

# Refresh my memory: Spontaneous reinstatements from episodic memory alter the content of working memory

Abigail N. Hoskin<sup>1</sup>, Aaron M. Bornstein<sup>2</sup>, Kenneth A. Norman<sup>1,2</sup>,  
Jonathan D. Cohen<sup>1,2</sup>

1. Department of Psychology, Princeton University, Princeton, NJ, USA

2. Neuroscience Institute, Princeton University, Princeton, NJ, USA

## Abstract

Does episodic memory (EM) reinstatement influence working memory (WM) maintenance? Previous research has shown that explicit errors in delayed recall or recognition tasks reflect the use of EM when WM is overloaded or disrupted. We hypothesized that EM affects WM even absent interference, by supporting maintenance. Using novel behavioral and neural signatures of EM, we show that delay-period EM reinstatement slows down accurate responses by reinstating task-irrelevant associations (*context*) present during initial encoding. The first two experiments establish that encoding context is evident in errors (Experiment 1) and implicitly, in response slowing in the absence of errors (Experiment 2). Experiment 3 shows that fMRI evidence of EM reinstatement during the delay predicts response slowing on each trial. Modeling responses using a Drift-Diffusion Model (DDM) that draws samples from WM, we show that model fit improves when trial drift rate varies with fMRI evidence of delay-period reinstatement, supporting the hypothesis that EM has a continual, condition-agnostic, effect on the content of WM.

## General Introduction

Our memories do not exist in isolation, and neither do the neural circuits that carry them. Experiences may produce transient records in working memory, a temporary store for information to be maintained and manipulated over delays of seconds (Baddeley 1992; Baddeley & Hitch 1974; Repovš & Baddeley 2006). Experiences can also simultaneously lay down more lasting traces as episodic memories, available to be recalled at a later time (beyond minutes), allowing us to relive specific, personally experienced events tied to the time and place of their occurrence (Tulving, 1983).

The neural structures that support episodic memory are often active during periods of rest (Buckner & Carroll, 2007; Buckner 2010), during which time they appear to be reinstating recent

experiences (Tambini et al 2010) or imagining potential future scenarios, constructed on the basis of past experiences (Addis et al 2007). These reinstatements can occur even without explicit recall, or a direct link to current goals, involve coordinated activity patterns across the entire brain (Logothetis et al 2012; Maingret et al 2016), and have been observed to reliably occur during even during brief lapses in external stimulation — such as those typically used as maintenance periods in working memory experiments. This lead us to ask the question: Do these spontaneous reinstatements from episodic memory affect the content of working memory?

While earlier models of working memory offered the hypothesis that working and long-term memory operated wholly in parallel (Shallice & Warrington, 1970), more recent models propose that they may support each other (Baddeley, 2000). Evidence for the dissociation between working memory and episodic memory largely came from lesion studies, which found damage to the medial temporal lobe (MTL) caused severe episodic memory deficits (Cave & Squire, 1992; Squire, 1992), while working memory, associated with the prefrontal cortex (D'Esposito et al. 2000), remained intact (Drachman & Arbit, 1966). However, there is accumulating evidence that episodic memory, and its neural correlate, the MTL, is also engaged in working memory tasks (Ranganath et al. 2004, 2005; Ranganath & Blumenfeld 2005; Axmacher et al. 2007), suggesting these memory systems do not operate entirely independently of one another.

Neuroimaging studies of healthy participants show MTL recruitment during working memory tasks, implicating the MTL in maintaining neural representations of novel stimuli in working memory (Ranganath & D'Esposito, 2001; Stern et al., 2001; Newmark et al., 2013). The MTL also appears to mediate short-delay performance when working memory is disrupted (Rose et al., 2014; Lewis-Peacock et al., 2016; Zanto et al., 2016).

Yet, it is possible these MTL activations reflect episodic memory processes irrelevant to working memory maintenance, such as the encoding of novel stimuli used in those experiments, as opposed to episodic memory supporting working memory maintenance. For example, Ranganath et al. (2005) observed MTL delay period activity when participants performed a delayed match to sample (DMS) task with a 7–13 second retention interval. While this activity predicted performance on a post-task recognition test, ceiling DMS performance prevented assessment of whether MTL delay period activity also impacted performance on a working memory task. Further, while it is critical to use short delay periods to study working memory maintenance, studies using delay periods of 13 seconds or less often found MTL activity was highest at the beginning of the delay period, raising the possibility that the activity may be driven by encoding the initial stimulus, as opposed to reinstating it in working memory during retention or at the time of response (Ranganath & D'Esposito, 2001; Ranganath et al., 2005; Nichols et al., 2006). Olsen and colleagues (2009) addressed these issues by presenting participants with two familiarized target faces to maintain over a 30 second, distraction free delay. The researchers

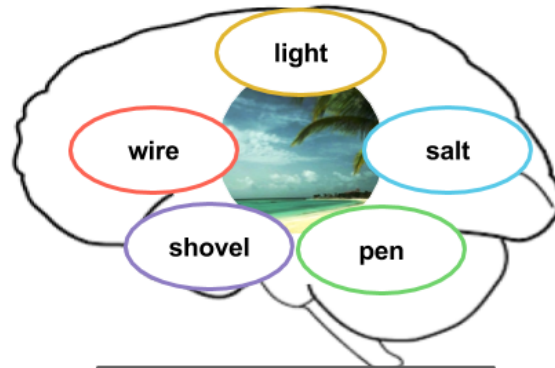
found the MTL exhibited greater delay period activity on high-confidence correct trials than on incorrect trials, suggesting episodic memory could indeed support working memory through reinstatement of working memory representations.

Here, we examine the mechanisms by which episodic memory might support working memory maintenance. We leverage the fact that retrievals from episodic memory carry with them temporal and associative context (Howard & Kahana, 2002) such that recalling a given context can cause the subsequent, involuntary recall of other memories sharing that context (Hupbach et al., 2009). This can occur even at the short delays typically associated with WM rather than EM (Hannula et al., 2006). We show that such information affects choices and response times, and that these behavioral effects scale with the amount of neural evidence for its retrieval.

In Experiment 1, we establish a new signature of EM involvement in short-term memory decisions. Specifically, we build on a classic short-term retention task with varying levels of distraction during the delay. We show that as distraction increases — and episodic memory is more strongly engaged — substitution errors in free recall responses were more likely to come from the encoding context of the target words, a signature of retrieval from EM. In Experiment 2, we build on this signature. Specifically, we use a delayed match to sample task with familiarized stimuli to show that, even without distraction, the influence of encoding context can be observed in reaction times to recognition cues. Retrieving context from episodic memory can speed or slow response times, depending on the recognition probe. In Experiment 3, we used the same task as Experiment 2, while participants were scanned using functional magnetic resonance imaging (fMRI). Using trial-by-trial neural evidence of episodic memory reinstatement, we show that episodic memory reinstatement during the delay period can either speed or slow responses to the probe, depending on both the type of probe word and the information reinstated. We fit response times on this task with a drift-diffusion model (DDM) in which evidence is sampled from working memory contents set by episodic memory. Critically, we show that the DDM fits are significantly improved by using neural reinstatement evidence to set the drift rate on each trial.

Together, these results establish a mechanism by which working memory maintenance over even relatively brief intervals is supported by covert retrievals from episodic memory, and that these retrievals can carry with them information not previously present in WM, a signature of EM's involvement that can help or hinder decision-making, depending on context (Figure 1).

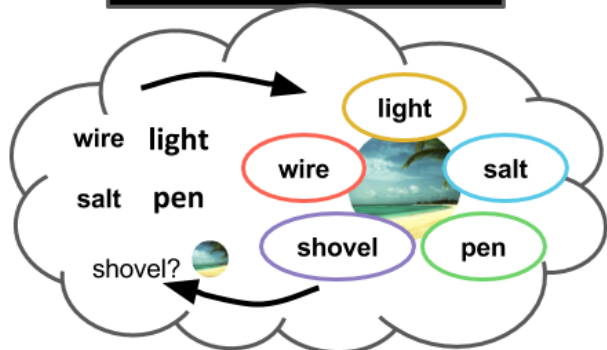
**A.** Episodic memory encodes items, such as the words “light”, “salt”, “pen”, “shovel”, and “wire”, along with the context in which they were learned—in this example, a picture of the beach.



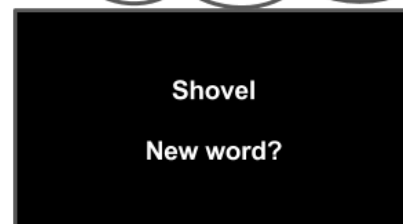
**B.** Example target words to remember over a delay period



**C.** Episodic memory helps maintain targets over a delay, but retrieving items from episodic memory brings the other items originally associated with the targets into working memory.



**D.** Remembering which items were targets could consequently be more difficult due to task-irrelevant items in working memory.



**E.** However, decision making could also be facilitated if context information from episodic memory provides additional evidence to support making the correct decision.



**Figure 1. Episodic memory can inject incidental information into working memory.**

Episodic memory encodes items along with the context in which they were learned (A). When presented with items to maintain over a delay period (B), WM maintenance may periodically refresh from EM. These reinstatements may carry *context*: other associations formed with the desired information at the time of encoding (C). This context information, now in working memory, can affect subsequent behavior. Decision making is impeded when this context information provides evidence for making an incorrect decision (D), but decision making could also be facilitated when context information provides evidence in

favor of making the correct decision (E).

## Experiment 1: A signature of episodic memory use in a short-term retention task.

### Introduction

In Experiment 1, we modified a classic short-term free-recall task to permit measure of the influence of episodic memory on responses. In this task, participants are given a list of words, followed by a short delay, then asked to recall the words they had just seen. Previous studies using this task have reported both behavioral and neuroimaging evidence indicating that distraction during the delay causes participants to rely on episodic, rather than working memory (Brown, 1958; Peterson & Peterson 1959; Rose et al., 2014; Lewis-Peacock et al., 2016; Zanto et al., 2016). Here, we use multiple levels of distraction, together with manipulation of encoding context, to more precisely index the engagement of episodic memory, including in the no distraction condition.

### Methods

**Stimuli.** The experiment used six context pictures, one for each set of 12 words with which it was a uniquely associated. The pictures were color photographs of outdoor scenes. The words were concrete nouns drawn from the Medical Research Council Psycholinguistic Database (Wilson, 1988). All words had a maximum of two syllables, Kucera-Francis written frequency of at least 2, a familiarity rating of at least 200, a concreteness rating of at least 500, and an imageability rating of at least 500. Pictures consisted of famous landmarks without any people in them. The words used in each set and the image associated with each set were randomized across participants.

**Participants.** 15 Princeton psychology students (9 females; ages 18 to 22) completed the study for course credit. All participants had normal or corrected-to-normal vision and provided informed consent. The study protocol was approved by the Princeton University IRB.

