Active Removal of Information from Working Memory

By

Jiangang Shan

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Psychology)

at the

UNIVERSITY OF WISCONSIN-MADISON

2025

Date of final oral examination: 5/08/2025

The dissertation is approved by the following members of the Final Oral Committee:

Bradley R. Postle, Professor, Psychology & Psychiatry
Yuri B. Saalmann, Professor, Psychology
Timothy T. Rogers, Professor, Psychology
C. Shawn Green, Professor, Psychology
Larissa Albantakis, Assistant Professor, Psychiatry

Acknowledgements

Completing this PhD has been one of the most challenging and rewarding experiences of my life, and I owe a great deal of gratitude to the many people who supported me throughout this journey.

First and foremost, I would like to express my deepest thanks to Brad, my PhD advisor, for his unwavering support, guidance, and patience over the years. Your mentorship has been instrumental in shaping not only my research but also the way I approach problems, face challenges, and grow as a scientist. I am incredibly grateful for the freedom you gave me to explore my ideas, while always being there to offer clarity and direction when I needed it most.

I also want to sincerely thank the other members of my defense committee — Yuri, Shawn, Tim and Larissa — for their valuable time, feedback, and perspectives, which have helped refine my thinking and sharpen my scientific perspective. Your thoughtful questions and insights pushed me to think more deeply about my work and helped improve the quality of my research.

To my labmates and friends in the Postle lab — thank you for creating such a positive and collaborative environment. Whether it was troubleshooting experiments, brainstorming ideas, or just sharing thoughts about the daily life, your camaraderie has made this journey not only manageable but truly enjoyable. I couldn't have asked for a better group of people to work alongside.

I would also like to thank my collaborator, Nicholas Myers, for the stimulating discussions and insightful perspectives we've shared during our work together. Although our collaborative project is not included in this thesis, it has significantly enriched my broader research experience and helped me grow as a scholar.

Finally, and most importantly, I want to thank my parents and friends in China. Your unconditional love, encouragement, and belief in me have been my foundation throughout this entire process. I am forever grateful for the comfort and laughter you have brought to me.

Thank you all — this accomplishment is as much yours as it is mine.

Table of Contents

Abstract

The ability to control the content maintained in working memory (WM) and to remove it out of WM when it becomes useless is critical for the functioning of WM because of its limited capacity. Compared to its importance, however, how information is removed from WM is an understudied question. While the default strategy that people use under the majority of scenarios is simply withdrawing one's attention from the to-be-removed item, such passive removal strategy cannot erase the outdated information completely and leads to the interference between old and new information. In this thesis, I investigated whether the removal of information from WM can be done in a more active manner and the neural mechanism behind the active removal.

In Chapter 1, I briefly reviewed the literatures in this field and introduced a novel behavioral task that we designed to evoke active vs. passive removal. I discussed about the behavioral findings we got with this task, and proposed the hijacked adaptation model, which we believe is the neural mechanism behind active removal. The hijacked adaption model posits that the active removal is achieved by a top-down control signal modulating the perceptual circuits to reduce the gain of sensory channels tuned to the to-be-remove item, in a way reminiscent of the sensory adaption.

In Chapters 2 and 3, predictions of the hijacked adaptation model were tested in an electroencephalography (EEG) study and a functional magnetic resonance imaging (fMRI) study. Consistent with our model, we found neural evidence for the top-down control recruited for active removal, manifested as stronger central-midline ERP and stronger anterior-to-posterior traveling waves. By investigating the neural responses to a visual "ping", we found neural evidence for reduced excitability of perceptual circuits when active removal was conducted, which supports the idea that active removal is accompanied by a down-modulation of gain. Using multivariate analyses on fMRI signal, we further found, with active removal but not passive removal, the location information of the to-be-removed item was actively suppressed. This active suppression had the effect of erasing the synaptic traces of the to-be-removed information.

Chapter 1

Introduction

One way that the cognitive system compensates for the capacity limitations of working memory (WM) is via the flexible updating of the contents of WM and the control of priority among those contents. For example, a memory item that is needed to guide current behavior is prioritized and represented in a different state relative to items that are not immediately relevant, but might be needed in the future (Lewis-Peacock and Postle, 2012; Rose et al., 2016; Yu et al., 2020). Additionally, when a memory item loses relevance, it can be removed from WM (Fulvio and Postle, 2020). This thesis is focused on assessing the mechanisms whereby an item in WM can be strategically removed.

The extensive history of the study of proactive interference in memory, including WM, documents the fact that subjects do not routinely employ removal of information from WM in an active way. For example, in a WM task using a recognition procedure in which subjects memorize a set of letters and then, after a delay period, judge whether the probe presented at the end of the trial was in the memory set, "recent-negative" probes not in the memory set of the current trial but that were in the memory set of the previous trial are rejected with lower accuracy and longer RTs than "nonrecent-negative" probes (Monsell, 1978). Whereas infrequent recent-negative probes recruit reactive control (Burgess and Braver, 2010; Braver, 2012) that is supported by phasic activity in inferior prefrontal cortex (PFC; (D'Esposito et al., 1999; Feredoes et al., 2006)), high levels of recent-negative probes recruit dorsal PFC-supported proactive control (Burgess and Braver, 2010; Braver, 2012). Importantly, however, these examples of control are not necessarily related to the removal of information: proactive control could prompt stronger encoding of trial-specific context, whereas reactive control would influence the recognition decision (resolution of

the conflict between the familiarity of the recent negative item versus the recollection of the memory set (Feredoes and Postle, 2010)).

Visual WM tasks using a recall procedure (a.k.a. delayed estimation), also reveal evidence of less-than-complete loss of information from trial to trial. On these tasks, subjects first encode the critical feature of the sample stimulus (e.g., the orientation of a Gabor patch) and then, after a delay period, they report that feature (e.g., recreate the remembered orientation with a response dial). On this type of WM task, the sample shown on each trial is typically chosen at random, and so on these tasks it is also assumed that the representation of the sample is "dropped" from WM once the response is made. Nonetheless, a common observation on this type of task is that the recall report on the current trial is attracted toward the value of the memory item from the previous trial (e.g., (Fischer and Whitney, 2014; Bliss et al., 2017)). (E.g., if the orientation of the to-be-recalled sample on the current trial is 90°, and the orientation of the sample on the previous trial had been 120°, there is a tendency for recall to be biased toward 120° (i.e., a recall value of 92° is more likely than a recall value of 88°).) This attractive influence of the previous trial is called serial dependence, and it indicates that stimulus information from a trial is typically not totally removed at the end of the trial. Indeed, two recent studies have found direct evidence for this. With electroencephalography (EEG) data collected during delayed recall for orientation, Bae and Luck (2019) were able to decode the orientation of the previous trial's sample after the onset of the current trial's sample. For delayed recall of location, reactivation of an activity-silent representation of the previous trial's sample location was observed near the end of the intertrial-interval (ITI) in PFC in nonhuman primates, as was a similar reactivation in whole-scalp EEG in humans (Barbosa et al.,

2020). These effects were modelled as the consequence of the filtering of a "nonspecific anticipatory signal" by a residual trace of the representation of the sample "imprinted in neuronal synapses as a latent activity-silent trace". The dynamics of the bump attractor model of Barbosa et al. (2020) explicitly capture our intuition of what it means to assume a passive loss of information from WM (a.k.a., decay-based forgetting): it is modeled by simply removing activation from the elements that represent the information that had been held during that trial, and the bump of elevated activity recedes to baseline. We note that the same dynamics have also been assumed for the within-trial loss of information from WM, such as during a task when new information leads to the "reallocation" of resources away from a newly irrelevant item (Chatham and Badre, 2013).

In contrast to the idea of a "passive loss" of information from WM, as reviewed up to this point, there are also theoretical reasons to postulate, and empirical evidence for, an active removal mechanism. These derive from tasks in which more than one item is held in WM simultaneously, and so competition is occurring between stimulus representations being held simultaneously (rather than between items from different trials). In visual WM, it is well-established that performance declines as a monotonic function of load (e.g., (Ma et al., 2014)). Performance improves, however, when a retrodictive cue ("retrocue") that appears during the delay indicates which item from the memory set is the one that will be tested at the end of the trial (e.g., (Lepsien and Nobre, 2006)). So too does the strength of the neural representation of the retrocued item (Sprague et al., 2016). There is also behavioral evidence for it from a slightly more complicated task – dual serial retrocuing (DSR) task. In DSR, there are two memory probes, each one preceded by a retrocue. After the offset of the trial's two sample stimuli, the first retrocue indicates which item will be tested by the first

probe. However, the uncued item can't yet be dropped, because a second retrocue will then indicate which of the two will be tested by the second memory probe, and both memory items are equally likely candidates for this second retrocue. Thus, the DSR requires subjects to control the priority status of items held in WM, with the first retrocue designating which item is the prioritized memory item (PMI) and which the unprioritized memory item (UMI). Although functional magnetic resonance imaging (fMRI) and EEG studies indicate that evidence for an active representation of the UMI can decline to baseline levels (Lewis-Peacock and Postle, 2012; LaRocque et al., 2013), or transform into a different representational format (Wan et al., 2020; Yu et al., 2020), a pulse of TMS during the delay following the first retrocue has two effects: it evokes a brief reactivation of the representation of the UMI; and it boost the rate of false alarms when the UMI is used as a recognition probe (Rose et al., 2016). The evidence for an active removal process comes from the absence of comparable reactivation effects when TMS is delivered following the second retrocue on DSR trials (Rose et al., 2016), or following the sole retrocue on single-retrocue trials (Fulvio and Postle, 2020).

What might be the mechanism that implements the active removal of information from WM? Different theoretical frameworks provide different answers to this question. In a bump-attractor model, removal would be accomplished by a nonspecific burst of activation broad and strong enough to swamp the stimulus-representing bump and to saturate the residual synaptic trace. In the interference model (Oberauer and Lin, 2017), active removal of an item from WM is accomplished by breaking the association between an item and its context (Lewis-Peacock et al., 2018). A third, intuitive, possibility for active removal is the suppression of the neural representation of the to-be-removed item.

To study whether people can remove information from working memory in a more active way, and to investigate the mechanism behind such active removal, in a previous study (Shan and Postle, 2022), we designed a novel behavioral task called the ABC-retrocuing task. The ABC-retrocuing task aims to evoke active and passive removal of outdated information from WM in different conditions. I'll describe the general procedure of the task below (Fig. 1.1):

Subjects were first presented with two memory items (oriented gratings, A and B) simultaneously at two locations on screen. After their offset, a retrocue (a circle presented at either the location of A or B) indicated which item might be tested at the end of the trial by its location on screen, and subjects were explicitly informed that the uncued item (by convention, item B) would not be tested (thereby making it an "irrelevant memory item", IMI). Next, a third item (C) was shown, followed by a delay, and finally a response dial, the (unpredictable) location of which indicated the item to be recalled (i.e., A or C, with equal chance). The key manipulation, which was blocked, was whether item C was shown at the location that had been occupied by IMI (the "overlap" condition), or at a location that had been occupied by neither item A nor B (the "no-overlap" condition). Subjects were explicitly informed about this arrangement before the start of each block.

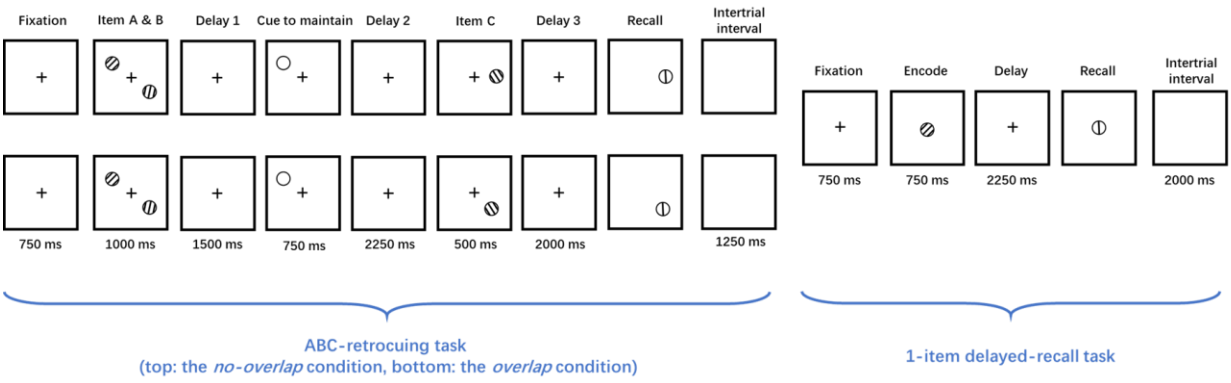

Figure 1.1. The behavioral task used in (Shan and Postle, 2022). Each of the ABC-retrocuing-task trial (which was designed to evoked active vs. passive removal of information from WM) was followed by a 1-item delayed-recall trial (which was used to test the fate of the removed information from the previous trial).

The logic was that, because the overlap condition featured higher cue conflict between the IMI and item C, subjects would be motivated to actively remove the IMI from WM, as soon as it was so designated by the retrocue. For the no-overlap condition, in contrast, because item C was always presented at a new location, the potential interference from IMI would be weaker, and so subjects might just use the default strategy of passive removal of item B. To assess the removal operation that was employed on these trials, each trial of the ABC-retrocuing task was followed by a trial of 1-item delayed recall, and the serial dependence exerted by the IMI from the preceding ABC-retrocuing trial on the report in the following 1-item trial was measured. The expectation was in the no-overlap condition, an attractive bias indicating the incomplete removal of the IMI should be found, and in overlap condition there should be no serial bias, as a result of the complete removal of the IMI achieved by the active removal.

Replicating the pattern from many previous studies (Fischer and Whitney, 2014; Bliss et al., 2017; Samaha et al., 2019) and consistent with a passive removal account, in the no-overlap condition we found an attractive bias induced by the IMI from the ABC-retrocuing trial on the report in the subsequent 1-item delayed-recall trial. In the overlap condition, in contrast, the IMI was found to exert a repulsive bias on the 1-item recall (Shan and Postle, 2022). This reversal of the sign of the serial dependence effect indicates that the IMI was processed differently during the two trial types (Fig. 1.2), and suggests that active removal

may entail the transformation of IMI into a "reversed" version of itself, one that also reverses its influence on subsequent WM processing.

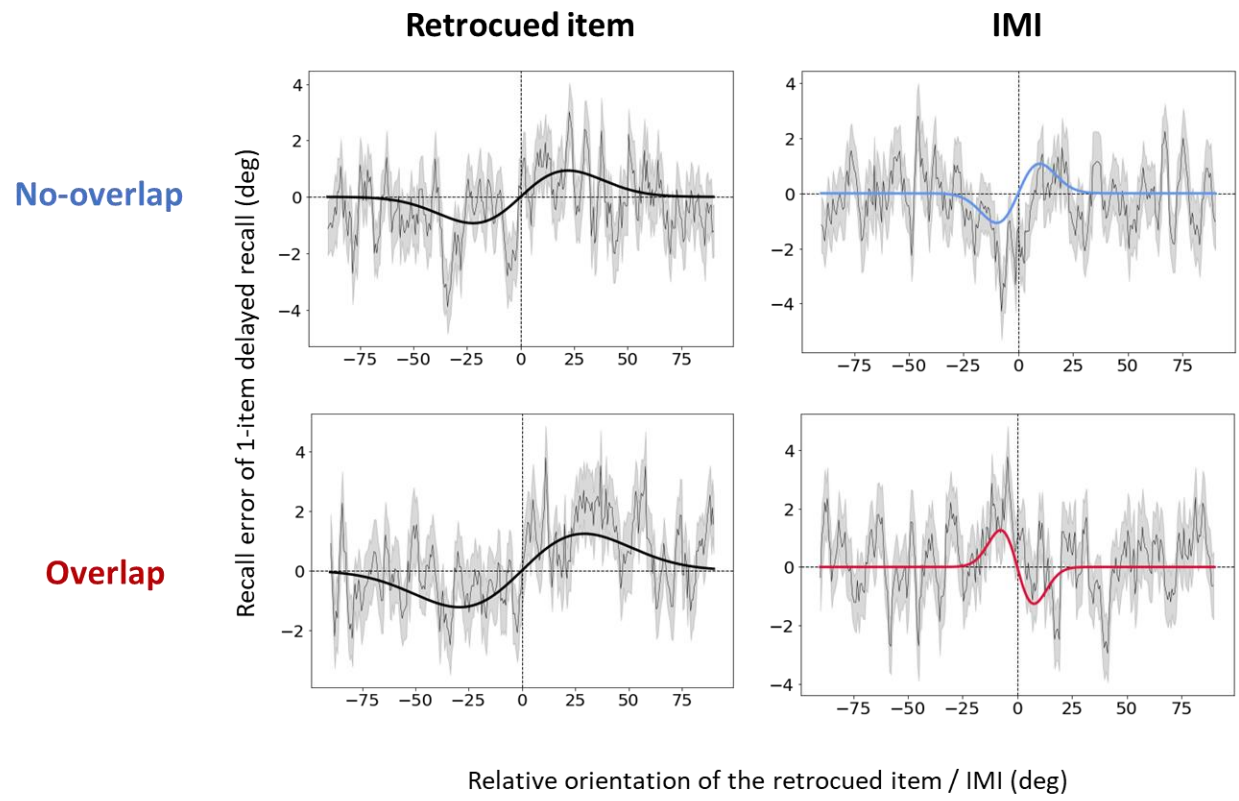

Figure 1.2. Serial dependence results in (Shan & Postle, 2022). The (presumably passively removed) IMI in the no-overlap condition caused an attractive bias on the next-trial report, and the (presumably actively removed) IMI in the overlap condition caused a repulsive serial bias. In comparison, the retrocued item caused the classical attractive bias on the next trial in both conditions.

A second study using a dual WM+discrimination task has observed a similar phenomenon: On low-conflict trials when the memorandum and discriminandum match one another, the discriminandum exerted an attractive bias on recall on the subsequent trial; however, on high-conflict trials when it would be advantageous to actively remove any trace of the

discriminandum, it exerted a repulsive bias on recall on the subsequent trial (Teng et al., 2022).

Understanding the cause of the repulsive serial bias observed during high-conflict conditions in WM may contribute to understanding the mechanism(s) recruited for the active removal of information from WM. A first step is to consider that serial dependence is believed to arise from a two-stage process with opposite effects (Fritsche et al., 2017; Fritsche et al., 2020): during the encoding of the sample item for the current trial, the encoded representation is biased repulsively from the previous trial's sample, due to the sensory adaptation; later in the trial, however, the memory-guided decision is biased attractively toward the previous trial's sample, in a manner consistent with Bayesian-inference. This model, therefore, raises the possibility that the repulsive effects associated with active removal (Shan and Postle, 2022; Teng et al., 2022) may implicate an adaptation-like mechanism. This, in turn, leads us to consider the results from afunctional magnetic resonance imaging (fMRI) study of WM with a retrocuing task (Lorenc et al., 2020). In findings conceptually similar to Shan and Postle (2022) and Teng et al. (2022), the authors found that the multivariate representation of an item was reversed after it acquired the status of IMI. Follow-up modeling analysis showed this result was best explained by a mechanism in which the gain of perceptual circuits tuned to the this item was down-modulated, a mechanism reminiscent of the one that underlies sensory adaptation (Jin et al., 2005). Taken these together, these results lead us to hypothesize that the active removal of information in WM may be accomplished by a top-down modulation of the perceptual circuits, such that the gain of sensory channels tuned to the to-be-removed item is reduced. Because this hypothesized mechanism implicates circuit-level changes that are

similar to those responsible for sensory adaptation, but differs from true adaptation in that it is putatively triggered by a top-down control signal, we refer to it as the "hijacked adaption" model of active removal from WM.

In the work presented in Chapters 2 and 3, we tested the behavioral and neural predictions of the hijacked adaptation using the ABC-retrocuing task with electroencephalogram (EEG) recording (Chapter 2) and functional magnetic resonance imaging (fMRI, Chapter 3). Specifically, in Chapter 2 we used ERP and traveling-wave analyses to reveal the EEG correlates of the triggering and the consequence of active removal. These, along with the behavioral evidence obtained from the target confusability competition (TCC) model, supported our hijacked adaptation model. In Chapter 3, we applied multivariate analyses on blood-oxygenation-level–dependent (BOLD) signal patterns and showed that the active removal was achieved by an active suppression of the location of the to-be-removed information, such that it became less likely to be reactivated by a non-specific "ping" stimulation. The work presented in this thesis significantly adds to our understanding of how people flexibly handle the outdated information in WM in the face of different levels of interference and shed light on the neural mechanisms behind this function.

Chapter 2

EEG correlates of active removal from working memory

Jiangang Shan, Bradley R. Postle

Submitted to *Journal of Neuroscience*

Abstract

The removal of no-longer-relevant information from visual working memory (WM) is important for the functioning of WM, given its severe capacity limitation. Previously, with an "ABC-retrocuing" WM task, we have shown that removing information can be accomplished in different ways: by simply withdrawing attention from the newly irrelevant memory item (IMI; i.e., via "passive removal"); or by or "actively" removing the IMI from WM (Shan and Postle, 2022). Here, to investigate the neural mechanisms behind active removal, we recorded electroencephalogram (EEG) signals from human subjects (both sexes) performing the ABC-retrocuing task. Specifically, we tested the hijacked adaptation model, which posits that active removal is accomplished by a top-down-triggered down-modulation of the gain of perceptual circuits, such that sensory channels tuned to the to-be-removed information become less sensitive. Behaviorally, analyses revealed that, relative to passive removal, active removal produced a decline in the familiarity landscape centered on the IMI. Neurally, we focused on two epochs of the task, corresponding to the triggering, and to the consequence, of active removal. With regard to triggering, we observed a stronger anterior-to-posterior traveling wave for active versus passive removal. With regard to the consequence(s) of removal, the response to a task-irrelevant "ping" was reduced for active removal, as assessed with ERP, suggesting that active removal led to decreased excitability in perceptual circuits centered on the IMI.

