. *Cerebral Cortex*, *23*(10), 2467–2478. https://doi.org/10.1093/cercor/bhs244

Durrant, S. J., Taylor, C., Cairney, S., & Lewis, P. A. (2011). Sleep-dependent consolidation of statistical learning. *Neuropsychologia*, *49*(5), 1322–1331. https://doi.org/10.1016/j.neuropsychologia.2011.02.015

Elliott Wimmer, G., & Büchel, C. (2019). Learning of distant state predictions by the orbitofrontal cortex in humans. *Nature Communications*, *10*(1), Article 1. https://doi.org/10.1038/s41467-019-10597-z

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1456), 815–836. https://doi.org/10.1098/rstb.2005.1622

Galea, J. M., Albert, N. B., Ditye, T., & Miall, R. C. (2010). Disruption of the dorsolateral prefrontal cortex facilitates the consolidation of procedural skills. *Journal of Cognitive Neuroscience*, *22*(6), 1158–1164. https://doi.org/10.1162/jocn.2009.21259

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science*, *9*(6), 641–651.

Gershman, S. J. (2018). The successor representation: Its computational logic and neural substrates. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *38*(33), 7193–7200. https://doi.org/10.1523/JNEUROSCI.0151-18.2018

Gershman, S. J., Moore, C. D., Todd, M. T., Norman, K. A., & Sederberg, P. B. (2012). The successor representation and temporal context. *Neural Computation*, *24*(6), 1553–1568. https://doi.org/10.1162/NECO_a_00282

Gilboa, A., & Moscovitch, M. (2021). No consolidation without representation: Correspondence between neural and psychological representations in recent and remote memory. *Neuron*, *109*(14), 2239–2255. https://doi.org/10.1016/j.neuron.2021.04.025

Gui, W.-J., Li, H.-J., Guo, Y.-H., Peng, P., Lei, X., & Yu, J. (2017). Age-related differences in sleep-based memory consolidation: A meta-analysis. *Neuropsychologia*, *97*, 46–55. https://doi.org/10.1016/j.neuropsychologia.2017.02.001

Hindy, N. C., Ng, F. Y., & Turk-Browne, N. B. (2016). Linking pattern completion in the hippocampus to predictive coding in visual cortex. *Nature Neuroscience*, *19*(5), Article 5. https://doi.org/10.1038/nn.4284

Hobson, J. A., & Friston, K. J. (2012). Waking and dreaming consciousness: Neurobiological and functional considerations. *Progress in Neurobiology*, *98*(1), 82–98. https://doi.org/10.1016/j.pneurobio.2012.05.003

Hsieh, L.-T., & Ranganath, C. (2015). Cortical and subcortical contributions to sequence retrieval: Schematic coding of temporal context in the neocortical recollection network. *NeuroImage*, *121*, 78–90. https://doi.org/10.1016/j.neuroimage.2015.07.040

Janacsek, K., & Nemeth, D. (2012). Predicting the future: From implicit learning to consolidation. *International*

Journal of Psychophysiology, *83*(2), 213–221. https://doi.org/10.1016/j.ijpsycho.2011.11.012

Kalm, K., Davis, M. H., & Norris, D. (2013). Individual sequence representations in the medial temporal lobe. *Journal of Cognitive Neuroscience*, *25*(7), 1111–1121. https://doi.org/10.1162/jocn_a_00378

Kalm, K., & Norris, D. (2014). The representation of order information in auditory-verbal short-term memory. *Journal of Neuroscience*, *34*(20), 6879–6886. https://doi.org/10.1523/JNEUROSCI.4104-13.2014

Kim, R., Seitz, A., Feenstra, H., & Shams, L. (2009). Testing assumptions of statistical learning: Is it long-term and implicit? *Neuroscience Letters*, *461*(2), 145–149. https://doi.org/10.1016/j.neulet.2009.06.030

Kóbor, A., Janacsek, K., Takács, Á., & Nemeth, D. (2017). Statistical learning leads to persistent memory: Evidence for one-year consolidation. *Scientific Reports*, *7*(1), Article 1. https://doi.org/10.1038/s41598-017-00807-3

