# Long-Term Effects of Working Memory Retrieval From Prioritized and Deprioritized States

Frieda Born[1], Bernhard Spitzer[1]

[1]Research Group Adaptive Memory and Decision Making, Max Planck Institute for Human Development, Berlin, Germany

Correspondence to:
born@mpib-berlin.mpg.de
spitzer@mpib-berlin.mpg.de

# Abstract

Which factors determine whether information temporarily held in working memory (WM) is transferred to long-term memory (LTM)? Previous work has shown that retrieving ("testing") memories from LTM can benefit their future LTM recall. Here, we examined the extent to which a benefit for subsequent LTM may also occur after retrieval from WM, depending on whether the WM contents were retrieved from a prioritized or deprioritized state. In three experiments, we combined variants of a novel visual WM paradigm with a subsequent surprise LTM recall test. We found a LTM benefit of WM testing both for prioritized and deprioritized WM contents, which, interestingly, was stronger for the deprioritized information. This pattern showed similarly across experiments with different priority manipulations. Subsequent LTM benefits generally occurred after WM testing with a recall-like test format (continuous report), but not after simple WM comparisons against a probe. The surprisingly larger LTM benefit for deprioritized WM contents may reflect enhanced encoding of the participants' own subjective WM report – as opposed to the originally presented sample information – into LTM.

# Introduction

Of the myriads of information we experience in our daily lives, only a small fraction can later still be remembered. It is commonly assumed that momentary experiences are processed in Working Memory (WM), a strictly capacity-limited system that maintains information only as long as is needed for immediately upcoming tasks (D'Esposito, 2007; D'Esposito & Postle, 2015; Hitch & Baddeley, 1976; Miller, 1956). While the contents of WM typically persist only for seconds or less, some of the information may be committed to storage in long-term memory (LTM) and potentially persist for a lifetime (Atkinson & Shiffrin, 1968; Baddeley, 2012). But which factors determine whether WM contents become durable in LTM or are eventually forgotten? Previous research has shown that LTM formation is facilitated, for instance, by "deep" semantic encoding of the stimulus materials (*Levels of Processing*; Craik & Lockhart, 1972), by their emotional salience (e.g., *flashbulb memories*; Brown & Kulik, 1977), and by directing top-down attention to them (Khader et al., 2010; Sundby et al., 2019). Other work has established that LTM storage is furthermore consolidated by "testing", that is, by attempting to remember ("retrieval practice") the information again at a later point in time (Roediger & Butler, 2011; Rowland, 2014). From a WM perspective, these factors pertain to how information is encoded into WM, and whether the information is encoded from the environment or is internally generated (the latter typically benefits subsequent LTM; for review, see Bertsch et al., 2007).

Beyond encoding, WM function is thought to include purposeful maintenance, updating, and retrieval processes (e.g., Ranganath et al., 2004; Cowan, 1999; McElree, 2006; for review, see Bledowski et al., 2010). Whether and how specific (sub)processes within WM affect subsequent LTM is subject to ongoing research. Focusing on maintenance, several studies with verbal materials found that longer WM retention periods were associated with improved subsequent LTM recall (Jarjat et al., 2018; Madigan & McCabe, 1971; Souza & Oberauer, 2017; for review, see Hartshorne & Makovski, 2019). These findings suggest an LTM-benefit of subvocal rehearsal in WM, although similar results have been reported for non-verbal (e.g., visual materials as well, see Hartshorne & Makovski, 2019). A number of studies have also investigated how subsequent LTM is affected by attentional (de-)prioritization of WM contents during maintenance, e.g., via retrospective cueing (*retro-cues*; Griffin & Nobre, 2003). While many of these studies found that prioritization improved subsequent LTM (Fan & Turk-Browne, 2013; Jeanneret et al., 2023; LaRocque et al., 2015; Reaves et al., 2016; Strunk et al., 2019; Wang & Van Ede, 2024), others found no such effect (Bartsch et al., 2018; Mao Chao et al., 2023) or even found superior LTM when attention was diverted from the WM information (Rose et al., 2014). Interestingly, the latter result appears consistent with a body of work by McCabe and colleagues (Loaiza & McCabe, 2012, 2013; McCabe, 2008),

which showed that intermittent distraction during word list learning (“complex span” task) impaired the words’ immediate WM recall, but paradoxically improved their subsequent LTM recall. This counterintuitive finding has been explained in terms of “covert retrieval” of the WM items back into the focus of attention (Loaiza & McCabe, 2012; McCabe, 2008), based on the idea that during distraction, WM information is temporarily maintained in “activated long-term memory” (see also Beukers et al., 2021; Cowan, 1999; Oberauer, 2002). A possible corollary of this view is that retrieval processes in WM might foster subsequent LTM in quite similar ways as the well-established ”testing” effects in the LTM literature, i.e., through retrieval from an LTM-like storage format, especially when the WM information was unattended.

However, unlike LTM effects attributed to covert WM processing (i.e., during retention), the role of overt WM retrieval during WM testing has thus far received only relatively little attention (but see Tozios & Fukuda, 2024; Xie & Reuter-Lorenz, 2024). One reason for this might be that WM testing typically involves some form of reexposure to the WM information when it is presented as a recognition probe (e.g., in delayed-match-to-sample tasks) or when it is reproduced by the participant themselves (e.g., in WM recall). An overall LTM benefit of WM testing might thus be trivially expected if the WM test provides an additional learning opportunity for subsequent LTM. Studies of the well-known testing effect in the LTM literature typically control for reexposure by including matched "restudy" conditions, in which no retrieval is required (Rowland, 2014). Creating similar conditions in a WM-task context is difficult because without the expectation of a WM test, there would be no reason for participants to engage in WM maintenance at all (Baddeley, 2012; Postle & Oberauer, 2022). Perhaps unsurprisingly, thus far only few WM studies have examined how the LTM consequences of overt WM testing are affected by task factors such as distraction (Loaiza et al., 2021; McCabe, 2008) or cueing (LaRocque et al., 2015), mostly with negative results. In other words, while the idea that LTM benefits from presumed “covert” WM retrieval processes has been increasingly established, less is known about whether and how LTM is affected by overt WM retrieval during explicit WM testing.

Here, we used a novel approach to investigate how active retrieval from WM (“WM testing”) affects the transfer of information to LTM, depending on whether the information was retrieved from a prioritized or a deprioritized WM state. Unlike previous studies which reported no differential LTM-benefit of word-list recall after distraction in a ‘complex’ span WM task (Loaiza et al., 2021; McCabe, 2008), our WM tasks required participants to maintain visual information, specifically, the orientations of one or two rotated objects. Further, whereas previous studies of the long-term consequences of attentional (de-)prioritization typically used recognition tests to probe WM, we asked participants to provide continuous orientation reports, both as WM tests (Experiments 1 & 2) and when probing participants’ subsequent LTM in a later surprise test. Not only did we hypothesize that continuous reporting promotes

active WM retrieval (and hence a sizable WM-'testing' effect), it also allowed us to examine the extent to which subsequent LTM reports reflected the participants' own (subjective) WM-test reports (instead of the WM samples' original orientation), and whether such "testing bias" varied with the attentional state of the WM information (prioritized vs. deprioritized). Lastly, in Experiment 3, we asked whether the LTM consequences of WM retrieval indeed hinged on the format of WM testing (continuous reports vs. delayed comparison).

Our results showed clear subsequent LTM benefits of WM testing with continuous reports and further revealed that participants' LTM reporting was biased towards their WM reports. Interestingly, these effects were more pronounced when the WM information was retrieved from a deprioritized state, both when we manipulated priority via testing order (Experiment 1) and using retro-cues (Experiment 2). Thus, although deprioritization expectedly reduced immediate WM test accuracy, it paradoxically improved subsequent LTM accuracy, in line with a stronger WM-testing effect for unattended WM contents. Across the WM and LTM tests in both experiments, we further observed a pattern of within-trial 'primacy' effects to suggest that the role of episodic factors in accessing the memoranda increased from prioritized over deprioritized to long-term storage. Lastly, when using a delayed comparison WM test format (binary choice) instead of continuous reports (Experiment 3), we obtained starkly different results, with no effects of priority in neither WM nor LTM tests, and substantially lower LTM performance overall. Together, our findings highlight a critical and multi-faceted role of explicit WM testing in understanding the link(s) between short- and long-term storage in human memory.

# Results

We report the results of three experiments in which randomly oriented pictures of real-world objects were used as sample stimuli in a WM task (**Fig. 1a**). After completion of the WM task, a short distractor task ensued (mental arithmetic), followed by a surprise LTM test (**Fig. 1b**) in which participants were asked to recall the orientations of the previously encountered WM samples (**Fig. 1c**).

## Experiment 1

On WM task trials in Exp. 1 (n = 187), either one or two randomly oriented sample stimuli were to be maintained over a short delay period (**Fig. 1a**). When probed after

the delay, the sample object reappeared in a random orientation and participants were asked to re-rotate it to its previous orientation (continuous report). In half of the two-sample trials (randomly varied), only one of the two samples (randomly selected) was probed. On the remaining two-sample trials, after the first WM test (Test 1), also the orientation of the other, previously unprobed sample was probed (Test 2). Thus, participants had to maintain the orientation of both WM samples until Test 1, during which the unprobed sample can be assumed to be deprioritized for the remainder of the trial. Importantly, the continuous report procedure used for WM testing in Exp. 1 provided no information about the samples' original orientations beyond the participants' own WM reports.