**Procedure. Word-context learning trials.** On each of 72 learning trials, participants were shown four words drawn from the same set, arranged vertically in the center of the screen, alongside the photograph of either a face or a scene that was associated with that set, and that was displayed either on the left or right-hand side of the screen (Figure 2A). The picture served as an *encoding context*. To help participants encode the 12 words associated with the same

picture as all belonging to the same context, each word was presented three times along with three other words randomly sampled from the same list and always displayed in the same context (i.e., with the picture associated with that list. On each trial, the four words and the picture associated with

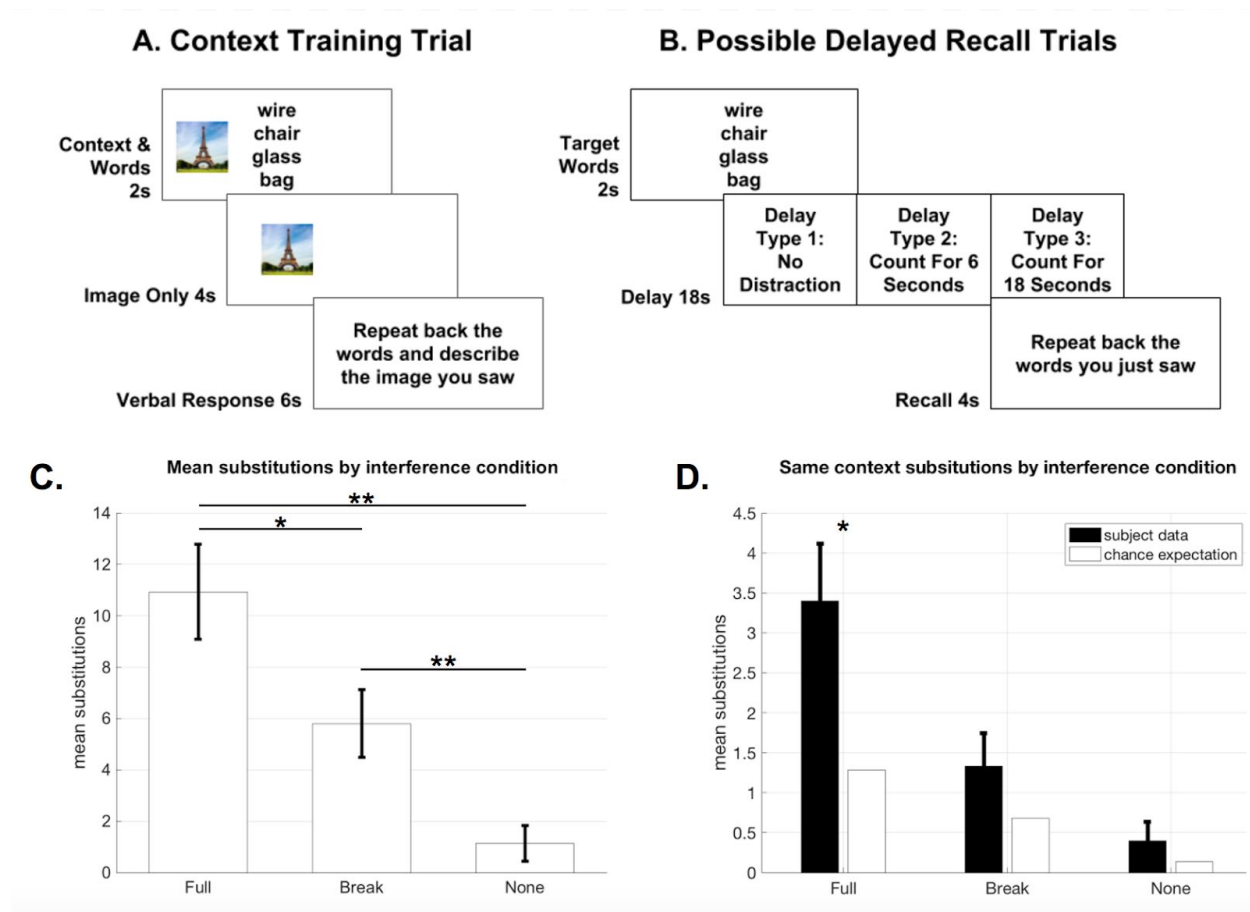
those words were presented for two seconds before the words disappeared and the picture remained on-screen. Four seconds later, the context picture was replaced by a prompt asking participants to vocally repeat back the four words just shown, and to then briefly describe the picture they had just seen. Participants were given six seconds to respond. Trials were of fixed length, regardless of participant's responses.

**Free recall phase.** After the learning block was completed, participants performed 54 trials of a short-term retention task. On each trial, participants were shown four *target* words, arranged vertically.

The four target words were all drawn from the same context. No picture was presented alongside the words. Words remained on the screen for two seconds and were followed by an 18 second delay.

There were three types of delay (Figure 2B). Delay trial types were randomly intermixed, with 18 trials of each type. In the *no distraction* condition, participants were shown a fixation cross, in the center of the screen, for the entirety of the 18 second delay. In the *break distraction* condition, participants were shown a fixation cross in the center of the screen for six seconds. After six seconds, participants were shown a three-digit number, in the center of the screen. The number served as a prompt to count down out loud by sevens, starting at that number. After six seconds of counting, participants were again shown a centered fixation cross, for six more seconds. In the *full distraction* condition, participants were shown a three-digit number at the start of the delay period, and instructed to count backwards out loud by sevens from the number for the entire delay period.

In all conditions, participants were given eight seconds after the delay period to vocally recall the words shown at the beginning of the trial. These responses were recorded and scored for the number of words correctly recalled (zero through four). Mistakes were categorized as one of three types: 1) other words from the same context, 2) words from the previous free recall trial, or 3) other words learned during the experiment but not in categories 1 or 2. (No substitutions were made using words not learned during the experiment.)



**Figure 2. Experiment 1: Free recall task with added context.** A. Participants studied lists of words as part of contexts distinguished by different, lateralized pictures of a face or scene. B. We then probed how these contexts would affect performance on a short term recall task under three conditions: 1. when working memory was not disrupted, 2. briefly disrupted, or 3. completely disrupted. C. Participants ( $n = 15$ ) made more errors in the full distraction than the break distraction conditions ( $t(14) = 3.2756$ ;  $p < .01$ ; paired t-test), and more errors in the full distraction versus no distraction conditions ( $t(14) = 6.4526$ ,  $p < .001$ ;  $p < .01$ ; paired t-test). Participants also made more errors in the break distraction condition than the no distraction condition ( $t(14) = 4.4852$ ,  $p < .001$ ;  $p < .01$ ; paired t-test). D. When working memory is disrupted by sustained distraction, participants ( $n = 15$ ) make errors that reflect the influence of reinstated context. Specifically, participants make substitution errors during recall that reflect the encoding context of the target set, or *same context* errors, at a higher rate than would be expected if they were randomly substituting words previously learned in the experiment. If substitutions were uniformly distributed among the 68 possible words, only 8/68 of the errors made in each interference condition should be *same context* substitutions. Instead, the proportion of *same context* substitutions was higher than what would be expected by chance when working memory was disrupted by sustained interference ( $t(14) = 2.9529$ ,  $p = .01$ ; one sample t-test), suggesting context information from episodic memory enters working memory when working memory is overloaded. Error bars reflect SEM. \* signifies  $p < .05$ , \*\* signifies  $p < .01$ .

## Results

We expected to see increasing numbers of substitutions as the demands on working memory increased; therefore we predicted the lowest number of substitutions following delays with no distraction, some substitutions following break distraction, and the most substitutions following full distraction.

Consistent with our predictions, participants made more errors in the full distraction condition than in the break distraction condition ( $t(14) = 3.2756$ ;  $p < .01$ ; paired t-test), or the no distraction condition ( $t(14) = 6.4526$ ,  $p < .001$ ; Figure 2C), and more errors in the break distraction condition than the no distraction condition ( $t(14) = 4.4852$ ,  $p < .001$ ; Figure 2D).

We also predicted that distraction would increase reliance on episodic memory, and thus that substitution errors would reflect information retrieved from episodic memory. To evaluate this hypothesis, we marked errors as belonging to one of three categories. First, as has been shown previously (e.g., Glanzer, 1972), we expected recently-experienced words — in particular, the four words from the trial immediately previous — to be most accessible in episodic memory, and therefore likely to be recalled and brought into working memory and be present as mistaken substitutions. We refer to these as *previous-target* substitutions. Second, we expected that *context* would serve as a signature of episodic memory reinstatements (Howard & Kahana, 2002; Gershman et al., 2013). As a result, substitutions should include one of the eight words that were associated with the same context as the target words, but that were not part of the target words. We refer to these as *same context* substitutions. Contexts changed each trial, making previous-target and same context substitutions mutually exclusive. Substitutions using one of the 56 remaining words from the task that were neither targets nor *previous-target* or *same context* errors, were referred to as *other* errors.

By categorizing errors in this way, we could compare the number of each kind of error to the number that would be expected were the errors drawn at random from the 68 possible non-target words. While all three kinds of words should be retained in episodic memory, recency and context lead us to predict that words from Previous-target and same context errors should be overrepresented relative to Other errors.

Specifically, if substitution errors were uniformly distributed among the 68 possible words, only 8/68 of the errors made in each interference condition should be *same context* substitutions. Instead, the proportion of *same context* substitutions was greater than what would be expected by chance on full interference trials ( $t(14) = 2.9529$ ,  $p = .01$ ; one sample t-test), suggesting that



context information was indeed affecting decision making when working memory was overloaded (Figure 2D). Same context substitutions were not greater than what would be expected by chance on break ( $t(14) = 1.5870, p = .13$ ; one sample t-test) and no interference trials ( $t(14) = 1.1346, p = .28$ ; one sample t-test), hinting that context could also be affecting working memory even in the absence of distraction, a possibility further investigated in Experiments 2 and 3.

## Experiment 1 Discussion

Participants completed a short term retention task with three distraction conditions. When there was no distraction during the retention delay, participants made almost no errors, consistent with the idea that they were able to easily use working memory to complete this task. Errors increased when participants were made to perform a distractor task midway through the delay, and increased again when the distractor task encompassed the entire retention interval. These errors took the form of substituting other words from the experiment in place of the current trial's target words.

A disproportionate number of substitutions were made using words from the same encoding context as the target words, despite the fact that these kinds of words represented only a small fraction of the words used on the task. This distribution of substitutions is consistent with previous observations that, when working memory maintenance is interrupted, participants use episodic memory to maintain information over short delays (Lewis-Peacock, Cohen & Norman, 2016; Zanto et al., 2016; Rose et al., 2014). Critically, our results establish that the context-based nature of errors can serve as a signature of episodic memory use in a short-term retention task.