Key Words: ***working memory; active removal; target confusability competition (TCC) model; EEG; traveling wave***

Significance Statement

The removal of no-longer-relevant information from working memory is critical for the flexible control of behavior. However, to our knowledge, the only explicit accounts of this operation describe the simple withdrawal of attention from that information (i.e., "passive removal"). Here, with measurements of behavior and electroencephalography (EEG), we provide evidence for a specific mechanism for the active removal of information from WM–hijacked adaptation–via the top-down triggering of an adaptation-like down-regulation of gain of the perceptual circuits tuned to the to-be-removed information. These results may have implications for disorders of mental health, including rumination, intrusion of negative thoughts, and hallucination.

Introduction

A hallmark property of working memory (WM) is that it is rapidly updatable, in that new information can be added "on the fly" as required by moment-to-moment changes in the environment and/or in behavioral goals. And because of the severe capacity limitation of WM, this updating process is commonly assumed to also entail the removal of no-longer-relevant information. Despite this, however, empirical evidence indicates that removal is often not absolute. For example, the recall (e.g., of the orientation of a grating) of the current trial's item is often biased toward the sample from the previous trial (i.e., an attractive serial bias, e.g., (Fischer and Whitney, 2014; Samaha et al., 2019)). On the neural level, there is also considerable evidence for the incomplete removal of the no-longer-relevant information, with information from the previous trial persisting into the current

trial (Bae and Luck, 2019; Barbosa et al., 2020; Zhang and Lewis-Peacock, 2024). All these findings suggest that, under many circumstances, updating may involve a default strategy of "passive removal" (i.e., the simple withdrawal of attention from no-longer-relevant information), such that some residual trace of this information remains in the cognitive system, and can interfere with subsequent behaviors.

There is, however, evidence that updating can involve the active removal of information from WM. In a behavioral study using an "ABC-retrocuing" WM task, whereas passive removal resulted in an attractive serial bias (replicating, e.g., (Fischer and Whitney, 2014; Bliss et al., 2017; Samaha et al., 2019) active removal had the opposite effect: a repulsive serial bias. Similarly, in a dual WM+discrimination task, the condition encouraging passive removal produced attractive serial dependence, but the condition encouraging active removal produced repulsive serial dependence (Teng et al., 2022). These reversals of the sign of the serial dependence effect suggest that active removal may entail the transformation of the IMI into a "reversed" version of itself, one that also reverses its influence on subsequent WM processing. This idea gains support from a functional magnetic resonance imaging (fMRI) study of WM with a retrocue (Lorenc et al., 2020), in which the multivariate representation of a no-longer-relevant item was reversed after it acquired this status. Computational modeling indicated that this result was best explained by a mechanism in which the gain of perceptual circuits tuned to this item was down-modulated (Lorenc et al., 2020), a mechanism reminiscent of the one underlying sensory adaptation (Jin et al., 2005).

The repulsive bias of serial dependence is believed to arise during the encoding of the current trial's item, due to sensory adaptation (Fritsche et al., 2017; Pascucci et al., 2019; Fritsche et al., 2020; Sheehan and Serences, 2022). Together, these findings have led us to speculate that the repulsive effects associated with active removal may implicate an adaptation-like mechanism. We refer to this mechanism as the "hijacked adaption" model of active removal from WM.

The current study was carried out to test behavioral and neural predictions of the "hijacked adaptation" model, whereby the active removal of information in WM is accomplished by a top-down modulation of the perceptual channels tuned to that information. To do this we recorded EEG activity from subjects performing an ABC-retrocuing WM task and focused on two epochs in the trial to assess the triggering of the active removal process, and, later in the trial, the consequence(s) of its deployment.

Materials and Methods

Subjects

Twenty-seven subjects from the University of Wisconsin–Madison community participated in the study and were compensated monetarily. All subjects provided informed consent approved by the University of Wisconsin–Madison Health Sciences Institutional Review Board. One was removed from analyses for failing to follow task instructions; another was removed due to poor performance (mean absolute recall error more than 2 SD higher than

the group average). Thus, data from 25 subjects (19 females, 6 males, mean age = 23.8, range = 19 - 30) were included in all analyses.

Stimuli

Subjects were seated in a dimly lit room at a viewing distance of 50 cm from the monitor. A chin rest was used to help keep the head stable during the task. Memory-sample stimuli were oriented gratings (radius = 3°, spatial frequency = 1circle/°, contrast = 0.5, random phase) presented at six possible locations (30°, 90°, 150°, 210°, 270°, 330°) on an imaginary circle with radius of 7°. Stimuli were white (RGB = 255,255,255) appearing on a gray (128,128,128) background. Ping stimuli were white bullseyes presented at all six locations (radius = 3°, spatial frequency = 1circle/°, contrast = 1, random phase). The ping was intended to provide strong but orientation-neutral stimulation of circuits involved in visual perception. Retrocues were white circles with the same radius as the samples, and the probes were black response dials (unfilled black circles with a black line corresponding to the diameter of the circle) with the same radius and random starting orientation. A white fixation dot was presented at the center of the screen throughout each block of trials.

Experimental design

Subjects completed 6 blocks of ABC-retrocuing task (Fig. 2.1) in one session, 3 blocks of "no-overlap" trials followed by 3 blocks of "overlap" trials. Each block had 120 trials and lasted for ~27 min. Subjects were asked to take a self-paced pause every 40 trials (i.e., ~ 9 min) and were required to keep their head stable during the pause. Each trial started with central fixation (750 ms), after which two sample gratings (items *A* and *B*) were presented at different locations (1000 ms). Subjects were to remember these two samples across an

initial delay (Delay 1; 1500 ms), after which a retrocue appeared at the location occupied by one of the two gratings (750 ms), thereby indicating that this item might be tested at the end of the trial. (Subjects were explicitly informed that the uncued item would not be tested, thereby making it an IMI.) After a second delay (Delay 2.1; 2000 ms), the ping was presented for 250 ms, followed by Delay 2.2 (1000 ms), then another sample grating (item *C*, 500 ms), Delay 3 (1000-ms), and finally a probe that appeared at the location that had been occupied by the retrocued item or by item C (with equal chance). The probe was displayed for 3000 sec, during which the subject was to recall the orientation of the probed item by adjusting the orientation of the response dial with a computer mouse, followed by feedback displaying the number of degrees of error (1000 ms). The inter-trial interval varied randomly from 500 to 700 ms.

The orientations of items *A*, *B*, and *C* were randomly drawn, with replacement, from a pool of 6 base orientations (20°, 50°, 80°, 110°, 140°, 170°), with a random jitter of -3 to 3° added. The locations of *A* and *B* were randomly selected from two of the six possible locations. Importantly, for the first three blocks (the no-overlap condition), item *C* was randomly presented at one of the four locations that had not been occupied by *A* or *B*, a fact that was specified during pre-experiment instructions. For the final three blocks (the overlap condition item), *C* was always presented at the same location as had been the IMI. The logic of this manipulation was that, because the overlap condition featured higher cue conflict between the IMI and item *C*, subjects would be motivated to actively remove the IMI from WM as soon as it was so designated by the retrocue. For the no-overlap condition, in contrast, because item *C* was always presented at a new location, the potential interference from IMI would be weaker, and so subjects might just use the default strategy of passive

removal of IMI. Note that there was no explicit instruction about using active or passive removal. The location-related difference between the overlap and the no-overlap condition was only explained to subjects after they had completed the first three blocks of the experiment, a procedural detail intended to reduce the likelihood that subjects would engage an active removal strategy on no-overlap trials.

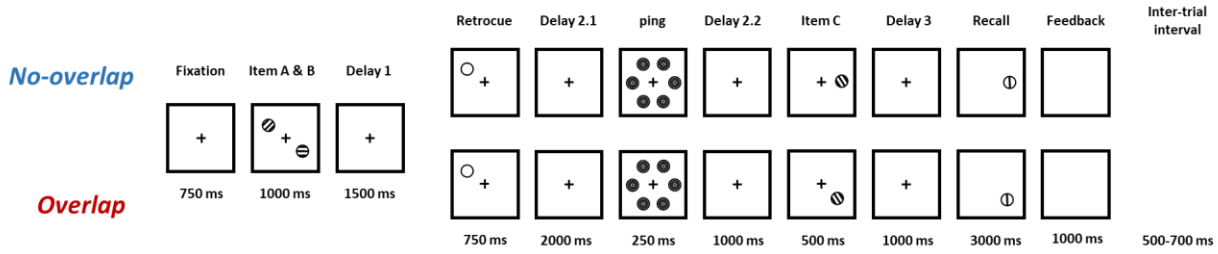

Figure 2.1. The experimental procedure.

Behavioral analyses

Mean absolute recall error was calculated for each subject and each condition. A two-tailed paired t-test was conducted to test whether performance differed across conditions.

We applied a variant of the target confusability competition (TCC) model to track the fate of the IMI (Schurgin et al., 2020). With TCC models, the subject's report on a trial is assumed to be based on an aggregated familiarity landscape that reflects the influence of all of the items currently in WM (see Fig. 2.2A for an example illustration). TCC estimates the memory strength of an item with a single parameter, d', which represents the magnitude of the signal corresponding to that item. Whereas the original formulation of the TCC model was restricted to the items currently in WM, and d' was bounded to be non-negative, for the present study we extended this model to include an estimate of d' for the putatively

removed item (c.f., (Zhang and Lewis-Peacock, 2023)), and so we needed to modify it in order to allow d' to take on negative values. In particular, for the ABC-retrocuing task, although the passive removal of the IMI was predicted to leave a small but positive d' (corresponding to an incompletely removed memory trace), the active removal of the IMI was predicted to produce a negative d'. This is because, according to the hijacked adaptation model, the active removal of the IMI is accomplished by down-modulating the gain of sensory channels tuned to its orientation. Thus, the subsequent encoding of item *C* should be influenced such that the final familiarity landscape would be reduced for orientations close to the IMI (see Fig. 2.2B). A negative IMI d' in the overlap condition would be consistent with this model.

The TCC model incorporates the psychological similarity between stimulus items, and posits that a graded familiarity signal is generated by each item in WM (Schurgin et al., 2020). For orientation, a 0° memorandum would boost familiarity for all possible orientations ranging from -90° to 90°, with the magnitude of this boost for any one value determined by its similarity to 0° (e.g., Fig. 2.2A). To apply the TCC model, we used the psychometric similarity function for orientation estimated in a previous study (Cai et al., 2022). To assess whether active removal of the IMI influenced the processing of item *C*, we focused on the 50% of trials in which subjects were probed to recall item *C*. The familiarity landscape for each trial was modeled to be the sum of the familiarity of the probed item (i.e., item *C*, estimated with probed d') and the signal corresponding to the IMI (estimated with IMI d'). Both probed d' and IMI d' were allowed to take a negative value. To calculate d' values, we combined the data from all subjects into a "super subject" and conducted all TCC analyses on this super subject. The TCC model was fitted to data from the overlap condition

and the no-overlap condition separately, using Markov Chain Monte Carlo (MCMC function in MemToolbox; (Suchow et al., 2013)). 15,000 post-convergence samples were taken to calculate the 95% confidence interval (95% CI) of the parameter estimates. To investigate the difference between conditions, the differences in the estimated parameters from the 15,000 samples in each condition were calculated and the 95% CI was generated from them. *p* values were also calculated from each pool of samples by calculating the proportion of samples higher or lower than 0 and multiplying the result by 2 to make it two-tailed.

Although we do not have an explicit prediction about the results of fitting TCC on report-A trials (because item A was encoded before the IMI was removed, and, indeed, before it was known which item was the IMI), we also conducted this analysis for completeness.

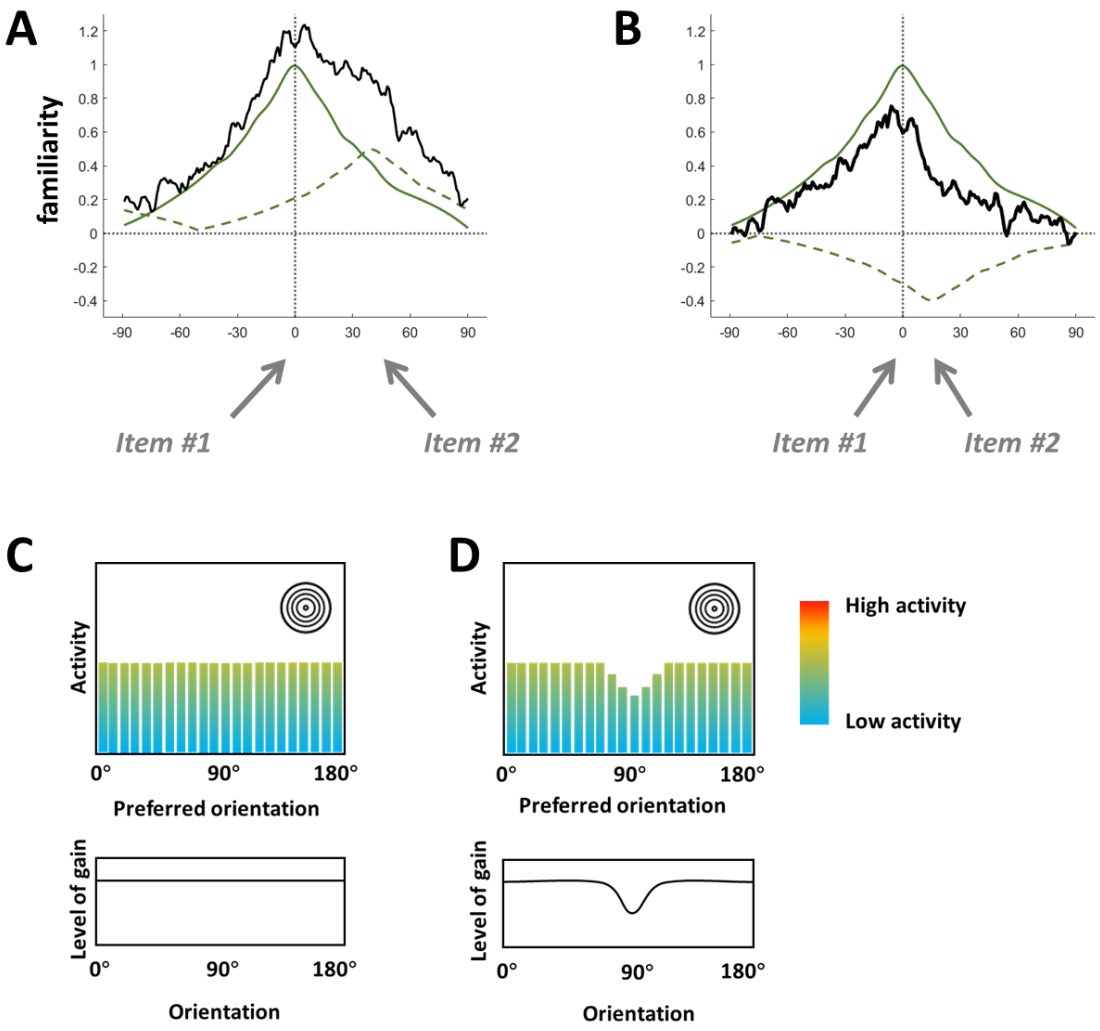

Figure 2.2. Illustrations of predictions of the hijacked adaptation model. A. Hypothetical behavioral data from a single trial, fit to TCC model, in which the familiarity signals for item #1 (solid green line) and for item #2 (dashed green line, are strong and weak, respectively. The solid black curve is the resultant familiarity landscape that is accessible to the cognitive system, generated by taking the sum of two green curves and adding some noise (note the "shoulder" corresponding to the weaker item #2). B. Hypothetical behavioral data from a trial in which item #2 has been actively removed, the down-modulation of its gain resulting in negative familiarity centered on its orientation (note the consequent reduction in the corresponding portion of the familiarity landscape). C. On trials in which neither item has been actively removed, the gain of all orientation-tuned sensory channels is

comparable (lower plot), and so a nonspecific visual "ping" has the effect of activating each by the same amount (upper plot; each bar illustrates the level of activity of an orientation-tuned sensory channel). D. On trials in which an item (90°) has been actively removed, the gain of sensory channels centered on that item's orientation has been reduced. Consequently, the response of these sensory channels to a nonspecific visual "ping" is lower.

EEG recording and preprocessing

Scalp EEG was recorded from 60 electrodes with an actiCHamp Plus system (Brain Products), at a sampling rate of 1,000 Hz. The position of electrodes was based on the extended 10-20 international system. The recording was referenced online to a frontal electrode (FCz). Preprocessing and analysis were conducted with EEGLAB (Delorme and Makeig, 2004), Fieldtrip (Oostenveld et al., 2011) and custom MATLAB scripts. Data were first down-sampled to 500 Hz and then a bandpass filter of 1 – 100 Hz was applied. Bad channels were detected and removed using the pop_rejchan function of EEGLAB. Then, the data were re-referenced to the average of all remaining electrodes. Continuous EEG data were segmented into 13.25-sec epochs (from 1.25 sec before sample onset to the offset of the feedback) and epochs containing artifacts were detected and rejected using the pop_autorej function of EEGLAB. After reducing the dimensionality of data to 32 with principal component analysis, and applying an independent component analysis, components containing eye-movement artifacts, muscle artifacts and line noise were detected and removed using ICLabel (Pion-Tonachini et al., 2019). Finally, bad channels were interpolated.

Logic of EEG analyses

As a mechanism of top-down control, one would expect hijacked adaptation to have two discrete sets of neural correlates: one associated with the triggering of the control signal, tightly timelocked to the prioritization cue, and one associated with the consequences of the deployment of this control signal, which would be revealed in ping-evoked activity. Thus, we focused on activities timelocked to these two events for our EEG analyses.

In activity timelocked to the onset of the retrocue, we planned to look for evidence of the top-down control signal that is hypothesized to implement hijacked adaptation. We hypothesized that one potential manifestation of this signal might be traveling waves, the spatial propagation of neural oscillations across the cortex that have been shown to support communication between brain regions, and that have been proposed to subserve cognitive processing (Muller et al., 2018; Xu et al., 2023; Luo and Ester, 2024). In particular, forward (posterior-to-anterior) and backward (anterior-to-posterior) traveling waves have been shown to play an important role in bottom-up and top-down processing of information, respectively (Alamia and VanRullen, 2019; Mohan et al., 2024). Thus, analyses time-locked to the retrocue were predicted to reveal a stronger backward traveling wave on overlap versus on no-overlap trials.

Regarding the consequence of active removal, the down-regulated gain of sensory channels tuned to the to-be-removed orientation was predicted to reduce the overall excitability of the perceptual circuit responsible for its processing (see Fig. 2.2D for an illustration). To assess this predicted effect, we delivered a strong orientation-nonspecific visual ping during the delay period that followed the retrocue, so as to trigger a robust evoked

response and thereby detect the change in excitability of the critical perceptual circuit. Specifically, we predicted that the ping-evoked response would be weaker following active removal of the IMI (Fig. 2.2D), relative to passive removal (Fig. 2.2C), because only the former would be accompanied by a gain reduction in orientation-tuned sensory channels corresponding to the actively removed item. For traveling waves timelocked to the ping, we expected to see decreased forward traveling waves on overlap versus on no-overlap trials, corresponding to the decreased bottom-up processing of information with hijacked adaptation.

ERP analyses

We studied the ERPs evoked by the onset of the retrocue and the ERPs evoked by the ping. For all ERP analyses the EEG was baseline-corrected using the 750 ms before the start of the trial. To test the difference of retrocue-evoked ERPs across two conditions, we applied the cluster-based permutation test in Fieldtrip. First, a paired t-test was conducted and all data points with a p-value lower than 0.05 were identified. Temporally and spatially adjacent significant data points were taken into the same cluster and the summed t value for each cluster was calculated. Finally, the summed t values were compared to the largest summed t values generated from 1,000 permutated data with randomly permuted labels of conditions, and significant clusters (cluster forming p-value threshold = 0.05) were detected with a two-tailed alpha of 0.05.

For the ping-evoked response, we focused on posterior sensors (P7, P8, P5, P6, PO7, PO8, PO3, PO4, POz, O1, O2, Oz), because the hijacked adaptation model predicts a change in the sensory circuits responsible for encoding stimulus orientation. To analyze the statistical

significance, temporal-cluster-based permutation tests were conducted with a cluster forming p-value threshold of 0.05, 1,000 random permutations and a two-tailed alpha of 0.05.

After the results were initially analyzed, we carried out an additional analysis to rule out the potential concern that a difference between ERPs may have been driven by a practice effect, because in our design all three no-overlap blocks always proceeded the three overlap blocks. For this additional analysis we analyzed retrocue-evoked and ping-evoked ERPs in each subject and in each block. To quantify the effects for retrocue-locked ERPs we took the averaged voltage from the time window in which a significant across-condition difference had been found (i.e., 208 to 456 ms after retrocue onset, see Results section below). These averaged voltages were tested with a repeated-measures ANOVA, with block number as the factor. Post-hoc tests between all possible pairs of blocks were also conducted. For the ping-evoked ERPs, in the addition to the analysis on the averaged voltage in the time window (220 to 356 ms after ping onset, see Results), we took the peak-to-peak distance of each block (defined as the difference in the negative and positive peaks in the ERP, see Results below for details) and used repeated-measures ANOVAs to test for a main effect of block number.