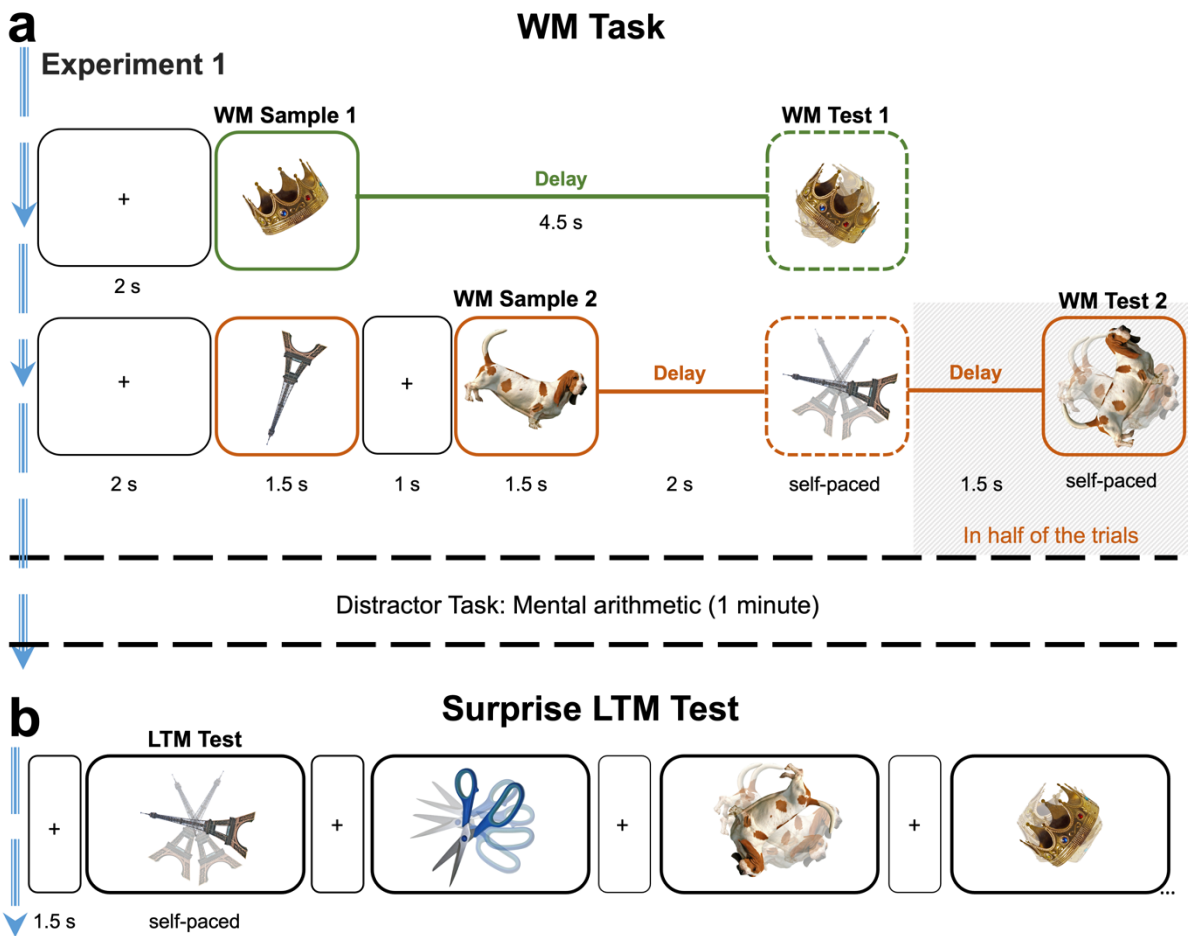



**Figure 1. Task layout of Experiment 1. a,** WM task; examples of one-sample (top, green) and two-sample trials (bottom, orange). Participants were presented with one or two randomly oriented objects as WM Samples. When probed after the delay, the sample object reappeared in a random orientation and participants were asked to re-rotate it to its previous orientation (continuous report). In half of the two-sample trials (randomly varied), only one of the two WM samples was probed (Test 1). On the remaining two-sample trials, also the other, previously unprobed WM sample was probed (Test 2). After the WM task, participants performed a short distractor task, in which they were asked to solve simple math problems. **b**, Example LTM test trials. After the distractor task, participants were asked to report the orientation of all previously seen WM samples another time. Each object appeared again in

a new random orientation and participants were asked to re-rotate it to the orientation it had when presented as a WM sample (see ***a**, left*).

## WM performance

**Figure 2a** shows the error (absolute angular difference from the sample orientation; note inverted y-axis) of participants' reports in the WM task. As expected, WM accuracy was significantly higher (i.e. smaller errors) on one-sample trials (M = 10.05°, SE = 0.49°) compared to two-sample trials [Test 1 and 2 combined; M = 17.42°, SE = 0.73°; $t(186) = -14.53$, $p < 0.001$]. Further, in the two-sample trials, performance on Test 2 (M = 20.36°, SE = 0.96°) was significantly reduced compared to Test 1 [M = 15.95°, SE = 0.71°; $t(186) = 6.50$, $p < 0.001$], as was expected by deprioritization of the second tested sample during and after Test 1.

In the two-sample trials, we also examined the extent to which WM accuracy was modulated by sample position. A 2 x 2 repeated measures ANOVA with the factors Sample Position (1/2) and WM Test (1/2) showed a main effect of Sample Position [$F(1,186) = 4.836$, $p = 0.029$, $\eta2 = 0.003$], indicating that first-presented samples were remembered better ("primacy" effect), and a main effect of WM Test [$F(1,186) = 39.358$, $p < 0.001$, $\eta2 = 0.026$], reflecting the lower performance on Test 2 (see above). There also was a significant interaction between the two factors [$F(1,186) = 11.369$, $p < 0.001$, $\eta2 = 0.007$], indicating that the primacy effect was stronger on Test 2 than on Test 1 (**Fig. 2a**). Post-hoc tests confirmed a significant primacy effect on Test 2 [M = 18.41°, SE = 1.05° vs. M = 22.15°, SE = 1.1°; $t(186) = -3.239$, $p = 0.001$], but not on Test 1 [M = 16.347, SE = 0.776 vs. M = 15.523, SE = 0.786; $t(186) = 1.208$, $p = 0.229$]. Together, the results from two-sample trials are in line with earlier findings of reduced WM recall after deprioritization (Emrich et al., 2017; Rerko & Oberauer, 2013; Souza et al., 2016; Souza & Oberauer, 2016). In addition, the results showed a WM "primacy" effect (e.g., Gorgoraptis et al., 2011; Hurlstone et al., 2014), which occurred only for the second-tested (deprioritized) samples.

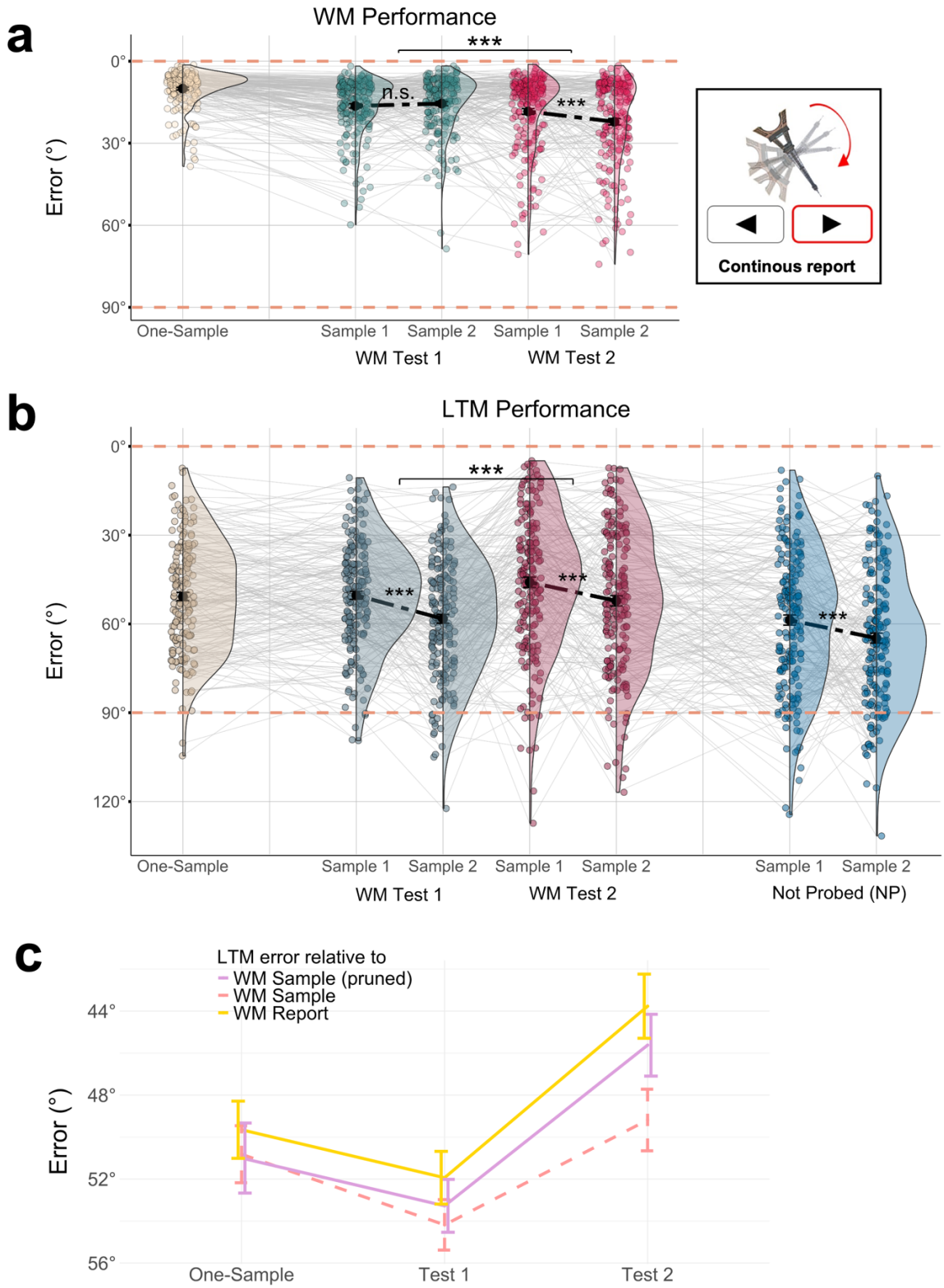



**Figure 2. WM and LTM performance in Experiment 1. a,** *Left*, WM task performance. Black dots, means; colored dots, individual participants. Black error bars show the standard error of the mean (SEM) and half-violin outlines illustrate the distribution over participants using a kernel density estimation. Dashed horizontal lines mark ceiling (0°) and chance-level performance (90°). WM accuracy on two-sample trials was significantly lower than on one-sample trials, and was significantly reduced after deprioritization (Test 2) compared to Test

1. **a** *Right*, in Exp. 1, a continuous report format was used in both the WM and LTM tests. **b**, LTM test performance. Same plotting conventions as in *a.* In contrast to WM performance, subsequent LTM performance was *increased* for Test 2 samples compared to Test 1 samples. **c**, Purple: LTM performance for WM samples that have been 'pruned' for equal WM performance levels across conditions (see *Results* and *Methods*). For comparison, the similarity of the LTM reports to the original WM sample orientations is shown (dashed red), which corresponds to the LTM performance measure shown in *b*. Yellow: similarity (in terms of absolute difference in degrees, note inverted y-axis) between the LTM- and WM-test reports. See *results* for details.

## LTM performance

In the subsequent surprise LTM test, the participants were asked to report the orientations of all sample objects that had been presented in the WM task, including those that were not probed in a WM test. Focusing on the probed samples, as expected, participants' LTM reports (**Fig. 2b**) were considerably less accurate (M = 51.23°, SE = 0.79°) than their previous reports in the WM task [t(186) = 37.28, p < 0.001]. LTM performance for samples from one-sample WM trials appeared descriptively better (M = 50.73°, SE = 1.37°) than for samples from two-sample trials (Test 1 and 2 combined; M = 52.31°, SE = 1.20°) but the difference was not significant [t(186) = -1.544, p = 0.124]. Interestingly, focusing on the samples from two-sample trials, LTM accuracy was significantly higher for samples that had been probed second (i.e. after deprioritization) in the WM task (WM Test 2, M = 49.00°, SE = 1.48°), compared to samples that had been probed first [WM Test 1; M = 54.18°°, SE = 1.20°, t(186) = -4.319, p < 0.001]. Thus, whereas the WM accuracy for deprioritized samples was expectedly reduced (see WM results above), their subsequent LTM recall was surprisingly improved compared to samples from WM Test 1, and was on par with the LTM recall of samples from one-sample WM trials [see **Fig. 2b**; t(186) = -1.374, p = 0.171][1,2].

A 2 x 2 ANOVA (specified analogously as above) of the LTM performance for the samples from two-sample WM trials showed a main effect of Sample Position [1/2;

[1] By closer inspection, samples with the shortest distance between presentation and WM test (Sample 2, Test 1) were recalled significantly worse in the LTM test than samples from one-sample trials [t(186) = 5.51, p < 0.01]. Interestingly, however, samples with the longest distance (Sample 1, Test 2), which had been retrieved from a deprioritized WM state (see **Fig. 1a**), were recalled significantly *better* even than the samples from one-sample WM trials [t(186) = 3.127, p = 0.009].