This finding leaves open two questions pertinent to our line of inquiry. First, does episodic memory support working memory absent external distraction? While substitutions in the no distraction condition were numerically biased towards being from the same encoding context as the target words, there were too few errors, of any kind, to support meaningful inference. The second question is what role episodic memory retrieval plays in supporting performance on this task: Are episodic memories used to support maintenance, or simply retrieved at the time of response? Answering this question will require a measure of activity during the retention interval. We use the context signature established in Experiment 1 to address these questions in Experiments 2 and 3.

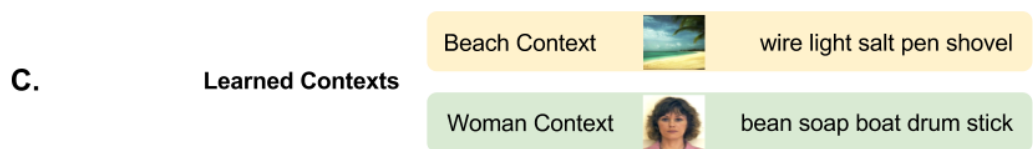
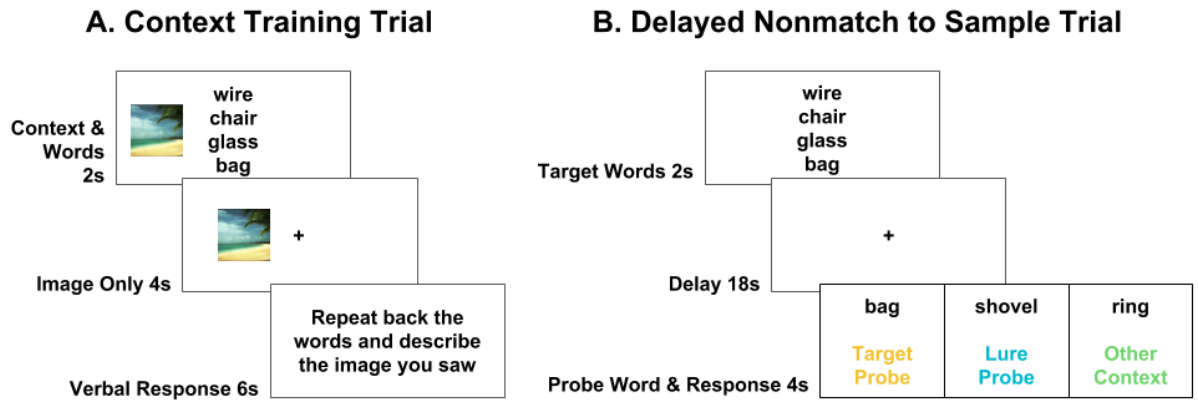
## Experiment 2: Context information from episodic memory affects response times in a working memory task without distraction.




### Introduction

Because participants in Experiment 1 made so few errors in the no-distraction condition, we could not evaluate whether episodic memory contributed to working memory maintenance on trials absent distraction. We reasoned that response time (RT), a more sensitive measurement, might reveal such an influence.

In Experiment 2, participants performed a delayed non-match to sample task (DNMS; Figure 3B), where four items had to be maintained over an 18 second delay without distraction. The lack of distraction and length of delay period were chosen so that participants could presumably complete the task using only working memory. Previous evidence suggests episodic memory and its presumed neural correlate, the hippocampus (Squire, 1992), are not necessary to perform DNMS tasks with short delays; research with non-human primates has found selective damage to the hippocampus does not eliminate the ability to perform DNMS tasks (Nemanic et al., 2004), even with maintenance intervals as long as 30, 60, or 120 seconds (Murray & Mishkin, 1998). Rodent research has also found intact DNMS performance following selective hippocampal damage, even with maintenance intervals of up to 60 seconds (Aggleton et al., 1986).

In Experiment 2, participants learned words as belonging to one of four contexts (Figure 3A), as in Experiment 1, but in the test phase they were now asked whether a *probe* word matched one of the four *target* words. The test phase thus had three types of trials, ones in which the probe: was one of the targets (*target trials*); was a word from the same encoding context as the targets (*lure trials*); or was from another encoding context (*other context trials*; Figure 3B).



| DNMS Trial Type     | Targets                      | Reinstatements of Target Context During Delay   | Probe  | Reaction Time Prediction   |
|---------------------|------------------------------|---|--------|--|
| Target Probe        | wire<br>light<br>salt<br>pen | wire light salt pen shovel<br>  | wire   | Probe in WM, facilitates recognizing it was a target<br><b>FASTER RT</b> |
| Lure Probe          | wire<br>light<br>salt<br>pen | wire light salt pen shovel<br> | shovel | Probe in WM, misleadingly seems like a target<br><b>SLOWER RT</b>        |
| Other Context Probe | wire<br>light<br>salt<br>pen | wire light salt pen shovel<br> | bean   | Probe not in WM, no misleading information                               |

**Figure 3. Experiment 2: DNMS task with added context.** A. In the *learning* phase, participants learned forty eight words that were split into four lists. Each list was paired with a unique *context picture* of a face or scene. The context picture was consistently displayed on either the left or the righthand side of the screen. B. In the *testing* phase, participants performed a delayed non-match to sample (DNMS) task, in which they remembered four *target* words across an 18 second delay. After the delay, they were shown a single *probe* word and asked whether that word was *not* one of the four they had just seen. Response times were recorded and used as a measure of whether the participants' performance had been affected by context information reinstated from episodic memory. C. Subsets of two example contexts are presented for illustrative purposes. D. We hypothesized that the contents of working memory are influenced by reinstatements from episodic memory. These reinstatements activate working memory representations of trial irrelevant items that were linked to the reinstated target items at first encoding. We predicted that when the probe word was one of the target words, participants would be fastest to respond since the target probe should clearly match the content of working memory, allowing the search process to terminate quickly. For non-target probe trials, we predicted participants would respond slower since they need to exhaustively search through the contents of working memory to decide to reject the probe. Within non-target probe trials, we predicted participants would be slowest to respond to lure probes, since these probes would match the context information in working memory elicited by the target words but mismatch with the actual target words in working memory. Since this confusing, conflicting evidence is not present in other context probe trials--the probe word does not match the context information or target words in working memory--we predicted participants would be less impaired on these trials.

We measured RTs to making a match or mismatch decision, predicting that RTs would reflect the influence of encoding context. Specifically, we hypothesized that the contents of working memory are influenced by reinstatements from episodic memory and that these reinstatements should carry with them the other words linked with the target words' context at first encoding (Figure 3C, 3D). We modeled response selection as arising from an evidence accumulation process that draws sequential samples from working memory to infer whether the word on the screen is among the target set.

We predicted that participants would respond fastest to target probes, as the probe word would clearly match the contents of working memory. In contrast, non-target probe trials, in which the probe word does not match any of the targets, would be slower because they require an exhaustive search of the contents of working memory to decide on rejection (irrespective of whether search is considered to be serial or parallel -- Sternberg, 1969; Ratcliff, 1979). Within non-target probe trials, we predicted that participants would be slower to respond to lures than other context probes. This is because, if context reinstatements from episodic memory during the delay period activate trial irrelevant items from the same context as the target words, lure words would become activated in working memory during the delay period; this would cause lure probes to match the context information in working memory while also being a mismatch with

the target words in working memory, leading to confusion and RT slowdowns.

## Methods

**Participants.** 33 Princeton students (20 females; ages 18 to 22; native English speakers) completed the study for course credit. All participants had normal or corrected-to-normal vision and provided informed consent. The Princeton University IRB approved the study protocol. One participant was excluded from analyses as accuracy scores were more than 2 standard deviations below the mean, leaving 32 participants reported here.

**Procedure.** The task was a delayed non-match to sample task (DNMS) that consisted of 3 phases. In the *learning phase*, participants studied four different word lists, each containing 12 words that were drawn from the same words used in Experiment 1. Each list of words was paired with a unique context picture — one of two faces or two scenes. The face pictures were emotionally neutral and of non-famous individuals, taken from the Psychological Image Collection at Stirling. The scene pictures were of two natural scenes of non-famous places. One of the faces and one of the scenes were always displayed on the left side of the screen; the other face and other scene were always displayed on the right side of the screen. Thus, each list was associated with one of the following context stimuli: a face on the left, a face on the right, a scene on the left, or a scene on the right. The paired words and orientation of each context picture were randomly assigned across participants.

The 48 trials of context training, 12 trials for each context, followed the same timing as the context training trials in Experiment 1 (Figure 2A, Figure 3A). Each word was presented three times along with three other words randomly sampled from the same context. On each trial, the four words and the picture associated with those words were presented for two seconds before the words disappeared while the picture remained on-screen. Four seconds later, the context picture was replaced by a prompt asking participants to vocally repeat back the four words just shown and describe the picture. Participants were given six seconds to respond. Trials were of fixed length, regardless of participant's responses.

In the *testing phase* participants performed 60 trials of a DNMS task using the words learned in the Learning phase (Figure 3B). On each trial, four *target* words were selected at random from one randomly-selected context. These words were shown on the screen together for two seconds. When the words disappeared, they were replaced by a centered fixation cross, displayed for 18 seconds. Before the experiment, participants were instructed to use this delay to remember the four words they had just seen. After the delay period, participants were shown a *probe* word for five seconds, and asked to respond *yes* if the given word was not one of the four they had just

seen on this trial, or *no* if it was one of the four target words. The keys used to signify yes and no — the left and right arrows — were counterbalanced across participants. A successful response was indicated by a green fixation cross while an unsuccessful response (incorrect response or time-out) was indicated with a red fixation cross.

Probe words could be one of three types: *target* probes were drawn from the four-word target set presented on the current trial; *lure probes* were drawn from the same context list as the target words, but, critically, not one of the target words; *other context* words were drawn from one of the three contexts other than the one from which the target words were drawn. Participants were not signaled as to which kind of probe was being used on that trial. There were 20 trials of each probe type, randomly intermixed.

