# Refresh my memory: Episodic memory reinstatements intrude on working memory maintenance

Abigail N. Hoskin[1], Aaron M. Bornstein[2], Kenneth A. Norman[1,2], Jonathan D. Cohen[1,2]

1. Department of Psychology, Princeton University, Princeton, NJ, USA
2. Neuroscience Institute, Princeton University, Princeton, NJ, USA

**Abstract**

A fundamental question in memory research is how different forms of memory interact. Previous research has shown that when working memory (WM) is overloaded or maintenance is interrupted in short-term memory tasks, humans and animals can rely on episodic memory (EM) to support performance. Furthermore, episodic memory reactivation appears also to occur on its own (i.e., irrespective of demand), even during the short delays typically used in WM experiments. Based on these observations, we hypothesized that EM reinstatements would affect WM, even in the absence of any interference. Using novel behavioral and neural signatures of the effect of EM on WM, we show that EM introduces additional information into WM by reinstating incidental associations (*context*) present during initial encoding. The first two experiments establish that the influence of encoding context is evident both in errors (Experiment 1) and in slowing of responses (Experiment 2). Experiment 3 shows that fMRI evidence of EM reinstatement during the delay predicts response times on each trial. Modeling WM search using a Drift-Diffusion Model (DDM), we show that fits improve when trial drift rate varies with fMRI evidence for reinstatement during that trial's delay period. These results expose a previously hidden interaction between WM maintenance and EM replay, and raise new questions about the adaptive nature of the interplay between these mechanisms.

## Introduction

Our memories do not exist in isolation, and neither do the neural circuits that represent 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; Repov & Baddeley 2006). Experiences can also simultaneously lay down more lasting traces as episodic memories, available to 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).

Early models proposed that working memory and long-term memory operated wholly in parallel (Shallice & Warrington, 1970). 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). More recent models propose that they support each other (Baddeley, 2000). There is accumulating evidence that episodic memory, and its neural substrates in the MTL, are engaged during short-term memory tasks that also engage working memory (Ranganath, 2005; Ranganath et al. 2004, 2005; Ranganath & Blumenfeld 2005; Axmacher et al. 2007), suggesting these memory systems do not operate entirely independently of one another.

Experiments testing for an interaction between these two types of memory have largely focused on the hypothesis that episodic memory is used to support working memory when maintenance is disrupted, leading to errors that reflect features of episodic memory. For instance, participants show proactive interference from recently studied stimuli when working memory is disrupted for 18 seconds (Wickens et al., 1976). Findings of this sort, showing that episodic memory contributes if working memory is disrupted, raise the question of whether episodic memory only contributes when working memory is disrupted, or whether it contributes more ubiquitously. This is the question we set out to address.

A growing number of studies indicated that, during periods of rest, the neural structures that support episodic memory are often active (Buckner, 2010), and appear to be reinstating recent experiences (Tambini et al., 2010) or activating potential future scenarios constructed on the basis of past experiences (Buckner & Carroll, 2007). These reinstatements trigger coordinated activity patterns across a broad swath of cortical regions, including those presumably involved in working memory maintenance such as prefrontal cortex (Miller & Cohen, 2001). This widespread activation is reliably present even during brief lapses in external stimulation (Logothetis et al., 2012), such as those typically used as maintenance periods in working memory experiments. These observations lead us to ask the question: Do these ongoing reinstatements from episodic memory affect the content of working memory, even when the latter is not being disrupted? Though such an effect had not previously been observed using measures of accuracy (e.g., substitution errors), we reasoned that an influence on working memory search might nevertheless be revealed in analyses of response time.

Here, we leverage the fact that retrievals from episodic memory carry with them temporal and associative context (Howard & Kahana, 2002), such that triggering the recall of one memory from a given context can cause the subsequent, involuntary recall of other memories sharing that context (Hupbach et al. 2009; Bornstein & Norman, 2017). This can occur even at the short delays typically associated with WM (Hannula et al., 2006). Therefore, we reasoned that — if reinstatements from episodic memory occurred during WM maintenance — these reinstatements would likely be of memories that shared an encoding context with the target stimuli. Even if these reinstated memories do not lead to overt errors, they may compete with other, task-relevant representations being maintained in working memory, and thereby affect search and response times.

We present three experiments testing the hypothesis that context reinstated from episodic memory intrudes on working memory maintenance. In Experiment 1, we test whether participants intrude same-context items in response to interference in a classic short-term delayed recall task. In Experiment 2, we examine whether the influence of reinstated context is evident in response times, even when accuracy is at ceiling. In Experiment 3, we repeat the task from Experiment 2 in an fMRI scanner, and use multivariate pattern analysis (MVPA) to generate a trial-by-trial neural measure of how likely it was that participants were recalling a specific past context. We used this neural index of reinstatement to predict the degree of response time bias on a given trial. Finally, we model working memory search as a thresholded evidence accumulation process (Ratcliff, 1978), and test whether model fit is improved by allowing drift rate — reflecting the consistency of information in working memory — to vary with reinstatement evidence on each trial.

Together, the results of these experiments reveal a new aspect of the interaction between episodic and working memory: Ongoing reinstatements from episodic memory influence the content of working memory. Critically, these reinstatements carry incidental associations from the context of the encoded targets, introducing new items that were not previously present in working memory, and so raise new questions about the interplay of working memory maintenance and episodic memory reinstatement.

# Results

## Experiment 1

15 participants participated in the study, which had two phases. The goal of the initial *context learning* phase was to associate words with distinct encoding contexts. In this phase, participants learned to associate six non-overlapping sets of 12 words with a unique picture (one per set); each combination of 12 words and a picture comprised a single encoding context. In the second phase, participants performed a short-term delayed-recall task in which (on each trial) they saw a set of four target words that they needed to recall after an 18-second delay; we manipulated whether the delay was unfilled (*no distraction*), or participants did 6 seconds of backwards counting at the outset of the delay (*break distraction*, as in: we were briefly "breaking up" their ability to maintain information in working memory), or participants counted backwards throughout the full 18-second delay (*full distraction*). Previous studies using this kind of short-term recall test have found that distraction during the delay causes participants to rely on episodic rather than working memory, as evidenced by the fact that errors are primarily words substituted from recent trials (Brown, 1958; Peterson & Peterson, 1959; Rose et al., 2014; Lewis-Peacock et al., 2016; Zanto et al., 2016). Here, we tested whether these substitutions can be biased by the encoding context of the target words. Specifically, if the four target words are sampled from one of the 12-word encoding contexts established at the outset of the experiment, does this lead to substitution of other (non-target) words from the same context? The logic of the study is shown in Figure 1, and examples of the initial context learning and delayed recall trials are shown in Figures 2A and 2B, respectively.