Traveling wave analyses

Forward (posterior-to-anterior) and backward (anterior-to-posterior) traveling waves were assessed in EEG epochs timelocked to retrocue onset and timelocked to ping onset. Traveling waves were estimated with a procedure based on 2D fast Fourier transform (2D-FFT). We used four anterior-posterior axes linking frontal and occipital electrodes (Fig.

2.5A; [PO7, P5, CP5, C5, FC5, F5, AF3], [O1, PO3, P3, CP3, C3, FC3, F3], [O2, PO4, P4, CP4, C4, FC4, F4], and [PO8, P6, CP6, C6, FC6, F6, AF4]). For each axis of electrodes, a 500 ms time window was used and the voltage from electrodes was taken to construct a 7 (channels)-by-250 (timepoints) image. Then, a 2D-FFT was applied to this image. The upper and lower quadrants of the resulting spectra quantify the magnitude of the backward (anterior-to-posterior) and forward (posterior-to-anterior) traveling waves, respectively (Alamia et al., 2023; Luo and Ester, 2024), with the x axis corresponding to the temporal frequency of the traveling waves and the y axis corresponding to their propagating speed. Then, for each temporal frequency, the maximum value in each quadrant was extracted. This resulted in a vector of traveling wave power for this time window. Then, the time window was slid by 50 ms and the procedure repeated. As the result, a time-by-frequency matrix of traveling wave power was generated for each axis of electrodes and each direction (forward versus backward).

The power matrix was corrected with a baseline matrix. The baseline matrix was constructed by randomly shuffling electrodes and generating time-by-frequency matrices with the procedure described above, repeating this process 50 times, then averaging the results. Lastly, traveling wave power was calculated as a decibel ratio between the observed and baseline power:

$$W = 10 * \log_{10} \frac{observed\ power}{baseline\ power}$$

Forward and backward waves were estimated separately from epochs timelocked to retrocue onset and timelocked to ping onset. (Note that these analyses allow for forward and backward waves to be present along the same axis during the same temporal epoch.)

For the comparison between conditions, traveling wave power from the four axes of electrodes was averaged and cluster-based permutation testing (with 1,000 random permutations, cluster forming p-value threshold of 0.05, and two-tailed alpha of 0.05) was used to detect significant clusters.

Alpha lateralization

One possible concern about our task design was that because in the overlap condition item C would always appear at the same location as the IMI, subjects might allocate spatial attention to the IMI location in anticipation of item C appearing at this location. For this reason, any difference found in the retrocue-locked or ping-locked EEG activity between the overlap and no-overlap conditions could be attributable to a difference in attentional allocation, rather than to the hypothesized active removal process. To address this concern, we directly assessed whether subjects allocated spatial attention to the IMI location, following the retrocue, by assessing alpha lateralization (Thut et al., 2006; Pietrelli et al., 2022). Specifically, we tested whether alpha-band power contralateral to the IMI location was greater relative to ipsilateral alpha on overlap trials. To carry out this analysis we applied a complex Morlet wavelet decomposition (1–50 Hz, 0.35 Hz frequencies bins, 3–10 cycles) to the cleaned EEG data using Fieldtrip. The resulting power values were averaged across the alpha band (8-14 Hz). The power in the time window (-750 to -300 ms relative to A&B onset) was used as the baseline to correct for alpha band power in the time window of interest (500 ms before retrocue onset to the time of probe onset) and to convert it to the decibel (dB) scale. The averaged power over left (P7, PO7, O1) and right (P8, PO8, O2) parieto-occipital electrodes was taken to assess alpha power contralateral and ipsilateral to

the location of IMI. To test whether there was a significant difference between contra- and ipsilateral alpha power, we conducted a cluster-based permutation test (cluster forming p-value threshold = 0.05, number of permutations = 100,000, two-tailed alpha = 0.05) on the time courses of contra- and ipsilateral alpha power.

Statistical Analyses

Details about the statistical test used for each analysis appear in preceding subsections of the Methods. In brief, a two-tailed paired t-test was used to test whether there was a difference in behavioral performance across conditions; for the TCC model analysis, the 95% CI of the parameter estimates was calculated based on the Markov Chain Monte Carlo post-convergence samples and the statistical significance of difference between conditions was also obtained from these samples; for ERP analyses, traveling wave analyses, and the alpha lateralization analysis, cluster-based permutation tests were run to detect significant differences; repeated-measures ANOVA was used to assess the block-wise effect of ERPs.

Results

Behavioral Results

Mean absolute recall error indicated compliance with task instructions (M±SD = 12.540±3.069 deg), with recall error in the overlap condition (11.960±3.117 deg) significantly lower than in the no-overlap condition (13.120±3.195 deg; $t(24) = 3.8452$, Cohen's $d = 0.769$, $p < 0.001$). Breaking out performance as a function of which item was probed, in report-A trials (15.440±4.075 deg overall), performance in the overlap condition

(15.199±4.448 deg) was similar to performance in the no-overlap condition (15.681±4.049 deg; t(24) = 0.9703, Cohen's d = 0.194, $p = 0.3416$); for report-C trials (9.640±2.443 deg), performance in the overlap condition (8.721±2.248 deg) was significantly better than in the no-overlap condition (10.559±2.789 deg; t(24)=6.7580, Cohen's d = 1.352, $p < 0.001$).

Applying the TCC model to trials in which subjects were probed to recall item *C* indicated that this item evoked a strong familiarity signal in both conditions (Fig. 2.3; no-overlap probed d' = 2.532, 95% CI = [2.462, 2.604], $p < 0.0001$; overlap probed d' = 2.864, 95% CI = [2.785, 2.946], $p < 0.0001$). The probed d' was significantly higher in the overlap condition than in the no-overlap condition (Difference CI = [0.226, 0.448], $p < 0.0001$). In the no-overlap condition the IMI d' was small but significantly higher than zero (Fig. 2.3; no-overlap IMI d' = 0.252, 95% CI = [0.159, 0.341], $p < 0.0001$), suggesting that a residual trace of IMI persisted at the end of the trial. In the overlap condition, IMI d' was numerically negative (overlap IMI d' = -0.096, 95% CI = [-0.208, 0.016], $p = 0.0888$), and significantly lower than the no-overlap IMI d' (difference CI = [-0.4925, -0.2046], $p < 0.0001$). For reference, in report-A trials, there was not a significant difference in either probed d' (overlap probed d' = 1.7934, no-overlap probed d' = 1.7462, difference CI = [-0.0267, 0.1235], $p = 0.2136$) or IMI d' (overlap IMI d' = 0.0131, no-overlap IMI d' = -0.0587, difference CI = [-0.0191, 0.1645], $p = 0.1111$) across conditions.

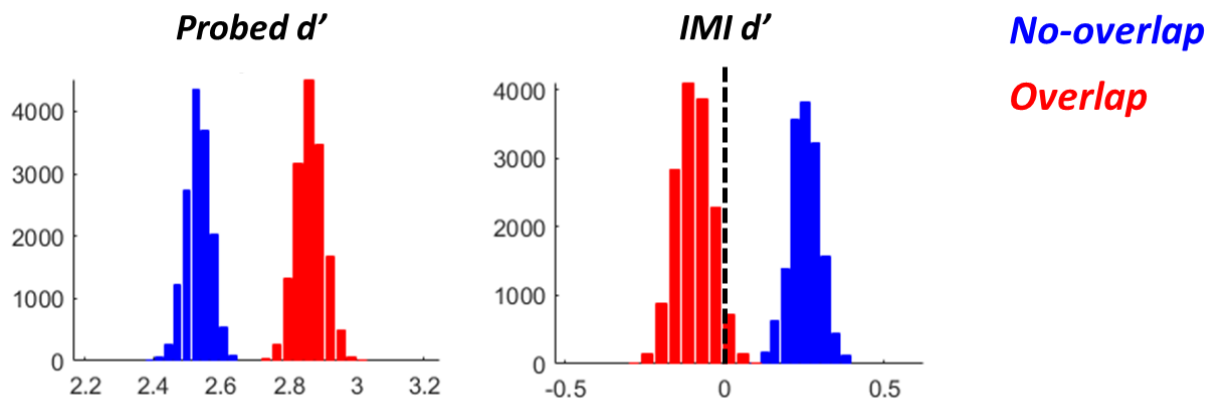

Figure 2.3. Histogram of parameter estimates from the 15,000 post-convergence samples with TCC model, of trials on which item C was probed. Note the difference, for item C ("probed d'") vs. for the IMI ("IMI d'"), in values on the horizontal axis.

EEG results

Retrocue-evoked ERPs. In both conditions, scalp topographies showed a large negativity at central midline electrodes. Comparison of ERPs averaged over these 8 central midline electrodes (Fz, F1, F2, FC1, FC2, C1, Cz, C2) revealed that the magnitude of the ERP was larger in the overlap condition, the two conditions diverging beginning during the initial negative-going deflection following retrocue onset and this difference persisting for roughly 250 msec (Fig. 2.4A; 208 to 456 ms after retrocue onset, $p = 0.002$).

To rule out the possibility that this across-condition difference was driven by a practice effect that gradually emerged across blocks - rather than a genuine difference between the two conditions - we analyzed the ERP broken out for each of the six blocks. The averaged voltage in the time window in which the across-condition difference was significant (i.e., 208 to 456 ms after retrocue onset) was calculated and tested with a repeated-measure

ANOVA, with block number as the factor, which revealed a significant main effect (Fig. 2.4B; F = 17.931, $p < 0.001$, eta2 = 0.428, with Greenhouse-Geisser correction). Consequently, post-hoc tests between all possible pairs of blocks were conducted, and among all pairs of adjacent blocks only the difference between the 1st and 2nd blocks and the difference between the 3rd and 4th blocks were significant (1st vs 2nd: t = 3.463, Cohen's d = 0.420, p = 0.016; 2nd vs 3rd: t = 1.114, Cohen's d = 0.097, p = 1; 3rd vs 4th: t = 3.247, Cohen's d = 0.237, p = 0.021; 4th vs 5th: t = 0.808, Cohen's d = 0.062, p = 1; 5th vs 6th: t = -0.850, Cohen's d = -0.074, p = 1; all p values Holm corrected). We suggest that the most plausible interpretation of these results is for an initial practice effect (block 1-to-2), and a step-like difference between conditions (block 3-to-4).

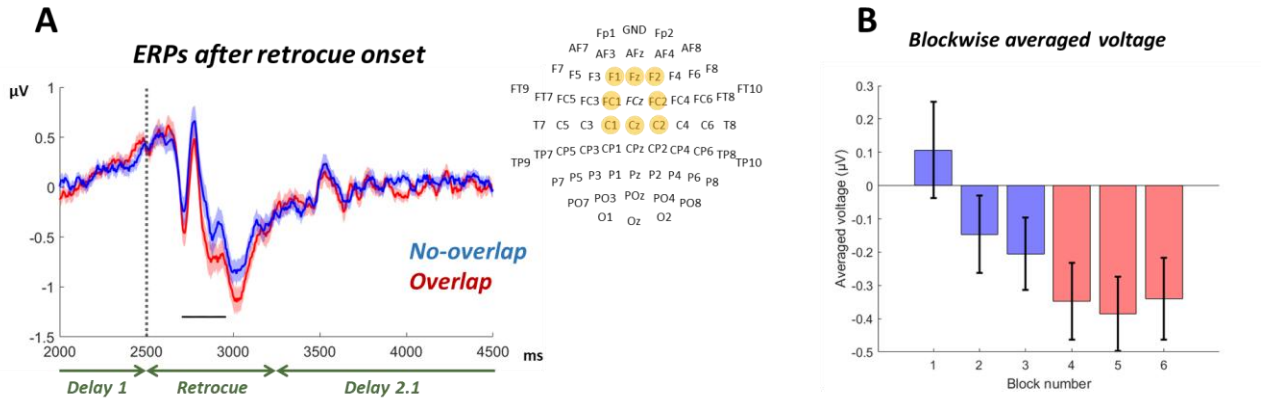

Figure 2.4. A. ERPs from frontal midline electrodes (see the insert plot), timelocked to retrocue onset. Shaded area indicates SEM across subjects. Horizontal black bar shows timepoints with significant difference across conditions. B. The averaged voltage of ERP (in the time window with the significant across-condition difference) for each of the six blocks. Error bars indicate SEM.

Traveling waves timelocked to the retrocue. The forward traveling wave analyses revealed evidence for tonically elevated forward wave at low frequencies (< 10Hz), and a

tonically suppressed forward wave in the beta band (20-26 Hz), in each condition, and for each spanning the duration of Delay 1, the retrocue, and Delay 2.1. In both conditions there was a brief increase in the magnitude of the low-frequency forward wave that was centered on the offset of the retrocue (i.e., at time 3.25 sec). The two conditions only differed for a brief period starting from ~800 ms after retrocue onset, when the tonically suppressed forward traveling in the beta-band dipped lower in the no-overlap condition relative to the overlap condition (Fig. 2.5D left, $p = 0.01$).

Backward traveling-wave analyses revealed similar patterns in both conditions of persistently elevated magnitudes in two frequency bands: one spanning high-alpha/low-beta and one, less strong, at higher frequencies in the beta band, spanning from ~20-25 Hz (Fig. 2.5B right and 2.5C right). (Note that we cannot rule out the possibility that the higher-frequency of these two backward traveling waves may be a harmonic artifact of the lower-frequency one.) In both conditions these backward traveling waves were prominent prior to retrocue onset. The power of these backward waves was stronger in the overlap condition both during the epoch surrounding the retrocue (in alpha/low-beta from ~700 ms before retrocue onset until ~2 sec after retrocue onset, Fig. 2.5D right, $p = 0.002$) and in higher beta beginning shortly after the retrocue onset (~200-750 ms after retrocue onset, Fig. 2.5D right, $p = 0.03$).

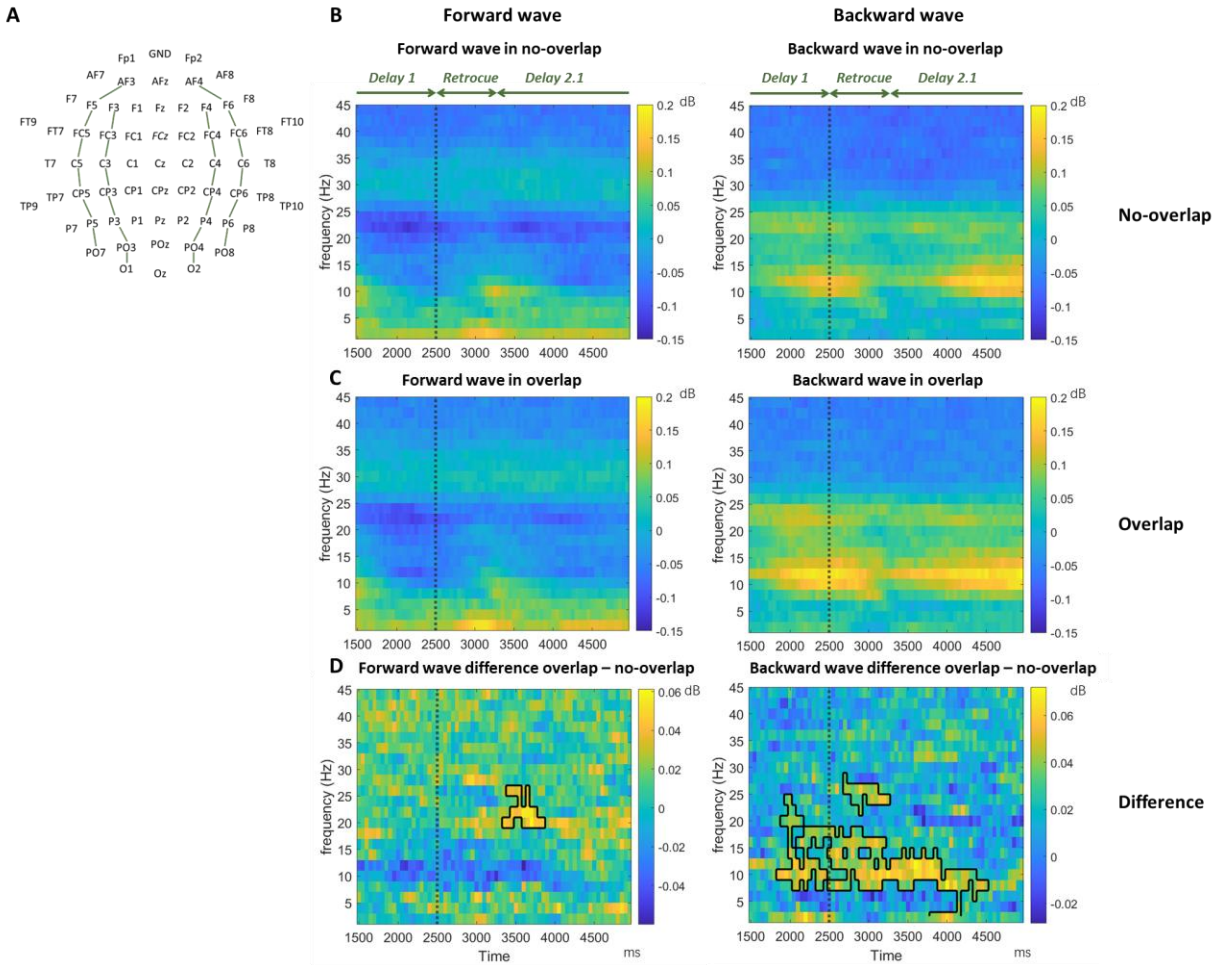

Figure 2.5. Forward and backward traveling waves in EEG, assessed from the average over four axes of electrodes (A), and timelocked to the retrocue onset (dashed line at 2.5 sec), in no-overlap (B) and overlap conditions (C). D shows the difference between conditions with significant clusters marked with black contours. The left column is forward waves and the right column shows backward waves.

Ping-evoked ERPs. For both conditions the ping evoked a large response in posterior electrodes, with a larger negative-going amplitude for the no-overlap condition beginning with the first negative-going deflection and persisting for roughly 135 msec (Fig. 2.6A; 220 to 356 ms after ping onset, $p = 0.028$). To assess whether this difference is best

characterized as a larger amplitude ping-evoked response or a DC shift between the two conditions, we calculated, for each subject, the peak-to-peak distance between the initial negative-going deflection and the subsequent large-amplitude positive-going deflection, reasoning that if the difference were due entirely to a DC shift, the peak-to-peak distances for the two conditions would be the same. For this analysis, the negative peak was defined as the lowest voltage in a 200-msec time window centered on the group-average negative peak and the positive peak was defined as the highest voltage in a 200-msec time window centered on the group-average positive peak. Statistical comparison of peak-to-peak distance between conditions indicated that the difference approached, but did not achieve, the threshold for significance ($t(24) = -1.9789$, $p = 0.0594$). Although this result was equivocal, it approached the level at which one would say that this difference was due, at least in part, to a larger-amplitude ping-evoked response.

To assess the possibility of a practice effect driving the ping-evoked ERP difference across conditions, we conducted a block-wise ERP analysis. There are two ways to assess the ERP in each block: First, similar to the blockwise retrocue-locked ERP, one can calculate the averaged voltage within the time window where the across-condition difference was found (i.e., 220 to 356 ms after ping onset). With this measure, the repeated-measure ANOVA showed a main effect of block number (Fig. 2.6B; $F = 24.823$, $p < 0.001$, eta2 = 0.508, with Greenhouse-Geisser correction). Consequently, we conducted post-hoc tests between all possible pairs of blocks. Among all pairs of adjacent blocks, significant differences were found between the 1st and 2nd blocks and between the 2nd and 3rd blocks (1st vs 2nd: $t = -6.457$, Cohen's $d = -0.386$, $p < 0.001$; 2nd vs 3rd: $t = -3.277$, Cohen's $d = -0.164$, $p = 0.029$; 3rd vs 4th: $t = -1.238$, Cohen's $d = -0.069$, $p = 0.976$; 4th vs 5th: $t = 0.411$, Cohen's $d = 0.018$,

p = 0.976; 5th vs 6th: t = -1.332, Cohen's d = -0.068, p = 0.976; all p values Holm corrected). This suggests that a practice effect might contribute to the difference across conditions. To further assess this, we conducted the blockwise analysis in the second way, with the peak-to-peak distance, which was calculated for each block in the same way as for the across-condition analysis. Repeated-measure ANOVA did not show a significant main effect of block number (Fig. 2.6C; F = 2.437, *p* = 0.067, eta2 = 0.092, with Greenhouse-Geisser correction), which suggests that it is unlikely that the ERP difference between conditions was driven by a practice effect.