[2] An alternative explanation for higher LTM performance for Test 2 vs. Test 1 samples could be that Test 2 was the last event in the WM trial episode, which might have rendered it more memorable, whereas Test 1 was the last event only in 50% of cases (**Fig. 1**). However, a control analysis showed no difference in LTM performance between Test 1 samples that were followed by a Test 2 and those that were not [i.e., where Test 1 was the last event in the WM trial; t(186) = 0.13, p = 0.90], which speaks against an explanation in terms of WM-test recency.

$F(1,186) = 35.805$, $p < 0.001$, $\eta 2 = 0.026$], indicating better LTM recall of samples that had been presented first in the WM trial (i.e., primacy), and a main effect of Test [1/2; $F(1,186) = 20.253$, $p < 0.001$, $\eta 2 = 0.015$], reflecting the improved LTM recall of deprioritized samples that had been probed on WM Test 2 (see above). There was no interaction between the two factors [$F(1,186) = 0.619$, $p < 0.433$, $\eta 2 = 0.004$], and post-hoc tests showed the primacy effect on LTM performance to be significant both for samples probed first [WM Test 1; $t(186) = -5.909$, $p < 0.001$] and second [WM Test 2; $t(186) = -3.488$, $p < 0.001$].

We next examined for comparison the LTM performance for samples that had been presented but not probed (NP) during the WM trials. NP samples occurred on 50% of the two-sample trials and can be assumed to have been deprioritized after the first WM probe (WM Test 1), like those samples that had been probed on WM Test 2. The NP samples thus provided a baseline to quantify the LTM benefit of WM retrieval on Test 2. We indeed found that the LTM recall of the NP orientations was significantly less accurate ($M = 62.16°$, $SE = 1.42°$) compared to those probed on WM Test 2 [$t(186) = 9.745$, $p < 0.001$], and also compared to those probed on the other WM Tests [WM Test 1: $t(186) = 6.908$, $p < 0.001$; one-sample: $t(186) = 8.594$, $p < 0.001$]. Thus, WM probing and/or -retrieval appeared to generally benefit subsequent LTM recall. Interestingly, the NP samples also showed a primacy effect in LTM: those presented first in the WM trial were subsequently recalled better than those presented second [$M = 58.700$, $SE = 1.721$ vs. $M = 64.80°$, $SE = 1.67°$; $t(186) = -3.308$, $p = 0.001$], just as was the case for the other (probed) samples (see above). In other words, the primacy effect on LTM performance occurred independent of WM retrieval and was more likely attributable to differences in encoding (or maintaining) the first vs. second WM sample.

## Comparing WM vs. LTM performance

To compare the WM and LTM results directly, we additionally performed a 2 x 2 x 2 repeated measures ANOVA with the factors Task (WM/LTM), WM Sample Position (1/2), and WM Test (1/2). The analysis showed anticipated main effects of Task [WM/LTM; $F(1,186) = 970.059$, $p < 0.001$, $\eta 2 = 0.468$] and Sample Position [1/2; $F(1,186) = 35.085$, $p < 0.001$, $\eta 2 = 0.014$], as well as a significant Task x Sample Position interaction [$F(1,186) = 19.805$ $p < 0.001$, $\eta 2 = 0.006$], indicating that primacy effects were generally stronger in the LTM than in the WM tests (cf. **Fig. 2a** and **2b**). Furthermore, the Task x WM Test interaction was significant [$F(1,186) = 69.396$ $p < 0.001$, $\eta 2 =0.018$], reflecting the opposite effects of WM priority on WM vs. LTM recall performance (see above). We also found a significant three-way interaction [Task x Sample Position x WM Test; $F(1,186) = 7.589$ $p = 0.006$, $\eta 2 = 0.002$)], which likely reflects the absence of primacy (or alternatively, a recency benefit for Sample 2, see *Discussion*) on WM Test 1, whereas all other tests (WM and LTM) showed primacy

(see **Fig. 2a** and **2b**). No other effects were significant [WM Test, $F(1,186) = 0.342$, $p = 0.560$, $\eta 2 = 0.0001$; Sample Position x WM Test, $F(1,186) = 1.132$ $p = 0.289$, $\eta 2 = 0.0004$].

## Pruning for equivalent WM performance

Conditions that differ in WM performance (like our one-sample vs. two-sample conditions) may be expected to differ trivially also in subsequent LTM, for example, due to information loss having occurred already during WM processing. To account for these differences, we pruned the data post-hoc such that the average WM performance in the one-sample, Test 1, and Test 2 conditions were equalized: for each participant and condition, we successively removed trials with extreme (high or low) WM reporting error until the differences in WM accuracy between the conditions were minimized (see *Methods, Pruning*). We then repeated the subsequent LTM analysis using only the remaining WM samples (**Fig. 2c**; note that a subset of participants (N = 14) whose data were incomplete after pruning were excluded from this analysis). Pruning increased the LTM benefit of WM testing after deprioritization: we now found significantly better LTM performance for the deprioritized WM samples [Test 2; M = 45.14°, SE = 1.54°] compared even to the one-sample condition [M = 51.28°, SE = 1.68°; $t(172) = -3.824$, $p < 0.001$]. Pruning also improved the LTM performance for Test 1 samples (M = 52.63°, SE = 1.32°), which were now on par with the one-sample condition [$t(172) = -0.917$, $p = 0.360$]. Thus, after accounting for differences in WM performance, we observed even clearer LTM benefits for samples that had been retrieved from a deprioritized WM state.

## LTM recall of WM sample vs. WM report

Although participants' task in the LTM test was to recall the orientation of the originally presented WM sample (**Fig. 1**), their LTM reports may have been biased towards the orientations they had reported at the WM test (i.e., with WM reporting error). To examine this possibility, in the unpruned data, we inspected the similarity (in terms of absolute angular difference in °) of the LTM reports to the WM reports (**Fig. 2c**). In fact, the LTM reports were overall more similar to the WM reports than to the original WM sample orientations [M = 48.50°, SE = 0.812° vs. M = 51.20°, SE = 0.79°; $t(186) = -10.770$, $p < 0.001$)]. This bias was evident for each of the WM-task conditions [One-sample, Test 1, Test 2; all $t(186) < -3.424$, all $p < 0.001$], but appeared most pronounced for the Test 2 condition [M = 44.10°, SE = 1.51°; $t(186) = -8.709$, $p < 0.001$). A repeated measures ANOVA confirmed that the increase in bias across conditions (One-sample, Test 1, Test 2) was significant [$F(1,301) = 25.247$, $p < 0.001$, $\eta 2 = 0.002$]. That the bias was strongest for WM Test 2 suggests a particularly strong long-term memory of the WM-*testing* episode (i.e., of the participant's own response)

after the sample information had been deprioritized. It may seem counterintuitive that in the Test 2 condition, the bias towards recalling the subjective WM reports (which include WM error) was increased in tandem with objective LTM accuracy (**Fig. 2b**), given that this condition showed the largest WM reporting error (**Fig. 2a**). However, the result can be explained when considering that the WM errors were generally much smaller than the LTM errors (cf. **Fig 2a** and **b**). A relatively stronger bias towards the subjective WM report may thus reduce the objective LTM error to be smaller, even if the WM error was relatively larger than in other conditions.

For completeness, we also inspected whether the LTM reports were additionally biased by the (random) orientations in which the WM-test probes first appeared on screen (i.e., before the participants re-rotated them). However, the LTM reports' similarity to these probe orientations did not differ from chance level (90°), in none of the conditions [One-sample: $t(186) = -0.485$, $p = 1.000$; Test 1: $t(186) = 2.010$, $p = 0.138$; Test 2: $t(186) = 0.420$, $p = 1.000$].

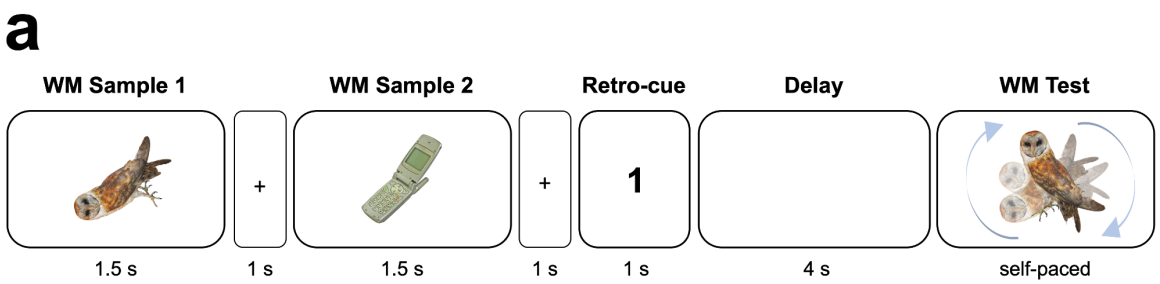
To summarize, while our deprioritization manipulation in Exp. 1 expectedly reduced WM-task performance, it increased the accuracy of subsequent LTM reports. The results appear consistent with a pronounced WM-"testing" effect for deprioritized materials, where participants formed a particularly strong long-term memory of the orientations they had reported at the WM test. An alternative explanation could be that the LTM performance for Test 2 items benefitted, regardless of their deprioritization, from having been maintained in WM for a longer period of time (**Fig. 1a**; e.g., Souza & Oberauer, 2017; Jarjat et al., 2018). To address this possibility, in Experiment 2, we manipulated WM priority using retro-cueing (Griffin & Nobre, 2003; for reviews, see Souza & Oberauer, 2016; Van Ede & Nobre, 2023), which holds the time between sample presentation and WM test constant.

## Experiment 2

**T**he WM task we used in Exp. 2 (N = 89) is illustrated in **Figure 3a**. After the presentation of two WM samples, a visual retro-cue ("1" or "2") indicated which of the two orientations was more likely to be probed after the WM delay. The retro-cue was valid in 75% or 83.33% of the trials (see *Methods for details*). The rationale behind this manipulation was that the cued sample should be maintained with higher priority in WM, while the uncued sample (which is considerably less likely to be tested) should be deprioritized (Griffin & Nobre, 2003). The WM testing procedure in Exp. 2 was otherwise identical to that in Exp. 1 (continuous reports), except that only a single item (cued or uncued) was probed on each trial. The WM task was again followed by a distractor task and a surprise LTM test analogous to Exp 1.

## WM performance

As expected based on previous work (Linde-Domingo & Spitzer, 2024; Oberauer, 2020; Oberauer & Hein, 2012), the WM accuracy for the retro-cued orientations (M = 17.19°, SE = 1.10°) was significantly higher than for the uncued orientations [M = 22.51°, SE = 1.87°; t(88) = -2.959, p = 0.004, **Fig. 3b**]. A 2 x 2 ANOVA with the factors Cueing (cued/uncued) and Sample Position (1/2), showed a main effect of Cueing [F(1,88) = 7.470, p = 0.008, η2 = 0.022], but no effects of Sample Position [main effect: F(1,88) = 0.165, p = 0.686, η2 = 0.0002; Cueing x Sample Position: F(1,88) = 1.157, p = 0.285, η2 = 0.002]. Thus, unlike in Exp. 1, there was no significant primacy effect on WM task performance in Exp. 2 (see *Discussion*). However, the probabilistic cueing did induce the anticipated retro-cue effect, indicating that the priority manipulation was successful.