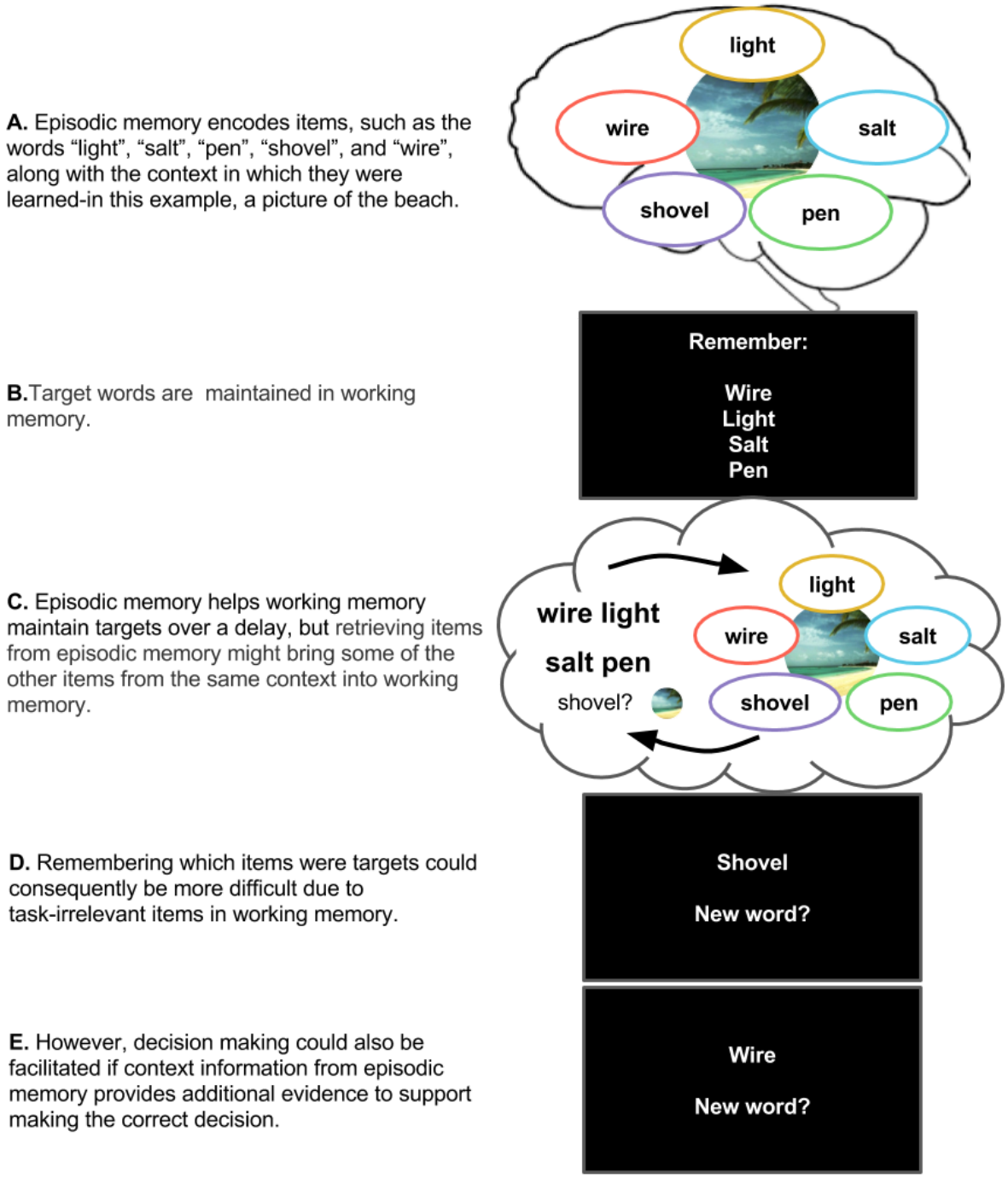


**Figure 1. Episodic memory can inject incidental information into working memory. A.** Episodic memory encodes items along with the context in which they were learned. **B.** When presented with target items to maintain over a delay period, working memory maintenance may be periodically influenced by reinstatements from episodic memory. **C.** These reinstatements may contain other items

sharing the encoding context of the target items. **D.** These items might affect subsequent behavior, by impeding decision making when these items support the incorrect decision; **E.** and/or by facilitating decision making when they support the correct decision.

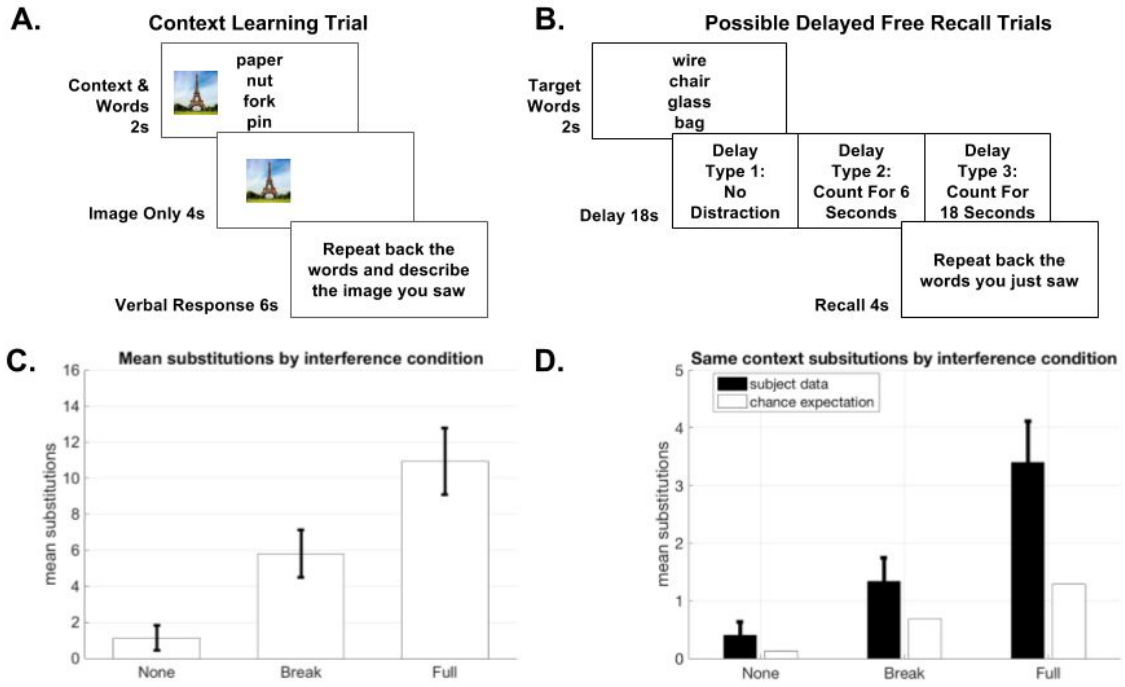


**Figure 2. Experiment 1: Free recall task with added context. A.** Participants (n = 15) studied lists of words in contexts distinguished by different, lateralized pictures. On each context learning trial, 4 words were linked to the picture. Across context learning trials, each picture was linked to 12 words (paired three times each to that picture, leading to strong item-context associations). **B.** We probed how these contexts 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 condition ($t(14) = 3.2756$; $p < .01$; paired, two-sided t-test), and more errors in the full distraction versus no distraction condition ($t(14) = 6.4526$, $p < .001$; $p < .01$; paired, two-sided 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, two-sided t-test). **D.** In all three conditions, participants made errors that reflected the influence of reinstated context. Specifically, participants made substitution errors during recall that reflected 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. This suggests that context information from episodic memory entered working memory, even when working memory was not overloaded. Error bars reflect SEM. * signifies $p < .05$, ** signifies $p < .01$, *** signifies $p < .001$.

We expected to see increasing numbers of substitutions as the demands on working memory increased; therefore, we predicted participants would make the fewest substitutions following delays with no 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, two-sided t-test) and 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 2C).

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: *previous-target substitutions*, *same-context substitutions*, and *other errors*. These categories were motivated by the following considerations. First 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 therefore mistakenly invoke an a target response. We refer to these as *previous-target* substitutions. Second, we expected that maintaining target words in working memory would trigger episodic memory reinstatement of the context in which these words were studied (Howard & Kahana, 2002; Gershman et al., 2013). If this occurs, we should see an elevated substitution rate for the eight words that were studied in the same context as the target words, but that were not part of the current trial's target set. We refer to these as *same context* substitutions. The context from which the target words were drawn changed with each trial, ensuring that previous-target and same context substitutions were mutually exclusive possibilities. Finally, we refer to substitutions from one of the 56 remaining words learned in the experiment, that were neither targets, *previous-target* or *same context* errors, 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 if the errors were drawn at random from the 68 possible non-target words. While all three kinds of words should be retained in episodic memory, the effects of recency and context lead us to predict that words from *previous-target* and *same context* errors should be overrepresented relative to *other* errors.