Kok, P., Failing, M. F., & de Lange, F. P. (2014). Prior expectations evoke stimulus templates in the primary visual cortex. *Journal of Cognitive Neuroscience*, *26*(7), 1546–1554. https://doi.org/10.1162/jocn_a_00562

Kok, P., Jehee, J. F. M., & de Lange, F. P. (2012). Less is more: Expectation sharpens representations in the primary visual cortex. *Neuron*, *75*(2), 265–270. https://doi.org/10.1016/j.neuron.2012.04.034

Kok, P., & Turk-Browne, N. B. (2018). Associative prediction of visual shape in the hippocampus. *Journal of Neuroscience*, *38*(31), 6888–6899. https://doi.org/10.1523/JNEUROSCI.0163-18.2018

Krenz, V., Alink, A., Sommer, T., Roozendaal, B., & Schwabe, L. (2023). Time-dependent memory transformation in hippocampus and neocortex is semantic in nature. *Nature Communications*, *14*(1), Article 1. https://doi.org/10.1038/s41467-023-41648-1

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Lee, C. S., Aly, M., & Baldassano, C. (2021). Anticipation of temporally structured events in the brain. *eLife*, *10*, Article e64972. https://doi.org/10.7554/eLife.64972

Lifanov, J., Linde-Domingo, J., & Wimber, M. (2021). Feature-specific reaction times reveal a semanticisation of memories over time and with repeated remembering. *Nature Communications*, *12*(1), Article 1. https://doi.org/10.1038/s41467-021-23288-5

Liu, H., Forest, T. A., Duncan, K., & Finn, A. S. (2023). What sticks after statistical learning: The persistence of implicit versus explicit memory traces. *Cognition*, *236*, Article 105439. https://doi.org/10.1016/j.cognition.2023.105439

Liu, Y., Mattar, M. G., Behrens, T. E. J., Daw, N. D., & Dolan, R. J. (2021). Experience replay is associated with efficient nonlocal learning. *Science*, *372*(6544), Article eabf1357. https://doi.org/10.1126/science.abf1357

Lutz, N. D., Wolf, I., Hübner, S., Born, J., & Rauss, K. (2018). Sleep strengthens predictive sequence coding. *Journal of Neuroscience*, *38*(42), 8989–9000. https://doi.org/10.1523/JNEUROSCI.1352-18.2018

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, *102*, 419–457. https://doi.org/10.1037/0033-295X.102.3.419

Michelmann, S., Hasson, U., & Norman, K. A. (2023). Evidence that event boundaries are access points for memory retrieval. *Psychological Science*, *34*(3), 326–344. https://doi.org/10.1177/09567976221128206

Momennejad, I. (2020). Learning structures: Predictive representations, replay, and generalization. *Current Opinion in Behavioral Sciences*, *32*, 155–166. https://doi.org/10.1016/j.cobeha.2020.02.017

Momennejad, I., & Howard, M. W. (2018). *Predicting the future with multi-scale successor representations*. bioRxiv. https://doi.org/10.1101/449470

Momennejad, I., Russek, E. M., Cheong, J. H., Botvinick, M. M., Daw, N. D., & Gershman, S. J. (2017). The successor representation in human reinforcement learning. *Nature Human Behaviour*, *1*(9), Article 9. https://doi.org/10.1038/s41562-017-0180-8

O'Keefe, J., & Nadel, L. (1979). Précis of O'Keefe & Nadel's *The hippocampus as a cognitive map*. *Behavioral and Brain Sciences*, *2*(4), 487–494. https://doi.org/10.1017/S0140525X00063949

Rauss, K., & Born, J. (2017). A role of sleep in forming predictive codes. In N. Axmacher & B. Rasch (Eds.), *Cognitive neuroscience of memory consolidation* (pp. 117–132). Springer. https://doi.org/10.1007/978-3-319-45066-7_8

Robin, J., & Moscovitch, M. (2017). Details, gist and schema: Hippocampal–neocortical interactions underlying recent and remote episodic and spatial memory. *Current Opinion in Behavioral Sciences*, *17*, 114–123. https://doi.org/10.1016/j.cobeha.2017.07.016

Romano, J. C., Howard, 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. https://doi.org/10.1080/09658211003742680