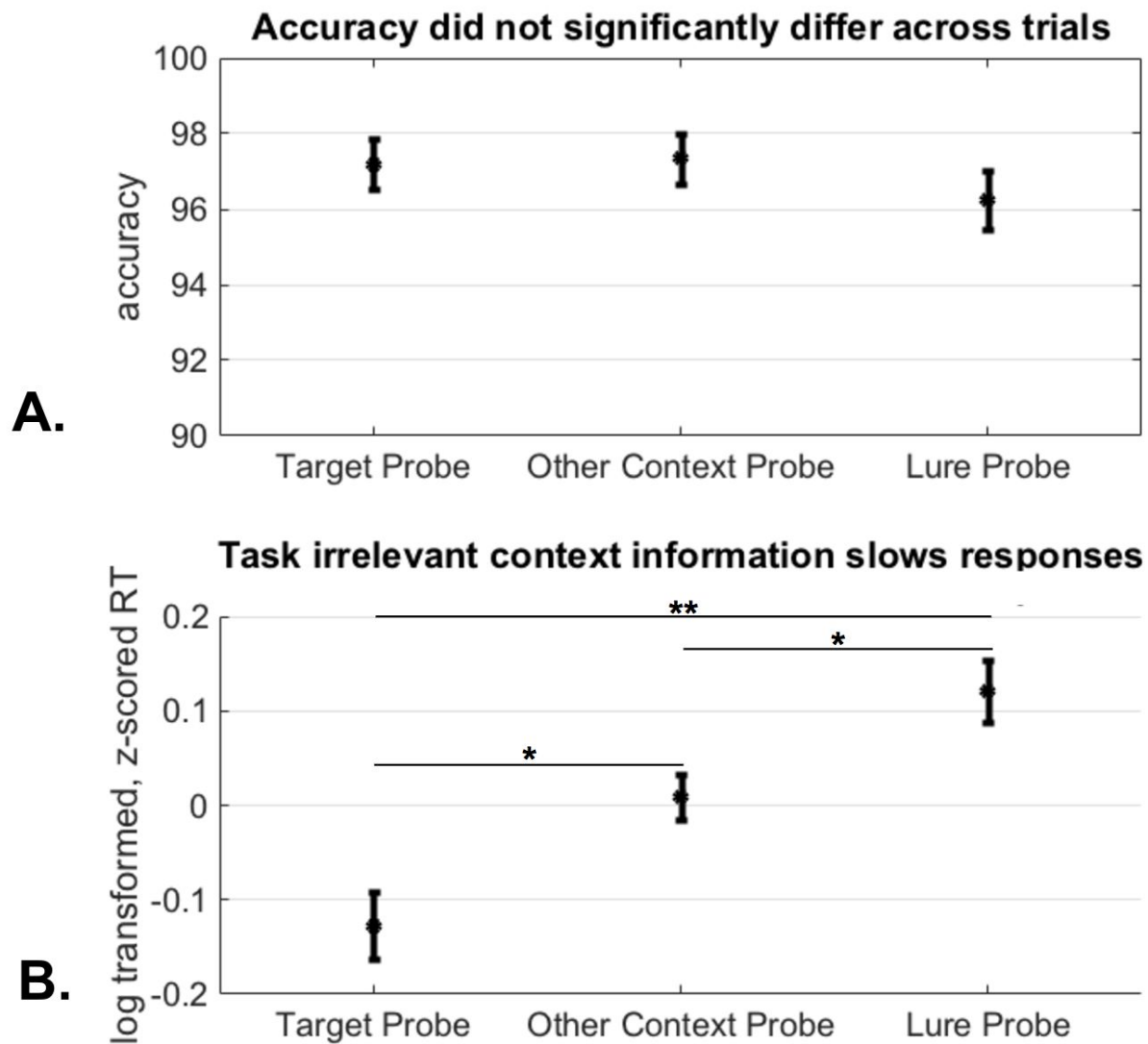
We recorded response times (RTs) to probes. Participants' RTs were log transformed and individually Z-scored, to compare the relative slow down or speed up effects of the different probe types.

## Results

**Accuracy.** As expected, due to the lack of distraction, accuracy was high across all three conditions (mean = 96.25%, SEM = 0.89%), with no significant difference in accuracy between *target* (mean = 97.03%, SEM = 0.70%), *lure* (mean = 95.00%, SEM = 1.19%), or *other context* trials (mean = 96.72%, SEM = 0.76%) (all  $p > .2$ ; Figure 4A). Because participants made so few errors (incorrect hits or rejections; mean number of errors = 2.25, SEM = .41), we exclude inaccurate trials from the RT analyses.

**Reaction times.** Our analyses focused on RTs, on the premise that these would serve as a more sensitive measure of the influence of how context influenced responses.

RTs were log transformed and z-scored within-subject. Using paired t-tests, we found that participants responded faster to *target* probes (mean zRT = -0.15, SEM = .04) than to *lure probes* (mean zRT = .12, SEM = .03;  $t(31) = -3.8616$ ,  $p < .001$ ) or *other context* probes (mean zRT = .01, SEM = .02;  $t(31) = -2.6602$ ,  $p < .01$ ; Figure 4B). Critically, participants responded slower to *lure probes* than to *other context* probes ( $t(31) = 2.5250$ ,  $p = .02$ ; Figure 4B), despite how the only difference between these two kinds of probe is whether the probe word was learned in the same context as the target.



**Figure 4. Response times reflect influence of study context.** A. Accuracy was high across all three conditions (mean = 96.25%, SEM = 0.89%), with no significant difference in accuracy between *target* (mean = 97.03%, SEM = 0.70%), *other context* (mean = 96.72%, SEM = 0.76%), or *lure* (mean = 95.00%, SEM = 1.19%) trials (all  $p > .2$ ). B. Only accurate trials are reported for RT analyses. On average, participants ( $n = 32$ ) responded faster to *target* probes — words drawn from the 4 word target set — relative to *lure* probes — non-target words drawn from the same context as the target words — ( $p < .001$ ; paired, two-tailed t-test) and relative to *other context* probes — words drawn from a different context as the target words ( $p < .05$ ; paired, two-tailed t-test). Critically, participants responded slower to *lure* probes than to *other context* probes ( $p < .05$ ; paired, two-tailed t-test), despite the fact that the only difference between these two kinds of probes was whether their encoding context (from the Learning phase) matched that of this trial's target set. Error bars reflect SEM. \* signifies  $p < .05$ , \*\* signifies  $p < .01$ .

## Experiment 2 Discussion

In Experiment 2, participants performed a delayed non-match to sample task using study words learned in one of four separate lists. Response times showed an effect of encoding context, even in the absence of distraction. Specifically, responses to target probes were faster than responses to lure and other-context probes, while responses to other-context probes were faster than those to lure probes. Consistent with this result, Bramao and Johansson (2016) found non-diagnostic context information led to a decrease in memory performance and an increase in neural measures of memory distraction (Hellerstedt & Johansson, 2014) from task irrelevant, but shared context memories.

If participant responses were simply biased towards the more prevalent response, they should be faster to respond to lure or other-context probes ( $\frac{2}{3}$  of trials), rather than target probes ( $\frac{1}{3}$  of trials). Similarly, if they were simply matching the contents of working memory to the words on the screen, one might expect a uniform response across trial types.

The observed effects are consistent with a role for episodic memory reinstatement in working memory maintenance during the delay. If the target words are refreshed from episodic memory, these refreshes will carry context, including the other words learned along with the target set. The presence of the lure word in working memory could slow responses to lure probes. Similarly, any incidental reinstatement of other contexts could slow responses to other-context trials; that these types of reinstatements should be less likely than the intended reinstatements, of the target context, could explain why responses to other-context probes are faster, on average, than are responses to lure probes.

## Experiment 3: Neural evidence for context reinstatement predicts RT slowing on a trial-by-trial basis.

### Introduction

Experiment 3 directly tested the key prediction of our hypothesis that working memory maintenance is affected by reinstatement from episodic memory. Specifically, we measured reinstatement during the delay period on each trial, and use it to predict response times to the probe on that trial.



To perform this test, we ran 36 additional participants on the same DNMS task as in Experiment 2, in an fMRI scanner. We used pattern classifiers to generate, for each trial, evidence that each context was reinstated during the delay. We divided these evidence into those that matched the context of the probe word on that trial, and the average of the other contexts (non-probe), and entered each into a regression on trial RTs. Our key test is the effect of context reinstatement on lure trial RTs. We predicted that probe context reinstatements would slow responses, by strengthening the representation of the lure word in working memory, and thus increasing the conflict between the match and mismatch responses. Symmetrically, we predicted that reinstatement of non-probe contexts would speed responses, by strengthening the representation of other words (Figure 5).



Next, we fit participant responses using a Drift-Diffusion Model (DDM; Ratcliff 1978). We focused on mismatch trials, or trials where the probe word was not one of the targets, in order to explore the cognitive mechanism underlying the behavioral slowdown observed in Experiment 2. We further focused our analysis on the drift rate, which reflects the coherence of evidence used to make the response. In this experiment, evidence coherence corresponds to how consistently the contents of working memory are congruent with the correct response. We therefore fit two variants of the model: one in which drift rate varied by subject and a second where drift rate varied by subject and also set each trial's drift rate according to the classifier evidence for context reinstatement on that trial. Following the regression predictions, we expected probe-context reinstatements to lower the drift rate, and non-probe context reinstatements to raise the drift rate.

## Example Contexts

**Target Context** Beach Context  wire light salt pen shovel ...

**Other Context** Woman Context  bean soap boat drum stick ...

## Predictions for Mismatch Trials

| Targets  | Context Reinstatements<br>During Delay   | Probe             | Reaction Time<br>Prediction   |
|--|--|-------------------|---|
| <div> <div>wire</div> <div>light</div> <div>salt</div> <div>pen</div> </div> | <div> <div>wire light</div> <div>salt pen</div> <div>shovel ...</div>  </div> | <div>shovel</div> | Probe in WM,<br>misleadingly seems<br>like a target<br><b>SLOWER RT</b> |
| <div> <div>wire</div> <div>light</div> <div>salt</div> <div>pen</div> </div> | <div> <div>bean soap</div> <div>boat drum</div> <div>stick ...</div>  </div> | <div>shovel</div> | Probe isn't in WM, no<br>misleading<br>information<br><b>FASTER RT</b>  |

**Figure 5. Reinstatements during delay should affect comparison process at probe.** We hypothesized that the contents of working memory are influenced by periodic reinstatements from episodic memory. These reinstatements carry with them other items that were linked to the reinstated items at first encoding. At probe, participants use an evidence accumulation process to decide whether the word on the screen is contained in working memory. Using a classifier trained to recognize reinstatements of our four different contexts, we sought to identify the kind of context reinstatement during the delay and relate it to behavior at probe. On mismatch trials (lure and other context probe trials), we predicted participants would be slowest to respond after they reinstated context information that brought the probe word to mind. This is because on mismatch trials the probe should be rejected as being part of the (recently viewed) target set; however, by reinstating the context associated with the probe, it may be re-activated in working memory, thus degrading the evidence in favor of rejecting it as a target. By the same logic, reinstating non-probe contexts would not produce this effect, and thus not degrade the evidence supporting a (correct) rejection, that should lead to faster reaction times.

## Methods

**Participants.** 40 healthy participants (26 females; ages 18 to 30) were recruited. All participants had normal or corrected-to-normal vision and provided informed consent. The Princeton University IRB approved the study protocol. Exclusion criteria for recruitment included the presence of metal in the body, claustrophobia, neurological diseases or disorders, tattoos above the waist, pregnancy, not speaking English as a native language, and left-handedness. 4 participants were excluded from the final analyses for the following reasons: excessive movement in the scanner — defined as maximal instantaneous displacement larger than 3 mm across any individual scanner run (2 participants), or numerically below-chance accuracy on the DNMS task (2 participants). Data is reported for the remaining 36 participants.