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, on full interference trials, the proportion of *same context* substitutions was greater than what would be

expected by chance (as computed by a bootstrap analysis; $p = .0001$). This suggests that context information was indeed affecting decision making when working memory was overloaded (Figure 2D). Same context substitutions were also greater than what would be expected by chance in the break condition ($p = .0001$). Critically, although the quantity of substitutions participants made on the no working memory interference trials was small (mean = 1.13; Figure 2C), when they did occur, they were biased towards coming from the same context as the target words ($p = .0025$).

## 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, and historically taken as evidence that episodic memory did not influence working memory function. Errors increased when participants were made to perform a distractor task midway through the delay, and were further increased when the distractor task spanned 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 et al., 2016; Zanto et al., 2016; Rose et al., 2014). 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, especially when retention in working memory is subject to interference.

This finding leaves two unanswered questions. First, does episodic memory support working memory in the absence of external distraction? While substitutions in the no distraction condition were significantly biased toward being from the same encoding context as the target words, there were too few errors (of any kind) to support meaningful interpretation. Second, when during the task does episodic memory retrieval occur, and how does it influence performance? Are episodic memories retrieved during the delay, either incidentally and/or to support maintenance, or strictly at the time of response? We use the signature of context effects established in Experiment 1 to address these questions in Experiments 2 and 3.

# Experiment 2

# Results

In Experiment 2, we tested whether episodic memory influenced working memory maintenance even in the absence of distraction. We predicted that response time (RT) would reveal the intrusion of non-target items into working memory, even if these did not measurably influence accuracy.

32 participants performed a context-learning session as in Experiment 1 (Figure 3A), followed by a two alternative forced-choice, delayed non-match to sample task (DNMS; Figure 3B). On each trial of the DNMS task, participants were asked to maintain four *target* items in memory over an uninterrupted delay of 18 seconds. They were then required to decide whether a given *probe* word was distinct from all 4 target items (in which case they responded "yes", indicating a non-match) or whether it matched one of the target items (in which case they responded "no"). Critically, the four target words were all chosen from the same encoding context. The test phase (Figure 3B) thus had three types of trials, in which: 1) the probe was one of the targets (*target trials)*; 2) the probe came from the same encoding context as the targets (*lure trials*); or 3) the probe came from another encoding context (*other context trials*).

***Accuracy.*** As expected, given the absence of distraction, accuracy was high across all three conditions (mean = 96.25%, SEM = 0.89%) with no difference in accuracy between *target* (mean = 97.03%, SEM = 0.70%), *lure* (mean = 95.00%, SEM = 1.19%), and *other context* trials (mean = 96.72%, SEM = 0.76%) ($p > .2$ by paired t-test for all pairwise comparisons; 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.

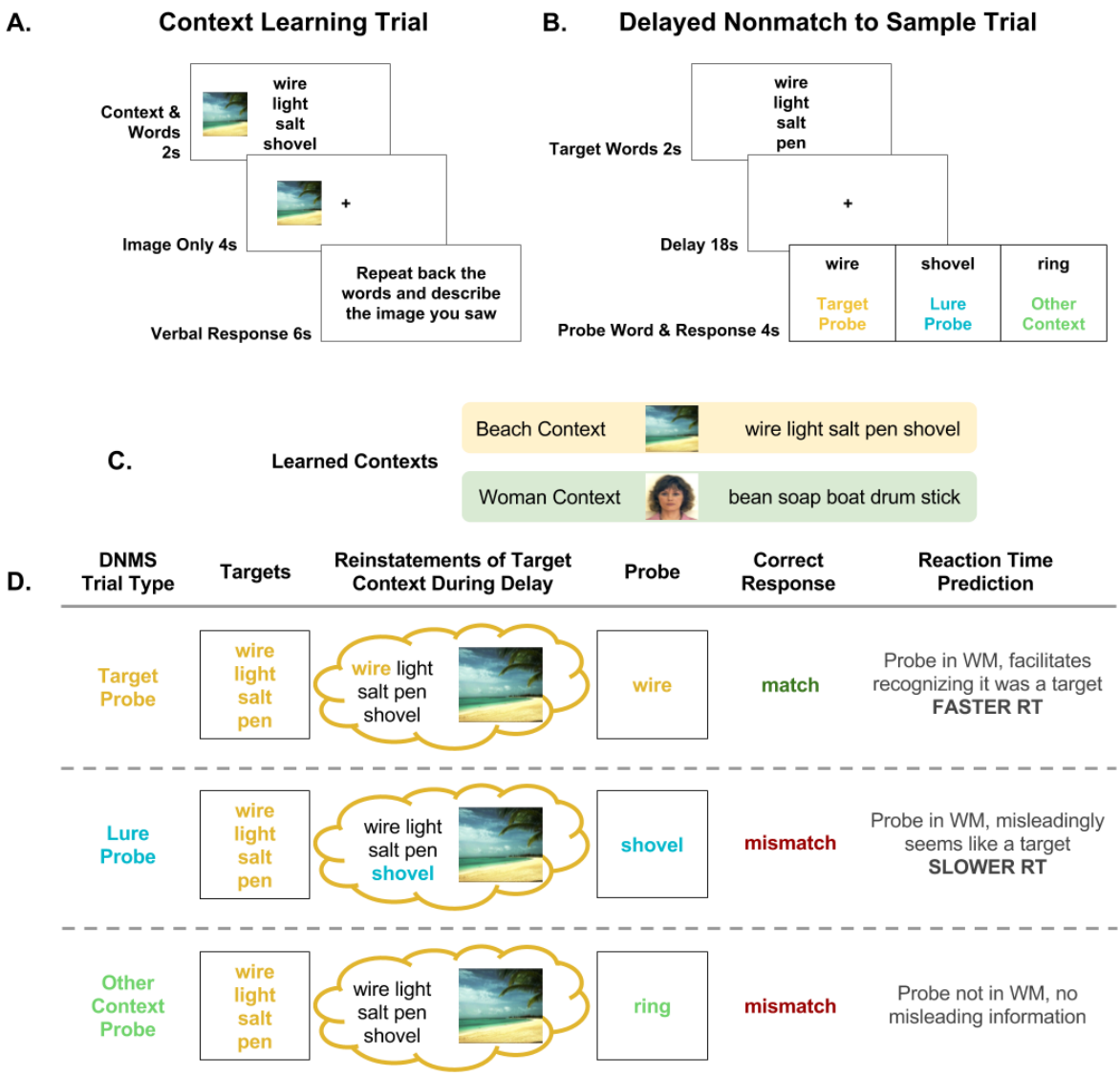
A.
Context Learning Trial
Context & Words 2s
wire
light
salt
shovel
+
Image Only 4s
Repeat back the words and describe the image you saw
Verbal Response 6s
B.
Delayed Nonmatch to Sample Trial
Target Words 2s
wire
light
salt
pen
+
Delay 18s
wire
shovel
ring
Target Probe
Lure Probe
Other Context
Probe Word & Response 4s
C.
Learned Contexts
Beach Context
wire light salt pen shovel
Woman Context
bean soap boat drum stick
D.
DNMS Trial Type
Targets
Reinstatements of Target Context During Delay
Probe
Correct Response
Reaction Time Prediction
Target Probe
wire light salt pen
wire light salt pen shovel
wire
match
Probe in WM, facilitates recognizing it was a target
FASTER RT
Lure Probe
wire light salt pen
wire light salt pen shovel
shovel
mismatch
Probe in WM, misleadingly seems like a target
SLOWER RT
Other Context Probe
wire light salt pen
wire light salt pen shovel
ring
mismatch
Probe not in WM, no misleading information

**Figure 3. Experiment 2: DNMS task with added context. A.** In the *context learning* phase, participants studied 48 words that were split into four sets of 12 words. Each set was paired with a unique *context picture* of a face or scene. The context picture was consistently displayed on either the left or the right 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 during the context learning phase. 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 the actual target words. Since this 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 *other probe* trials.