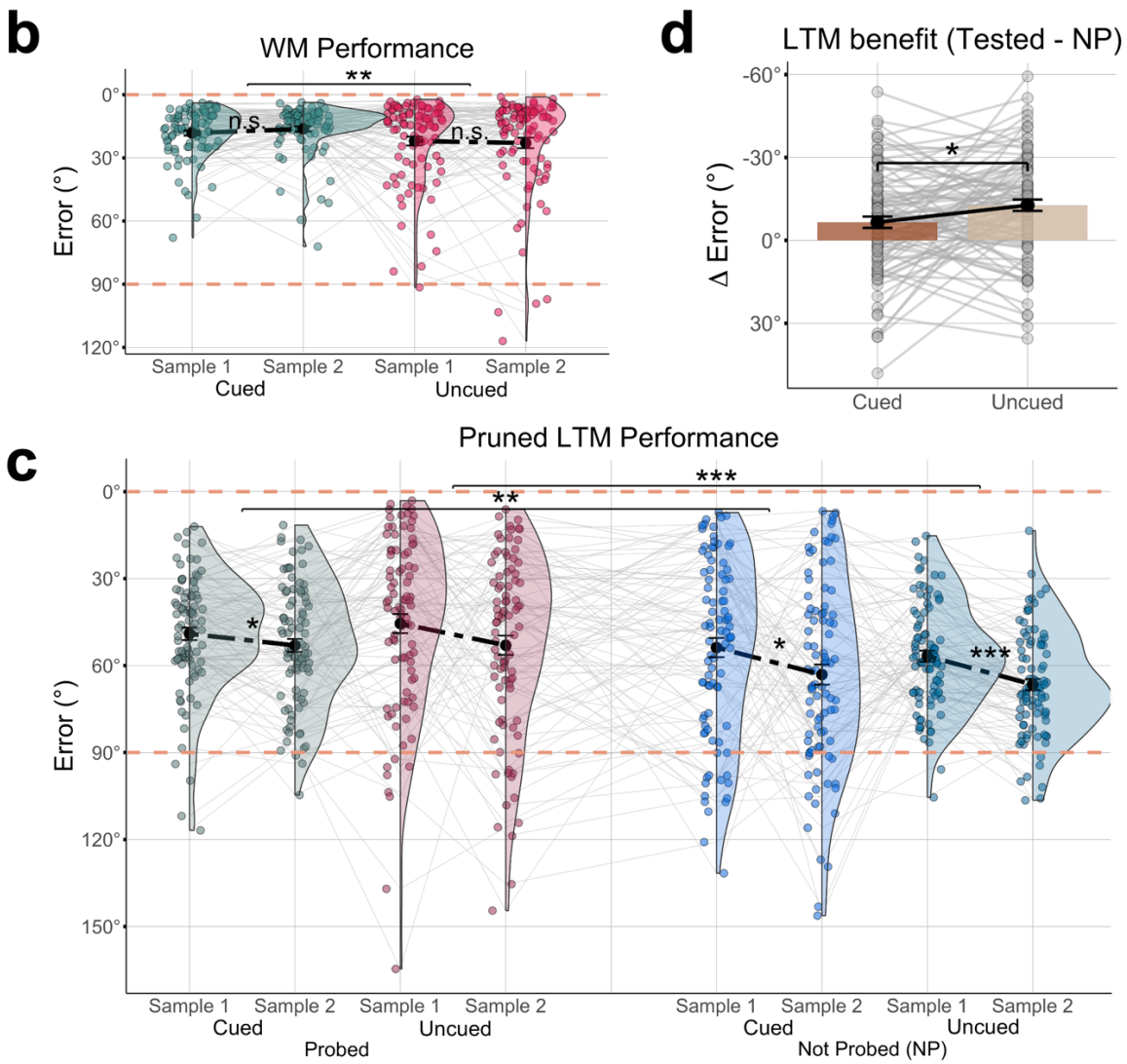



**Figure 3. Experiment 2. a**, WM task. After the presentation of two randomly oriented objects (WM Sample 1 and 2), a retro-cue ("1" or "2") indicated which of the two sample orientations was most likely to be probed after the delay. In the WM test, participants were probed to recall (continuous report) the orientation of the cued sample or, in a smaller fraction of trials, the uncued sample. **b**, WM task performance (same plotting conventions as **Fig. 2a**). WM accuracy for the uncued information was significantly lower than for the cued information. **c**, Subsequent LTM test performance after pruning for equal WM performance in the cued/uncued conditions. LTM accuracy for samples that had been probed (tested) in the WM task was significantly higher (smaller errors) than for samples that had not been probed (NP, blue) **d**. Benefit of WM retrieval (tested vs. NP) for subsequent LTM accuracy, plotted separately for cued and uncued samples. The LTM benefit of WM retrieval was significantly larger for uncued samples.

## LTM performance

Participants' overall accuracy in the LTM test of Exp. 2 (M = 56.11°, SE = 1.101°) was at similar levels as in Exp. 1 [M = 53.95°, SE = 1.14°; t(164.08) = 1.017, p = 0.311; Welch's t-test]. In Exp. 2, we again observed substantially higher LTM performance for WM samples that had been tested in the WM task (M = 51.98°, SE = 1.570), compared to unprobed (NP) samples [M = 60.23°, SE = 1.48°; t(88) = -7.51, p <0.001; **Fig. 1c**]. This LTM benefit of WM retrieval was evident both for cued and uncued samples [t(88) = -3.621, p = 0.002 and t(88) = -4.619, p < 0.001].

Turning to the LTM consequences of WM cueing, we first inspected the full data (i.e., without pruning) irrespective of the differences in WM performance between cued and uncued samples. A 2 x 2 ANOVA with the factors WM Testing (tested vs. NP) and Cueing (cued/uncued) showed a main effect of Testing [F(1,88) = 27.974, p < 0.001, η2 = 0.040], reflecting the overall LTM benefit of WM retrieval, but no significant effects of Cueing [main effect: F(1,88) = 1.758, p = 0.188, η2 = 0.002; Testing x Cueing: F(1,88) = 0.703, p = 0.404, η2 = 0.0007].

Next, we repeated the analysis after pruning the data (see Exp.1 and *Methods*) to warrant equivalent WM performance for cued and uncued samples. After pruning, the LTM results showed a significant interaction of WM Testing and Cueing [F(1,88) = 5.826, p = 0.01, η2 = 0.006; main effect of Testing: F(1,88) = 36.164, p < 0.001, η2 = 0.053; main effect of Cueing: F(1,88) = 0.352, p = 0.555, η2 = 0.0005], which indicates a greater WM-testing benefit for uncued than for cued samples. **Figure. 3d** shows the magnitude of the WM-testing benefit (tested vs. NP) which was significantly larger for the uncued than the cued samples. In other words, in terms of long-term memorability, deprioritized samples benefited more from WM retrieval than prioritized samples that had been retrieved (from WM) with equivalent accuracy. Further inspection of the LTM results with a 2 x 2 x 2 ANOVA (Sample Position x Testing x Cueing) showed a significant main effect of Sample Position [F(1,88) = 24.613, p < 0.001, η2 = 0.021] indicating a primacy effect on LTM recall (see also Exp. 1), but no additional new interactions [all F < 1.256, all p > 0.266, all η2 < 0.001].

Together, Exp. 2 confirmed a stronger LTM benefit of WM testing after deprioritization, even when the duration of WM maintenance was controlled for. While the effects of probabilistic retro-cueing were more subtle (both in terms of WM and LTM performance, **Fig. 3**) compared to the priority manipulation in Exp. 1 (cf. **Fig. 2**), they corroborate a role of WM priority for the magnitude of subsequent LTM benefits, over and above potential effects of maintenance duration.

# Experiment 3

An important aspect of the WM tests in Exp. 1 and 2 (continuous reports) was that the WM probes appeared in a quasi-random orientation (**Fig. 1a**, see *Methods)*, which provided no opportunity to 'restudy' the sample information. In other words, at the WM tests, participants could only possibly have ('re')studied the object orientation they had subjectively remembered and reproduced on screen themselves from WM. In that sense, our results appear reminiscent of a "generation effect" (for a review, see Bertsch et al., 2007; Karpicke & Zaromb, 2010; the finding that self-generated information is particularly memorable; Serra & Nairne, 1993), which may have been more pronounced after temporary deprioritization. In Experiment 3 (n=107), we explored whether another common type of visual WM testing (delayed comparison), which does not involve active reproduction of the WM information, may induce subsequent LTM benefits as well.

Except for the difference in WM testing and minor changes to the WM trial timings (see *Methods*), the design of Exp. 3 was identical to Exp. 1. The key difference was that the WM probes in Exp. 3 differed only slightly (+/- 14°) from the original sample orientation, and participants were asked to indicate with a single button press whether the difference was clockwise or counterclockwise (cw/ccw; **Fig. 4a**, *right*). Thus, whereas the WM probes in Exp. 1 and 2 were uninformative about the original sample orientation, the probes in Exp. 3 did repeat (approximate) information about the sample's orientation in 360° space. The WM task was again followed by a distractor task and a surprise LTM test (with continuous reports) analogous to Exp. 1 and 2.

## WM performance

Unlike in Exp. 1, the WM performance in Exp. 3 was not significantly modulated by load or priority (**Figure 4a**). Descriptively, the percentage of correct responses was highest in one-sample trials (M = 70.00%, SE = 1.10%), followed by Test 1 and Test 2 on two-sample trials (M = 69.00%, SE = 0.008 and M = 67.30%, SE = 1.20%), but the differences were not statistically significant [one-sample vs. Test 1: $t(106) = 0.901$, $p = 0.452$; Test 1 vs Test 2: $t(106) = 1.217$, $p = 0.452$]. Focusing on the two-sample trials, a 2 x 2 repeated measures ANOVA with the factors WM Sample Position (1/2) and Test (1/2) yielded no significant main effects [Sample Position: $F(1,106) = 0.309$, $p = 0.580$, $\eta2 = 0.0007$; Test: $F(1,106) = 1.164$, $p = 0.283$, $\eta2 = 0.003$] and no interaction [$F(1,106) = 0.213$, $p = 0.645$, $\eta2 = 0.0005$]. Thus, albeit WM performance in Exp. 3 was significantly above chance [$t(106) = 28.771$, $p < 0.001$]), it was hardly modulated by task factors (for similar null-results using a recognition test, see LaRocque et al., 2015).