**Stimuli.** The *Fixation* phase used scrambled scene pictures and scene pictures that were not used in any other phase of the experiment. In the *Learning* and *Test phases*, the scene pictures and words used were the same as in Experiment 2. The *Localizer* phase used a different set of scene pictures, along with scrambled scene pictures, neutral faces, and object pictures. All picture stimuli across all tasks were color photos scaled to the same size (500 x 500 pixels), equalized for overall brightness, and were displayed 7 degrees from the right or 7 degrees from the left of fixation.

**Procedure.** Prior to the fMRI session, participants practiced the tasks outside of the MRI scanner for about 10 minutes. Practice consisted of self-paced reading of written explanations of the fixation, context learning, DNMS, and localizer tasks in addition to a fixed number of practice trials of each task. Participants were encouraged to ask questions in case they needed any

instruction clarification. After participants felt comfortable with the instructions, they completed another practice trial of the context learning task and DNMS task in the scanner.

After practice in the scanner, participants were given 5 minutes of fixation training during which pictures appeared 7 degrees from the right or left of fixation. The goal of this training was to ensure participants perceived the context pictures as lateralized, rather than turning their gaze directly to the picture. We used an Eyelink 1000 eyetracker (SR Research, Ontario, Canada) to give participants real time feedback; if participants looked away from fixation, the images would disappear and an “X” would appear in the center of the screen until fixation was re-established.

After fixation training, participants completed the context list learning and DNMS tasks described in Experiment 2. Trials in which participants did not respond before the 4 second deadline were excluded from analyses, since there was no response time for these trials.

In the final, *localizer* phase, participants performed a localizer task that was used to discriminate regions of cortex that preferentially process left- and right- lateralized face and scene pictures. In this task, pictures were presented one at a time, and participants were asked to press a key indicating whether the currently presented picture was the same as the one immediately preceding. Pictures were presented in mini-blocks of 10 presentations each. Eight of the images in each block were trial-unique, and two were repeats. Stimuli in each mini-block were chosen from a large stimulus set of pictures not used in the main experiment, and each belonged to one of four categories - faces, objects, scenes or phase-scrambled scenes. and were presented on either the left or right side of the screen. Thus, there were eight different kinds of mini-block: left-face, right-face, left-object, right-object, left-scene, right-scene, left-scrambled, and right-scrambled. Pictures were each presented for 500 ms, and followed by a 1.5 second ITI. Participants completed a total of 24 mini-blocks (three blocks per four picture categories presented on either side of the screen), with each mini-block separated by a 12 second inter-block interval.

Finally, after the scanned portions of the experiment had completed, participants remained in the scanner to complete a memory task. Participants were shown each of the 48 words from context training, one at a time, above all four context pictures, and asked to report both which context was correct and their confidence about that judgement, between one (low confidence) and four (high confidence).

**Imaging methods. Data acquisition.** Functional magnetic resonance images (fMRI) were acquired during Phases 2, 3, and 4: context learning, DNMS test, and localizer. Data were acquired using a 3T Siemens Prisma scanner (Siemens, Erlangen, Germany) with a 64 channel volume head coil, located at the Princeton Neuroscience Institute. Stimuli were presented using a rear-projection system (Psychology Software Tools, Sharpsburg, PA). Vocal responses were recorded using a fiber optic noise cancelling microphone (Optoacoustics, Mazor, Israel), and manual responses were recorded using a fiber-optic button box (Current Designs, Philadelphia, PA). A computer running Matlab (Version 2012b, MathWorks, Natick, MA) controlled stimulus presentation.

Functional brain images were collected using a T2\*-weighted gradient-echo echo-planar (EPI) sequence (44 oblique axial slices, 2.5 x 2.5 mm inplane, 2.5 mm thickness; echo time 26 ms; TR 1000 ms; flip angle 50°; field of view 192 mm). To register participants to standard space, we collected a high-resolution 3D T1-weighted MPRAGE sequence (1.0 x 1.0 x 1.0 mm voxels).

**fMRI data preprocessing.** Preprocessing was performed using FSL 5.0.6 (FMRIB's Software Library, [www.fmrib.ox.ac.uk/fsl](http://www.fmrib.ox.ac.uk/fsl)). The first 8 volumes of each run were discarded. All images were skull-stripped to improve registration. Images were aligned to correct for participant motion and then aligned to the MPRAGE. The data were then high-pass filtered with a cutoff period of 128 seconds. 5 mm of smoothing was applied to the data.

**Region of interest definition.** Our anatomical regions of interest were fusiform gyrus, parahippocampal gyrus, and lingual gyrus, based on previous reports of visual category-selective patches of cortex — faces (Kanwisher et al., 1997) and scenes (Epstein & Kanwisher, 1998). We created a bilateral mask combining these three regions that was used for all pattern classifier analyses.

**Multivariate pattern analysis.** We extracted the time series of BOLD signal in our anatomical regions of interest during the localizer task and labeled each TR according to the category miniblock to which it belonged. These labeled time series were used to train an L2-regularized multinomial logistic regression classifier (Polyn et al., 2005), to predict the four class labels (left face/right face/left scene/right scene). In our classifier, the probability that each class is present do not sum to 1 because we don't want to assume that the categories are mutually exclusive, or that the presence of *left face* evidence necessarily indicates *right face* absence (Lewis-Peacock & Norman, 2014).

To examine how context reinstatements during the DNMS task affected RTs, we divided DNMS trials into 3 time periods: the 2 seconds in which the target words were presented (*target*

*presentation*), the 18 second delay period during which participants only saw a fixation cross (*delay period*), and the 4 seconds during which participants saw the probe word and had to respond (*probe presentation*). The trained classifier was then applied to each volume of activity during these three periods of each trial of the DNMS task. The classifier provided a readout of the probability that BOLD signal during that volume corresponded to a left face, right face, left scene, or right scene image; we will refer to this real-valued number (bounded between 0 and 1) as *left/right face/scene evidence*.

**Behavior analysis. RT regression models.** We used multiple linear regression to examine the relationship between trial-by-trial fMRI evidence for context picture reinstatement and response time. Our effect of interest is how reinstatement evidence alters responses to non-target probes, so our analysis focused exclusively on mismatch trials. All regression models contained variables reflecting the sum, over TRs, of classifier evidence for context reinstatement during each of the three trial epochs: a) the target display period (2 seconds), 2) the delay period (18 seconds), and the probe period (4 seconds).

For each DNMS trial, we computed separately the classifier evidence for the *probe* context, and the other three *non-probe* contexts summed across the delay period. We performed two regression analyses: one for *probe* context reinstatement and the other for *non-probe* reinstatement. For each model, we defined the three variables of interest as the summed classifier evidence for the context of interest, separately calculated over the target presentation, the delay period, and the probe presentation.

**Diffusion-model fits.** We modeled DNMS responses as resulting from an inference process that draws successive samples from working memory until reaching a decision. Specifically, we used the Diffusion Decision Model (DDM; Ratcliff 1978). We used a hierarchical Bayesian model fitting procedure (hDDM; Wiecki et al., 2013) to simultaneously estimate participant and group-level parameters. Our behavioral model of interest had the following free parameters: the rate of accumulation, or *drift rate*,  $v$ , trial-by-trial gaussian noise in the drift rate  $sv$ , the distance between response thresholds  $a$ , starting point of the drift  $z$ , trial-by-trial gaussian noise in the starting point of the drift  $sz$ , the *non-decision time* describing the components of stimulus perception and response preparation that are not part of the accumulation process  $t$ , and trial-by-trial gaussian noise in this non-decision time  $st$ . This parameterization follows the *extended DDM* formulation widely used to model responses in two-choice tasks (Ratcliff & Rouder, 1998).

In our models, the evidence that drives the accumulation process, the drift rate, is derived from the representation of the probe word in WM. The *higher quality* or more coherent the evidence in working memory, the larger the drift rate toward the appropriate decision boundary, and the faster the response. Here, we focus on mismatch trials (*lure* probe and *other context* probe trials) to determine whether evidence for reinstatement over the delay period sets the drift rate on each trial.

Specifically, we predicted that reinstating the probe word context in mismatch trials should result in less coherent evidence in working memory; when the probe word is not one of the targets, reinstating the context associated with the probe should activate representations of the probe word in WM. Subsequently, when participants sample from WM to decide if the probe word was a target, the activated probe in WM should cause the match response to compete with the mismatch response. This competition should diminish the strength of the drift towards the correct mismatch response.

In contrast, we predicted that reinstating the non-probe word context in mismatch trials should result in more coherent evidence in working memory; when the probe word is not one of the targets, reinstating the contexts not associated with the probe should activate representations of words other than the probe in WM. Subsequently, when participants sample from WM to decide if the probe word was a target, there should be no competition between the match response and mismatch response, allowing a relatively fast drift towards the correct response.

We also predicted that context reinstatements would account for changes in the drift rate better than the drift rate noise parameter included in the extended DDM.

To test our predictions, we used hDDM's regression functionality to estimate the relationship between trial-by-trial context evidence (computed as described in the regression analyses) and drift rate. We compared the fit of this reinstatement-driven model to the behavior-only model described above that included noise in the drift rate and starting point. The reinstatement-driven model did not include noise in the drift rate or starting point since we predicted that context reinstatements would account for variance better than noise.

In addition to the models of interest, additional model variants with other combinations of free parameters are included in the Supplementary Materials.