***Reaction times.*** We predicted that participants would on average respond fastest to *target probes*, as the probe word would most reliably match the contents of working memory (Figure 3D). 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 (a prediction that follows from both serial and parallel models of working memory search — Sternberg, 1969; Ratcliff, 1978).

On *non-target probe* trials, which included *lure* and *other context* probes, participants had to make the same response: to reject the probe word as one of the targets. Thus, any difference in RT between these two trial types could not be attributed to the valence of the required response.

Within *non-target probe* trials, we predicted that participants would be slower to respond to *lure* than *other context* probes. This is because, if context reinstatements from episodic memory activate trial irrelevant items from the same context as the target words, lure words would become activated in working memory. If lures are activated in working memory, lure probes would match the context information in working memory while also being a mismatch with the target words in working memory, leading to confusion and slower RTs. *Other context* probes

would not create this confusion since they would neither match the targets nor be activated in working memory.

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). Supporting our hypothesis, participants responded slower to *lure probes* than to *other context* probes ($t(31) = 2.5250$, $p = .02$; Figure 4B), despite the fact that the only difference between these two kinds of probe is whether the probe word was learned in the same context as the target.

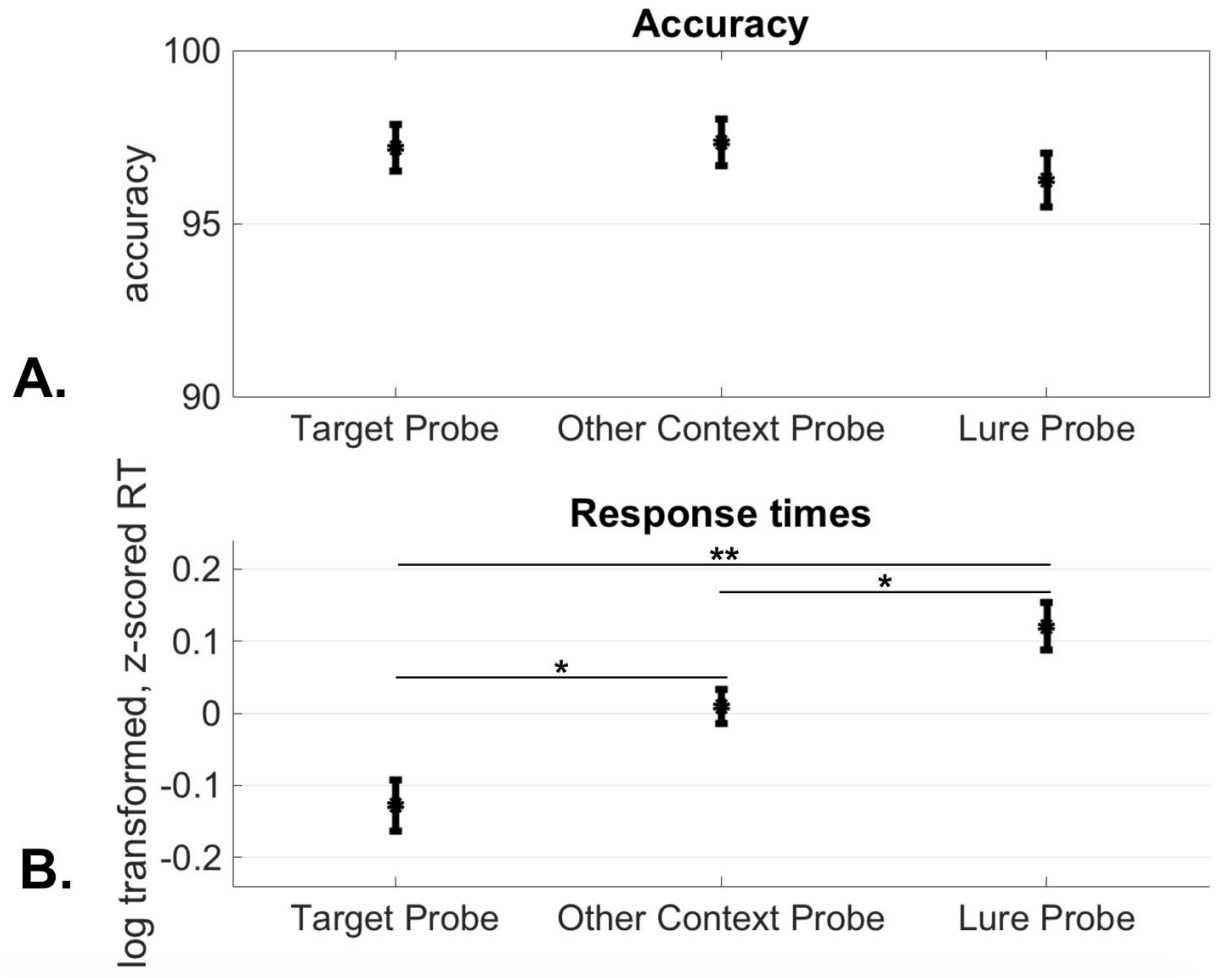


**Figure 4. Response times reflect influence of study context. A.** Accuracy was high across all three conditions (mean = 96.25%, SEM = 0.89%), with no significant difference in accuracy between *target* (mean = 97.03%, SEM = 0.70%), *other context* (mean = 96.72%, SEM = 0.76%), or *lure* (mean = 95.00%, SEM = 1.19%) trials (paired, two-sided t-tests, 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-sided t-test) and relative to *other context* probes — words drawn from a different context as the target words ($p < .05$; paired, two-sided t-test). Critically, participants responded slower to *lure* probes than to *other context* probes ($p < .05$; paired, two-sided 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 $p < .01$.

## Discussion

In Experiment 2, participants performed a DNMS task using study words learned in one of four separate contexts. The lack of distraction and the relatively short (18 second) delay period were chosen to make it easy for participants to use working memory to perform the task. While accuracy was near ceiling across all trial types, response times revealed an effect of encoding context, even in the absence of distraction. Specifically, while responses to target probes were faster than responses to both kinds of non-target probes, responses to lure probes — those sharing an encoding context with the target — were slower than responses to probes from any of the other three contexts.

This result is particularly striking because it is in the opposite direction of what would be expected if responses were simply biased towards the more prevalent response type (mismatch). If this were the case, then participants should be faster to respond to lure or other-context probes (⅔ of trials), rather than target probes (⅓ of trials). Instead, these results support the idea that responses proceed from integrative inference about the contents of working memory, an inference process slowed by context reinstatements that cause the intrusion of lure words into working memory. These reinstatements need not catastrophically interfere with maintenance; rather than occupying discrete "slots" in working memory, they may instead simply reduce the fidelity of the representation of the target set (Ma et al., 2014), leading to a slowed inference but not an incorrect response.

Note that the same logic should apply regardless of whether the probe is a lure or an other-context probe — if the correct response is "mismatch", but (during the delay) participants mentally reinstate the context matching the probe, this should lead to slower RTs to that probe. These reinstatements should be much more likely for the target context than for other contexts, which explains why responses to lure probes (from the target context) are slower, on average, than are responses to other-context probes.

# Experiment 3

## Results

Experiment 2 demonstrated that encoding context has an effect on responses following a delay. We interpret this result as following from putative episodic memory reinstatements during the delay period. We reasoned that this effect, observed in Experiment 2 as an average across trials,

should be determined on a trial-by-trial basis by whether episodic memory reinstatement of the probe context occurred on that trial, as well as which memories were reinstated. In Experiment 3, 36 additional participants performed the DNMS task from Experiment 2 while being scanned using functional magnetic resonance imaging (fMRI), which allowed us to use multivariate pattern analysis (MVPA) to measure memory reinstatement on each trial.