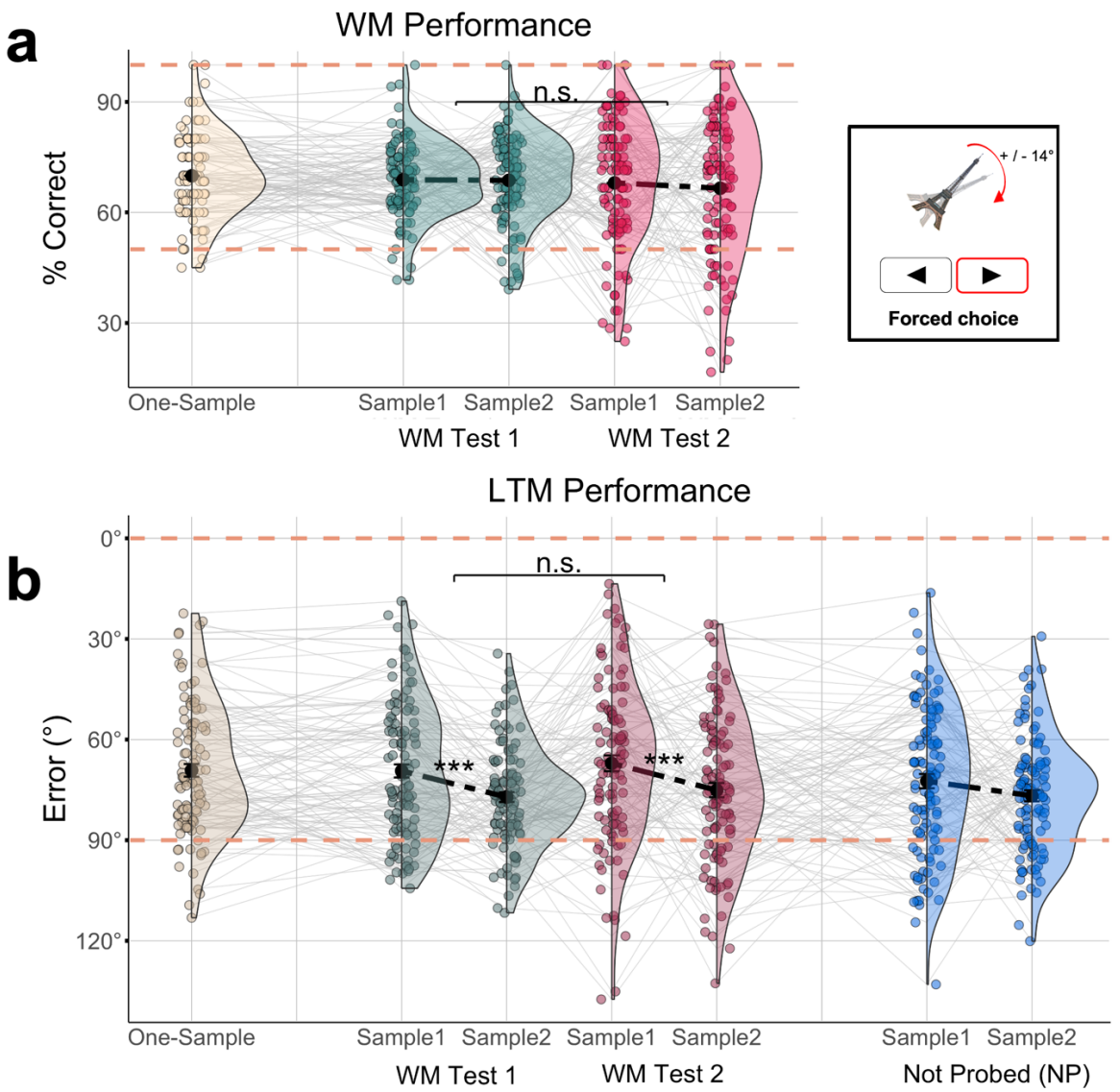



**Figure 4. WM and LTM performance in Experiment 3. a,** *Left,* WM task performance. Accuracy is shown as percentage correct responses, otherwise same conventions as in **Fig. 2a**. We observed no significant differences between conditions (see *Results* for details). *Right,* In the WM tests in Exp. 3, participants indicated whether the WM probe orientation was changed (+/-14°) clockwise (cw) or counterclockwise (ccw) relative to the WM sample. **b**, LTM test performance (continuous report), same conventions as in **Fig. 2b**. While the results showed significant load- and primacy effects, there was no benefit of WM testing (tested vs. NP) and no effect of WM priority (WM Test 2 vs. 1).

## LTM performance

The subsequent LTM test procedure in Exp. 3 was identical to that in Exps. 1 and 2. Compared to Exp. 1, the overall LTM accuracy in Exp. 3 was significantly lower [M = 72.14°, SE = 1.137°, t(236.6) = 10.168, $p < 0.001$]. Furthermore, unlike the previous experiments, Exp. 3 showed only a weak WM-testing benefit relative to NP items

[**Fig. 4b**; M = 70.95°, SE = 1.88° vs. M = 74.68°, SE = 1.45°; t(106) = -2.00, p = 0.048; paired t-test comparing Test 2 vs. NP samples, uncorrected]. Comparing the testing effects in Exps. 1 and 3 directly, a mixed-effects ANOVA with the between-subjects factor Experiment (Exp. 1 vs. 3) and the within-subjects factor WM-testing (Tested vs. NP) showed significant main effects for both factors [Experiment: F(1,292) = 76.202, p < 0.001, η2 = 0.171; Testing: F(1,292) = 51.062, p < 0.001, η2 = 0.035] as well as a significant interaction [F(1,292) = 18.606, p < 0.001, η2 = 0.013], which confirms that the delayed-comparison WM testing in Exp. 3 had less benefits for subsequent LTM than the continuous-report WM tests in Exp. 1.

Across conditions in Exp. 3, participants were slightly more accurate in recalling the orientations from one-sample WM trials (M = 68.83°, SE = 1.87°) compared to two-sample WM trials [Tests 1 and 2 combined, M = 72.34°, SE = 1.54°; t(106) = -2.38, p = 0.019; **Fig. 4b**]. However, focusing on two-sample trials, unlike in Exp.1, we found no significant LTM benefit for samples probed in WM Test 2 (M = 70.95°, SE = 1.88°) compared to WM Test 1 [M = 73.02°, SE = 1.55°; t(106) = -1.492, p =0.139]. A 2 x 2 ANOVA (specified analogously as above) showed a main effect of Sample Position [1/2; F(1,106) = 25.056, p < 0.001, η2 = 0.035], indicating a primacy effect, but no main effect of WM Test [1/2, F(1,106) = 3.056, p = 0.083, η2 = 0.003] and no interaction between the two factors [F(1,106) = 0.002, p = 0.969, η2 < 0.001]. Post-hoc tests showed the primacy effect to be significant both for samples probed first and second in WM [Test 1 and Test 2; t(106) = -4.553, p < 0.001 and t(106) = -5.069, p < 0.001; **Fig. 3d**].

Together, while the LTM results of Exp. 3 replicated a primacy effect, the different WM testing procedure eliminated the LTM benefit for deprioritized WM information that we found in the previous experiments. In fact, unlike in Exps. 1 and 2, WM testing in Exp. 3 barely had an LTM benefit at all. These observations corroborate that the LTM benefits for deprioritized WM information in Exp. 1 and 2 were likely mediated by retrieval and/or self-generation processes associated with continuous reporting. At the same time, Exp. 3 showed that the repeated presentation of (approximate) sample information during the WM tests was not sufficient to induce a substantial LTM benefit, for neither the deprioritized nor the deprioritized WM samples.

# Discussion

To summarize our main findings, using a novel visual WM-LTM paradigm, we found that although attentional deprioritization reduced immediate WM recall accuracy, it

increased subsequent LTM recall performance. This pattern was observed both when WM priority was manipulated through testing order (Exp. 1) or retro-cues (Exp. 2). More specifically, deprioritization appeared to enhance the LTM benefit of WM retrieval, and led to a stronger long-term memory of the information that had been remembered at the WM test (via continuous report). In contrast, no clear LTM enhancement was observed when we tested WM with a simpler (binary choice) delayed comparison task (Exp. 3). In addition, the LTM results in all experiments showed a 'primacy' pattern, such that items that had occurred at the beginning of a WM-trial episode were later recalled better. A similar primacy effect was also evident in WM recall, but only for deprioritized information. Together, our findings highlight various aspects in which WM retrieval of deprioritized information—as opposed to prioritized information—resembles retrieval from episodic LTM.

It is well-established that temporary deprioritization of WM content reduces its accuracy (Bae & Luck, 2018; Emrich et al., 2017; Oberauer, 2002), a finding we also replicated here (Exp. 1 and 2). By intuition, one might assume that if information is deprioritized in WM, it is also less likely to be encoded into a durable long-term memory. Indeed, several recent studies, mostly using recognition tests, found that unprioritized WM contents were later remembered less well than prioritized ones (Fan & Turk-Browne, 2013; Jeanneret et al., 2023; LaRocque et al., 2015; Reaves et al., 2016; Strunk et al., 2019; Wang & Van Ede, 2024). Here, using a more recall-like WM testing format (continuous reports), we found the opposite: deprioritization in WM paradoxically improved subsequent LTM recall. At face value, this counterintuitive result is reminiscent of previous work on the "McCabe effect" (Loaiza et al., 2023; Loaiza & McCabe, 2012, 2013; McCabe, 2008), where an intermittent distractor task during word-list learning impaired the words' immediate (WM-like) recall, but improved their later (LTM-like) recall after a longer delay. The McCabe effect has been explained in terms of 'covert' retrieval of the WM information back into the focus of attention (Loaiza & Halse, 2019; McCabe, 2008) after it had temporarily been stored in 'activated LTM' (e.g., Cowan, 1999; Oberauer, 2002). Our present experiments and materials differed in important ways from the word-span tasks used by McCabe and colleagues. A possibility exists that covert retrieval processes (i.e. after temporary deprioritization) may have contributed to the LTM results in Exp. 1, where the added WM delay (Delay 2) could have provided additional opportunity for covert retrieval. However, as outlined below, the entirety of our results across experiments indicates that the LTM benefits for deprioritized WM contents arose largely from *overt* WM testing, specifically with continuous reports.

We found subsequent LTM benefits after WM deprioritization not only when priority was manipulated through WM-testing order (Exp. 1), but also when using retrospective cues in Exp. 2. With the latter experiment design, the effect showed directly as a stronger WM-testing *benefit* (relative to untested/NP samples) for deprioritized information. This result cannot be easily explained by differences in covert retrieval and/or WM delay length, but can be attributed to the (overt) WM

testing proper. In line with this interpretation, in Exp. 1 and 2, participants' LTM reports were more similar to their own previous WM reports than to the original WM sample information, and this 'bias' was increased after WM deprioritization. In other words, after WM deprioritization, the participants appeared to show a stronger (overt) "generation effect" (Jacoby, 1978; Slamecka & Graf, 1978; for review, see Bertsch et al., 2007), which further underscores the role of overt testing/reporting in explaining our results. Notably, although LTM bias towards the (imperfect) WM reports in principle reflects a source of error, the WM errors were small enough (relative to the LTM error) for such bias to still go along with an objective LTM benefit for the deprioritized WM materials. An alternative explanation in terms of cardinal orientation biases (Bae, 2021; Linde-Domingo & Spitzer, 2024; Taylor & Bays, 2018) was not supported by our present data (see **Supplementary Fig. 1**). Finally, we observed no LTM boost for deprioritized WM information—and hardly any WM-testing benefits at all—in Exp. 3, which was near-identical to Exp. 1 but used a different WM testing procedure that relied less on active retrieval and/or self-generation. These aspects of our results underscore how overt WM testing may affect subsequent LTM, and show that the long-term consequences of WM testing can depend—in seemingly counterintuitive ways—on the attentional state of the WM information.

The present WM-testing effects, particularly for deprioritized information, show notable parallels to classic (LTM-)testing or "retrieval practice" effects in the episodic (long-term) memory literature. LTM-testing effects are known to be more pronounced if successful retrieval practice of the material is more difficult (Butler & Roediger, 2007; Glover, 1989; for review, see Rowland, 2014). In a similar vein, the present WM-testing effects were strongest for those materials that were hardest to remember in the WM task (i.e., the deprioritized materials). Further in line with a "retrieval-effort" account, LTM-testing effects are typically larger with cued/free recall than with recognition testing (Bjork, Robert A., 1975; Pyc & Rawson, 2009; Roediger & Karpicke, 2006), and we likewise observed greater benefits with a recall-like WM test (Exp. 1 & 2) than with simple (binary) sample-probe judgments (Exp. 3). Lastly, LTM-testing effects are typically shown relative to a "restudy" baseline where the memory material is presented again without retrieval requirements (Roediger & Karpicke, 2006; Rowland, 2014). Our WM experiments did not include dedicated restudy conditions, however, the WM probes in Exp. 3 did reshow the sample information in reasonable approximation (**Fig. 4a**) to allow for restudying it. The lack of clear testing effects in Exp. 3 thus renders it unlikely that the robust LTM benefits in Exp. 1 & 2 would also have occurred under restudy conditions. Taken together, the long-term consequences of WM testing in our tasks showed many of the hallmarks of classic (episodic) retrieval-practice and align well with existing accounts of LTM-testing effects (e.g., retrieval effort theories; Rowland, 2014).