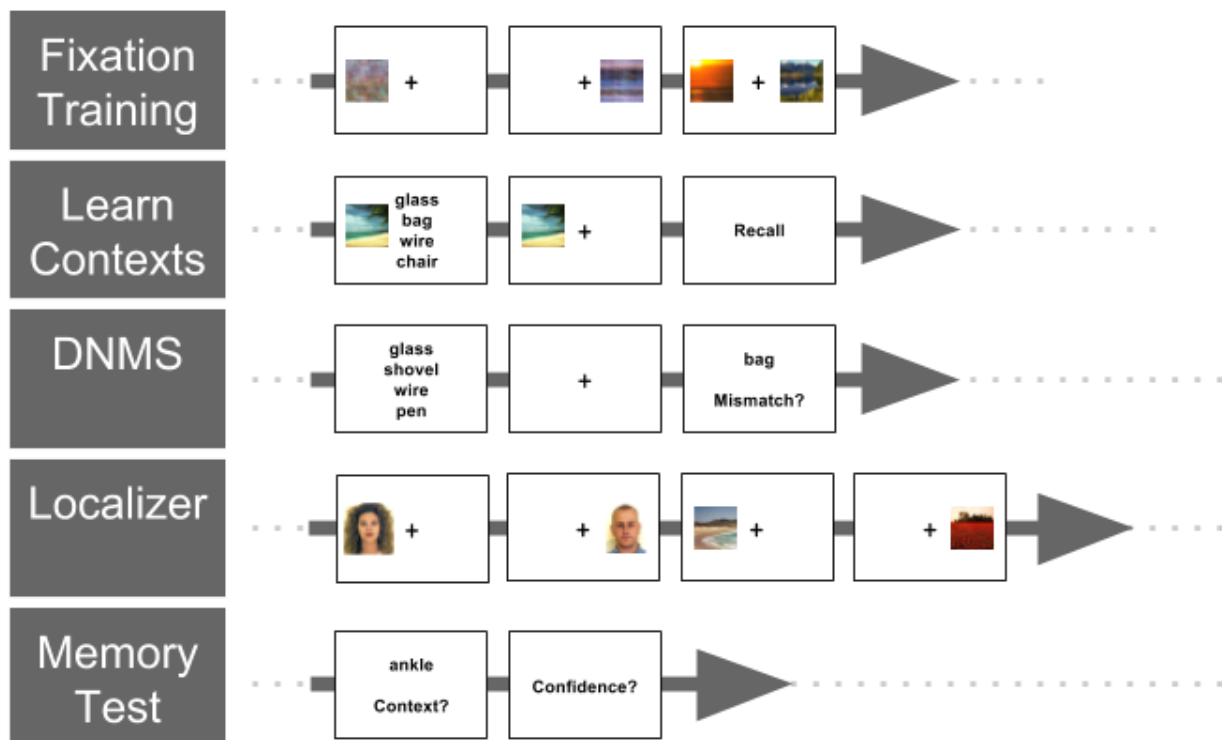
In all cases, model parameters were estimated using a Markov Chain Monte Carlo procedure (MCMC) using 100,000 samples, the first 50,000 samples treated as burn-in. Models were compared using the Deviance Information Criterion (DIC; Spiegelhalter et al., 2002), which penalizes models with higher complexity while taking account of the uncertainty in parameter

estimates. Lower DIC scores reflect better model fits. By convention, differences in DIC score greater than 3 constitute meaningful evidence (Spiegelhalter et al., 2002). For illustrative purposes, estimated  $p$ -values are reported using the formula given for AIC scores (Burnham & Anderson, 2002):

$$e^{\frac{DIC_1 - DIC_2}{2}}$$

where  $DIC_1$  is the DIC score of the superior model.

## fMRI Experiment Timeline





**Figure 6. Experiment 3 timeline.** We first trained participants to fixate on the center of the screen to ensure that they encoded pictures presented on the left side of the screen as *on the left* and pictures presented on the right side of the screen as *on the right*. This allowed us to teach participants to associate lists of words with 4 pictures: a face presented on the left, a face presented on the right, a scene presented on the left and a scene presented on the right. The order in which faces/scenes were displayed on the left/right was randomized across participants. Participants then performed the DNMS task from Experiment 2, before performing a one-back localizer task involving blocks of face, scene, object, and scrambled scene images presented on the left/right. Images used during the localizer were distinct from the task stimuli. Finally, participants reported the context with which they thought each word was associated.

## Results

**Raw response times and accuracy.** Accuracy for all participants was above chance (chance = 66.66%, mean accuracy = 85.87%, SEM = 3.68%). Accuracy did not differ between the three trial types (Target: mean = 84.44%, SEM = 3.73%; Other-context: mean = 86.25%, SEM = 3.82%; Lure: mean = 87.22%, SEM = 3.76%; all  $p > 0.2$ ).

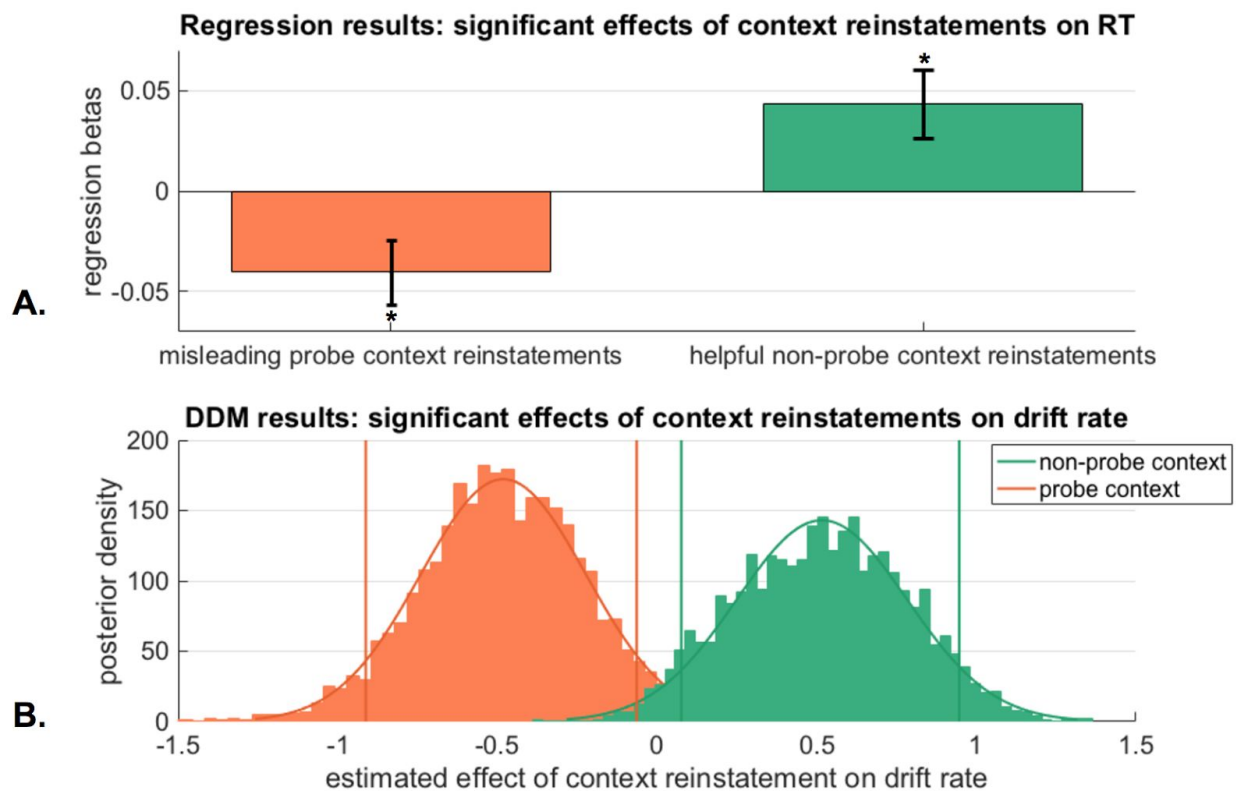
As in Experiment 2, we restricted our RT analyses to correct trials only. In contrast to Experiment 2, there was no RT difference between the two mismatch (Other-context and Lure) probe conditions ( $t(35) = -0.4702$ ,  $p = 0.64$ ), possibly reflecting an increase in the reinstatement of non-target contexts during the delay period.

Due to time restrictions, four participants were not able to complete the post-task item/context memory test. The 32 participants completed the test performed above chance (chance = 25%, mean accuracy = 41.71%, SEM = 3.40%,  $t(31) = 2.1380$ ,  $p = .04$ ).

**fMRI.** Using activity during the learning phase, we trained a pattern classifier (MVPA; Polyn & Norman 2006) to distinguish between each of the four encoding contexts. We then measured the evidence for reinstatement of each context during the delay period on each trial, computed as the sum of the classifier evidence scores for each delay-period TR. We split these into two variables: evidence for the context of the probe word, and the average of the evidence for the other three contexts. For each variable, we ran a separate multiple linear regression to predict response times. To control for any effect of context reinstatement at times other than during the delay period, we also entered into the regression the summed evidence scores from the target period (the 2 seconds during which the target were on the screen) and the probe period (the 4 seconds during which the probe word was on the screen). We ran the regression separately for each participant, and evaluated the regression weights across the population.

While in Experiment 2, we focused on the slowing of responses on lure trials vs. other context trials, our hypothesis about the effects of context-reinstatement implies that participants who are negatively affected by distracting context information could also be positively affected when they reinstate context information that helps them reach the correct decision. While reinstating the probe context on mismatch trials introduces distracting information into working memory, reinstating the non-probe contexts should increase the amount of evidence in WM that supports correctly rejecting a mismatch probe. Therefore, we predicted that participants would respond more quickly on trials where they reinstated the non-probe context (Figure 5).

When delay-period evidence for reinstatement of the probe context was higher, response times were slower (mean  $\beta = .042$ , SEM = .0148,  $t(35) = 2.7267$ ,  $p < .01$ ; Figure 7A). When evidence for delay-period reinstatements of non-probe contexts was higher, response times were faster (mean  $\beta = -.043$ , SEM .016,  $t(35) = -2.6810$ ,  $p < .05$ ; Figure 7A).



**Figure 7. Context reinstatements during the delay period predict reaction times.** **A.** Across all 36 participants, regression analyses of neural data suggest when evidence for delay-period reinstatements of non-probe contexts was higher, response times sped up (two-tailed, one sample t-test,  $p = .015$ ). Error bars reflect SEM. \* signifies  $p < .05$ . When delay-period evidence for reinstatement of the probe context was higher, response times slowed down (two-tailed, one sample t-test,  $p = .018$ ). Error bars

reflect SEM. **B.** Consistent with the regression results, the estimated effect on drift rate was different for models that included reinstating probe versus non-probe contexts; mean effect of reinstating non-probe context  $\beta = -0.52$ , 95% CrI =  $[-0.95 -0.076]$ , mean effect of reinstating probe context  $\beta = 0.48$ , 95% CrI =  $[.06 .91]$ . Vertical bars reflect 95% CrI.

Conversely, evidence for probe-context reinstatement during target presentation speeded response times (mean  $\beta=0.1347$ , SEM=.0566,  $t(35)=2.3826$ ,  $p=0.03$ ); while greater evidence for reinstatement of the non-probe context lead to a trend towards slower responses (mean  $\beta=-.1126$ , SEM=.0578,  $t(35)=-1.9464$ ,  $p=0.06$ ), possibly reflecting how well participants attended to the target set or how well they remembered the targets as learned during encoding. During the probe period, neither form of reinstatement evidence had an effect on response times (probe evidence: mean  $\beta=0.0123$ , SEM=.0448,  $t(35)=0.2752$ ,  $p = 0.78$ ; non-probe evidence: mean  $\beta=-.0037$ , SEM=.0419,  $t(35) = -0.0889$ ,  $p=0.93$ ).

**DDM Results.** We tested whether a DDM that used classifier evidence to set drift rate (the *neural DDM*) was a superior explanation of response times when compared to a standard DDM fit to behavior only (*behavior DDM*). The models were matched on all parameters, except that in the neural model, classifier evidence for memory reinstatements set the drift rate. We repeated the analysis using each kind of evidence: probe context, and non-probe context.