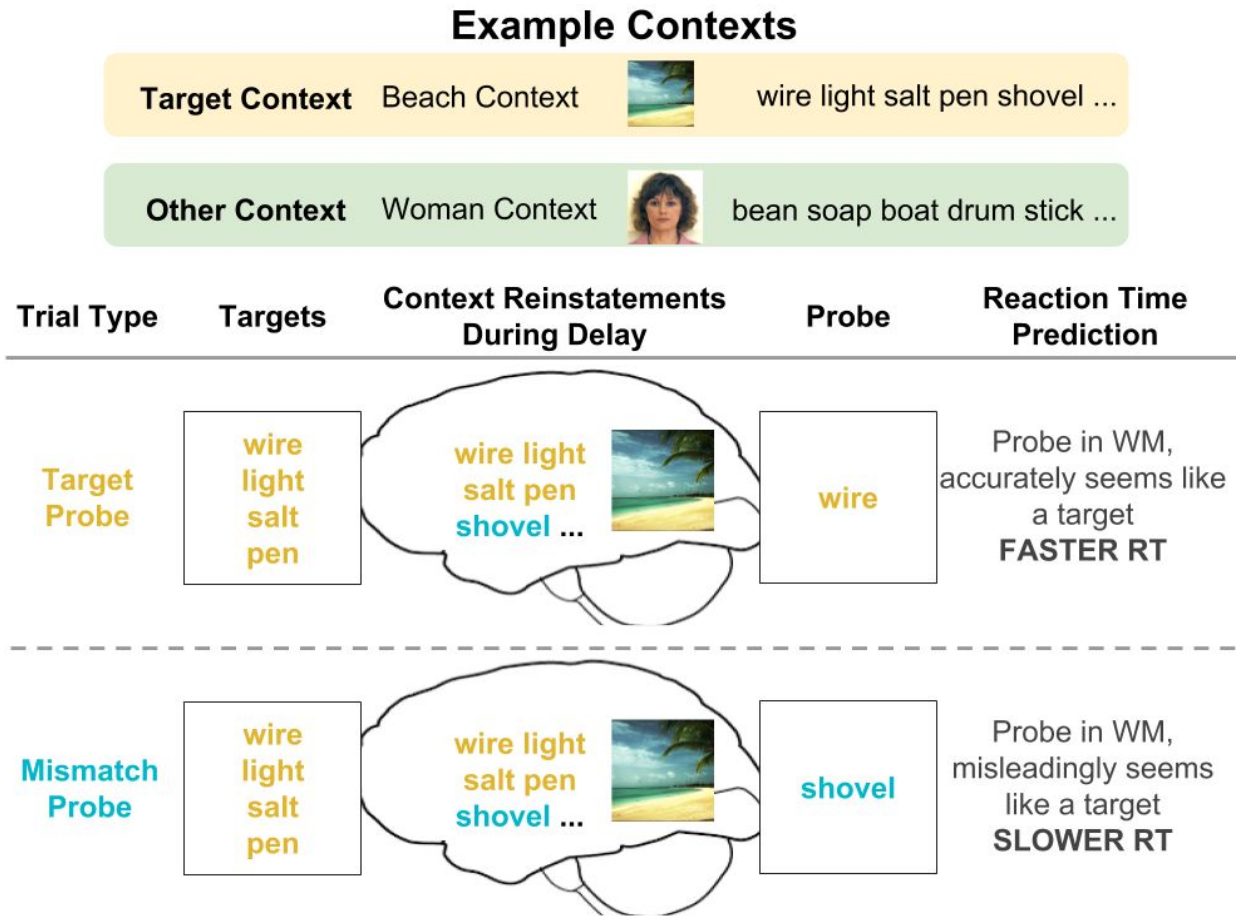

**Figure 5. Reinstatements during delay should affect comparison process at probe.** In our hypothesized process, working memory is influenced by ongoing, occasional reinstatements from episodic memory. Given that the target words should trigger a reinstatement of the context in which the target words were encoded, subsequent reinstatements are likely to follow from that same context. Using MVPA of fMRI data, we measured classifier evidence for the reinstatement of each of the four different contexts during the delay period on each trial, and related this evidence to responses to the probe word on that trial. Modeling response selection as an evidence accumulation process that infers the contents of working memory, we predicted that increased classifier evidence for the context of the probe word would slow responses on mismatch trials (lure and other-context probes), by reducing the consistency of evidence in working memory.

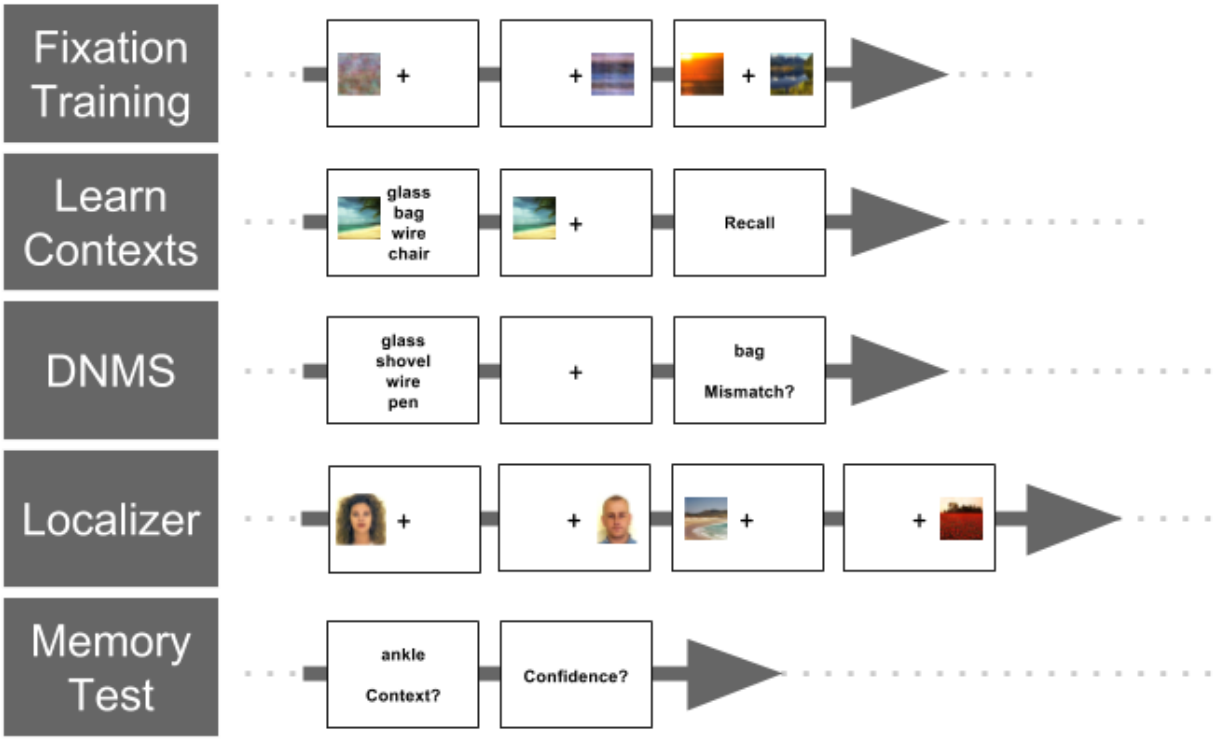


**Figure 6. Experiment 3 timeline.** We first trained participants to fixate on the center of the screen to ensure that they encoded pictures presented on the left side of the screen as *on the left* and pictures presented on the right side of the screen as *on the right.* Next, participants associated each of four "contexts" (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) with a unique set of 12 words. 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, after which they performed 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 which context they thought each word

was associated with during the initial context-learning phase, in addition to confidence in their report.

***Behavior.*** 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%; paired, two-sided t-tests, all $p > 0.2$).

Due to time restrictions, three participants were not able to complete the post-task item/context memory test. The 33 participants who completed the test performed above chance, as a group (chance = 25%, mean accuracy = 41.20%, SEM = 3.33%, $t(32) = 4.8648$, $p < .0001$, two-sided, one sample compared to chance t-test), and for 25/33 participants individually (proportion $p < .0001$ by binomial test).