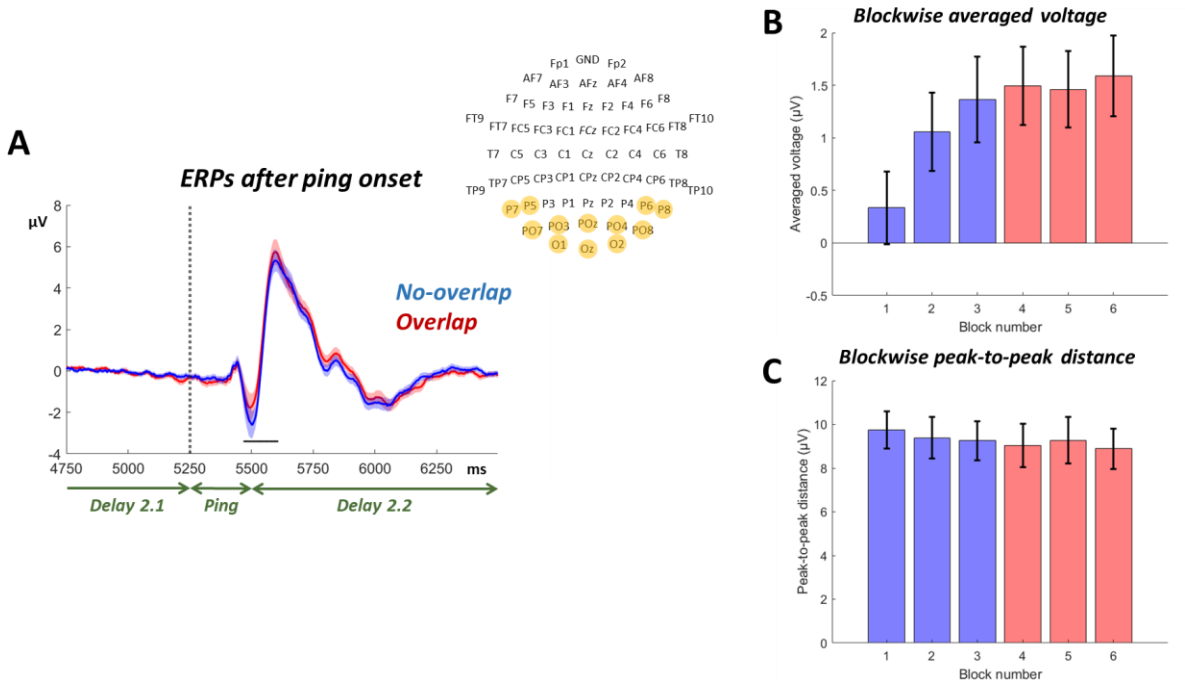

Figure 2.6. Ping-evoked ERPs from posterior electrodes (see the insert plot). Shaded area indicates SEM across subjects. Horizontal black bar shows timepoints with a significant difference across conditions. B. The averaged voltage of ERP (in the time window with the significant across-condition difference) for each of the six blocks. Error bars indicate SEM. C. The peak-to-peak distance of ERP for each of the six blocks. Error bars indicate SEM.

Traveling waves timelocked to the ping. Forward traveling waves, when timelocked to the onset of a visual stimulus, are believed to correspond to bottom-up processing of the perceptual input (Alamia and VanRullen, 2019). Consequently, the hijacked adaptation model predicted that the forward traveling wave triggered by the ping would be reduced during blocks that encouraged active removal. Consistent with this prediction, the visual ping evoked a strong forward traveling wave in the theta band (~4-8 Hz), starting from ~250 ms after ping onset in both conditions (Fig. 2.7A left and 2.7B left). This traveling wave was numerically weaker in the overlap condition relative to the no-overlap condition (Fig. 2.7C left).

A tonically elevated backward traveling wave spanning high-alpha/low-beta frequencies (~10-15 Hz) and extending in time from Delay 2.1 until well into Delay 3, with a pause briefly following the ping, was prominent in both conditions (Fig. 2.7A and B). This was likely a continuation of the backward traveling wave observed in the analyses timelocked to the retrocue (Fig. 2.5B and C). The power of this backward waves was stronger in the overlap condition during the epoch surrounding the ping (in alpha/low-beta starting from ~650 ms before ping onset until ~750 ms after ping onset, Fig. 2.7C right, $p = 0.002$). A significant difference was also observed in high beta (~23-28 Hz) during the time corresponding to the onset of item C, which was at 6.5 sec Fig. 2.7C right, $p = 0.016$).

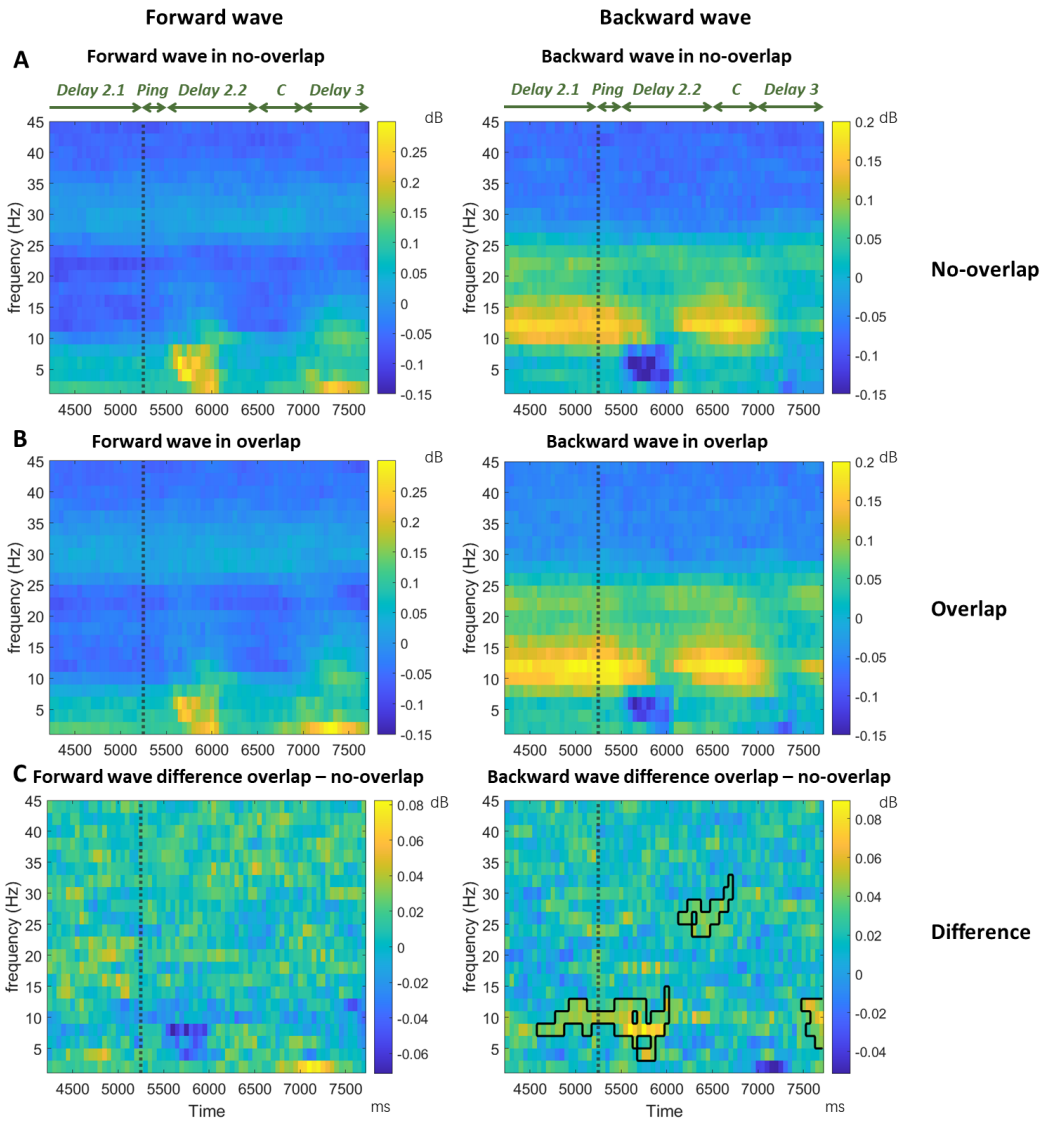

Figure 2.7. Forward and backward traveling waves in EEG timelocked to the ping onset in the no-overlap (A) and overlap conditions (B). Row C shows the difference between conditions. The left column shows forward waves and the right column backward waves. Note that the scale is different for forward and backward traveling waves. The significant difference clusters around or after 6.5 sec in time were likely caused by the onset of item C, which was at 6.5 sec. Plotting conventions are the same as figure 2.5.

Alpha lateralization. A possible concern about our design is that because in the overlap condition item C was always presented at the same location as the IMI, subjects may have preferentially allocated attention to this location. By this alternative, the difference of the neural responses (both ERP and traveling waves) between the two conditions may be a consequence of a difference in attentional allocation rather than of an active removal process. To assess this alternative possibility, we analyzed lateralized alpha (8 – 14 Hz) power from the time of retrocue onset to the time of probe onset in the overlap condition. Here, we used alpha lateralization relative to the IMI location to assess whether subjects were differentially allocating attention to the IMI location, retrocue, in the overlap condition. This analysis revealed two time windows with significantly different contra- vs. ipsilateral alpha power. First, shortly after retrocue onset, alpha power contralateral to the location of the IMI was higher than ipsilateral power (200 to 838 ms after retrocue onset, $p = 0.0055$), a finding that rules out the alternative possibility that subjects may have preferentially allocated spatial attention to the location of IMI during this period of time. The second epoch with significant lateralization was found after the onset of item C (210 to 1012 ms after item C onset, $p = 0.0065$): Here, as expected, alpha power contralateral to the IMI location (i.e., the location of item C) was weaker than ipsilateral alpha, indicating that subjects attended the location of item C when it was presented (presumably while encoding item C into WM). Importantly, no differences were found after the ping and before the onset of item C. To summarize, the alpha lateralization analysis showed that the differences in neural responses across conditions that are the effects of primary interest were unlikely to have been driven by the preferential allocation of spatial attention to the location of IMI in the overlap condition.

Discussion

Although it is widely assumed that updating the contents of working memory can include the operation of actively removing no-long-relevant contents (e.g., (Jonides et al., 1997; Postle et al., 2001)), the mechanisms that might accomplish this operation have received far less attention than other aspects of working memory. We approached this question with the assumption that there may be (at least) two qualitatively different ways whereby information leaves working memory: passively, as a consequence of the withdrawal of attention (Chatham and Badre, 2013; Barbosa et al., 2020; Tsubomi et al., 2024), or actively, via the application of cognitive control. (For alternative accounts, see (Kim et al., 2020; Beukers et al., 2021).) We operationalized these two scenarios with a single task in which we manipulated the level of cue conflict between a no-longer-relevant item and a newly added item, reasoning that the low-conflict ("no-overlap") condition would encourage passive removal, whereas the high-conflict ("overlap") condition would encourage active removal. In previous work, using this, and similar, designs, we have generated indirect evidence that the no-longer-relevant item is, indeed, processed differently in the two conditions. More specifically, the fact that, in the high-conflict condition, this item exerts a repulsive serial bias is consistent with the idea that active removal might be accomplished via a mechanism of hijacked adaption (Shan and Postle, 2022; Teng et al., 2022). The hijacked adaption model posits that the removal of information from WM is accomplished by a top-down control signal that down-modulates the gain of perceptual circuits tuned to that item. In the present study we have extended the previous work by finding direct

behavioral evidence for the suppression of the target item in the active-removal condition, and EEG effects consistent with the implementation of hijacked adaptation (at the time of the retrocue) and with a consequence of hijacked adaption (later in the trial, at the time of the ping).

The hijacked adaptation model posits that the active removal of an item in working memory (the irrelevant memory item; IMI) is accomplished by the down-modulation of gain in perceptual circuits tuned to that item. This particular mechanism has been posited, rather than some other form of inhibition, to account for the residual effects of active removal on the processing of subsequent items. For example, whereas the IMI exerts an attractive bias on recall on the subsequent trial when it has been passively removed, the sign of this serial dependence effect is reversed after active removal (Shan and Postle, 2022; Teng et al., 2022). Here we provide novel behavioral evidence for this account with a modified TCC analysis that allowed us to estimate the effect of active removal of the IMI during that same trial. TCC models have been previously used to study the memory strength of the memorandum (Schurgin et al., 2020) and the intrusion effect of a distractor (Zhang and Lewis-Peacock, 2024). Here we further extended the model in order to study the influence of the IMI, by including the IMI in the model and by allowing d' to take on negative values. This allowed us to make the novel observation that active removal of the IMI leads to a drop in the familiarity landscape centered on the orientation of the IMI, with d' for the IMI taking on a numerically negative value. In the no-overlap condition, in contrast, the lMI left a bump in the familiarity landscape, indicating it was not fully removed from WM, and the positive value of d' was significantly different from its value in the overlap condition. This pattern in the no-overlap condition was in line with our expectation

for passive removal, and consistent with many previous findings (Monsell, 1978; Fischer and Whitney, 2014; Bae and Luck, 2019; Samaha et al., 2019). Thus, these results represent novel behavioral evidence for active versus passive removal from WM.

An interesting possibility is that actively removing IMI may reduce working memory load, thereby releasing working memory resources. By this account, these freed-up resources could be allocated to the other two items, thereby benefiting recall performance for both. Inconsistent with this possibility, however, is the fact that both the mean absolute recall error of item A and the d' for item A (as estimated with the TCC model) remained similar across the two conditions, suggesting that the benefit of actively removing no-longer-relevant information was primarily reflected in the encoding of the subsequent item, rather than in the precision of items already in WM.

At the neural level, our design incorporated a visual ping to detect the predicted difference in perceptual circuits after subjects conducted active versus passive removal. Because the hijacked adaptation model predicts decreased gain in IMI-tuned perceptual channels, the overall response to the ping was expected to be reduced on overlap versus no overlap trials. The analysis of the ping-evoked ERP supported this prediction. The ERP evoked by the ping was significantly weaker in the overlap condition compared to the no-overlap condition, with the divergence in voltages beginning ~250 ms after ping onset, suggesting a reduction of perceptual-circuit excitability after subjects conducted active removal. Follow-up control analyses on this across-condition ERP difference showed mixed outcomes: while a practice effect seems to contribute to this ERP difference as assessed with averaged voltage within the time window (Fig. 2.6B), the peak-to-peak distance in the ping-evoked ERP was not

influenced by a practice effect (Fig. 2.6C). Complementing the ERP effect, a traveling-wave analysis also revealed a numerical reduction of the forward traveling wave, in the theta band, beginning at a time similar to the ERP divergence (i.e., ~250 ms after ping onset). Please note, however, that this observation should be qualified by the fact that this relative decrease of the forward traveling wave in the overlap condition did not survive the cluster-based permutation test. Forward traveling waves have been proposed to index the feedforward processing of visual inputs (Alamia and VanRullen, 2019; Mohan et al., 2024), possibly as a manifestation of the synchronization of neural oscillations between brain regions (Fries, 2015)). Together, these findings provide novel evidence that a mechanism for the active removal of information from WM, presumed to be initiated once the retrocue designated the IMI, had the effect of suppressing the visual processing of a task-irrelevant ping presented 2 sec later in the trial.

We acknowledge that a more direct evaluation of the hypothesized mechanism of hijacked adaptation would be via measurements of the neural representations of the information being removed from WM, such as via multivariate reconstruction methods like inverted encoding modeling (IEM; c.f., (Lorenc et al., 2020; Yu et al., 2020)). However, our attempts at reconstruction of stimulus orientation were unsuccessful, even for items actively maintained in WM (data not shown). This was likely due to the fact that our stimuli were presented at locations in the periphery, and because there were always two items in WM. (For a direct illustration of the effects on stimulus representation of hijacked adaptation, see (Teng et al., 2023))

Active removal via hijacked adaptation is assumed to be accomplished via top-down signaling that commands the down-modulation of gain in posterior perceptual circuits. In the ABC-retrocuing task, this control signal is assumed to be triggered by the retrocue, which designates (by implication) that trial's IMI. As with the ping, we carried out two sets of analyses timelocked to the onset of the retrocue: ERP, and traveling wave. For ERPs, we found that subjects showed a stronger negative-going deflection of the ERP at frontal-midline electrodes for the overlap condition relative to the no-overlap condition. For traveling waves, we observed widespread and prominent backward waves that were stronger in the overlap than the no-overlap condition. These backward traveling waves were strongest in a frequency band spanning alpha and low beta, started prior to retrocue onset, and persisted until 2 sec after the retrocue onset. Thus, the backward waves found in the current study could be reflecting a top-down signal that exerts cognitive control over posterior brain areas (c.f., (Alamia and VanRullen, 2019; Alamia et al., 2023; Luo and Ester, 2024)). At a general level, this aligns well with literatures showing the important role of alpha- and beta-band synchronization in feedback signaling (Bastos et al., 2015; Fries, 2015; Das and Menon, 2021, 2022). Although we hesitate to speculate too much about the specific function(s) of the pattern of backward traveling waves found in the current study, it may be that the stronger alpha/low-beta-band backward traveling waves in the overlap condition correspond to a preparatory control process that starts before retrocue onset and persists until removal of the IMI is completed. In a set of WM tasks, the magnitude of alpha backward traveling waves has been found to be modulated by the load of WM such that higher WM load was associated with weaker alpha backward waves (Zeng et al., 2024). The authors proposed that alpha backward waves may reflect top-down inhibitory gain control

(Zeng et al., 2024), which aligns well with our results in the current study. Oscillations in the beta band, on the other hand, are believed to be important for implementing domain-general inhibitory control (Wessel and Anderson, 2024), including for the inhibition of WM content (Lundqvist et al., 2024). Although our analyses also detected a significant difference, across conditions, in forward waves in retrocue-locked activity, it is unclear how to interpret this, because this effect was driven by greater suppression of forward traveling waves in the no-overlap condition (i.e., lower amplitude in comparison to baseline, see Methods).

To summarize, by using EEG recordings and TCC modeling of behavioral data, we have shown that active removal of information from WM is associated with a top-down control process manifesting at frontocentral electrodes. It results in reduced excitability at posterior electrodes (a hypothesized correlate of down-modulation of gain in posterior perceptual circuits) and behavioral evidence for a stimulus-specific reduction in the familiarity landscape from which the recognition decision is read out. Together, they add to growing evidence for hijacked adaptation as a mechanism for the active removal of information from WM.

Chapter 3

The neural mechanisms of active removal from working memory

Jiangang Shan, Bradley R. Postle

A Stage I-Accepted Preregistered Research Report at PLoS Biology

Abstract

The ability to frequently update the contents working memory (WM) is vital for the flexible control of behavior. What is the neural mechanism for the active removal of information from working memory, however, remains uncertain. In this Preregistered Report we tested the predictions of the hijacked adaptation model of active removal. We collected functional magnetic resonance imaging (fMRI) data while subjects perform a novel "ABC-retrocuing" task designed to elicit two modes of removal, active or passive, in two conditions designed to create a relatively high or low level of interference from the no-longer relevant item. The hijacked-adaptation model posits an adaptation-like modification of perceptual circuits combined with a weak activation of the to-be-removed item. Its predictions were assessed by using multivariate inverted encoding modeling (IEM) and photic "pings" to assay the state of feature-selective encoding channels and of putative activity-silent representations under active-removal versus passive-removal conditions. Although the IEM failed to show reliable reconstructions of the orientation of memoranda, with exploratory IEM reconstructions of the location of memory items, we explored the handling of the memory items across the two conditions. We found the IMI was processed differently across conditions: in the active-removal condition where the potential interference is high, the neural representation of the IMI's location was actively suppressed to a degree that is below baseline. No such suppression was found in the passive-removal condition. This further led to a complete removal of the location information of IMI out of WM, as assessed with the ping-evoked neural response, in the active-removal, but not the passive-removal, condition.

Introduction

A hallmark of working memory (WM) is that it is rapidly updateable, such that information that was relevant in the recent past can be easily replaced once circumstances change and different information has become of primary importance. One way that this is operationalized in the laboratory is with a block of stand-alone trials (e.g., of delayed recognition): Once trial *n* has been completed, subjects have little difficulty encoding a new memory set for trial *n + 1*. Because the set of items is randomly selected for each trial, the memory items for each trial lose their relevance at the end of that trial, and the common intuition is that they should be removed from WM. Despite this intuition, however, the phenomenon of proactive interference indicates that the assumed removal of no-longer-relevant information is often not complete. This is particularly notable for trials featuring "recent-negative" recognition probes that were not in the memory set of the current trial, but were in the memory set on the previous trial -- these lead to an increased false-alarm rate, and to longer reaction times (RTs), for correct rejections (Monsell, 1978). For visual WM tasks that test recall, the imperfect nature of removal manifests itself as serial dependence. For example, when the orientation of a Gabor patch is the feature to memorize and then recall, the reported orientation for trial *n* is commonly found to be biased toward the orientation of the item that had been shown on trial *n – 1* (e.g., (Fischer and Whitney, 2014; Bliss et al., 2017)). (This effect is commonly referred to as an "attractive bias," because it's as though the response on trial *n* is attracted toward the orientation from *n – 1*.)

There is also a growing body of neural evidence for the incomplete removal of information from WM. In an electroencephalography (EEG) study of delayed recall of orientation, Bae and Luck (2019) were able to decode the orientation of the previous trial's sample after the onset of the current trial's sample. For delayed recall of location, Barbosa et al. (2020) were able to decode the previous trial's sample location (from activity in the prefrontal cortex (PFC) of nonhuman primates) from late in the intertrial interval (ITI), just prior to the start of the next trial. Additionally, they observed a similar pattern of reactivation in whole-scalp EEG in humans. Simulations with a bump-attractor network model suggested that the reactivation of no-longer-relevant information may be due to "nonspecific" activation of a residual neural trace that is "imprinted in neuronal synapses as a latent activity-silent trace" (Barbosa et al., 2020). Tellingly, this model did not include an explicit mechanism for removal of no-longer-relevant information; rather, when an item was no longer relevant, activation was simply withdrawn from it, and the bump of activity representing it receded to baseline. Similarly, in a study using a different formal model of WM performance, the Prefrontal Basal Ganglia Working Memory (PBWM) model, the replacement of a no-longer-relevant item with a new one was accomplished via the "reallocation" of resources away from the former (Chatham and Badre, 2013). Thus, many frameworks assume that a default strategy for updating the contents of WM is to employ what we will refer to as the "passive removal" of no-longer-relevant information.