Sanchez, D. J., Gobel, E. W., & Reber, P. J. (2010). Performing the unexplainable: Implicit task performance reveals individually reliable sequence learning without explicit knowledge. *Psychonomic Bulletin & Review*, *17*(6), 790–796. https://doi.org/10.3758/PBR.17.6.790

Satterthwaite, F. E. (1941). Synthesis of variance. *Psychometrika*, *6*(5), 309–316. https://doi.org/10.1007/BF02288586

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. https://doi.org/10.1016/j.cub.2012.06.056

Schapiro, A. C., McDevitt, E. A., Chen, L., Norman, K. A., Mednick, S. C., & Rogers, T. T. (2017). Sleep benefits memory for semantic category structure while preserving exemplar-specific information. *Scientific Reports*, *7*(1), Article 1. https://doi.org/10.1038/s41598-017-12884-5

Sekeres, M. J., Bonasia, K., St-Laurent, M., Pishdadian, S., Winocur, G., Grady, C., & Moscovitch, M. (2016).

Recovering and preventing loss of detailed memory: Differential rates of forgetting for detail types in episodic memory. *Learning & Memory*, *23*(2), 72–82. https://doi.org/10.1101/lm.039057.115

Sekeres, M. J., Winocur, G., & Moscovitch, M. (2018). The hippocampus and related neocortical structures in memory transformation. *Neuroscience Letters*, *680*, 39–53. https://doi.org/10.1016/j.neulet.2018.05.006

Stachenfeld, K. L., Botvinick, M. M., & Gershman, S. J. (2017). The hippocampus as a predictive map. *Nature Neuroscience*, *20*(11), Article 11. https://doi.org/10.1038/nn.4650

Stickgold, R. (2005). Sleep-dependent memory consolidation. *Nature*, *437*(7063), Article 7063. https://doi.org/10.1038/nature04286

Summerfield, C., & Egner, T. (2009). Expectation (and attention) in visual cognition. *Trends in Cognitive Sciences*, *13*(9), 403–409. https://doi.org/10.1016/j.tics.2009.06.003

Tiganj, Z., Singh, I., Esfahani, Z. G., & Howard, M. W. (2022). Scanning a compressed ordered representation of the future. *Journal of Experimental Psychology: General*, *151*(12), 3082–3096. https://doi.org/10.1037/xge0001243

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological Review*, *55*, 189–208. https://doi.org/10.1037/h0061626

Tompary, A., & Davachi, L. (2017). Consolidation promotes the emergence of representational overlap in the hippocampus and medial prefrontal cortex. *Neuron*, *96*(1), 228–241.e5. https://doi.org/10.1016/j.neuron.2017.09.005

Tompary, A., & Davachi, L. (2022). *Consolidation-dependent behavioral integration of sequences related to mPFC neural overlap and hippocampal-cortical connectivity*. bioRxiv, 2022–10.

Townsend, J. T., & Ashby, F. G. (1978). Methods of modeling capacity in simple processing systems. In N. J. Castellan, Jr & F. Restle (Eds.), *Cognitive theory* (pp. 199–239). Psychology Press.

Walker, M. P., Brakefield, T., Morgan, A., Hobson, J. A., & Stickgold, R. (2002). Practice with sleep makes perfect: Sleep-dependent motor skill learning. *Neuron*, *35*(1), 205–211. https://doi.org/10.1016/S0896-6273(02)00746-8

Walker, M. P., Brakefield, T., Seidman, J., Morgan, A., Hobson, J. A., & Stickgold, R. (2003). Sleep and the time course of motor skill learning. *Learning & Memory*, *10*(4), 275–284. https://doi.org/10.1101/lm.58503

Wimmer, G. E., Liu, Y., McNamee, D. C., & Dolan, R. J. (2023). Distinct replay signatures for prospective decision-making and memory preservation. *Proceedings of the National Academy of Sciences, USA*, *120*(6), Article e2205211120. https://doi.org/10.1073/pnas.2205211120

Winocur, G., Moscovitch, M., & Bontempi, B. (2010). Memory formation and long-term retention in humans and animals: Convergence towards a transformation account of hippocampal–neocortical interactions. *Neuropsychologia*, *48*(8), 2339–2356. https://doi.org/10.1016/j.neuropsychologia.2010.04.016

Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., & Torralba, A. (2010, June). Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition* (pp. 3485–3492). IEEE.