As in Experiment 2, we restricted our RT analyses to correct trials only. In contrast to Experiment 2, there was no difference between average RTs in the two mismatch probe conditions (Other-context mean log-transformed, z-scored RT = .0311, std = .2313; Lure mean = .0321, std =.1780; $t(35) = -.0178$, $p = 0.9859$; paired sample, two-sided t-test), possibly reflecting an average increase in reinstatement of non-target contexts during the delay period, or a decrease in overall reinstatement, with respect to Experiment 2. Nevertheless, it remained possible that responses could have been speeded or slowed based on the nature of reinstatement on a trial-by-trial basis. We used neural evidence of reinstatement to address this possibility.

***fMRI.*** 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 same context during the initial context learning phase. 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 occurrence of probe-context reinstatement during the delay, and relate it to behavior at probe. On mismatch trials (lure and other context probe trials), we predicted that participants would be slower to respond if they reinstated context information that brought the probe word to mind — the more active the probe word is in working memory, the harder it will be to identify it as a mismatch. On target trials, where the probe word actually was one of the targets, we predicted that reinstating the probe word context would not slow participants.

To test these predictions, we measured fMRI evidence of reinstatement for each context during the delay period on each trial. We trained pattern classifiers to recognize the neural activity patterns distinguishing each context picture. These classifiers then produced a measure of

evidence for context reinstatements during each trial that we used to predict response times. For each trial, we computed classifier evidence for the context of the probe word (summed across the delay period), and we ran a multiple linear regression to predict response times using these trial-wise classifier evidence values. To test for any effect of context reinstatement at times other than during the delay period, we also entered into the regression the corresponding reinstatement evidence from the target period (the 2 seconds during which the target were on the screen) and the response 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.

Supporting our hypothesis, greater evidence for delay-period reinstatement of the probe context slowed responses on mismatch trials (mean $\beta$=-.042, SEM = .0148; t(35) = 2.4866, $p$ = .018; one-sample t-test compared to 0; Figure 7A). In contrast, greater evidence for delay-period reinstatement did not significantly affect RT on target trials (mean $\beta$=.018, SEM = .017; t(35) =1.0641, $p$ = .2946; one-sample t-test compared to 0; Figure 7A). Probe context reinstatements during mismatch trials affected RTs significantly differently than probe context reinstatements during target trials ($t$(35) = 2.5622, $p$ = .015; paired sample, two-sided, t-test).

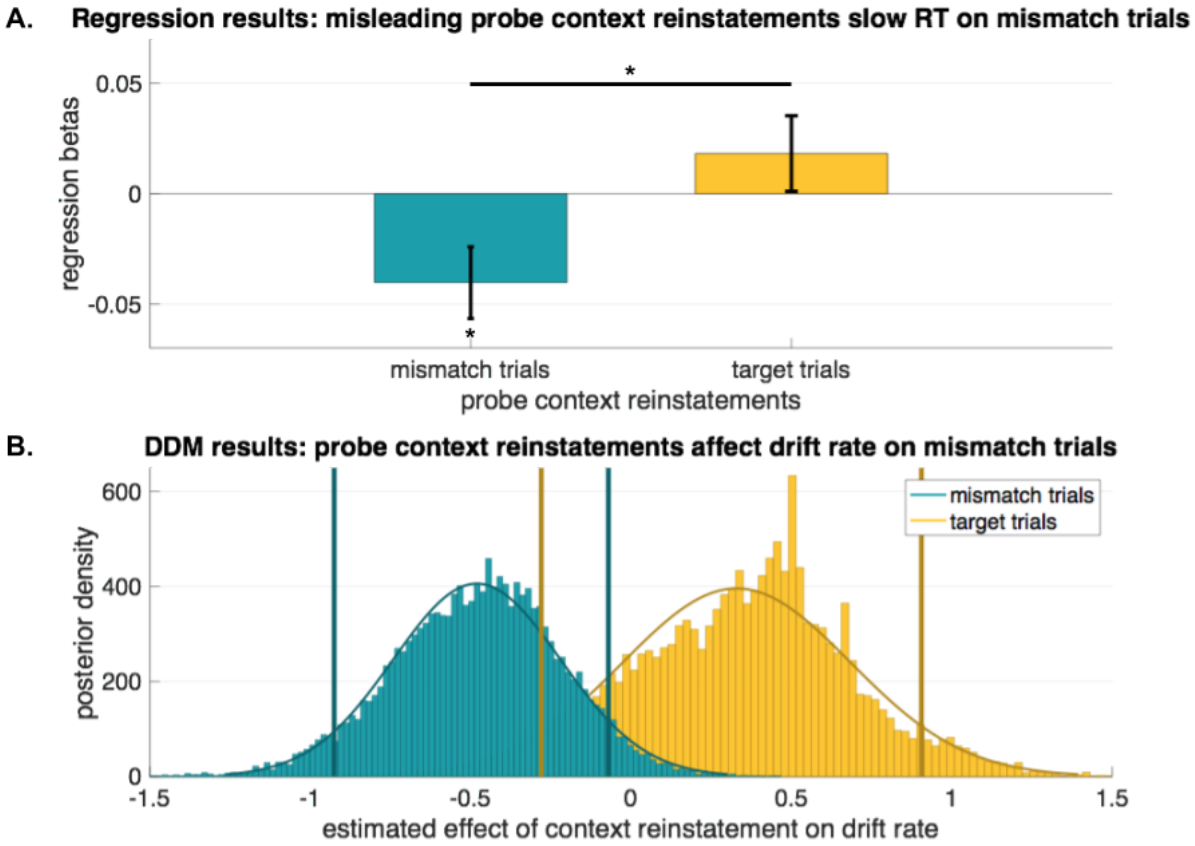


**Figure 7. Context reinstatements during the delay period predict reaction times. A.** Across all 36 participants, regression analyses of neural data indicate that on mismatch trials, when evidence for delay-period reinstatements of the probe context was higher, response times slowed down (two-sided, one sample t-test against 0, $p$ = .018). Probe context reinstatements during mismatch trials affected RTs significantly differently than probe context reinstatements during target trials ($t(35) = 2.5622$, $p$ = .015; paired sample, two sided t-test). Probe context reinstatements did not significantly affect response times on target trials. Error bars reflect SEM. * signifies $p < .05$. **B.** Consistent with the regression results, for mismatch trials, the estimated effect of probe context reinstatements on drift rate was negative for models that included reinstating the probe context; mean effect of reinstating probe context β = -0.48, 95% CI = [-.07 -.93]. In contrast, for target trials, the estimated effect of probe context reinstatements on drift rate overlapped with 0; mean effect of reinstating probe context β = .3292, 95% CI = [-.28 .95]. Vertical bars reflect 95% CrI.

These effects were unique to delay-period reinstatement. During target presentation, evidence for probe-context reinstatement *speeded* responses (the opposite of what was found during the delay; mean β=0.1347, SEM=.0566, t(35)=2.3826, $p$=.0228, two-sided, one-sample t-test compared to 0). This result could possibly reflect how well participants attended to the target set, with greater

attention during target presentation leading to a) more reinstatement of the target context and b) faster responses at probe, thereby inducing a correlation between these measures. During the response period, reinstatement of probe evidence had no effect on response times (mean $\beta$=0.0123, SEM=.0448, t(35)=0.2752, $p$=.7848, two-sided, one-sample t-test compared to 0).

Due to the slow nature of the BOLD response, it is possible that our measure of delay period reinstatement reflects some contribution of activity during target presentation or response period. To control for this possibility, we performed a post-hoc test, splitting the delay period into thirds, and repeated the above regression analysis on the summed context evidence from each third simultaneously. Our aim in this analysis was to examine whether the effect of context reinstatements on reaction time during the middle third of the delay period reflected our overall effect (since evidence from this interval was the least likely to be contaminated by activity from target presentation/encoding or probe presentation/response).