In addition to passive removal, there is also considerable evidence for an active removal mechanism, particularly during tasks that require the simultaneous maintenance of multiple items in WM. One example comes from dual serial retrocuing (DSR) tasks, in which subjects are first shown two stimuli to memorize, then a retrocue indicates which

will be tested first; after the first test, a second retrocue is shown to indicate (with equal probability) which of the two items will be tested in a second test. The first retrocue designates one of the two as a "prioritized memory item" (PMI), and the uncued item, by default, becomes an "unprioritized memory item" (UMI). Critically, the UMI can't be removed from WM, because it may be needed for the second test. After the second retrocue, the newly cued item takes on the status of PMI, and the uncued becomes irrelevant for the remainder of the task (i.e., an "irrelevant memory item," IMI). Thus, the DSR task creates three operationally different states for a memory representation: PMI; UMI; and IMI. In functional magnetic resonance imaging (fMRI) and EEG studies, the ability to decode the identity of a PMI during the delay period of a WM task is a hallmark of its active state. In contrast to this, in some studies using the DSR procedure, multivariate evidence for an active trace of the UMI can drop to baseline (e.g., (Lewis-Peacock et al., 2012; LaRocque et al., 2013)). Despite this, the fact that the UMI has not been removed from WM is inferred from that fact that a pulse of transcranial magnetic stimulation (TMS) has two effects: physiologically, it produces a transient reactivation of the active trace of the UMI (Rose et al., 2016); behaviorally, it produces an increase in false alarm responding when the UMI is used as an invalid memory probe (reminiscent of proactive interference effects; (Rose et al., 2016; Fulvio and Postle, 2020)). Turning at last to the IMI, evidence for its active removal is inferred from the absence of evidence for either TMS reactivation (Rose et al., 2016) or a TMS-related false alarm effect (Rose et al., 2016; Fulvio and Postle, 2020). It is important to emphasize that the labels "PMI," "UMI," and "IMI" refer to an item's state of operational relevance for the cognitive system, not to its presumed physiological state. Consider, for example, the fate of an item at the end of a trial. Although it is an IMI,

the fact that its identity can be decoded from the response of a pulse of TMS delivered late in the ITI is interpreted as evidence for a residual activity-silent trace of that item (Barbosa et al., 2020). In the DSR, in contrast, the absence of such a TMS-reactivation effect for the IMI is taken as evidence that an active removal mechanism has removed any activity-silent trace of the IMI (Rose et al., 2016; Fulvio and Postle, 2020).

Here we test a novel model of active removal – the "hijacked adaptation" model. This is a hypothesized top-down mechanism that works by combining an adaptation-like modification of perceptual circuits with a weak activation of the to-be-removed information. Its core function is two-fold: remove the active trace of the IMI and erase the activity-silent trace of the IMI (Fig. 3.2). The remainder of this introduction will be devoted to the empirical and theoretical contexts that motivate it:

Results from "ABC-retrocuing" provide behavioral evidence for an active removal mechanism

In a recent behavioral study, Shan and Postle (2022) designed a novel "ABC-retrocuing" task intended to engage active or passive removal of an IMI from WM. Each trial began with the simultaneous presentation of two sample oriented gratings (items "*A"* and *"B"*) in two of six possible locations. After a brief delay, a circle appearing at one of the two locations indicated that the corresponding item (for this example we'll say *A*) might be tested at the end of the trial, thereby designating *A* a PMI and *B* an IMI. After another brief delay, a third item ("*C*") was presented, and at the end of the trial recall of the orientation of either *A* or *C* was tested with a response dial appearing at the location of the to-be-recalled item. The

critical manipulation that was intended to encourage active versus passive removal was the location at which item *C* would be presented: In the *overlap* condition, item *C*'s location was always the same as that of the IMI (i.e., item *B*); and in the *no-overlap* condition item *C* always appeared at one of the locations that had not been occupied by either item *A* or *B*. Trials were blocked by condition, and subjects were explicitly informed about the condition prior to each block. The logic was that the *no-overlap* condition might encourage passive removal, just because this seems to be the default for many working memory tasks, as evidenced by the proactive interference and serial dependence effects reviewed above. For the *overlap* condition, however, subjects might be motivated to actively remove the IMI from WM, because otherwise its shared location with item *C* could lead to retrieval conflict when the response dial appeared at this shared location (i.e., "cue conflict"; (Oberauer and Lin, 2017)). A final element of the procedure is that each ABC-retrocuing trial was followed by a trial of simple 1-item delayed recall of orientation, with serial dependence of 1-item recall on the immediately preceding ABC-retrocuing trial used to index the fate of the IMI. (The elements of the ABC-retrocuing task are illustrated in Fig. 3.1, although the study by Shan and Postle (2022) differed from Fig. 3.1 in two respects: Shan and Postle (2022) did not include the "ping" illustrated in Fig. 3.1; and each trial of ABC-retrocuing in Shan and Postle (2022) was followed by a trial of 1-item delayed recall.)

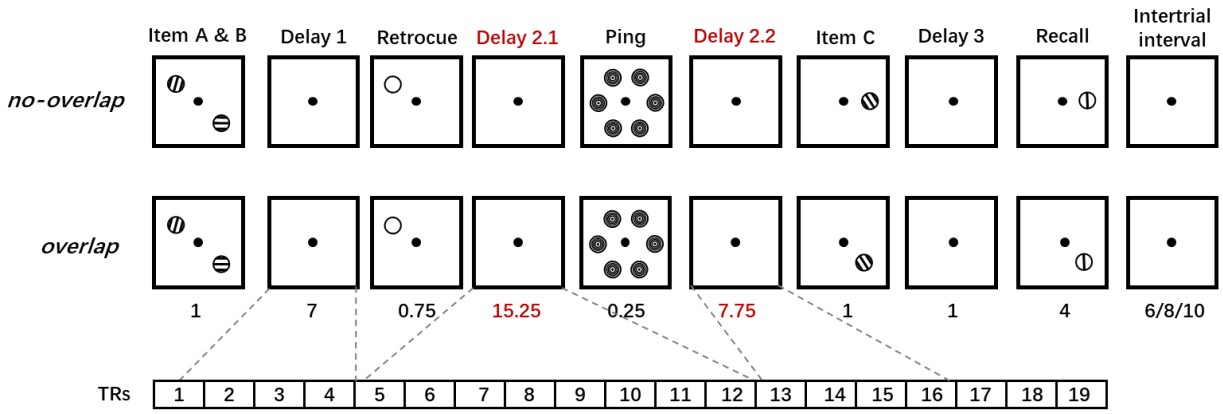

Figure 3.1. The ABC-retrocuing task. The top row illustrates a trial in the *no-overlap* condition, the bottom row a trial in the *overlap* condition. See text for details. Note that the design for the behavioral study from (Shan and Postle, 2022) was similar, with the exceptions that delay periods were shorter, there was no ping, and each trial of ABC-retrocuing was followed by a trial of 1-item delayed recall. The digits below each panel of the overlap trial correspond to elapsed seconds, and each TR in the timeline at the bottom of the figure corresponds to 2 sec.

Results from (Shan and Postle, 2022) revealed a striking difference between the *overlap* and *no-overlap* conditions. In the *no-overlap* condition, item *B* had an attractive serial bias on 1-item recall, consistent with the attractive serial bias observed in several previous studies that we assume were characterized by passive removal of no-longer-relevant stimulus information (e.g., (Fischer and Whitney, 2014; Bliss et al., 2017; Fritsche et al., 2020)). In the *overlap* condition, in contrast, item *B* had a repulsive serial bias on 1-item recall. That is, whereas in the *no-overlap* condition the responses on 1-item recall trials were biased toward the orientation of the IMI from the preceding trial of ABC-retrocuing, in the *overlap* condition the responses on 1-item recall trials were biased away from (hence, "repulsed by") the orientation of the IMI from the preceding trial of ABC-retrocuing. The fact that the serial bias from the IMI was flipped depending on condition suggested that the IMI was processed in a very different way during *overlap* vs. *no-overlap* trials. The

interpretation that the critical difference between the two conditions was active vs. passive removal of the IMI was reinforced by the fact that the serial bias exerted by item *A* was attractive in both conditions.

The logic underlying the hijacked-adaptation hypothesis: Common mechanisms may underlie repulsive serial dependence and active removal from WM

First, it is helpful to review some characteristics of serial dependence. Recent work from two independent groups has converged on the view that, in vision, serial bias effects arise from two different levels of processing, each producing an influence opposite in sign to the other (i.e., attractive versus repulsive). At the level of perception, adaptation to recent perceptual events produces repulsion from previous stimuli, whereas at the level of decision making, perceptual decisions are attracted toward previous decisions (Fritsche et al., 2020; Trapp et al., 2021). These opposing effects also differ with regard to strength of influence on behavior, and to time course. Decisional biases have a stronger influence on behavior, which explains why the serial bias that is most often reported in the literature is attractive. The influence of perceptual adaptation, however, is longer lasting. This accounts for the fact that whereas the dependency on an item from one or two trials previous is typically attractive, this effect flips for longer lags, such that, for example, the influence of the item from five trials previous is repulsive (Fritsche et al., 2020). Of critical relevance for the hijacked-adaptation model, one condition in the ABC-retrocuing results from (Shan and Postle, 2022) was at odds with this pattern: For the IMI in the *overlap* condition, the serial bias from the previous trial was repulsive. This raises a possibility that is at the heart of the

hijacked-adaptation model: active removal of an item in WM may be accomplished via a top-down mechanism that mimics the circuit-level adjustments that are the basis of perceptual adaptation (e.g., (Gibson and Radner, 1937; Barlow, 1961; Clifford et al., 2000; Jin et al., 2005)), but in a manner that is faster (effectively instantaneous) and more pronounced (such that its influence on the subsequent trial overcomes the attractive influence of the decision-making stage). It is important to note that we assume that the effects of perceptual adaptation can be modeled as reductions of gain in the perceptual circuits where this adaptation is taking place. If we start with the assumption that the perception of orientation is accomplished by passing visual signals through a bank of orientation-tuned filters, perceptual adaptation to, say, a 90° grating can be modeled as a decrease of the gain setting of the 90° filter and a smaller decrease of gain at adjacent filters (e.g., those centered on 60° and 120°). (In the framework of multivariate inverted encoding modeling (IEM), this putative bank of orientation-tuned filters is operationalized with a basis set of orientation-tuned "perceptual channels.") Next we consider evidence from a WM task (Lorenc et al., 2020) that is consistent with this idea.

Lorenc et al. (2020) carried out an fMRI study of DSR of oriented-grating stimuli, and one finding was that multivariate decoding evidence for an active trace of the IMI dropped to a level significantly below baseline. When the same data were analyzed with a multivariate inverted encoding model (IEM) the reconstruction of the IMI was "flipped" relative to its reconstruction as a PMI. (For other examples of priority-related "flipping" of IEM reconstructions, see (Sahan et al., 2020; Wan et al., 2020; Yu et al., 2020)). To better understand this effect, the authors carried out computational simulations comparing the effects of modifying the gain, the width, or the spacing (i.e. shifts in tuning profiles) of

orientation-tuned perceptual channels, combined with varying "memory strength," a factor that can be understood as the top-down attentional signal that maintains a WM representation in an active state. The empirical "flipping" effect was best modeled by a suppression of the gain of perceptual feature channels corresponding to the value of the IMI, combined with an intermediate level of memory strength.

Integrating across the findings from the perceptual decision-making (Fritsche et al., 2020; Trapp et al., 2021) and WM (Lorenc et al., 2020; Shan and Postle, 2022) literatures that we have reviewed here has given rise to the idea of hijacked adaptation: The active removal of information from WM may be implemented via a top-down "hijacking" of an adaptation-like modification of perceptual circuits, paired with a weak pulse of (top-down) activation. More specifically, this model posits that active removal from WM is accomplished by the co-occurrence of two events. The first is the adaptation-like modulation of the gain of the perceptual channels that were engaged by the encoding of the to-be-removed item. (This is illustrated by the dip in the "level of gain" in Fig. 3.2.) This putative operation is "adaptation-like" because it is triggered by the onset of the retrocue (not by the perceptual processing of the to-be removed item, which occurred at the beginning of the trial) and because it is greater in magnitude than is typical of perceptual adaptation (the repulsive effects of perceptual adaptation are typically weaker than the attractive influence of recent decisions). The second event, which is hypothesized to occur concurrently, is the brief, weak activation of this item (illustrated by the lower level of top-down "activation," relative to the PMI, in Fig. 3.2). (It is important to note that this hypothesized mechanism differs in important details from a different hypothesized mechanism that is not being investigated here, the nonmonotonic plasticity hypothesis, which predicts weakening and forgetting of

memories as a direct consequence of moderate reactivation (e.g., Norman et al., 2007; Lewis-Peacock and Norman, 2014; Wang et al., 2019). In highjacked adaptation, the construct of "weak activation" corresponds to the "memory strength" parameter in the simulation of (Lorenc et al., 2020), which combines with a decrease in a distinct "gain" parameter in the model (Lorenc et al., 2020). In our conceptualization of hijacked adaptation, these two effects are caused by two distinct top-down control signals, although a direct test of this possibility is outside the scope of the present work.)

We assessed this hijacked-adaptation model by collecting fMRI data while subjects performed an ABC-retrocuing task (Fig. 3.1) while high-contrast task-irrelevant visual stimuli were flashed to "ping" the visual system, so as to assay predicted consequences of this hypothesized mechanism for active removal. In the final subsection of the Introduction, we provide a narrative overview of predictions of the hijacked adaptation model:

Operationalizing tests of the hijacked adaptation model of active removal

Based on Shan and Postle (2022), we assume that the IMI undergoes active removal in the *overlap* condition of the ABC-retrocuing task, but passive removal in the *no-overlap* condition.

As diagrammed in Fig. 3.2, in the *overlap* condition, the hypothesized hijacked-adaptation operation is expected to produce a phasic "flipping" of the active representation of the IMI (operationalized as an IEM reconstruction of the IMI with a negative slope) during the first several seconds following the retrocue (i.e., early Delay 2.1; note that this would constitute a replication of (Lorenc et al., 2020)), followed by a disappearance of a detectable active

trace (i.e., an IEM reconstruction slope not different from 0). This will correspond to successful removal of the IMI. A longer-lasting consequence of active removal, however, will be the residual adaptation-like change to the gain of perceptual feature channels that correspond to the orientation of the IMI. This will be revealed in the filtering of the ping-evoked response (at TRs 15+16), which will also produce a transient flipped IEM reconstruction of the IMI. Note that the delay period after the retrocue and before the ping (i.e., Delay 2.1) is relatively long (15.25s), so as to be able to dissociate the endogenously generated flipped reconstruction of the IMI that is triggered by the retrocue (i.e., during early Delay 2.1) from the flipped reconstruction predicted to be evoked by the ping (at TRs 15+16). (Note that this is a novel prediction, in that, e.g., Lorenc et al. (2020) did not assess the state of representation of the IMI several seconds after the retrocue.) In the *no-overlap* condition, we predict that the withdrawal of attention will result in the disappearance of evidence for an active representation of the IMI during the first several seconds following the retrocue (i.e., early Delay 2.1). However, because the activity-silent trace of the IMI will not have been removed, the ping-evoked response will produce a conventional (i.e., not flipped) IEM reconstruction of the IMI (at TRs 15+16; c.f., (Barbosa et al., 2020)). This pattern of results (summarized in Fig. 3.2) would provide neural evidence that the active removal of information from WM can be accomplished via a mechanism of hijacked adaptation. It would also provide evidence relevant for accounts of the repulsive serial bias that is sometimes observed with perceptual discrimination tasks (e.g., (Fritsche et al., 2020; Trapp et al., 2021)). (We note that the viability of the hypothesis tests described here depends on the ability to track, with IEM, the simultaneous representation of two separate items held in WM. Our group has done this successfully in fMRI studies of DSR-with-

orientations (Yu et al., 2020) and DSR-with-direction-of-motion (Sahan et al., 2020), and of in an EEG study 2-back WM for orientations (Wan et al., 2020).)

(We note that the phenomenon of a flipped IEM reconstruction has also been described in studies that manipulate the priority of items held in WM (Wan et al., 2020; Yu et al., 2020). For example, in the DSR task when a retrocue designates an item a UMI, the IEM reconstruction of its orientation flips in early visual cortex (but not in IPS), and the IEM reconstruction of its location flips in IPS (but not in early visual cortex). When considered from the perspective of the levels-of-analysis framework of Marr and Poggio (1976), however, prioritization and active removal differ in fundamental ways. At the computational level, there are two discrete problems to be solved: holding an item in WM in a deprioritized state (Wan et al., 2020; Yu et al., 2020) versus actively removing an item from WM. At the algorithmic level, we believe it is also likely that the two differ profoundly. We have argued elsewhere that deprioritization is accomplished via a mechanism of "rotational remapping" (Wan et al., 2022), whereas here the mechanism that we are proposing for active removal is hijacked adaptation (the simultaneous suppression of the gain of perceptual feature channels corresponding to the value of the IMI, combined with an intermediate level of activation of that representation). Thus, it is only at the implementation level that priority-based remapping and hijacked-adaptation may both produce "flipped" IEM reconstructions. (For an in-depth consideration of caveats when inferring underlying physiological processes from IEM reconstructions, see (Liu et al., 2018; Sprague et al., 2018; Gardner and Liu, 2019; Sprague et al., 2019).)

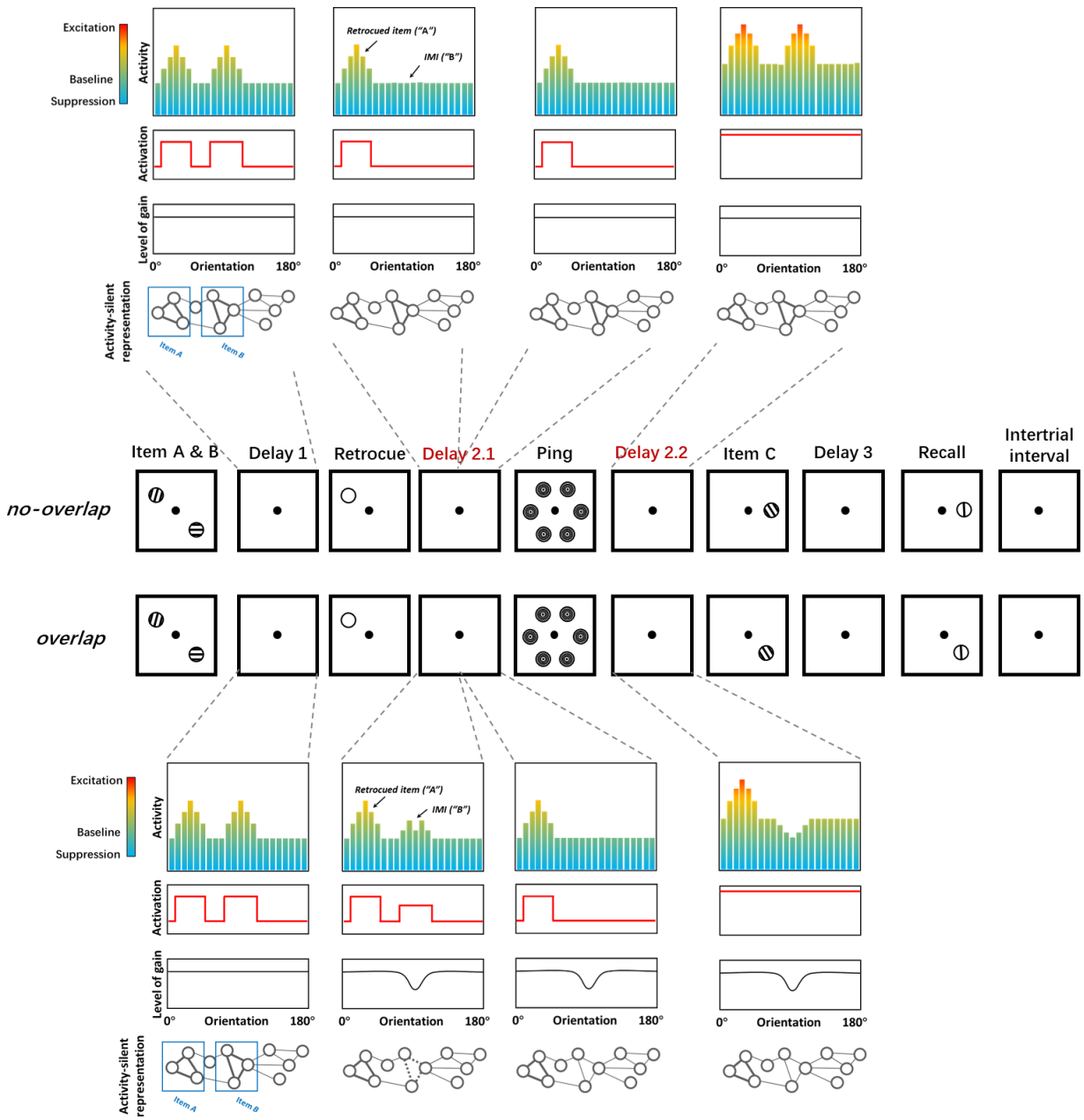

Figure 3.2. The mechanism of hijacked-adaptation, illustrated via the hypothesized states of perceptual circuits that encode and maintain stimulus information in WM during different epochs of the trial. The rows of panels above and below the timelines correspond to four elements of the model: 1) the colored bars represent the activation levels of hypothetical orientation-tuned perceptual channels; 2) the red lines represent the level of top-down activation allocated to each of the two memory items; 3) the black lines represent the level of gain of the perceptual channels, with the default value of each being 1.0, and hijacked-adaptation resulting in channel-specific

decreases in this value; 4) and the cartooned networks (Note that the level-of-gain in this figure only reflects the influence of top-down hijacked adaptation; the effects of true perceptual adaptation on channel gain are assumed to be too subtle to be detectable at the scale of the processes being illustrated here.) In the *no-overlap* trial (top five rows), each sample item is represented during Delay 1 with 1) elevated activity in the orientation channels corresponding to their value; 2) comparable levels of top-down activation; 3) baseline levels of gain at each channel; and 4) an activity-silent representation. During Delay 2.1, the representation of the retrocued item (i.e., *A*) remains elevated because its level of top-down activation remains unchanged. The active representation of the IMI (*B*), however, drops to baseline, because it is no longer receiving top-down activation. Importantly, however, the activity-silent representation of *B* remains. Early in Delay 2.2, the ping nonspecifically raises the activity level in every orientation channel. This produces a reactivation of *B*, because the activity from the ping is filtered through the activity-silent representation of *B.* In the *overlap* trial, the cuing of *A* prompts the active removal of *B* via highjacked adaptation: a coordinated decrease in the gain of the channels corresponding to *B* (illustrated by the orientation-specific dip in the gain field) plus a weak phasic activation of these channels (illustrated by the lower level of top-down activation, relative to the retrocued item) that occur during the early portion of Delay 2.1. The effect of these events is effectively instantaneous, and is two-fold: at the level of channel activity they produce an activity-based representation of *B* that is "flipped" (and labeled "IMI"); and at the level of activity-silent representation, the representation of *B* is removed (illustrated with dotted lines) due to synaptic weakening produced by the weak activation paired with the orientation-specific reduction of gain. Later during Delay 2.1, the modified gain field persists but this is not evident in the activity of the perceptual channels with only baseline levels of activity corresponding to the value of the IMI. Finally, early in Delay 2.2, responses to the ping, filtered through the modified gain field, produce a transient pattern of activity that is also a "flipped" version of *B*.