Another potential parallel between WM retrieval of deprioritized information and LTM retrieval in our tasks appeared evident in the extent to which the first or the second sample in a WM trial was remembered better. In the final LTM tests, in all our

experiments, we observed a clear ‘primacy’ benefit for the first-presented WM sample. As a possible explanation, the first sample marked the beginning of a new (WM-)task episode, which may have promoted its contextual encoding into episodic LTM. Of note, such within-trial primacy effects were clearly evident also for samples that were not probed (NP) in the WM task. This supports a view that the primacy effects reflected episodic/contextual encoding factors (Sederberg et al., 2006), unlike the retrieval-induced phenomena discussed in the previous paragraphs. Interestingly, in Exp. 1, a moderate primacy effect was evident also in WM recall, but only after deprioritization (WM Test 2). The WM recall of prioritized information (WM Test 1) in contrast showed, if anything, a (non-significant) recency effect, i.e., better recall of the last-presented sample. A similar but non-significant difference in primacy/recency was also seen in Exp. 2 which had a smaller participant sample. Albeit speculative, these observations may suggest that the role of episodic context factors, in terms of within-trial primacy, increased from prioritized WM over deprioritized WM to LTM recall, which adds to the apparent similarities between the latter two.

There exists a range of views on how unattended WM storage is implemented mechanistically in the brain (Beukers et al., 2021; Stokes, 2015; Van Loon et al., 2018; Wan, 2022; Wolff et al., 2017; Yu et al., 2020) and the extent to which the underlying processes are distinguished (or not) from episodic LTM remains debated (Beukers et al., 2023; Oberauer & Awh, 2022). A previous study found no evidence that unattended WM maintenance would improve subsequent LTM (LaRocque et al., 2015), and we likewise observed no LTM-benefits for deprioritized materials (see Exp. 2, NP items) unless the material was explicitly tested. We thus found no evidence that unattended WM contents would make stronger contact with LTM through unattended storage per se. Instead, the LTM benefits manifested only when the material was actively retrieved (i.e., tested) from its deprioritized state. The observed similarities to LTM retrieval are consistent in principle with a view that the deprioritized WM information may have been retrieved from a LTM-like storage state (Beukers et al., 2021, 2023). Alternatively, our results may indicate that retrieval from dedicated “unattended” storage formats within WM (e.g., Stokes, 2015; Yu et al., 2020) benefits LTM in similar ways as LTM retrieval, through yet unknown mechanisms. Future work using neural recordings will be needed to differentiate between these possibilities.

To conclude, factors that promote (or hinder) subsequent remembering are of central concern in basic memory research, but also in applied contexts such as the educational sector. Here, we showed that recalling information from WM can promote its long-term retention, particularly if the WM information was temporarily not in the focus of attention. Beyond resembling classic LTM-“testing” effects, our results join other findings that some memory operations (e.g., ‘replay’; Jafarpour et al., 2017; Schapiro et al., 2018) seem to favor weaker, or more distant memories (see also Antony et al., 2024) despite them potentially being less accurate. An intriguing

question for future work is how WM retrieval of deprioritized information intersects with processes thought to underlie long-term memory and learning on the neural level.

# Methods

## Experiment 1

**Participants.** Participants (N = 199, 58 female, 130 male, mean age = 26.99; missing demographic information for N = 11) were recruited online via Prolific Academic (https://www.prolific.ac/), with informed consent obtained via the Qualtrics platform (https://www.qualtrics.com). The eligibility criteria were that participants had to be between 18 and 35 years old, fluent in English, have a normal or corrected-to-normal vision, and have a minimum approval rate of 95% on Prolific. The experiment lasted approximately 40 minutes. Participants were reimbursed with £6.75 for completing the experiment. Partial payments were made if the experiment was not completed due to technical issues (N = 4), failed attention checks (N = 5), or early termination by the participant (N = 2). One participant (N = 1) was excluded post-experimentally for failing to perform significantly above chance in the WM task ($p < 0.05$, t-test against 90° angular error, one-tailed). Thus, N = 187 participants remained for analysis (55 female, 121 male; mean age 27.2 years; missing demographic information for N = 11). The experiment was approved by the Internal Review Board (IRB) of the Max Planck Institute for Human Development.

**Stimuli**. The experimental stimuli consisted of 110 pictures of animate and inanimate objects. For each participant, 100 pictures were randomly selected for the experiment, while 10 were designated for practice trials. An additional 3 pictures, identical across all participants, were used for instruction. All pictures were selected from the Bank of Standardized Stimuli (BOSS) database (Brodeur et al., 2010). The stimuli were presented rotated in the experiment **(Fig. 1**). The WM-sample orientations were selected randomly and independently from a set of 16 equidistant angles from 11.25° to 348.75° in steps of 22.5°, which excluded the cardinal axes (0°, 90°, 180°, and 270°).

**Task(s).** The experiment consisted of 3 phases: a WM task (60 trials), a distractor task (approx. 1 minute), and a surprise LTM test (100 trials). In the WM task, participants were asked to briefly remember the orientation of one or two WM samples. Each trial started with a fixation cross (2 s), at the screen center. In one-sample trials, the WM sample was presented for 1.5 s followed by a delay period (empty screen of 4.5 s, after which the participants were asked to remember the WM sample's orientation (see below). In two-sample trials, two WM samples were sequentially presented (with a 1 s inter-stimulus interval during which the fixation cross was shown). After a WM delay of 2 s, the orientation of one of the two objects (randomly selected) was probed (Test 1). After Test 1, in 50% of the two-sample trials (randomly varied), the experiment continued with the next trial. On the remaining two-

sample trials, Test 1 was followed by another delay period (1.5s) and participants were probed to also remember the orientation of the other, previously unprobed sample (Test 2). The assignment of sample stimuli to the one-sample, Test 1, and Test 2 conditions was counterbalanced across participants using a Latin square approach, such that on average, the exact same sample stimuli were used in each of these conditions. In all WM tests, the sample orientation in question was probed by the WM object reappearing in a new orientation (random, but at least 22.5° different from the original orientation). Participants were asked to re-rotate the probe to the remembered orientation using the left and right arrow keys (continuous report; **Fig. 1**) and to submit the result by pressing the space key. Trials in which participants failed to submit a response within a generously allotted time window (15 s) were excluded from the analysis (2.01% of trials on average; min = 0.00%, max = 7.93%).

After the WM task, participants performed a short distractor task (approximately one minute) in which they were asked to solve a series of simple math problems (e.g., 100 - 7 = ?) using mental arithmetic and entering the solutions via the computer keyboard.

In the subsequent surprise LTM test, participants were asked to recall the orientations of each of the previously encountered WM samples again. Each test trial started with a fixation cross (1.5 s) after which one of the previous WM sample objects appeared in a new orientation (fully random) as LTM probe. Participants were asked to reproduce the objects' original orientation (i.e., the orientation it had as a WM sample), using the same response procedure as in the WM tests (continuous reporting; see above). Each WM sample was probed once (in random serial order) across the LTM test trials.

**Procedure.** Participants were given written instructions about the WM task and could practice the continuous reporting procedure (i.e., re-rotating stimuli via arrow keys) prior to the experiment. They were free to repeat the instructions until they felt confident to perform the task. Participants first performed six practice trials of the WM task (two per trial type: one-sample trials, and two-sample trials with and without Test 2). Thereafter, each participant performed a total of 60 WM trials (20 one-sample trials and 40 two-sample trials, in random serial order). After 14 WM trials, a brief attention check task was performed (6 trials). For this, a number word (e.g., "three") was presented at the screen center, surrounded by 4 different number symbols. Participants were asked to pick the correct number symbol (e.g., "3") via arrow keys. When a participant failed this check on more than 2 of 6 trials, the experiment was aborted (see *Participants*). After completing the WM task, participants performed the distractor task (mental arithmetic, approx,1 minute). After this, they were informed about the surprise LTM test and received short instructions about its procedure. Participants then performed 100 LTM test trials in which they were asked to recall all of the sample orientations they had encountered in the WM task (including those WM samples that had not been probed in the WM task, i.e., on two-sample trials with a

single WM test). Participants could take a break (self-paced for up to 2 minutes) after 34 trials of the WM task.

## Experiment 2

**Participants.** For Exp. 2, we recruited another sample of N = 101 participants online (46 female, 42 male, and 1 diverse, mean age 25.3 years; missing demographic information for 12 participants), with the same modalities of recruitment, informed consent, ethics approval, reimbursement, and experiment duration as in the previous experiments. For N = 4 participants, the data was not saved due to technical problems. Two further participants were excluded due to failed attention checks, N = 1 used paper and pencil to solve the task, N = 1 started the experiment more than once, N = 3 participants did not enter any data, and N = 1 completed the task but experienced other technical problems leading to an exclusion. Thus, N = 89 participants remained for analysis (41 female, 42 male, 1 diverse; mean age 25.0 years; missing demographic information for 5 participants).

**Stimuli, Task, and Procedure**. The stimulus material for Exp. 2 was extended to 112 objects from the BOSS database and was otherwise identical to Exps. 1 and 3. Exp. 2 differed from the previous experiments only in WM task design. Each trial in the WM task started with the presentation of two WM samples, like the two-sample trials of the previous experiments. However, after a short delay (0.5 s fixation cross and 0.5 s blank screen), a retro-cue (“1” or “2”) was displayed (1 s) which indicated which of the two WM samples would be more likely to be probed at the WM test. The retro-cue was followed by a WM delay (4 s, empty screen), after which the WM probe appeared and participants were asked to re-rotate it using the same WM-test procedure (continuous report) as in Exp. 1. We initially tested N = 55 participants (N = 47 after exclusions) with a cue validity of 75% (i.e., in 25 % of trials, the uncued sample was probed). After inspecting the preliminary WM task data, we increased the cue validity to 83.33% for the remaining N = 46 participants (N = 42 after exclusions). For counterbalancing reasons, participants in the former group performed 56 trials (with 112 sample objects) and participants in the latter group performed 48 trials (with 96 sample objects) in the WM task. Within each group, the stimulus material used in the different conditions (cued/uncued x probed/unprobed) was counterbalanced across participants using a Latin square design. Trials in which participants failed to respond within the allotted time were excluded from the analysis M = 0.33% of trials min = 0.00%, max = 5.36%). After the WM task, participants performed a distractor task (mental arithmetic) and a surprise LTM test analogous to Exps. 1 and 3.

## Experiment 3

**Participants.** For Exp. 3, we recruited a new sample of 155 participants online (44 female, 100 male, diverse = 1, mean age 27.4 years; missing demographic information for N = 10). The modalities of recruitment, eligibility criteria, informed consent, ethics approval, and reimbursement, were the same as in Exp. 1. For N = 5 participants, the experiment was terminated prematurely due to failed attention checks, and N = 5 participants had to be excluded due to technical problems. Of the remaining participants, N = 38 were excluded because they failed to perform above chance level in the WM task ($p < 0.05$, Binomial test against 60% correct responses, one-tailed), leaving N = 107 participants (27 female, 50 female, mean age 27.4 years; demographic information missing for 30 participants) for analysis.