Probe context evidence from the first and middle thirds were both reliable predictors of the reaction time effect (first third: mean $\beta$=0.0644, SEM=0.0256, $p$=.049; middle third: $\beta$=-0.0896 SEM=0.0304, $p$=.0171; last third: $\beta$=-0.0439, SEM=0.0317, $p$=.5253; all $p$-values Bonferroni-corrected for post-hoc multiple comparisons). Consistent with our hypothesis, probe context reinstatements during the middle of the delay period slowed reaction times. The first third of the delay period had an opposite effect than the middle third of the delay period, suggesting reinstatements during the first third of the delay period might reflect something about the encoding process (e.g., attention to target encoding, as discussed above).

***DDM Results.*** To provide a formal test of our hypothesis that reinstatements affected the coherence of information in working memory, we modeled responses using a Drift-Diffusion Model (DDM; Ratcliff 1978) that has previously been used to capture inference about the contents of working memory. We predicted that, when the probe word was not in the target set, reinstatement of memories associated with the probe would introduce conflicting evidence into working memory, thus reducing the effective drift rate on that trial.

To test these predictions, we evaluated whether a DDM that used classifier evidence to set drift rate (the neural DDM) provided a superior fit to response times compared with a standard DDM fit to behavior only (behavioral DDM). The models were matched on all parameters except for those affecting drift rate. In the behavioral DDM, drift rate fluctuated trial-by-trial according to a normal distribution centered around the mean fit drift rate. The variance of this normal distribution was fit using a parameter, sv, following the standard formulation of the model

(Ratcliff & McKoon, 2008; Ratcliff & Smith, 2004; Ratcliff et al., 1999). In contrast, drift rate in the neural DDM was set on each trial based on the quantity of probe context reinstatement evidence measured by our 4-way classifier, and thus fluctuated non-parametrically a trial-by-trial basis. The relationship between probe context evidence and drift rate was set by a linear regression, with fitted slope and intercept parameters. Therefore, the models were matched on the number - though not the function - of parameters.

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; model comparison methods elucidated in methods section). Consistent with our hypothesized mechanism, and with the response time results above, probe-context reinstatement reduced the drift rate on mismatch (mean $\beta$=-0.4808, std = .261; $p$=0.026; Figure 7B), but not on target trials (mean $\beta$=.3292, std = .3549; $p$=0.184; Figure 7B).

The behavioral model contains an additional free parameter, *sv*, that models trial-by-trial gaussian variation in the drift rate; this parameter was omitted from the neural model on the premise that variation in the drift rate is explainable as the effect of memory reinstatement on the fidelity of working memory content. Consistent with this proposal, the fit of the neural model did not change when the *sv* parameter was added (DIC(neural+drift rate noise)=1135; estimated $p$=.76).

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

# Discussion

By maintaining a high-fidelity record of recent information, working memory allows us to perform tasks that require accurate storage over short periods of time. However, interference can compromise the value of this form of storage:  The presence of distraction or the need to focus on a new task can compromise that record and impair performance. Episodic memory complements these characteristics, by storing memories over a longer term, at the cost of reduced fidelity and the risk of retrieval failure (e.g., Cohen & O'Reilly, 1996).

While the identification and study of these distinct systems has benefited by efforts to isolate

them, it is unlikely that they operate in an isolated or entirely independent fashion under natural conditions. Regions associated with episodic memory that are engaged during memory task performance have been observed to be active even during rest, reflecting ongoing reinstatement of episodic memories (Wilson & McNaughton, 1994; Carr et al., 2011; Jadhav et al., 2012). These memory reinstatements can lead to the incidental reinstatement of the context in which the memories were experienced (Bornstein & Norman, 2017). These reinstatements have also been observed to involve coordinated activity across the entire brain, including prefrontal areas thought to underlie working memory maintenance (Miller & Cohen, 2001). Thus, in a manner analogous to externally-driven stimuli, internally-driven reinstatements from episodic memory may also also impact representations stored in working memory.

Over a series of three experiments, we tested the hypothesis that episodic memory reinstatement influences working memory maintenance, even in the absence of external interference. In Experiment 1 we showed that substitution errors in a delayed recall task, a classic hallmark of the contribution of episodic memory when working memory maintenance is disrupted, reflect the influence of context present at the time of initial encoding. Absent interference, accuracy was close to ceiling.

Experiment 2 revealed that, even when accuracy is near-ceiling, other measures of performance can reflect the influence of encoding context. 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 which 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 was predictable 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 determine the DDM drift rate on each trial, consistent with the hypothesis that episodic memory reinstatement during the delay period influenced the evidence used to make responses at the time of probe.

## The function of replay during working memory maintenance

We have shown that reinstatement of recent experiences from episodic memory has specific,

measurable influence on the contents of working memory, even over short delay periods, and even in the absence of explicit interference. Why is working memory influenced by episodic memory reinstatement? The effect of episodic memory contents on working memory could simply be a side effect, or it could indicate that laboratory tests of working memory maintenance obscure key features of the way that working memory is used in more naturalistic environments. One suggestion is that working memory selectively recruits episodic memory in order to "refresh" decaying or disrupted representations. This flexibility may, however, leave it vulnerable to disruption from other information present in episodic memories. Our findings thus present a new puzzle: on what basis, if any, is this intrusion rational? One possibility is that, while some of these reinstatements may be strategically directed recalls, in service of maintaining decaying working memory representations, others may instead be ongoing replay of the sort associated with resting-state activity or forward planning (Foster & Wilson, 2006; Tambini et al., 2010; Deuker et al., 2013). On this view, the ability to interact with working memory may be an adaptive feature of resting-state replay from episodic memory — in other words, it may not just sustain, but also transform working memory representations, by integrating information in working memory with that from recent events. That these reinstatements include contextually-related events implies that such an interaction could support rapid, goal-relevant generalizations (Kumaran & McClelland, 2012). The mechanism outlined here both constrains, and expands, that proposal, with potentially broad impacts for the study of memory-guided decision-making.

# 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. A.N.H. was supported by a National Defense Science and Engineering Grant. A.M.B., A.N.H., K.A.N., and J.D.C. acknowledge support from the Templeton Foundation and the Intel Corporation. The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the John Templeton Foundation.

# Contributions

A.N.H., A.M.B., and J.D.C. conceived the experiment; A.N.H., A.M.B., J.D.C., and K.A.N. designed the experiments 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

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, 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

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

Bornstein, A. M. & Norman, K. A. (2017) Reinstated episodic context guides sampling-based decisions for reward. Nature Neuroscience. https://doi.org/10.1038/nn.4573

Brown, J. (1958). Some tests of the decay theory of immediate memory. Quarterly Journal of Experimental Psychology, 10, 12-21.