Methods

Preregistered hypotheses

Three primary hypotheses of the hijacked adaptation model were tested in this Preregistered Report.

Hypothesis 1a: In the *overlap* condition, the reconstruction of the orientation of the IMI during early Delay 2.1 (TR 7), with an IEM trained on the retrocued item at TR 7, will have a significantly negative slope. *(Rationale: This pattern of a "flipped" IEM reconstruction, replicating (Lorenc et al., 2020), is hypothesized to be a consequence of hijacked adaptation.)*

Hypothesis 1b: In the *no-overlap* condition, the reconstruction of the orientation of the IMI during early Delay 2.1 (TR 7), with an IEM trained on the retrocued item at this time (i.e., TR 7), will have either a small positive slope (smaller than the retrocued item) or a slope not different from 0. (*Rationale: Because this condition is assumed to involve passive removal, from the perspective of the hijacked adaptation model no correlate of active removal is expected. Thus, for this model, the critical prediction is that the reconstruction of the IMI will not have a negative slope.)*

Hypothesis 1c: The slopes from 1a and 1b will differ. *(Rationale: If the "flipped" IEM reconstruction is specific to active removal (predicted by hijacked-adaptation model), the slopes of IEM reconstructions from the two conditions should differ. Confirmation of this hypothesis would provide quantitative evidence that the two conditions differ in terms of the processing of the IMI (active vs. passive removal). This outcome is a necessary precondition for the subsequent hypotheses about predicted consequences of hijacked activation to be valid.)*

Hypothesis 2a: In the *overlap* condition, if an IEM can be successfully trained to reconstruct the retrocued item during late Delay 2.1 (i.e., at TR 12), the reconstruction of the orientation of the IMI at this same time point, with that same IEM, will be unsuccessful (i.e.,

slope not different from 0). *(Rationale: This is a sanity check for the hijacked-adaptation model, which predicts that active removal will have removed any active trace of the IMI.)*

Hypothesis 2a' (if needed): In the *overlap* condition, if an IEM cannot be successfully trained to reconstruct the retrocued item during late Delay 2.1 (i.e., at TR 12), the reconstruction of the orientation of the IMI at this same time point with an IEM trained on the retrocued item during early Delay 2.1 (i.e., at TR 7) will be unsuccessful (i.e., slope not different from 0). *(Rationale: This is an alternative way to carry out the same sanity check from* Hypothesis 2a*, if* 2a *cannot be tested.)*

Hypothesis 2b: In the *no-overlap* condition, if an IEM can be successfully trained to reconstruct the retrocued item during late Delay 2.1 (i.e., at TR 12), the reconstruction of the orientation of the IMI at that time point, with that same IEM will be unsuccessful (i.e., slope not different from 0). *(Rationale: This is merely a statement of the expectation that there will no longer be a detectable active trace of no-longer-relevant item (the IMI) at the end of Delay 2.1 (i.e., 14 sec after the retrocue designated it the IMI).)*

Hypothesis 2b' (if needed): In the *no-overlap* condition, if an IEM cannot be successfully trained to reconstruct the retrocued item during late Delay 2.1 (i.e., at TR 12), the reconstruction of the orientation of the IMI during late Delay 2.1 with an IEM trained on the retrocued item from early Delay 2.1 (i.e., TR 7) will be unsuccessful (i.e., slope not different from 0). *(Rationale: This is an alternative way to confirm the same expectation as described for* Hypothesis 2b*, if* 2b *cannot be tested.)*

Hypothesis 3a: In the *overlap* condition, if an IEM can be successfully trained to reconstruct the retrocued item from the ping-evoked response (i.e., at TRs 15+16), the reconstruction of the orientation of the IMI from the ping-evoked response (i.e., at TRs 15+16), with that same IEM, will have a significantly negative slope. *(Rationale: This is a key prediction of the hijacked-adaptation model, which is that the persistence of the pattern of channel-specific gain modification (resultant from the application of this mechanism to effect active removal of the IMI) – the same phenomenon hypothesized to be responsible for the repulsive serial bias effect* (Shan and Postle, 2022) – *will be revealed when signals from the ping are filtered through these perceptual channels.)*

Hypothesis 3a' (if needed): In the *overlap* condition, if an IEM cannot be successfully trained to reconstruct the retrocued item from the ping-evoked response (i.e., from TRs 15+16), the reconstruction of the orientation of the IMI from the ping-evoked response (i.e., from TRs 15+16) with an IEM trained on the retrocued item at TR 7 will have a significantly negative slope. *(Rationale: This is an alternative way to test the prediction of* Hypothesis 3a*, if* 3a *cannot be tested.)*

Hypothesis 3b: In the *no-overlap* condition, if an IEM can be successfully trained to reconstruct the retrocued item from the ping-evoked response (i.e., from TRs 15+16), the reconstruction of the orientation of the IMI from the ping-evoked response (i.e., from TRs 15+16) with that same IEM will have a significantly positive slope. *(Rationale: Because passive removal is assumed to leave the (putative) activity-silent representation of the IMI intact (c.f. (Bae and Luck, 2019; Barbosa et al., 2020)), when signals from the ping interact with this activity-silent representation of the IMI (e.g., are filtered through it (Rose et al.,*

2016) or, in the sonar metaphor, "bounce off it" (Wolff et al., 2015; Wolff et al., 2017)) the ping-evoked response will reveal this residual activity-silent representation of the IMI (c.f., (Wolff et al., 2015; Wolff et al., 2017).)

Hypothesis 3b' (if needed): In the *no-overlap* condition, if an IEM cannot be successfully trained to reconstruct the retrocued item from the ping-evoked response (i.e., from TRs 15+16), the reconstruction of the orientation of the IMI from the ping-evoked response (i.e., from TRs 15+16), with an IEM trained on the retrocued item from early Delay 2.1 (i.e., from TR 7) will have a significantly positive slope. *(Rationale: This is an alternative way to test the prediction of* Hypothesis 3b*, if* 3b *cannot be tested.)*

Hypothesis 3c: The slopes from *3a* and *3b* will differ. *(Rationale: If the persistence of the pattern of channel-specific gain modification is specific to active removal, the slopes of IEM reconstructions from the two conditions should differ. Confirmation of this hypothesis would provide quantitative evidence that the persistent effects of active vs. passive removal differ in the way predicted by the hijacked-adaptation model.)*

Hypothesis 3c' (if needed): The slopes from 3a' and 3b' will differ. *(Rationale: This is an alternative way to test the prediction of* Hypothesis 3c*, if* 3c *cannot be tested.)*

Subjects

30 subjects who are 18-35 years in age with normal or corrected-to-normal vision and report no history of neurological disease were recruited from the University of Wisconsin–Madison community. Informed consent was obtained. All experimental procedures for the

Preregistered Research Article have been approved by the University of Wisconsin–Madison Health Sciences Institutional Review Board (protocol ID 2017-0344).

Power analysis. Using data from (Yu et al., 2020), in which a negative slope of the IEM reconstruction of the UMI in a DSR-of-orientation task has been observed, power analysis of the 2-tailed one sample *t*-test shows we need data from 30 subjects to achieve 90% power to detect a significantly negative slope for the reconstruction of orientation of the UMI (Cohen's d = 0.617), and data from 26 subjects to detect a significantly positive slope for the reconstruction of orientation of the PMI (Cohen's d =0.675).

To the best of our knowledge, there is no established way to perform power analysis for bootstrapping, which we used in the current study to test for the predicted positive and negative slopes of reconstructions. We used data from (Yu et al., 2020) to simulate the *p*-values obtained from *t*-tests versus from bootstrapping with different sample sizes. Because this sample had data from 13 subjects, we generated estimates ranging from $N = 8$ to $N = 12$, by randomly drawing N subjects from the sample, without replacement, and conducting a *t*-test and a bootstrap analysis on these data. For the *t*-tests, we collapsed over channel responses on both sides of the target channel, averaged them, and calculated the slope of the averaged UMI reconstruction of each subject with linear regression. The slopes were then compared to 0 with a 2-tailed one sample *t*-test. For bootstrapping, the method was the same as specified in the *Statistical Analyses* subsection of fMRI Analyses section of the *Methods*. This process was repeated 10 times at each N to get 10 (different) sets of subjects and 10 *p*-values for each test. For $N=13$, one *p*-value was obtained from each test. Across sample sizes, the bootstrapping was generally more sensitive than the *t*-test (Fig.

3.3). It has been shown by other researchers that the bootstrap consistently outperformed the t-test in a more systematic way (Ahad et al., 2012). Based on this, we reason that the sample size estimated by the power analysis for a *t*-test provides a conservative estimation of the sample size required in the current study (because we were using the more sensitive bootstrapping procedure). In the current study we use a sample size of 30 subjects.

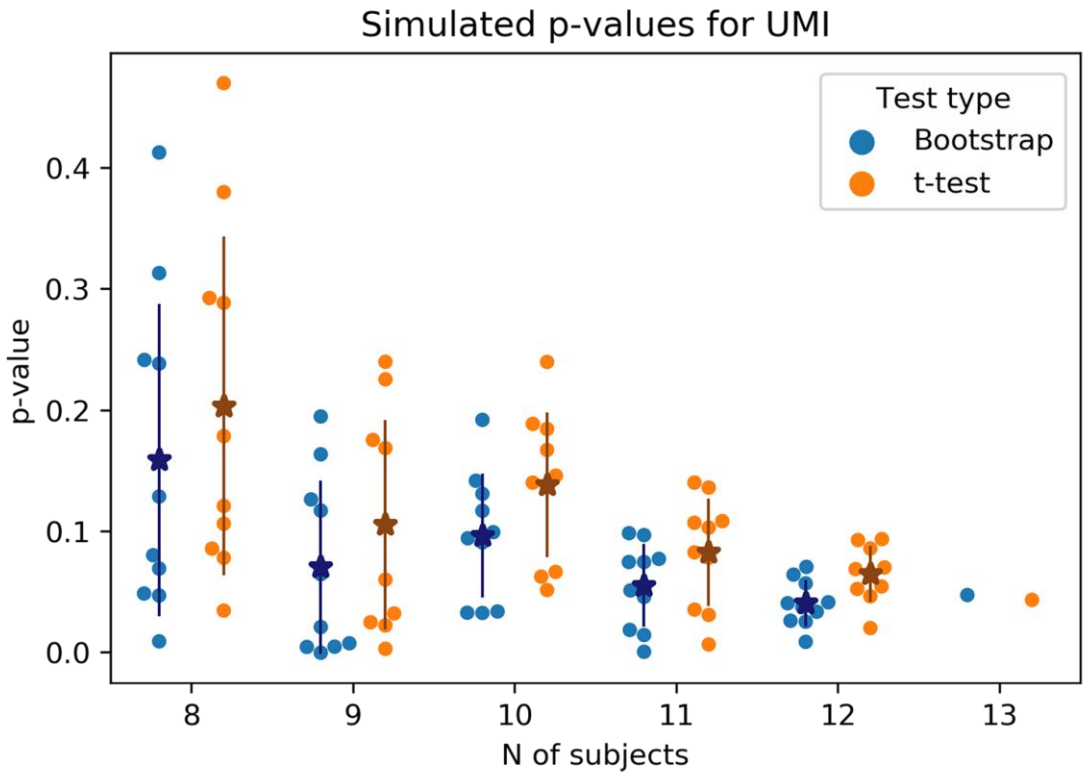

Figure 3.3. The *p*-values obtained from bootstrapping tests and *t*-tests of subsets of subjects from Yu, Teng, and Postle (Yu et al., 2020). The darker stars overlaid on the dots represent the mean and the error bars show the standard deviation of each set of *p*-values.

Stimuli and procedure

The stimulus presentation and response collection were implemented with MATLAB (MathWorks, Natick, MA, USA) with the Psychtoolbox-3 extensions (Brainard, 1997; Pelli, 1997). The display was projected into the scanner and onto a mirror mounted on the head coil at 60-Hz (Avotec Silent Vision 6011 projector; Avotec, Stuart, FL, USA). The viewing distance was roughly 69 cm and the screen width was 33.02 cm. The sample stimuli were grayscale sinusoidal gratings (radius = 3˚; contrast = 0.6; spatial frequency = 1 cycles/˚; random phase angle) presented on gray background (L= 52, a = 0, and b = 0 in CIEL*ab space). There were six possible sample orientations: 20°, 50°, 80°, 110°, 140°, 170°; with a random jitter of ±0°- 3° added with each presentation. These and all ensuing stimuli appeared at any of six possible locations on an imaginary circle centered on fixation (radius of 8°, locations centered at each of these polar angles: 30°, 90°, 150°, 210°, 270°, 330°). The retrocue was a white circle (thickness=0.08°) with the same radius as the sample stimuli. Ping stimuli were high contrast concentric circles with the same radius and spatial frequency as the gratings (contrast=1). The response dial was a black bar (thickness=0.08°) corresponding to the diameter of a black circle with the same radius as the gratings (thickness=0.08°). Subjects were instructed to adjust the orientation of the bar using an MR-compatible trackball (Current Designs, Philadelphia, PA, USA) and to report their response by pressing a button on the trackball when the orientation of the bar matches their memory for the probed sample. After the button was pressed by the subject, the black bar of the response dial became thicker (thickness=0.16°) to indict the response has been made and cannot be changed. A white fixation dot was present throughout each block (i.e., also during the ITI).

Each trial of ABC retrocuing started with the simultaneous presentation of two samples (*A* and *B*; 1 s) followed by *Delay 1* (7 s). Next the retrocue appeared for 0.75 s at the location that had been occupied by either *A* or *B*, thereby designating a PMI (which might be tested at the end of the trial) and, by implication, the IMI (no longer relevant for that trial). The retrocue was followed by *Delay 2.1* (15.25 s), which was followed by the simultaneous presentation (0.25 s) of ping stimuli at each of the six locations, then *Delay 2.2* (7.75 s), then sample item *C* (1 s), then *Delay 3* (1 s). Finally, the response dial appeared at the location that had been occupied by the retrocued item or by item *C*, prompting the recall of the orientation of that item (4-s response window). The inter-trial interval ITI varied randomly between 6, 8, and 10 s.

On each trial the orientation of items *A* and *B* was randomly selected, with replacement, from the pool of six possible values. The locations of item *A* and *B* were randomly selected from the six possible locations. To fully cross the orientations of item *A* and *B*, 21 unique trials are required. 252 trials (12 repetitions per unique trial) were used for each condition. The retrocuing of *A* or *B* was randomly determined on every trial. The orientation of item *C* was randomly selected from the pool of six possible values (i.e., independent of *A* and *B*), and its location depended on condition: in the *overlap* condition it appeared at the location that had been occupied by the uncued item; in the *no-overlap* condition it appeared in a location randomly selected from the four that had not been occupied by *A* or *B*. The retrocued item or the item *C* was probed for recall equiprobably.

Trials were blocked by condition (*overlap*, *no-overlap* condition), and subjects were explicitly informed of the condition before the start of each block. Each subject participated

in 4 scanning sessions. The first scanning session consisted of 6 runs, each run corresponding to a 14-trial block. The three remaining scanning sessions each consisted of 10 runs (each run corresponding to a 14-trial block). There were fewer runs in the first session due to acquisition of structural images. To facilitate the consistent use of active removal and passive removal, within each session the first 3 blocks (for the first session) or 5 blocks (for the last three sessions) were of one condition and the remaining blocks were of the other condition. The order of conditions within a session was counterbalanced across sessions and across subjects. In the first session, each subject first did two practice blocks (one block for each condition) outside the scanner and another practice block (with the same condition as the first real block) inside the scanner. An Avotec RE-5700 eye-tracking system (Avotec) was used to track eye position throughout each scanning session, and to assure that subjects' eyes are open during the ping.

Behavioral Data Analysis

The mean absolute error of recall across subjects was calculated for each condition separately. The performance across the two conditions was compared with a paired *t*-test.

fMRI Data Acquisition

Whole-brain images were acquired at the Lane Neuroimaging Laboratory at the University of Wisconsin–Madison HealthEmotions Research Institute (Department of Psychiatry) using a 3 Tesla GE MR scanner (Discovery MR750; GE Healthcare, Chicago, IL, USA). A high-resolution T1 image was acquired with a fast spoiled gradient recalled echo sequence (8.2 ms TR, 3.2 ms TE, 12˚ flip angle, 176 axial slices, 256 × 256 in-plane, 1.0 mm isotropic) for each session. Functional data were acquired with a gradient-echo echo-planar sequence (2

s repetition time [TR], 22 ms echo time [TE], 60˚ flip angle) within a 64 × 64 matrix (42 axial slices, 3 mm isotropic).

fMRI Data Preprocessing

fMRI data were preprocessed with the Analysis of Functional Neuroimages (AFNI) package (https://afni.nimh.nih.gov). To achieve a steady state of tissue magnetization, the first four TRs of each run were discarded. The data were then registered to the final volume of each scan and then to the anatomical images from the first session. Volumes were motion corrected with six nuisance regressors to account for mead motion artifacts. Linear, quadratic, and cubic trends were removed for each run and the z-scores of fMRI time series data were calculated within each run.

fMRI Analyses

Task-related activity. The fMRI data were fitted to a general linear model (GLM) with regressors for each epoch of the task -- *Encoding A&B* (2 s), *Delay 1* (6 s), *Delay 2.1* (16 s), *Delay 2.2* (8 s), *Encoding C* (2 s), *Recall* (4 s) – each convolved with a canonical hemodynamic response function, as well as nuisance covariates for between-trial and between-scan drift, and head motion.

ROI creation. Hypothesis tests were carried out in an early visual cortex ROI and an intraparietal sulcus (IPS) ROI. First, an anatomically defined ROI of early visual cortex was created from masks corresponding to V1 and V2 (merged, both hemispheres), and an anatomically defined ROI of IPS (comprising IPS0–5; merged, both hemispheres), both based on the probabilistic atlas of Wang and colleagues (Wang et al., 2015) and warped to each subject's structural scan in native space. Hypothesis testing was carried out in the 500

voxels within the anatomical early visual cortex ROI with have the strongest weights on the Encoding A&B regressor, which we refer to as the early visual ROI. For the IPS, hypothesis testing was carried out in the 500 voxels within the anatomical IPS ROI with have the strongest weights on the Delay 2.1 regressor.