**Stimuli, Task, and Procedure.** The design of Exp. 3 closely resembled Exp. 1. The main difference was that in Exp. 3, the WM tests were delayed comparisons, where the WM probe was rotated +/- 14° relative to the WM sample. Participants were asked to indicate with a single key press (right or left arrow key) whether the sample-probe difference was clockwise (cw) or counterclockwise (ccw). Given the expectedly faster WM testing procedure (compared to the continuous reports in Exp. 1), we decreased the response time window to 3 s and slightly changed the lengths of the WM delays: in two-sample trials, the first WM delay was shortened to 1 s, and the second delay was extended to 2 s. In one-sample trials, the WM delay was extended to 5 s to approximately match the time between the first sample and Test 2 in two–sample trials.

Participants could practice the binary choice test format before starting with the experiment. Trials in which participants failed to respond within the allotted time were excluded from analysis (M = 0.51% of trials, min = 0.00%, max = 8.33%). The WM task in Exp. 3 was followed by a distractor task (mental arithmetic) and a subsequent surprise LTM test, using the same procedures as in Exps. 1 and 2.

## Pruning for equivalent WM performance

To account for differences in WM performance when comparing LTM performance between conditions, in our experiments with continuous reports (Exp. 1 and 2), we used a pruning approach. For each participant and condition (e.g., one-sample, Test 1, and Test 2 in Exp. 1), we successively removed individual trials with extreme (high or low) WM reporting error until the WM accuracy in all conditions was maximally similar to the participant's overall mean WM accuracy (averaged over all conditions). We then repeated the LTM analysis using only the WM samples that remained after

this pruning. For a subset of participants in Exp. 1 (N = 14), pruning was not possible because their error distributions in the different conditions did not sufficiently overlap. In Exp. 2, all participants (N = 89) could be retained in the analysis even after pruning. For completeness, we also performed exploratory LTM analysis of Exp. 3 where we included only samples from WM trials in which the binary WM report was correct. However, the results from this analysis were qualitatively identical to those reported in **Fig. 4b** which included all trials.

### Statistical Analysis

Throughout the analyses of continuous report data (WM and LTM) we examined memory performance in terms of absolute angular error (or deviation) in degrees (°), where lower values indicate higher accuracy (note inverted y-axes in Figures). Unless stated otherwise, all reported pairwise comparisons (t-tests) were corrected for multiple testing using the Holm-Bonferroni method.

## Data Availability

The data that support the findings of this study are available on GIN: DOI: 10.12751/g-node.3p3ryv

## Code Availability

The analysis code and experiment code are available on GitHub and archived on Zenodo: DOI: 10.5281/zenodo.13867139 and DOI: 10.5281/zenodo.13867798

## Author Contributions

F.J.B.: conceptualization, data curation, formal analysis, investigation, project administration, validation, visualization, writing – original draft, writing – review & editing. B.S.: conceptualization, funding acquisition, methodology, project administration, resources, supervision, writing – original draft, writing – review & editing.

# Acknowledgement

We thank Maik Messerschmidt and Philip Jakob for technical assistance with online data collection, Stefan Appelhoff for advice regarding data-sharing and Open Science, Marit Petzka for help with statistical review and Aleksandra Zinoveva for help with stimulus selection and administration. We also thank Or Yizhar, Juan-Linde-Domingo, and Lukas Muttenthaler for helpful comments and discussions.

# Funding

This research was supported by European Research Council Consolidator Grant ERC-2020-COG-101000972 (B.S.) and Deutsche Forschungsgemeinschaft (DFG) Grant 462752742 (B.S.). The funders had no role in the study design, data collection and analysis, and decision to publish.

# Conflicts of Interest

None

# References


Antony, J., Liu, X. L., Zheng, Y., Ranganath, C., & O'Reilly, R. C. (2024). Memory out of context: Spacing effects and decontextualization in a computational model of the medial temporal lobe. *Psychological Review*. https://doi.org/10.1037/rev0000488

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In *Psychology of learning and motivation* (Vol. 2, pp. 89–195). Elsevier.

Baddeley, A. (2012). Working Memory: Theories, Models, and Controversies. *Annual Review of Psychology*, *63*(1), 1–29. https://doi.org/10.1146/annurev-psych-120710-100422

Bae, G.-Y. (2021). Neural evidence for categorical biases in location and orientation representations in a working memory task. *NeuroImage*, *240*, 118366. https://doi.org/10.1016/j.neuroimage.2021.118366

Bae, G.-Y., & Luck, S. J. (2018). Dissociable Decoding of Spatial Attention and Working Memory from EEG Oscillations and Sustained Potentials. *The Journal of Neuroscience*, *38*(2), 409–422. https://doi.org/10.1523/JNEUROSCI.2860-17.2017

Bartsch, L. M., Singmann, H., & Oberauer, K. (2018). The effects of refreshing and elaboration on working memory performance, and their contributions to long-term memory formation. *Memory & Cognition*, *46*, 796–808.

Bertsch, S., Pesta, B. J., Wiscott, R., & McDaniel, M. A. (2007). The generation effect: A meta-analytic review. *Memory & Cognition*, *35*(2), 201–210. https://doi.org/10.3758/BF03193441

Beukers, A., Buschman, T. J., Cohen, J. D., & Norman, K. A. (2021). Is Activity Silent Working Memory Simply Episodic Memory? *Trends in Cognitive Sciences*, *25*(4), 284–293. https://doi.org/10.1016/j.tics.2021.01.003

Beukers, A., Hamin, M., Norman, K. A., & Cohen, J. D. (2023). When working memory may be just working, not memory. *Psychological Review*.

Bjork, Robert A. (1975). Retrieval as a memory modifier: An interpretation of negative recency and related phenomena. In *Information processing and cognition* (1st Edition).

Bledowski, C., Kaiser, J., & Rahm, B. (2010). Basic operations in working memory: Contributions from functional imaging studies. *Behavioural Brain Research*, *214*(2), 172–179. https://doi.org/10.1016/j.bbr.2010.05.041

Brodeur, M. B., Dionne-Dostie, E., Montreuil, T., & Lepage, M. (2010). The Bank of

Standardized Stimuli (BOSS), a New Set of 480 Normative Photos of Objects to Be Used as Visual Stimuli in Cognitive Research. *PLoS ONE*, *5*(5), e10773. https://doi.org/10.1371/journal.pone.0010773

Brown, R., & Kulik, J. (1977). Flashbulb memories. *Cognition*, *5*(1), 73–99.

Butler, A. C., & Roediger, H. L. (2007). Testing improves long-term retention in a simulated classroom setting. *European Journal of Cognitive Psychology*, *19*(4–5), 514–527. https://doi.org/10.1080/09541440701326097

Cowan, N. (1999). An Embedded-Processes Model of Working Memory. In A. Miyake & P. Shah (Eds.), *Models of Working Memory* (1st ed., pp. 62–101). Cambridge University Press. https://doi.org/10.1017/CBO9781139174909.006

Craik, F. I., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, *11*(6), 671–684.

D'Esposito, M. (2007). From cognitive to neural models of working memory. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1481), 761–772. https://doi.org/10.1098/rstb.2007.2086

D'Esposito, M., & Postle, B. R. (2015). The Cognitive Neuroscience of Working Memory. *Annual Review of Psychology*, *66*(1), 115–142. https://doi.org/10.1146/annurev-psych-010814-015031

Emrich, S. M., Lockhart, H. A., & Al-Aidroos, N. (2017). Attention mediates the flexible allocation of visual working memory resources. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(7), 1454–1465. https://doi.org/10.1037/xhp0000398

Fan, J. E., & Turk-Browne, N. B. (2013). Internal attention to features in visual short-term memory guides object learning. *Cognition*, *129*(2), 292–308.

Glover, J. A. (1989). The 'testing' phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology*, *81*(3), 392–399. https://doi.org/10.1037/0022-0663.81.3.392

Gorgoraptis, N., Catalao, R. F. G., Bays, P. M., & Husain, M. (2011). Dynamic Updating of Working Memory Resources for Visual Objects. *Journal of Neuroscience*, *31*(23), 8502–8511. https://doi.org/10.1523/JNEUROSCI.0208-11.2011

Griffin, I. C., & Nobre, A. C. (2003). Orienting attention to locations in internal representations. *Journal of Cognitive Neuroscience*, *15*(8), 1176–1194.

Hartshorne, J. K., & Makovski, T. (2019). The effect of working memory maintenance on long-term memory. *Memory & Cognition*, *47*, 749–763.

Hitch, G. J., & Baddeley, A. D. (1976). Verbal Reasoning and Working Memory. *Quarterly Journal of Experimental Psychology*, *28*(4), 603–621. https://doi.org/10.1080/14640747608400587

Hurlstone, M. J., Hitch, G. J., & Baddeley, A. D. (2014). Memory for serial order across domains: An overview of the literature and directions for future research. *Psychological Bulletin*, *140*(2), 339.

Jacoby, L. L. (1978). On interpreting the effects of repetition: Solving a problem versus remembering a solution. *Journal of Verbal Learning and Verbal Behavior*, *17*(6), 649–667. https://doi.org/10.1016/S0022-5371(78)90393-6

Jafarpour, A., Penny, W., Barnes, G., Knight, R. T., & Duzel, E. (2017). Working Memory Replay Prioritizes Weakly Attended Events. *Eneuro*, *4*(4), ENEURO.0171-17.2017. https://doi.org/10.1523/ENEURO.0171-17.2017

Jarjat, G., Hoareau, V., Plancher, G., Hot, P., Lemaire, B., & Portrat, S. (2018). What makes working memory traces stable over time? *Annals of the New York Academy of Sciences*, *1424*(1), 149–160. https://doi.org/10.1111/nyas.13668

Jeanneret, S., Bartsch, L. M., & Vergauwe, E. (2023). To be or not to be relevant: Comparing short- and long-term consequences across working memory prioritization procedures. *Attention, Perception, & Psychophysics*, *85*(5), 1486–1498. https://doi.org/10.3758/s13414-023-02706-4

Karpicke, J. D., & Zaromb, F. M. (2010). Retrieval mode distinguishes the testing effect from the generation effect. *Journal of Memory and Language*, *62*(3), 227–239. https://doi.org/10.1016/j.jml.2009.11.010

Khader, P. H., Jost, K., Ranganath, C., & Rösler, F. (2010). Theta and alpha oscillations during working-memory maintenance predict successful long-term memory encoding. *Neuroscience Letters*, *468*(3), 339–343.

LaRocque, J. J., Eichenbaum, A. S., Starrett, M. J., Rose, N. S., Emrich, S. M., & Postle, B. R. (2015). The short- and long-term fates of memory items retained outside the focus of attention. *Memory & Cognition*, *43*(3), 453–468. https://doi.org/10.3758/s13421-014-0486-y

Linde-Domingo, J., & Spitzer, B. (2024). Geometry of visuospatial working memory information in miniature gaze patterns. *Nature Human Behaviour*, *8*(2), 336–348. https://doi.org/10.1038/s41562-023-01737-z

Loaiza, V. M., Doherty, C., & Howlett, P. (2021). The long-term consequences of retrieval demands during working memory. *Memory & Cognition*, *49*(1), 112–126. https://doi.org/10.3758/s13421-020-01079-5

Loaiza, V. M., & Halse, S. C. (2019). Where working memory meets long-term memory: The interplay of list length and distractors on memory performance. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(8), 1455–1472. https://doi.org/10.1037/xlm0000652

Loaiza, V. M., & McCabe, D. P. (2012). Temporal–contextual processing in working

memory: Evidence from delayed cued recall and delayed free recall tests. *Memory & Cognition*, *40*(2), 191–203. https://doi.org/10.3758/s13421-011-0148-2

Loaiza, V. M., & McCabe, D. P. (2013). The influence of aging on attentional refreshing and articulatory rehearsal during working memory on later episodic memory performance. *Aging, Neuropsychology, and Cognition*, *20*(4), 471–493. https://doi.org/10.1080/13825585.2012.738289

Loaiza, V. M., Oftinger, A.-L., & Camos, V. (2023). How does Working Memory Promote Traces in Episodic Memory? *Journal of Cognition*, *6*(1), 4. https://doi.org/10.5334/joc.245

Madigan, S. A., & McCabe, L. (1971). Perfect recall and total forgetting: A problem for models of short-term memory. *Journal of Verbal Learning and Verbal Behavior*, *10*(1), 101–106.