**Probe context evidence.** When using probe context evidence from the delay period, the neural DDM was a better fit to response times than was the behavior-only model (DIC(neural)=1136, DIC(behavior)=1144; estimated  $p=0.01$ ). Consistent with our hypothesized mechanism, probe-context reinstatement reduced the drift rate (mean  $\beta=0.418$ ;  $p=0.029$ ; Figure 7B).

The neural DDM fit did not change when trial-varying gaussian noise in the drift rate was removed compared to when it was included in the model (DIC(neural-drift rate noise)=1135; estimated  $p=.76$ ), consistent with the idea that what appears as noise in the drift rate is explainable in our task by the effect of memory reinstatement on the quality of evidence.

This result was exclusive to delay-period reinstatement evidence. When drift rate was set using either targets presentation period or response period reinstatement evidence, the neural model was not superior to the behavior model (targets presentation period: DIC(neural)=1145, estimated  $p=.82$ ; response period: DIC(neural)=1145, estimated  $p=.89$ ).

**Non-probe context evidence.** When using non-probe context evidence from the delay period, the neural DDM was a better fit to response times than was the behavior model (DIC(neural:non-probe)=1135, DIC(behavior)=1144; estimated  $p=0.01$ ). Symmetric to the probe-context result, non-probe context reinstatement increased the drift rate (mean  $\beta=-0.5164$ ;

$p=.02$ ; Figure 7B). The neural DDM fit was not improved by removing trial-varying gaussian drift rate noise compared to when it was included in the model (DIC(neural-drift rate noise)=1136; estimated  $p=.66$ ),

Again, this finding was constrained to the delay period. When we specified the drift rate using targets presentation period or response period reinstatement of non-probe context, the targets presentation model was worse than the behavioral DDM (targets presentation: DIC(neural)=1151,  $p=0.04$ ) and the response period model was not different than the behavioral model (response period: DIC(neural) 1144,  $p=.82$ ).

## General Discussion

Over a series of three experiments, we tested the hypothesis that episodic memory reinstatement influences working memory maintenance. In Experiment 1, using a delayed recall task, we showed that interfering with working memory maintenance caused participants to make substitution errors that reflected the encoding context of the target words. The longer was the interference (six seconds, or the entire 18 seconds), the greater was the proportion of errors, and the greater was the proportion of errors reflecting encoding context. Absent interference, performance was near ceiling, showing no apparent influence of episodic memory.

Experiment 2 revealed that even ceiling performance reflects the influence of episodic memory. On a delayed non-match to sample task (DNMS) with a distraction-free 18 second delay, participants were slowed in their responding to lure probes — words that shared an encoding context with the target set, but were not actually members of the target set.

Experiment 3 repeated the DNMS task from Experiment 2, using fMRI to measure evidence for episodic memory reinstatement during the delay period. This analysis revealed that the degree of response slowing on each trial resulted from the specific content of episodic memory reinstatement during the delay period on that trial. Model-based analysis using two variants of a Drift-Diffusion Model (DDM) revealed that the effect of reinstatement on response time could be captured by letting reinstatement evidence vary the DDM drift rate on each trial, consistent with the hypothesis that episodic memory reinstatement during the delay period sets the evidence used to make responses at the time of probe.

Together, these results establish a mechanism by which working memory maintenance is influenced by covert, spontaneous retrievals from episodic memory, even in the absence of distraction.

## Converging evidence from animal studies

Converging evidence from studies in rodents and monkeys supports the idea that neural machinery underlying episodic memory may influence working memory, even over 30 second or shorter delays, and even in the absence of overt distraction. In rats and macaques, entorhinal cortex is engaged during, and predicts behavior in, DMS tasks (Suzuki et al., 1997; Young et al., 1997). Also in rats, Farovik and colleagues (2010) showed that hippocampal lesions induce severe impairments on a non-spatial test of short-term memory for sequences of odors. Electrophysiological data from rodent hippocampus have revealed hippocampal activation over a 30 second delay period predicts accuracy (Deadwyler & Hampson, 2004). This converging evidence suggests the neural machinery underlying working memory and episodic memory may support each other, even over 30 second or shorter delays, and in the absence of overt distraction.

## Relevance to attractor models of WM maintenance

One framework for understanding WM maintenance is an *attractor* model. Attractors are dynamically stable patterns of neuronal activity, each of which represents the memory of a specific item with a unique firing pattern (Wang, 2001; Hopfield, 1982). Modeling work has demonstrated that during WM maintenance, multiple attractors can be simultaneously active, with each attractor maintaining a separate piece of information (Laing et al., 2002).

Attractor networks fluctuate over delay periods, tending to drift randomly as a diffusion process (Camperi & Wang, 1998; Compte et al., 2000). Single cell recordings from the prefrontal cortex in macaques suggests the extent of drift away from an attractor during WM maintenance was predictive of errors on a WM task (Wimmer et al., 2014). While the prefrontal cortex has traditionally been associated with WM, Kaminiski and colleagues (2017) replicated these findings in humans with cells in the MTL.

What causes neural trajectories to drift away from their attractors during WM maintenance? Theoretical work suggests attractor networks can drift due to proximity to other attractors in state space (Wang, 2001; Camperi & Wang, 1998). One direction for future research is to investigate whether context reinstatements from EM during the delay period affect the drift of an attractor network during WM maintenance. Do repeated context reinstatements over longer delay periods push neural trajectories towards other attractors associated with the reinstated contexts?

# Acknowledgements

The authors would like to thank Michael Shvartsman for help with model comparison analysis, Nicholas H. DePinto for technical support with the fMRI scanner and MR compatible eye tracker, and Michael J. Frank for helpful comments.

# Contributions

A.N.H. and A.M.B. conceived the experiment; A.N.H., A.M.B., J.D.C., and K.A.N. designed the experiment and analyses; A.N.H. wrote the experiment code; A.N.H. ran the experiment; A.N.H. and A.M.B. performed the analyses; A.N.H. and A.M.B. wrote the paper, with input from J.D.C. and K.A.N.

# References

- Aggleton, J. P., Hunt, P. R., & Rawlins, J. N. (1986). The effects of hippocampal lesions upon spatial and non-spatial tests of working memory. *Behavioural Brain Research*, 19(2), 133–146. [https://doi.org/10.1016/0166-4328\(86\)90011-2](https://doi.org/10.1016/0166-4328(86)90011-2)
- Axmacher, N., Mormann, F., Fernández, G., Cohen, M. X., Elger, C. E., & Fell, J. (2007). Sustained neural activity patterns during working memory in the human medial temporal lobe. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, 27(29), 7807–7816. <https://doi.org/10.1523/JNEUROSCI.0962-07.2007>
- Baddeley, A. (1992). Working Memory Alan Baddeley. *Science*, 255(5044), 556–559. <https://doi.org/10.1126/science.1736359>
- Baddeley, A. D., & Hitch, G. (1974). Working memory. *The Psychology of Learning and Motivation Advances in Research and Theory*. [https://doi.org/10.1016/S0079-7421\(08\)60452-1](https://doi.org/10.1016/S0079-7421(08)60452-1)
- Baddeley, A. D., & Hitch, G. J. (2000). Development of Working Memory: Should the Pascual-Leone and the Baddeley and Hitch Models Be Merged? *Journal of Experimental Child Psychology*, 77(2), 128–137. <https://doi.org/10.1006/jecp.2000.2592>
- Bramão, I., & Johansson, M. (2016). Benefits and Costs of Context Reinstatement in Episodic Memory: An ERP Study. *Journal of Cognitive Neuroscience*, 26(3), 1–13. [https://doi.org/10.1162/jocn\\_a\\_01035](https://doi.org/10.1162/jocn_a_01035)
- Camperi, M., & Wang, X. J. (1998). A model of visuospatial working memory in prefrontal cortex: Recurrent network and cellular bistability. *Journal of Computational Neuroscience*, 5(4), 383–405. <https://doi.org/10.1023/A:1008837311948>
- Cave, C. B., & Squire, L. R. (1992). Intact verbal and nonverbal short-term memory following damage to the human hippocampus. *Hippocampus*, 2(2), 151–163. <https://doi.org/10.1002/hipo.450020207>

- Compte, A., Brunel, N., Goldman-Rakic, P. S., & Wang, X. J. (2000). Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. *Cerebral Cortex*, 10(9), 910–923.
- D'Esposito, M., Postle, B. R., & Rypma, B. (2000). Prefrontal cortical contributions to working memory: evidence from event-related fMRI studies. *Experimental Brain Research*, 133(1), 3–11. <https://doi.org/10.1007/s002210000395>
- Deadwyler, S. A., & Hampson, R. E. (2004). Differential but complementary mnemonic functions of the hippocampus and subiculum. *Neuron*, 42(3), 465–476. [https://doi.org/10.1016/S0896-6273\(04\)00195-3](https://doi.org/10.1016/S0896-6273(04)00195-3)
- Drachman, D. A., & Arbit, J. (1966). Memory and the Hippocampal Complex. *Archives of Neurology*, 15, 52–61. <https://doi.org/10.1001/archneur.1964.00460160081008>
- Farovik, A., Dupont, L. M., & Eichenbaum, H. (2010). Distinct roles for dorsal CA3 and CA1 in memory for sequential nonspatial events. *Learning & Memory (Cold Spring Harbor, N.Y.)*, 17(1), 12–17. <https://doi.org/10.1101/lm.1616209>
- Hannula, D. E., Tranel, D., & Cohen, N. J. (2006). The Long and the Short of It: Relational Memory Impairments in Amnesia, Even at Short Lags. *Journal of Neuroscience*, 26(32), 8352–8359. <https://doi.org/10.1523/JNEUROSCI.5222-05.2006>
- Hellerstedt, R., & Johansson, M. (2014). Electrophysiological correlates of competitor activation predict retrieval-induced forgetting. *Cerebral Cortex*, 24(6), 1619–1629. <https://doi.org/10.1093/cercor/bht019>
- Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences of the United States of America*, 79(8), 2554–2558. <https://doi.org/10.1073/pnas.79.8.2554>
- Howard, M. W., & Kahana, M. J. (2002). A Distributed Representation of Temporal Context. *Journal of Mathematical Psychology*, 46(3), 269–299. <https://doi.org/10.1006/jmps.2001.1388>
- Hupbach, A., Gomez, R., & Nadel, L. (2009). Episodic memory reconsolidation: Updating or source confusion? *Memory (Hove, England)*, 17(5), 502–510. <https://doi.org/10.1080/09658210902882399>
- Kamiński, J., Sullivan, S., Chung, J. M., Ross, I.B., Mamelak, A.N., Rutishauser, U. (2017). Persistently active neurons in human medial frontal and medial temporal lobe support working memory. *Nature Neuroscience*, 20, 590–601. <https://doi.org/10.1038/nn.4509>
- Kliegl, O., Pastötter, B., & Bäuml, K.-H. T. (2015). The Contribution of Encoding and Retrieval Processes to Proactive Interference. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, 41(6), 1778–1789. <https://doi.org/10.1037/xlm0000096>
- Laing, C. R., Troy, W. C., Gutkin, B., & Ermentrout, G. B. (2002). Multiple bumps in a neuronal model of working memory. *Siam Journal on Applied Mathematics*, 63(1), 62–97. <https://doi.org/10.1137/S0036139901389495>
- Lewis-Peacock, J. A., Cohen, J. D., & Norman, K. A. (2016). Neural evidence of the strategic choice between working memory and episodic memory in prospective remembering. *Neuropsychologia*, 93, 280–288. <https://doi.org/10.1016/j.neuropsychologia.2016.11.006>
- Murray, E. A., & Mishkin, M. (1998). Object recognition and location memory in monkeys with excitotoxic lesions of the amygdala and hippocampus. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, 18(16), 6568–6582.
- Nemanic, S., Alvarado, M. C., & Bachevalier, J. (2004). The hippocampal/parahippocampal regions and