Inverted Encoding modeling. IEM analyses were performed with custom functions in MATLAB. In IEM, the responses of each voxel are assumed to be a weighted sum of responses of several hypothetical tuning channels. Six tuning channels of orientation (or location) were used and the tuning curve of each channel was defined as a half-wave–rectified sinusoid raised to the eighth power. We first computed the weight matrix W (v voxels × k channels) that projects the hypothesized channel responses C1 (k channels × n trials) to the measured voxel responses B1 (v voxels × n trials) with the training dataset to get the estimate of the weight matrix $\widehat{W}$. Then we used $\widehat{W}$ to reconstruct the channels responses $\widehat{C_2}$ from the voxel activities B2 of the testing dataset. The relationship between B1, W and C1 was characterized by

$$B_1 = WC_1$$

The least-squared estimate of the weight matrix ($\widehat{W}$) was calculated using linear regression:

$$\widehat{W} = B_1{C_1}^T({C_1{C_1}^T})^{-1}$$

The channels responses $\widehat{C_2}$ of the testing dataset was then calculated with the weight matrix ($\widehat{W}$) and the BOLD data (B_2):

$$\widehat{C_2} = (\widehat{W}^T\widehat{W})^{-1}\widehat{W}^T B_2$$

The IEMs were trained with the orientation (or location) of the retrocued item or the IMI and test on the orientation (or location) of the retrocued item or the IMI at each TR. We used a leave-one-run-out procedure in which the model was trained with data of all but one run and tested on the left-out run. This process was repeated until the reconstruction of all runs is acquired. The estimated channels responses were centered on the orientation (or location) of the tested item. The reconstruction was generated on all TRs but our pre-registered hypotheses focused on specific TRs: *Hypothesis 1a, 1b, and 1c* focused on TR 7; *Hypothesis 2a and 2b* focused on TR 12. For *Hypothesis 3a, 3b and 3c*, the averaged BOLD of TR 15 and TR 16 was used to train and test the model. In case we cannot get a reliable reconstruction of the retrocued item at TR 12 and/or TRs 15+16 due to the representation of the retrocued item shifts to an activity-silent state after a long time span since the presentation of the item, the retrocued item at TR 7 was used to train the model and TR 12 (for *Hypothesis 2a and 2b*) and/or TRs 15+16 (for *Hypothesis 3a, 3b and 3c*) was tested.

For all analyses of orientation IEM training collapsed across the location at which items appeared on the screen. This choice is justified for two reasons. First, in a previous study that presented orientations at only four different locations (and therefore collected enough data to train location-specific IEMs of orientation), location-specific IEMs were found to be only subtly numerically superior (i.e., higher reconstruction slopes) to location-nonspecific IEMs trained with the same number of trials (Cai et al., 2019). Second, in a previous study that presented orientations at eight different locations (and therefore did not collect enough data to train location-specific IEMs of orientation), location-nonspecific IEMs or orientation were robust (Yu et al., 2020). (Indeed, it is with data from Yu, Teng, and Postle (Yu et al., 2020) that we performed power calculations for this study.)

Exploratory IEM reconstruction of locations. For the reconstruction of the location of memory items, one possible concern is that because the item A and the IMI never share the same location in the current task, for an IEM that was trained on the item A and tested on the IMI, a negative reconstruction of the IMI may simply be driven by a positive representation of the item A combined with null information of the IMI. This is because when the estimated channels responses were centered on the IMI's location, the channel responses corresponding to the item A would increase the responses for all other channels except the center one, and thus lead to a negative reconstruction of the IMI even if there was no real reconstruction of it (i.e., the minimum distance problem). To address this, we used the averaged reconstruction of the four locations that were not occupied by item A or B as the baseline, as these locations should be equally influenced by this minimum distance problem. For all location reconstructions, we first subtracted the baseline reconstruction from the reconstruction of interest, then calculated their slopes with the method described below.

Statistical Analyses. The strength of IEM reconstructions of memory items was operationalized by their slope. We collapsed over channel responses on both sides of the target channel, averaged them, and calculated the slope of the reconstruction with linear regression for each subject separately.

We used bootstrapping to test the statistical significance of the group-average slope of each reconstruction (Ester et al., 2015; Ester et al., 2016). For each hypothesis test, we randomly sampled 30 reconstructions from the pool of 30 (one per subject), with replacement, and calculated the average of the channel responses. This process was repeated 10,000 times to

get 10,000 resampled group-average reconstructions, and the slopes of these reconstructions were calculated. Two-tailed *p*-values were computed as the proportion of positive or negative slopes, whichever is smaller, multiplied by 2. To test the difference between slopes, we calculated the difference between the 2 slopes of interest for each one of the 10,000 resampled data sets. Two-tailed *p*-values were the proportion of positive or negative differences, whichever is smaller, multiplied by 2.

Results

Behavioral results

Subjects' performance was at a similar level in the *overlap* condition (mean absolute error of recall, M±SD = 15.038±4.849 deg) and in the *no-overlap* condition (14.862±4.836 deg; $t(29) = 0.5321$, $p = 0.5987$).

Preregistered hypotheses

To test the preregistered hypotheses, we first tried to reconstruct the orientation of the retrocued item by training and testing the IEM with the retrocued item on a TR-by-TR basis. In the early visual ROI, the reconstruction of item A's orientation was only successful during TR 4 and 5 in the *no-overlap* condition (Fig. 3.4 top; $p = 0.0004$ and 0.0326, respectively, with bootstrapping tests), and it was not successful during any TR in the *overlap* condition. In the IPS ROI, the item A's orientation reconstruction was successful during TR 15 in the *overlap* condition (Fig. 3.4 bottom; $p = 0.0162$), but not elsewhere.

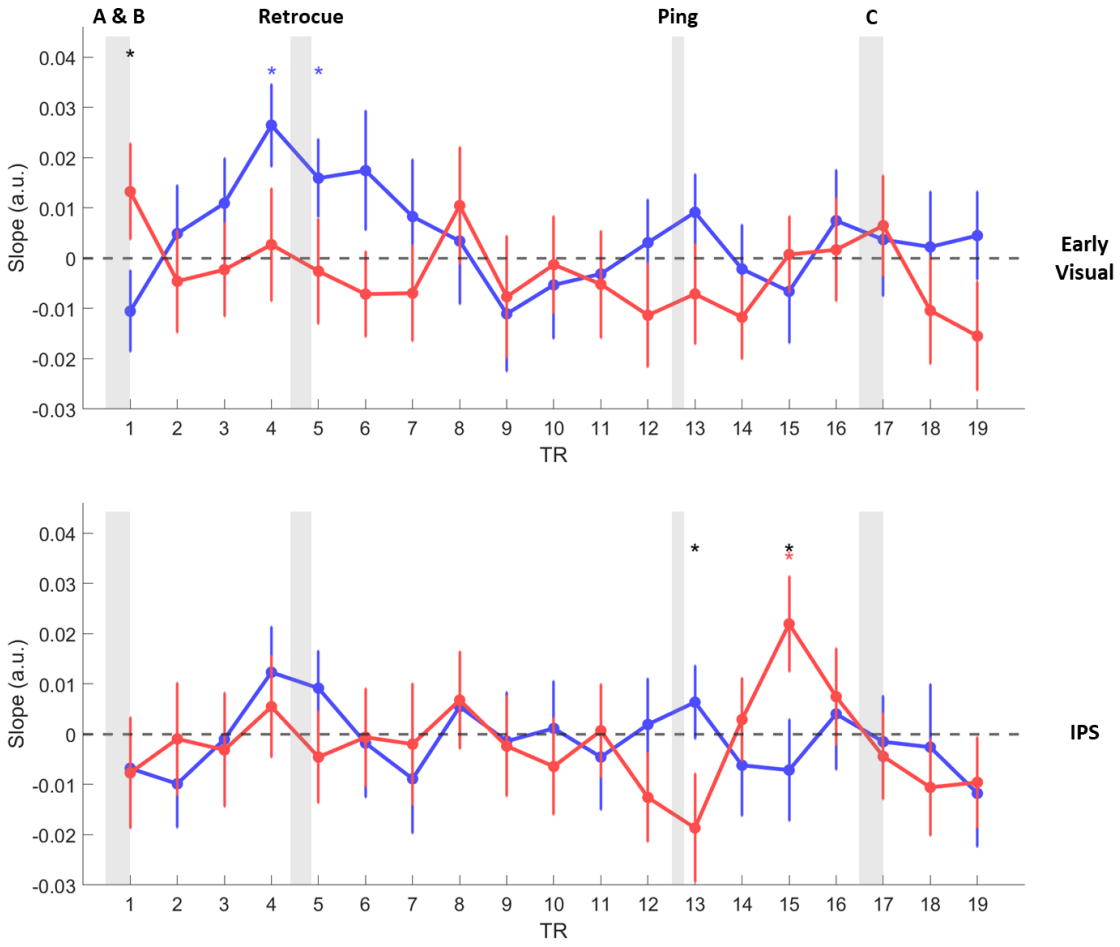

Figure 3.4. The slope of the reconstruction of the item A's orientation in the early visual ROI (top panel) and the IPS ROI (bottom panel) at each TR. The blue and red curves are for the *no-overlap* and the *overlap* conditions, respectively. Error bars indicate SEM across subjects. The significant positive or negative reconstruction for each condition is marked with blue or red asterisks ($p < 0.05$, bootstrapping test), and the significant difference between conditions is marked with black asterisks ($p < 0.05$, bootstrapping test). The time of important events in the trial is marked with gray shading areas with the label of the event shown above. Each TR corresponds to 2 sec.

The reliable reconstruction of the item *A*'s orientation is a prerequisite to test our preregistered hypotheses. Specifically, a successful reconstruction of the item *A* at TR 7

(Hypothesis 1), TR 12 (Hypothesis 2), and/or TRs 15+16 (Hypothesis 3) is needed before one can test the item-*A*-trained IEM on the orientation of the IMI. Because this prerequisite was not met, we were unable to carry out the preregistered analyses. However, because the neural representation of the location of a stimulus is known to have a much higher signal-to-noise ratio than the neural representation of its orientation, it is typically easier to decode stimulus location (c.f., (Yu et al., 2020)). Therefore, rather than abandoning this data set, we chose to carry out a set of exploratory analyses corresponding to the originally planned preregistered analyses, but using stimulus location as an proxy for the representation of the item itself. We reasoned that this would produce interpretable results because the results from several previous studies suggests that information about the location of a stimulus may be bound obligatorily to the representation of that item's identity and, indeed, successful performance on the ABC retrocuing task required the retention of stimulus location, because the location of the recall probe was the only information indicating which item was to be recalled at the end of the trial.

Neural representation of the memory items' locations

We started with reconstructing the location of the item *A* for each condition separately. The IEMs were trained and tested on the item *A*'s location on a TR-by-TR basis. In the early visual ROI, this reconstruction was successful for both conditions for all but the first and last TR of the trial, demonstrating the robustness of the neural representation of location information.

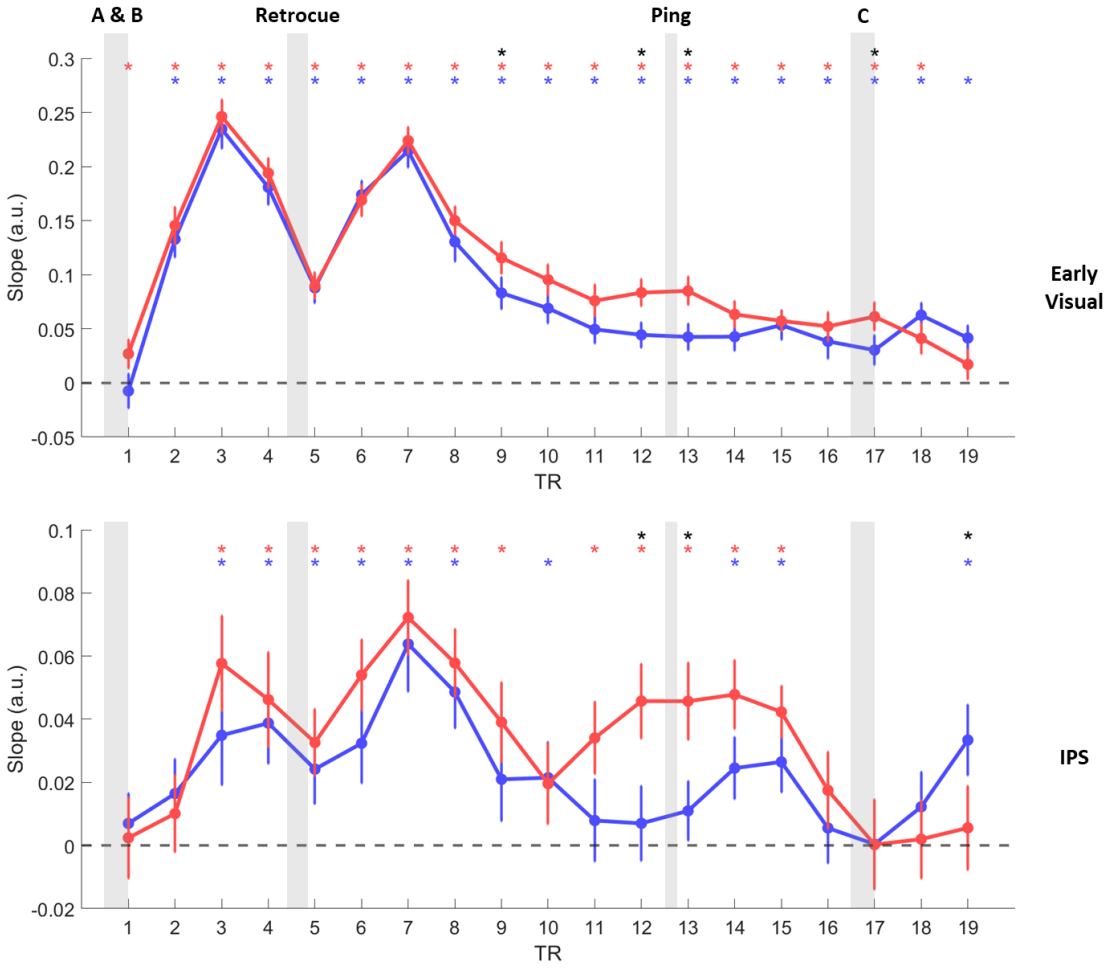

Figure 3.5. The slope of the reconstruction of the item *A*'s location with IEM trained with item A TR by TR in the early visual ROI (top panel) and the IPS ROI (bottom panel). Plotting conventions are the same as figure 3.4.

Note that the representation of the item *A*'s location was numerically stronger in the *overlap* condition comparing to the *no-overlap* condition from the late half of the Delay 2.1 to the onset of item C (which was significant during TR 9, 12, 13 and 17, all ps<0.05; Fig. 3.5 top). A similar effect was also found in the IPS ROI (Fig. 3.5 bottom): In the *overlap* condition, the neural representation of item *A*'s location remained elevated during the late

half of the Delayed 2.1 and in the early ping-evoked response (TR 11 to 15, all ps<0.05), but it was not significantly above the baseline in the *no-overlap* condition during this period (TR 11 to 13). A significant difference between conditions was found during TR 12 and 13 (p = 0.0346 and 0.03). To summarize, in the *overlap* condition where the subjects anticipated to face a higher level of interference, the information of the task-relevant item *A* was kept more firmly during the maintain period.

Next, we studied the neural representation of the IMI location, to see if and how it was handled differently in the *overlap* vs. *no-overlap* condition. In the early visual ROI, the information of the IMI location in the *overlap* condition was significantly above the baseline from TR 2 to TR 9 (i.e., from the onset of the item to the early half of Delay 2.1; Fig. 3.6 top), whereas in the *no-overlap* condition, the neural representation of the IMI location persisted longer (Fig. 3.6 top; TR 2 to 11, until the end of Delay 2.1). A significant difference across conditions was found at TR 10 (p = 0.0294). For the IPS ROI (Fig. 3.6 bottom), the information of IMI location, again, persisted longer in the *no-overlap* condition (TR 3 to 6) than in the *overlap* condition (TR 3 and 4).

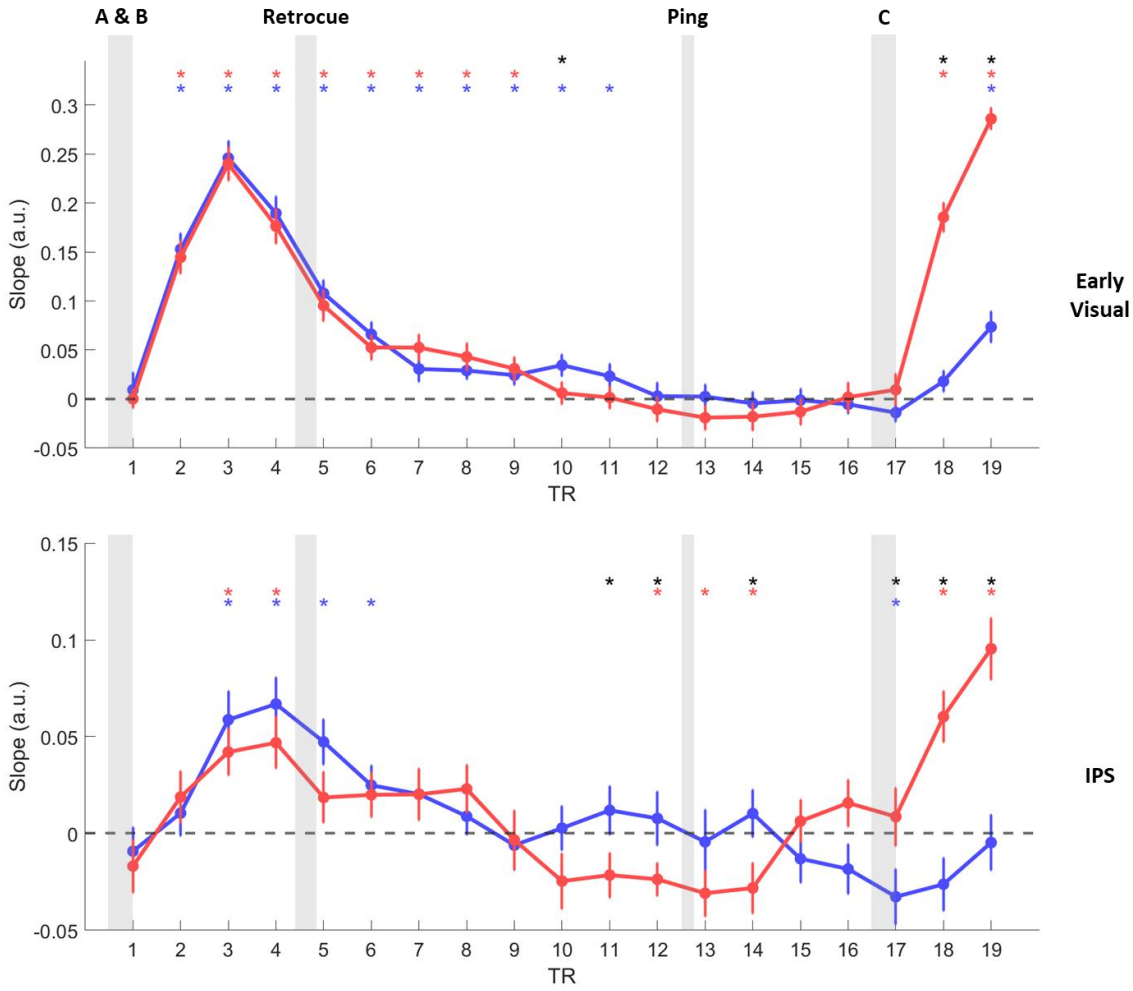

Figure 3.6. The slope of the reconstruction of the IMI's location with IEM trained with IMI TR by TR in the early visual ROI (top panel) and the IPS ROI (bottom panel). Plotting conventions are the same as figure 3.4.

During the late half of the Delay 2.1 and the early ping-evoked responses, in the IPS ROI, the IEM reconstruction of the IMI location showed a negative peak (Fig. 3.6 bottom; TR 12 to 14, all ps<0.05) in the *overlap* blocks. That is, there was a suppressed representation of this location comparing to the non-sample locations that were not occupied by either item *A* or *B* (Fig. 3.7 bottom). In contrast to this, in the *no-overlap* condition, the representation of the

IMI location did not differ from the baseline. The difference between conditions was significant during TR 11, 12 and 14 (all ps<0.05). For the early visual ROI (Fig. 3.6 top), the slope of the *overlap*-condition IMI location's reconstruction was also below zero during this period (TR 12 to 15, and Fig. 3.7 top), but this effect was not statistically significant.

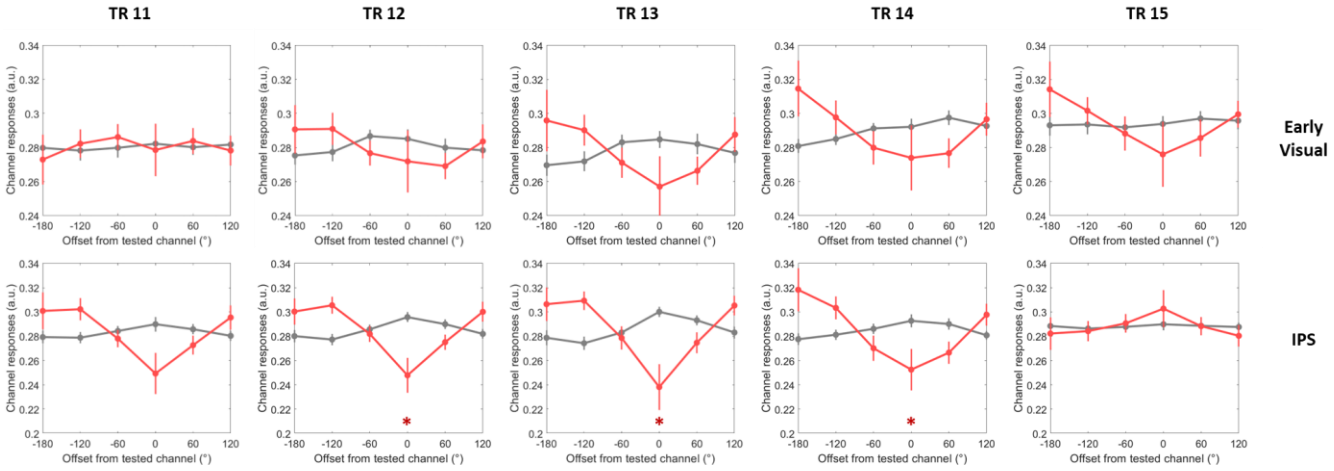

Figure 3.7. IEM reconstructions of the IMI's location in the *overlap* condition from TR 11 to 15, in the early visual ROI (top row) and the IPS ROI (bottom row). The red curves are the reconstructions of the IMI location. The gray curves are the averaged reconstructions of the four non-sample locations (locations that were not occupied by *A* or *B*. i.e., the baseline). Error bars indicate SEM across subjects. Red asterisks indicate the IMI reconstructions that are significantly different from the baseline ($p < 0.05$).