Mao Chao, C., Xu, C., Loaiza, V., & Rose, N. S. (2023). Are latent working memory items retrieved from long-term memory? *Quarterly Journal of Experimental Psychology*, 17470218231217723. https://doi.org/10.1177/17470218231217723

McCabe, D. P. (2008). The role of covert retrieval in working memory span tasks: Evidence from delayed recall tests. *Journal of Memory and Language*, *58*(2), 480–494. https://doi.org/10.1016/j.jml.2007.04.004

McElree, B. (2006). Accessing Recent Events. In *Psychology of Learning and Motivation* (Vol. 46, pp. 155–200). Elsevier. https://doi.org/10.1016/S0079-7421(06)46005-9

Miller, G. (1956). Human memory and the storage of information. *IEEE Transactions on Information Theory*, *2*(3), 129–137. https://doi.org/10.1109/TIT.1956.1056815

Oberauer, K. (2002). Access to information in working memory: Exploring the focus of attention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(3), 411–421. https://doi.org/10.1037/0278-7393.28.3.411

Oberauer, K. (2020). Towards a Theory of Working Memory: From Metaphors to Mechanisms. In K. Oberauer, *Working Memory* (pp. 116–149). Oxford University Press. https://doi.org/10.1093/oso/9780198842286.003.0005

Oberauer, K., & Awh, E. (2022). Is There an Activity-silent Working Memory? *Journal of Cognitive Neuroscience*, *34*(12), 2360–2374. https://doi.org/10.1162/jocn_a_01917

Oberauer, K., & Hein, L. (2012). Attention to information in working memory. *Current Directions in Psychological Science*, *21*(3), 164–169.

Postle, B. R., & Oberauer, K. (2022). *PostleAndOberauer_OxfordHandbook*. https://doi.org/10.31234/osf.io/963kf

Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language*, *60*(4), 437–447.

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. *Journal of Neuroscience*, *24*(16), 3917–3925.

Reaves, S., Strunk, J., Phillips, S., Verhaeghen, P., & Duarte, A. (2016). The lasting memory enhancements of retrospective attention. *Brain Research*, *1642*, 226–237.

Rerko, L., & Oberauer, K. (2013). Focused, unfocused, and defocused information in working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(4), 1075–1096. https://doi.org/10.1037/a0031172

Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, *15*(1), 20–27. https://doi.org/10.1016/j.tics.2010.09.003

Roediger, H. L., & Karpicke, J. D. (2006). Test-Enhanced Learning: Taking Memory Tests Improves Long-Term Retention. *Psychological Science*, *17*(3), 249–255.

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

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin*, *140*(6), 1432–1463. https://doi.org/10.1037/a0037559

Schapiro, A. C., McDevitt, E. A., Rogers, T. T., Mednick, S. C., & Norman, K. A. (2018). Human hippocampal replay during rest prioritizes weakly learned information and predicts memory performance. *Nature Communications*, *9*(1), 3920. https://doi.org/10.1038/s41467-018-06213-1

Sederberg, P. B., Gauthier, L. V., Terushkin, V., Miller, J. F., Barnathan, J. A., & Kahana, M. J. (2006). Oscillatory correlates of the primacy effect in episodic memory. *NeuroImage*, *32*(3), 1422–1431. https://doi.org/10.1016/j.neuroimage.2006.04.223

Serra, M., & Nairne, J. S. (1993). Design controversies and the generation effect: Support for an item-order hypothesis. *Memory & Cognition*, *21*(1), 34–40. https://doi.org/10.3758/BF03211162

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory*, *4*(6), 592–604. https://doi.org/10.1037/0278-7393.4.6.592

Souza, A. S., & Oberauer, K. (2016). In search of the focus of attention in working memory: 13 years of the retro-cue effect. *Attention, Perception, & Psychophysics*, *78*(7), 1839–1860. https://doi.org/10.3758/s13414-016-1108-5

Souza, A. S., & Oberauer, K. (2017). Time to process information in working memory

improves episodic memory. *Journal of Memory and Language*, *96*, 155–167. https://doi.org/10.1016/j.jml.2017.07.002

Souza, A. S., Rerko, L., & Oberauer, K. (2016). Getting more from visual working memory: Retro-cues enhance retrieval and protect from visual interference. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(6), 890–910. https://doi.org/10.1037/xhp0000192

Stokes, M. G. (2015). 'Activity-silent' working memory in prefrontal cortex: A dynamic coding framework. *Trends in Cognitive Sciences*, *19*(7), 394–405. https://doi.org/10.1016/j.tics.2015.05.004

Strunk, J., Morgan, L., Reaves, S., Verhaeghen, P., & Duarte, A. (2019). Retrospective Attention in Short-Term Memory Has a Lasting Effect on Long-Term Memory Across Age. *The Journals of Gerontology: Series B*, *74*(8), 1317–1325. https://doi.org/10.1093/geronb/gby045

Sundby, C. S., Woodman, G. F., & Fukuda, K. (2019). Electrophysiological and behavioral evidence for attentional up-regulation, but not down-regulation, when encoding pictures into long-term memory. *Memory & Cognition*, *47*, 351–364.

Taylor, R., & Bays, P. M. (2018). Efficient Coding in Visual Working Memory Accounts for Stimulus-Specific Variations in Recall. *The Journal of Neuroscience*, *38*(32), 7132–7142. https://doi.org/10.1523/JNEUROSCI.1018-18.2018

Tozios, C. J. I., & Fukuda, K. (2024). Decomposing the multiple encoding benefit in visual long-term memory: Primary contributions by the number of encoding opportunities. *Memory & Cognition*. https://doi.org/10.3758/s13421-024-01602-y

Van Ede, F., & Nobre, A. C. (2023). Turning Attention Inside Out: How Working Memory Serves Behavior. *Annual Review of Psychology*, *74*(1), 137–165. https://doi.org/10.1146/annurev-psych-021422-041757

Van Loon, A. M., Olmos-Solis, K., Fahrenfort, J. J., & Olivers, C. N. (2018). Current and future goals are represented in opposite patterns in object-selective cortex. *eLife*, *7*, e38677. https://doi.org/10.7554/eLife.38677

Wan, Q. (2022). *Priority-based transformations of stimulus representation in visual working memory*. https://doi.org/10.17605/OSF.IO/SGQVN

Wang, S., & Van Ede, F. (2024). *Tracking how attentional focusing in working memory benefits long-term memory*. https://doi.org/10.1101/2024.03.25.586271

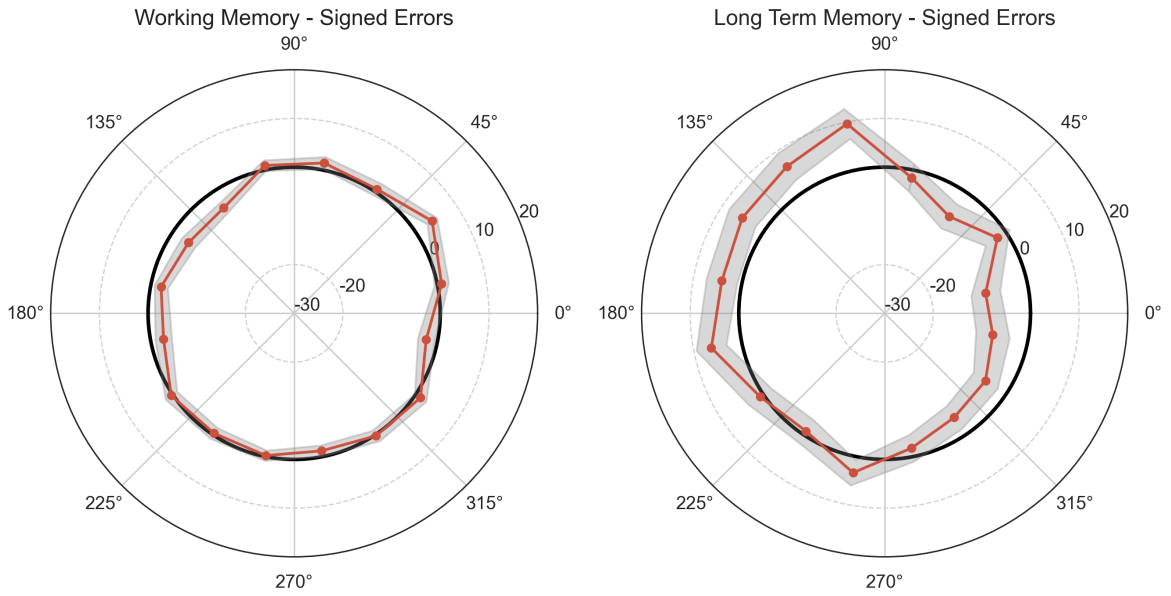
Wolff, M. J., Jochim, J., Akyürek, E. G., & Stokes, M. G. (2017). Dynamic hidden states underlying working-memory-guided behavior. *Nature Neuroscience*, *20*(6), 864–871. https://doi.org/10.1038/nn.4546

Xie, K. Y., & Reuter-Lorenz, P. A. (2024). The impact of working memory testing on long-term associative memory. *Memory & Cognition*. https://doi.org/10.3758/s13421-024-

01568-x

Yu, Q., Teng, C., & Postle, B. R. (2020). Different states of priority recruit different neural representations in visual working memory. *PLOS Biology*, *18*(6), e3000769. https://doi.org/10.1371/journal.pbio.3000769

# Supplementary Figure 1




**Supplementary Figure 1. Exploring cardinal repulsion bias in WM and LTM reports**. An alternative explanation for our finding that participants' LTM reports were biased towards their previous WM reports (**Fig. 2c**) could be that both reports (WM and LTM) exhibited canonical "cardinal" bias. (Repulsive) cardinal bias (Bae, 2021; Taylor & Bays, 2018) refers to the finding that behavioral reports of stimulus orientation (for example, of Gabor gratings) can be biased away from the cardinal (vertical and horizontal) axes. In our present experiments, such a phenomenon can only be examined for a subset of stimulus objects (n = 63) which had a clear real-world upright position (see examples in **Fig. 1**; other objects, such as scissors were excluded). The polar plots show the mean signed error in degrees (cw < 0 < ccw) for each sample orientation in Exp.1, where 90° refers to the objects' upright orientation. Cardinal repulsion bias would be evident if the response errors were consistently positive (ccw) to the ccw side, and negative (cw) to the cw side of the cardinal axes (0°, 90°, 180°, and 360°; cf. Linde-Domingo & Spitzer, 2024). However, there was no clear indication of such systematic patterns in the present data, and the patterns in the WM (*left*) and LTM (*right*) tests were dissimilar (if anything, the mean signed errors correlated even negatively, $r_{\text{Spearman}}$ = -0.41). Our findings in **Fig. 2c** can thus not be easily explained in terms of canonical cardinal bias.