- recognition memory: insights from visual paired comparison versus object-delayed nonmatching in monkeys. *J Neurosci*, 24(8), 2013–2026. <https://doi.org/10.1523/JNEUROSCI.3763-03.2004>
- Newmark, R. E., Schon, K., Ross, R. S., & Stern, C. E. (2013). Contributions of the hippocampal subfields and entorhinal cortex to disambiguation during working memory. *Hippocampus*, 23(6), 467–475. <https://doi.org/10.1002/hipo.22106>
- Nichols, E. A., Kao, Y. C., Verfaellie, M., & Gabrieli, J. D. E. (2006). Working memory and long-term memory for faces: Evidence from fMRI and global amnesia for involvement of the medial temporal lobes. *Hippocampus*, 16(7), 604–616. <https://doi.org/10.1002/hipo.20190>
- Olsen, R. K., Nichols, E. A., Hunt, J. F., Chen, J., Glover, G. H., Gabrieli, J. D. E., & Wagner, A. D. (2009). Performance-related sustained and anticipatory activity in human medial temporal lobe during delayed match-to-sample. *The Journal of Neuroscience*, 29(38), 11880. <https://doi.org/10.1523/JNEUROSCI.2245-09.2009>
- Paivio, A., & Okovita, H. W. (1971). Word imagery modalities and associative learning in blind and sighted subjects. *Journal of Verbal Learning and Verbal Behavior*, 10(5), 506–510. [https://doi.org/10.1016/S0022-5371\(71\)80021-X](https://doi.org/10.1016/S0022-5371(71)80021-X)
- Paivio, A., & Yuille, J. C. (1969). Changes in associative strategies and paired-associate learning over trials as a function of work imagery and type of learning set. *Journal of Experimental Psychology; Journal of Experimental Psychology*, 79(3, Pt.1), 458–463. <https://doi.org/10.1037/H0026929>
- Polyn, S. M., Natu, V. S., Cohen, J. D., & Norman, K. A. (2005). Category-Specific Cortical Activity Precedes Retrieval During Memory Search. *Science (New York, N.Y.)*, 310(5756), 1963–6. <https://doi.org/10.1126/science.1117645>
- Race, E., LaRocque, K. F., Keane, M. M., & Verfaellie, M. (2013). Medial temporal lobe contributions to short-term memory for faces. *Journal of Experimental Psychology. General*, 142(4), 1309–22. <https://doi.org/10.1037/a0033612>
- Ranganath, C. (2005). Working memory for visual objects: Complementary roles of inferior temporal, medial temporal, and prefrontal cortex. *Neuroscience*, 139(1), 277–289. <https://doi.org/10.1016/j.neuroscience.2005.06.092>
- Ranganath, C., & Blumenfeld, R. S. (2005). Doubts about double dissociations between short- and long-term memory. *Trends in Cognitive Sciences*. <https://doi.org/10.1016/j.tics.2005.06.009>
- Ranganath, C., Cohen, M. X., Dam, C., & Esposito, M. (2004). Inferior temporal, prefrontal, and hippocampal contributions to visual working memory maintenance and associative memory retrieval. *J Neurosci*, 24(16), 3917–3925. <https://doi.org/10.1523/JNEUROSCI.5053-03.2004>
- Ranganath, C., D'Esposito, M., Friederici, A. D., & Ungerleider, L. G. (2005). Directing the mind's eye: prefrontal, inferior and medial temporal mechanisms for visual working memory This review comes from a themed issue on Cognitive neuroscience Edited. *Current Opinion in Neurobiology*, 15, 175–182. <https://doi.org/10.1016/j.conb.2005.03.017>
- Ranganath, C., & D'Esposito, M. (2001). Medial temporal lobe activity associated with active maintenance of novel information. *Neuron*, 31(5), 865–873. [https://doi.org/10.1016/S0896-6273\(01\)00411-1](https://doi.org/10.1016/S0896-6273(01)00411-1)
- Ranganath, C., Yonelinas, A. P., Cohen, M. X., Dy, C. J., Tom, S. M., & D'Esposito, M. (2004). Dissociable correlates of recollection and familiarity within the medial temporal lobes. *Neuropsychologia*, 42(1), 2–13. <https://doi.org/10.1016/j.neuropsychologia.2003.07.006>
- Ratcliff, R., & Rouder, J. (1998). Modeling response times for two-choice decisions. *Psychological*



- Science*, 9, 347–356. <https://doi.org/10.1111/1467-9280.00067>
- Repovš, G., & Baddeley, A. (2006). The multi-component model of working memory: Explorations in experimental cognitive psychology. *Neuroscience*, 139(1), 5–21. <https://doi.org/10.1016/j.neuroscience.2005.12.061>
- Rose, N. S., Buchsbaum, B. R., & Craik, F. I. M. (2014). Short-term retention of a single word relies on retrieval from long-term memory when both rehearsal and refreshing are disrupted. *Memory & Cognition*, 42(5), 689–700. <https://doi.org/10.3758/s13421-014-0398-x>
- Schon, K., Hasselmo, M. E., Lopresti, M. L., Tricarico, M. D., & Stern, C. E. (2004). Persistence of parahippocampal representation in the absence of stimulus input enhances long-term encoding: a functional magnetic resonance imaging study of subsequent memory after a delayed match-to-sample task. *J Neurosci*, 24(49), 11088–11097. <https://doi.org/10.1523/JNEUROSCI.3807-04.2004>
- Shallice, T., & Warrington, E. K. (1970). Independent functioning of verbal memory stores: a neuropsychological study. *The Quarterly Journal of Experimental Psychology*, 22(2), 261–273. <https://doi.org/10.1080/0033557043000203>
- Squire, L. R. (1992). Memory and the Hippocampus : A Synthesis From Findings With Rats, Monkeys, and Humans. *Psychological Review*, 99(2), 195–231. <https://doi.org/10.1037/0033-295X.99.3.582>
- Stern, C. E., Sherman, S. J., Kirchhoff, B. A., & Hasselmo, M. E. (2001). Medial temporal and prefrontal contributions to working memory tasks with novel and familiar stimuli. *Hippocampus*, 11(4), 337–346. <https://doi.org/10.1002/hipo.1048>
- Suzuki, W. A., Miller, E. K., & Desimone, R. (1997). Object and place memory in the macaque entorhinal cortex. *J Neurophysiol*, 78(2), 1062–1081.
- Tulving, E. (1983). Elements of Episodic Memory. *Canadian Psychology*, 26(3), 351. <https://doi.org/http://dx.doi.org/10.1017/S0140525X0004440X>
- Unsworth, N., & Engle, R. W. (2007). Individual Differences in Working Memory Capacity and Retrieval: A Cue-Dependent Search Approach. *The Foundations of Remembering: Essays in Honor of Henry L. Roediger, III*.
- Wang, X. J. (2010). Attractor Network Models. In *Encyclopedia of Neuroscience* (pp. 667–679). <https://doi.org/10.1016/B978-008045046-9.01397-8>
- Wilson, M. (1988). MRC Psycholinguistic Database : Machine-usable dictionary , version 2 .00. *Behavior Research Methods, Instruments, & Computers*, 20(1), 6–10. <https://doi.org/10.3758/BF03202594>
- Wimmer, K., Nykamp, D. Q., Constantinidis, C., & Compte, A. (2014). Bump attractor dynamics in prefrontal cortex explains behavioral precision in spatial working memory. *Nature Neuroscience*, 17(3), 431–9. <https://doi.org/10.1038/nn.3645>
- Young, B. J., Otto, T., Fox, G. D., & Eichenbaum, H. (1997). Memory representation within the parahippocampal region. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, 17(13), 5183–5195.
- Zanto, T. P., Clapp, W. C., Rubens, M. T., Karlsson, J., & Gazzaley, A. (2016). Expectations of Task Demands Dissociate Working Memory and Long-Term Memory Systems. *Cerebral Cortex*, 26(3), 1176–1186. <https://doi.org/10.1093/cercor/bhu307>

## Supplement

**Supplemental Table 1: Drift diffusion model from only mismatch trials, all subjects.** “ $v \sim \text{context}$ ” signifies drift rate was set by context reinstatement evidence from the classifier. Behavior models did not include any neural evidence. Relevant comparisons in this table are the neural models that include the same parameters (each row) to the behavioral models with the same parameters (above each row).

|   | delay<br>period:<br>probe<br>context | delay<br>period:<br>not probe<br>context | targets<br>period:<br>probe<br>context | targets<br>period:<br>not probe<br>context | response<br>period:<br>probe<br>context | response<br>period:<br>not probe<br>context |
|---|--------------------------------------|--|--|--|---|---|
| behavior model, sv, st, sz, z           | 1144                                 |  |  |  |   |   |
| $v \sim \text{context}$ , sv, st, sz, z | 1136                                 | 1135                                     | 1145                                   | 1151                                       | 1145                                    | 1144  |
| behavior model, st, sz, z               | 1139                                 |  |  |  |   |   |
| $v \sim \text{context}$ , st, sz, z     | 1135                                 | 1136                                     | 1143                                   | 1148                                       | 1143                                    | 1146  |
| behavior model, sv, st, sz              | 1142                                 |  |  |  |   |   |
| $v \sim \text{context}$ , sv, st, sz,   | 1142                                 | 1136                                     | 1145                                   | 1147                                       | 1147                                    | 1146  |
| behavior model, st, sz                  | 1145                                 |  |  |  |   |   |
| $v \sim \text{context}$ , st, sz        | 1139                                 | 1139                                     | 1145                                   | 1143                                       | 1143                                    | 1145  |
| behavior model, sv                      | 1141                                 |  |  |  |   |   |
| $v \sim \text{context}$ , sv            | 1134                                 | 1131                                     | 1144                                   | 1143                                       | 1141                                    | 1141  |
| behavior model, no noise                | 1140                                 |  |  |  |   |   |
| $v \sim \text{context}$ , no noise      | 1133                                 | 1134                                     | 1138                                   | 1144                                       | 1142                                    | 1136  |