Buckner, R. L. (2010). The role of the hippocampus in prediction and imagination. Annual Review of Psychology, 61, 27–48, C1-8. http://doi.org/10.1146/annurev.psych.60.110707.163508

Buckner, R. L., & Carroll, D. C. (2007). Self-projection and the brain. Trends in Cognitive Sciences, 11,

49–57.
Burnham, K.P., Anderson, D.R. (2002). Model selection and multimodel inference: a practical information-theoretic approach, 2nd edn. Springer, New York.
Carr, M.F., Jadhav, S. P., & Frank, L.M. (2011). Hippocampal replay in the awake state: a potential substrate for memory consolidation and retrieval. Nature Neuroscience, 14, 147-153.
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
Cohen, J.D. & O'Reilly, R.C. (1996). A Preliminary Theory of the Interactions Between Prefrontal Cortex and Hippocampus that Contribute to Planning and Prospective Memory. M. Brandimonte, G.O.
Einstein & M.A. McDaniel (Eds) Prospective Memory: Theory and Applications, 267-296, Mahwah, New Jersey: Lawrence Erlbaum Associates.
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
Deuker, L., Olligs, J., Fell, J., Krantz, T.A., Mormann, F., Montag, C., Reuter, M., Elger, C.E., & Axmacher, N. Memory consolidation by replay of stimulus-specific neural activity. The Journal of Neuroscience, 33(49), 19373-19383. https://doi.org/10.1523/JNEUROSCI.0414-13.2013
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
Epstein, R. & Kanwisher, N. (1998). A cortical representation of the local visual environment. Nature, 392, 598-601.
Foster, D.J., & Wilson, M.A. (2006). Reverse replay of behavioural sequences in hippocampal place cells during the awake state. Nature, 440(7084), 680-683. https://doi.org/10.1038/nature04587
Gershman, S.J., Schapiro, A.C., Hupbach, A. & Norman, K.A. Neural context reinstatement predicts memory misattribution. J. Neurosci. 33, 8590–8595 (2013).
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
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
Jadhav, S.P., Kemere, C., German, P.W., Frank, L.M.. (2012) Awake hippocampal sharp-wave ripples support spatial memory. Science, 336, 1454–1458.
Kanwisher, N. McDermott, J., Chun, M.M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. Journal of Neuroscience, 17(11), 4302-4311.
Kumaran, D., & McClelland, J.L. (2012). Generalization through the recurrent interaction of episodic memories: A model of the hippocampal system. Psychological Review, 119(3), 573-616. https://doi.org/10.1037/a0028681
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

Lewis-Peacock, J.A. & Norman, K.A. (2014). Competition between items in working memory leads to forgetting. Nature Communications, 5(5768).

Logothetis, N. K., Eschenko, O., Murayama, Y., Augath, M., Steudel, T., Evrard, H. C., Besserve, M., Oeltermann, A. (2012). Hippocampal-cortical interaction during periods of subcortical silence. Nature, 491(7425), 547–53. http://doi.org/10.1038/nature11618

Ma, W.J., Husain, M., Bays, P.M. (2014). Changing concepts of working memory. Nature Neuroscience 17(3), 347-356.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal function. Annual Review of Neuroscience, 24, 167–202.

Peterson, L.R., & Peterson, M. J. (1959). Short-term retention of individual verbal items. Journal of Experimental Psychology, 58, 193-198.

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

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., & D'Esposito, M. (2004). Inferior temporal, prefrontal, and hippocampal contributions to visual working memory maintenance and associative memory retrieval. JNeurosci, 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

Ratcliff, R. (1978). A theory of memory retrieval. Psychological Review, 85(2), 59-108.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. Neural Computation, 20, 873–922.

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

Ratcliff, R., & Smith, P. (2004). A comparison of sequential sampling models for two-choice reaction time. Psychological Review, 111, 333–367.

Ratcliff, R., Van Zandt, T., & McKoon, G. (1999). Connectionist and diffusion models of reaction time. Psychological Review, 106, 261–300.

Repov, 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

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/10.1080/00335557043000203

Spiegelhalter, D., J., Best, N. G., Carlin, B.P., Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. J. R. Statist. Soc. B, 64(4), 583-639.

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

Sternberg, S. (1969). Memory-scanning: mental processes revealed by reaction time experiments. 57(4), 421-457.

Tambini, A., Ketz, N., & Davachi, L. (2010). Enhanced brain correlations during rest are related to memory for recent experiences. Neuron, 65(2), 280–290. http://doi.org/10.1016/j.neuron.2010.01.001

Tulving, E. (1983). Elements of Episodic Memory. Canadian Psychology, 26(3), 351. https://doi.org/http://dx.doi.org/10.1017/S0140525X0004440X

Wickens, D.D., Dalezman, R.E., Eggemeier, F.T. (1976). Multiple encoding of word attributes in memory. Memory and Cognition, 4(3), 307-310.

Wiecki, T.V., Sofer, I. & Frank, M.J. (2013). HDDM: Hierarchical Bayesian estimation of the Drift-Diffusion Model in Python. Front. Neuroinform. 7:14. doi: 10.3389/fninf.2013.00014

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

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

# Methods

## Experiment 1

***Stimuli.*** The experiment used six scene pictures, each of which served as "context" that uniquely linked one of six sets of 12 words. 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 48 learning trials, participants were shown four words drawn from the same set alongside the photograph associated with that set (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 set 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.

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 randomly generated 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, starting from the prompted 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. words from the same encoding context as the targets, 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.)

## Experiment 2

***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 on the basis of their accuracy scores being more than 2 standard deviations below the mean, leaving 32 participants reported here.

***Procedure.*** In the *Learning phase*, participants studied four different sets of words, each containing 12 words drawn from the same words used in Experiment 1. Each word set 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 University (PICS; http://pics.stir.ac.uk). The scene pictures depicted two natural, 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 set 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 anew for each participant. *Learning phase* trials followed the same procedure as in Experiment 1 (Figure 2A; Figure 3A), now over four contexts of 12 words each.

In the *Testing phase*, participants performed 60 trials of a DNMS task, in which targets were selected from the words learned in the learning phase (Figure 3B). On each trial, one context was selected at random, and then four target words were selected from within that context. These words were shown on the screen together for two seconds — critically, without the associated context image. When the words disappeared, they were replaced by a centered fixation cross, displayed for 18 seconds. 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 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 after four seconds) was indicated with a red fixation cross.

Probe words could be one of three types: 1. *target* probes were drawn from the four-word target set presented on the current trial; 2. *lure probes* were drawn from the same context list as the target words, but, critically, these probes were not one of the target words; 3. *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 each 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.

## Experiment 3

***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 are reported for the remaining 36 participants.

***Stimuli.*** *The Fixation* phase used scene and scrambled 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 reported that they felt they understood 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 m, 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 learning, 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). 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 do not assume the categories are mutually exclusive (e.g., we do not assume that the presence of left face evidence necessarily indicates right face absence; Lewis-Peacock & Norman, 2014). To establish the sensitivity of our classifier to the four categories of interest, we performed a leave-one-out cross-validation. First, we split the MRI data from the localizer phase into four runs by time. Then, we trained the classifier on three of the runs, and tested its performance on the fourth, repeating this procedure once using each run as the holdout set. The resulting average performance was significantly above chance (chance = 25.00%, mean = 66.99%, std = 18.30%, $t(35) = 14.1419$, $p < .001$; one-sample t-test compared to chance).

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 response period (4 seconds).

For each DNMS trial, we computed the classifier evidence for the probe context. We performed a regression analysis where we used probe context evidence (for each trial) to predict RT on that trial. For each regression model, we defined the three variables of interest as the summed classifier evidence for the given context during (1) target presentation, (2) the delay period, and (3) the response period.

**Diffusion-model fits.** We modeled DNMS responses as resulting from an inference process that draws successive samples from working memory until reaching a decision threshold. 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. We fit two main classes of models: the *behavioral* model, and the *neural* model.

The *behavioral* model was fit using 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). The *neural* model followed the same parameterization, with the exception that the *v* and *sv* parameters were removed, and replaced by using hDDM's regression functionality to set *v* as a function of trial-by-trial reinstatement evidence. By replacing *v* and *sv* in the *behavioral* model with the slope and intercept of the linear function relating to *v* in the *neural* model, both models ultimately had the same number of parameters.

In all cases, model parameters were estimated using a Markov Chain Monte Carlo procedure (MCMC) using 30,000 samples, the first 15,000 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. As DIC can be considered as an approximation of AIC with negligible priors, we report for illustrative purposes estimated *p*-values computed using the formula applied to AIC scores (Burnham & Anderson, 2002): $e^{\frac{DIC_1 - DIC_2}{2}}$, where $DIC_1$ is the (lower) DIC score of the superior model. By convention, differences in DIC score greater than 6 (yielding estimated *p* values less than .05) constitute meaningful evidence (Spiegelhalter et al., 2002).