The reconstructions reported above was conducted by training and testing on the same item and on the same TR. They showed whether there was information about the location of the item represented in the BOLD activity, but they do not care about the format or code that was representing this information (e.g., whether it was a "sensory code" or a "memory code"). To study the representation of the IMI in WM, for each condition, we trained one

IEM on the activity from TR 12 based on the label of the item *A*'s location, and used this IEM to reconstruct the IMI's location on all TRs. The TR 12 was the last TR before the onset of the ping, and it was the furthest away from the *A* & *B* presentation (TR 1) and the retrocue (TR 5), both of which should evoke a sensory signal for the location of the item *A*. Thus, the location information of the item *A* is expected to be maintained primarily in a memory code in this TR and should be less impacted by the sensory-driven signal.

This set of analyses were only conducted in the early visual ROI, but not the IPS ROI, as there was not a reliable reconstruction of the item *A*'s location at TR 12 in the IPS ROI (Fig. 3.5 bottom). With this memory-code IEM, there was, in general, a negative reconstruction of the IMI's location after the presentation of item *A* and *B* (i.e., TR 2 and 3; Fig. 3.8 top) under both conditions. This was arguably showing that the sensory-driven signal of the memory items was in stark difference from the memory signal representing the item *A*.

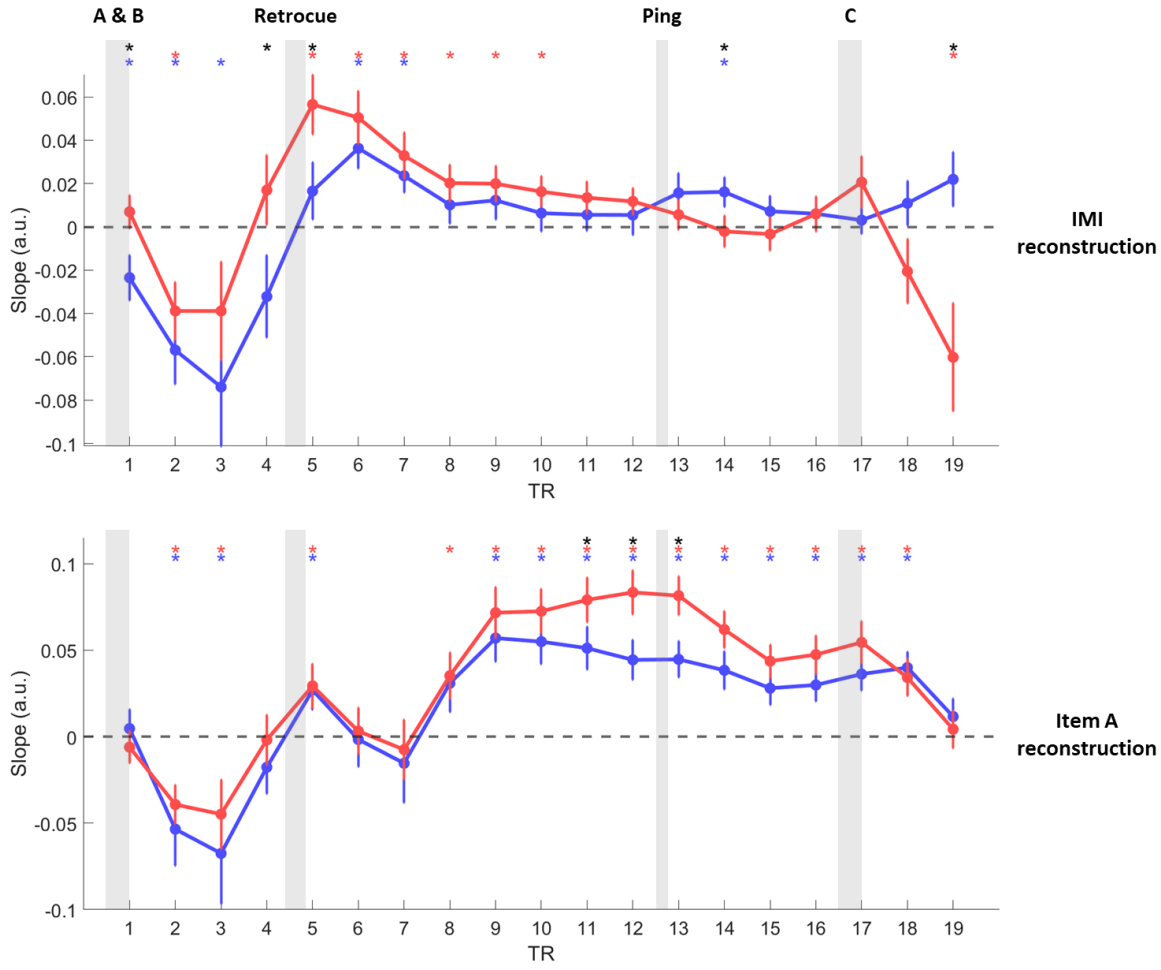

Figure 3.8. The slope of the reconstruction of the location of IMI (top panel) or item A (bottom panel) with the IEM trained with item A at TR 12. Plotting conventions are the same as figure 3.4.

The next important effect to notice is the ping-evoked activity: while there was a reactivation of the IMI's location in the *no-overlap* condition (Fig. 3.8 top; TR 14, p = 0.0204), there was no such reactivation in the *overlap* blocks, and the difference between conditions was significant (TR 14, p = 0.0072).

For completeness, we also reported the reconstruction of the item A's location with this memory-code IEM. The IEM was trained on the item *A*'s location at TR 12 and tested on the

item *A*'s location on all TRs. Similar to reconstruction of the IMI's location, there was a negative reconstruction of the item *A*'s location during TR 2 and 3 under both conditions (Fig. 3.8 bottom), showing the sensory-driven signal representing the item *A*. In the later Delay 2.1 and early ping-evoked response (Fig. 3.8 bottom; TR 11 to 13, all ps < 0.05), the memory representation of the item *A* was stronger in the *overlap* comparing to the *no-overlap* condition, replicating what was found with the TR-by-TR IEMs (Fig. 3.5).

Discussion

This Preregistered Report was aimed to test the predictions of the hijacked adaptation model via applying multivariate reconstruction of the to-be-removed orientation on ping-evoked BOLD signal. However, the IEM failed to reliably reconstruct the orientation of the retrocued item under both conditions during the early and late Delay 2.1, as well as in the ping-evoked response. Thus, the prerequisites of testing the preregistered hypotheses were not met. However, by taking advantage of the superior signal-to-noise ratio of location information in brain activities, we were able use the reconstruction of the memory item's location to study the processing of the retrocued item and the IMI in the *overlap* vs. *no-overlap* condition. We note that the current results on location reconstructions do not provide direct evidence for or against the hijacked adaptation model, as it concerns more about the gain of perceptual circuits representing the memory *content* (here, the orientation of the grating), instead of the *context* (i.e., the location of the grating). However, these results provided valuable insight into how the active removal from WM was implemented from another perspective.

The *overlap* condition is characterized with its higher level of potential interference: subjects were informed before the start of the block that the item *C* would be shown at the same location as the IMI. This shared location caused a potential for cue conflict (Oberauer and Lin, 2017) and may have increased the difficulty of this condition. However, at the behavioral level, the subjects' performance in the *overlap* blocks was similar to their performance in the, arguably easier, *no-overlap* blocks. This may be achieved, partially, by subjects holding on to the information of the item A more firmly in the *overlap* comparing to the *no-overlap* condition. This was evident during a relatively late part of the maintain period (late Delay 2.1 and early ping-evoked response, Fig. 3.5) with a stronger neural representation of the item A's location in the *overlap* condition than the *no-overlap* condition.

With regard to the IMI, which is the main focus of the current study, we found its location was maintained for a shorter period of time after the retrocue onset in *overlap* vs. *no-overlap* blocks (Fig. 3.5). This is consistent with the predictions of the context-breaking model of active removal, in which the active removal is accomplished by breaking the association between an item's *content* (e.g., the orientation of a Gabor patch) and its *context* (e.g., where this patch had appeared on the screen, or the ordinal position within the series in which it had appeared; (Oberauer and Lin, 2017; Lewis-Peacock et al., 2018)). This hypothesized context-breaking operation has the effect of removing the item from WM, and thus removing the active trace (of the *context* and the *content*) needed to successfully reconstruct it with IEM. Thus, the faster fading of the active representation of the IMI's location in the *overlap* blocks could be a result of the context breaking.

Importantly, later in the trial, during the late Delay 2.1 and the early ping-evoked response, we observed the representation of the IMI's location was suppressed below the baseline (as defined by the reconstructions of the non-sample locations; Fig. 3.6 and 3.7). This active suppression cannot be explained by the context-breaking model of active removal, as simply unbinding the *content* and *context* of the IMI should not cause a below-baseline suppression of the representation. This effect suggested a more active process may be underlying the active removal instead of / in addition to what is hypothesized with the context-breaking model.

What is the neural mechanism behind this negative reconstruction of the IMI's location? This cannot be explained simply by a stronger or weaker memory signal/activation of the IMI location alone, and may be accomplished with a modulation of perceptual circuits encoding this information. In a previous study in which a negative reconstruction for the orientation of the removed item was found, Lorenc et al. (2020) used computational modeling to simulate the result of IEM with various parameters for the perceptual circuits encoding the memory item. By changing the gain, the width and the centers of the tuning channels, they were able to find which combination of parameters could best explain the empirical IEM reconstructions they got from human BOLD signal. A similar method could be used on the location reconstructions in the current study to investigate the neural mechanism(s) behind this active suppression.

In previous studies, a pulse of TMS or visual pings has been used to study the activity-silent representation of WM (Wolff et al., 2015; Rose et al., 2016; Wolff et al., 2017). By stimulating the perceptual circuits with a non-specific input, researchers could reveal the

changes in synaptic weights between neurons representing certain information, and this activity-silent trace, if there was any, could be "reactivated" by such stimulation (Barbosa et al., 2020). In the current study, the hijacked adaptation model predicts an erase of the activity-silent representation of the to-be-removed item (Fig. 3.2) when active removal was conducted. This erase of activity-silent traces was supported by the failed reactivation of the IMI's location in the *overlap*, but not the *no-overlap*, condition with the memory-code IEM (Fig. 3.8).

Chapter 4

Conclusion

Despite its importance in the functioning of the cognitive system, how information is removed from WM when it becomes no longer relevant to the task is an often times overlooked question in the field. Most of the empirical and modeling works assumed the removal of information from WM is achieved by simply withdrawing one's attention from that information, and hence letting it passively decay over time (Barbosa et al., 2020; Tsubomi et al., 2024) or be overwritten by the new information (Chatham and Badre, 2013). However, such a passive process inevitably leads to incomplete removal of the outdated information, and thus, leads to interference with the encoding and recall of new information (manifested as, for example, proactive interference (Monsell, 1978) and serial dependence (Fischer and Whitney, 2014)). This problem is even more serious in task designs where the subjects needed to maintain or manipulate more than one memorandum (Rose et al., 2016; Fulvio and Postle, 2020).

In the works presented in the current thesis, I studied if people are able to conduct the removal of information from WM in a more active way, in addition to the default passive removal strategy. Such active removal could be beneficial to prevent the interference caused by the outdated information on the current task. Driven by this assumption, we designed the ABC-retrocuing task with two conditions aimed to evoke different removal strategies: an *overlap* condition with higher potential interference between the to-be-removed and the new information and a *no-overlap* condition with lower potential interference. In a behavioral study with the ABC-retrocuing task (Shan and Postle, 2022), we found that people indeed handled the to-be-removed IMI differently across the two conditions, and the active removal led to a reversed behavioral consequence comparing to the passive removal: The actively removed IMI caused a repulsive serial bias on the report

in the next trial, instead of an attractive one. This repulsive serial bias effect suggested the active removal may have been achieved by a neural mechanism that is reminiscent of the sensory adaptation effect, which is commonly believed to be underlying the repulsive serial dependence (Fritsche et al., 2017; Pascucci et al., 2019; Fritsche et al., 2020; Sheehan and Serences, 2022). Based on this idea, as well as the findings of Lorenc et al. (2020) in which the IEM reconstruction of a memorandum became reversed after it was removed from WM, we proposed the hijacked adaptation model of active removal of information from WM. The hijacked adaptation model posits that the active removal is achieved by a top-down control signal that modulates the gain of perceptual circuits responsible for encoding the memorandum, such that the gain of channels tuned to the to-be-removed information is reduced.

We tested the hijacked adaptation model in an EEG study (Chapter 2) and an fMRI study (Chapter 3). Specifically, we focused on two components of the hijacked adaptation model: the implementation of active model, which is assumed to be supported by a top-down control signal, and the consequence of it, which is the down modulation of gain centered on the to-be-removed information.

Regarding the implementation, consistent with our hypotheses, we found the active removal was accompanied by a stronger central-midline ERP and an alpha-beta band traveling wave that propagated in the anterior-to-posterior direction, supporting the top-down signaling. The active removal of the IMI was also accompanied with an active suppression of the neural representation of the IMI's location.

With regard to the consequence of active removal, we found reduced excitability of the perceptual circuits (as measured by the ERP and forward traveling waves evoked by an

orientation-neutral visual ping), as predicted by the hijacked adaptation model. Studying the neural representation of the memorandum's location in the ping-evoked response, there was a reactivation of the IMI's location when passive removal was used, but no such reactivation with active removal. This suggested the active removal successfully erased the activity-silent synaptic traces of the IMI's location. At last, the active removal also influenced the subjects' behavioral reports in a way predicted by the hijacked adaptation model, as assessed with TCC modeling.

To summarize, our works provided behavioral and neural evidence that people can conduct the removal of outdated information from WM in an active way. This is accomplished by active top-down control and results in modulations of perpetual circuits responsible for the encoding/maintaining of the memoranda.

References

Ahad NA, Abdullah S, Lai CH, Mohd Ali N (2012) Relative power performance of t-test and bootstrap procedure for two-sample. Pertanika Journal of Science & Technology 20:43-52.

Alamia A, VanRullen R (2019) Alpha oscillations and traveling waves: Signatures of predictive coding? PLOS Biology 17:e3000487.

Alamia A, Terral L, D'Ambra MR, VanRullen R (2023) Distinct roles of forward and backward alpha-band waves in spatial visual attention. eLife 12:e85035.

Bae G-Y, Luck SJ (2019) Reactivation of previous experiences in a working memory task. Psychological science 30:587-595.

Barbosa J, Stein H, Martinez RL, Galan-Gadea A, Li S, Dalmau J, Adam KCS, Valls-Solé J, Constantinidis C, Compte A (2020) Interplay between persistent activity and activity-silent dynamics in the prefrontal cortex underlies serial biases in working memory. Nature Neuroscience 23:1016-1024.

Barlow HB (1961) Possible principles underlying the transformation of sensory messages. In: Sensory Communication (W WR, Sensory Communication. Cambridge MMITPpDhdom, 9780262518420.003.0013, eds). Cambridge, MA: MIT Press.

Bastos André M, Vezoli J, Bosman Conrado A, Schoffelen J-M, Oostenveld R, Dowdall Jarrod R, De Weerd P, Kennedy H, Fries P (2015) Visual Areas Exert Feedforward and Feedback Influences through Distinct Frequency Channels. Neuron 85:390-401.

Beukers AO, Buschman TJ, Cohen JD, Norman KA (2021) Is Activity Silent Working Memory Simply Episodic Memory? Trends in Cognitive Sciences 25:284-293.

Bliss DP, Sun JJ, D'Esposito M (2017) Serial dependence is absent at the time of perception but increases in visual working memory. Scientific reports 7:14739.

Brainard DH (1997) The Psychophysics Toolbox. Spatial Vision 10:433-436.

Braver TS (2012) The variable nature of cognitive control: a dual mechanisms framework. Trends in cognitive sciences 16:106-113.

Burgess GC, Braver TS (2010) Neural mechanisms of interference control in working memory: effects of interference expectancy and fluid intelligence. PloS one 5:e12861.

Cai Y, Sheldon AD, Yu Q, Postle BR (2019) Overlapping and distinct contributions of stimulus location and of spatial context to nonspatial visual short-term memory. Journal of Neurophysiology 121:1222-1231.

Cai Y, Fulvio JM, Samaha J, Postle BR (2022) Context Binding in Visual Working Memory Is Reflected in Bilateral Event-Related Potentials, But Not in Contralateral Delay Activity. eneuro 9:ENEURO.0207-0222.2022.

Chatham CH, Badre D (2013) Working memory management and predicted utility. Frontiers in Behavioral Neuroscience 7:83.

Clifford CWG, Wenderoth P, Spehar B (2000) A functional angle on some after-effects in cortical vision. Proceedings of the Royal Society of London Series B: Biological Sciences 267:1705–1710.

D'Esposito M, Postle BR, Jonides J, Smith EE (1999) The neural substrate and temporal dynamics of interference effects in working memory as revealed by event-related functional MRI. Proceedings of the National Academy of Sciences 96:7514-7519.

Das A, Menon V (2021) Asymmetric Frequency-Specific Feedforward and Feedback Information Flow between Hippocampus and Prefrontal Cortex during Verbal Memory Encoding and Recall. The Journal of Neuroscience 41:8427.
Das A, Menon V (2022) Replicable patterns of causal information flow between hippocampus and prefrontal cortex during spatial navigation and spatial–verbal memory formation. Cerebral Cortex 32:5343-5361.
Delorme A, Makeig S (2004) EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. Journal of Neuroscience Methods 134:9-21.
Ester EF, Sprague TC, Serences JT (2015) Parietal and frontal cortex encode stimulus-specific mnemonic representations during visual working memory. Neuron 87:893-905.
Ester EF, Sutterer DW, Serences JT, Awh E (2016) Feature-selective attentional modulations in human frontoparietal cortex. J Neurosci 36:8188-8199.
Feredoes E, Postle BR (2010) Prefrontal control of familiarity and recollection in working memory. Journal of Cognitive Neuroscience 22:323-330.
Feredoes E, Tononi G, Postle BR (2006) Direct evidence for a prefrontal contribution to the control of proactive interference in verbal working memory. Proceedings of the National Academy of Sciences 103:19530-19534.
Fischer J, Whitney D (2014) Serial dependence in visual perception. Nature neuroscience 17:738-743.
Fries P (2015) Rhythms for Cognition: Communication through Coherence. Neuron 88:220-235.
Fritsche M, Mostert P, de Lange FP (2017) Opposite effects of recent history on perception and decision. Current Biology 27:590-595.
Fritsche M, Spaak E, De Lange FP (2020) A Bayesian and efficient observer model explains concurrent attractive and repulsive history biases in visual perception. Elife 9:e55389.
Fulvio JM, Postle BR (2020) Cognitive Control, Not Time, Determines the Status of Items in Working Memory. Journal of Cognition 3.
Gardner JL, Liu T (2019) Inverted encoding models reconstruct an arbitrary model response, not the stimulus. eNeuro 6.
Gibson JJ, Radner M (1937) Adaptation, after-effect and contrast in the perception of tilted lines. I. quantitative studies. J Exp Psychol 20:453–467.
Jin DZ, Dragoi V, Sur M, Seung HS (2005) Tilt Aftereffect and Adaptation-Induced Changes in Orientation Tuning in Visual Cortex. Journal of Neurophysiology 94:4038-4050.
Jonides J, Schumacher EH, Smith EE, Lauber EJ, Awh E, Minoshima S, Koeppe RA (1997) Verbal Working Memory Load Affects Regional Brain Activation as Measured by PET. Journal of Cognitive Neuroscience 9:462-475.
Kim H, Smolker HR, Smith LL, Banich MT, Lewis-Peacock JA (2020) Changes to information in working memory depend on distinct removal operations. Nature Communications 11:6239.
LaRocque JJ, Lewis-Peacock JA, Drysdale AT, Oberauer K, Postle BR (2013) Decoding attended information in short-term memory: an EEG study. Journal of cognitive neuroscience 25:127-142.

Lepsien J, Nobre AC (2006) Cognitive control of attention in the human brain: Insights from orienting attention to mental representations. Brain research 1105:20-31.
Lewis-Peacock JA, Postle BR (2012) Decoding the internal focus of attention. Neuropsychologia 50:470-478.
Lewis-Peacock JA, Norman KA (2014) Competition between items in working memory leads to forgetting. Nature Communications 5:5768.
Lewis-Peacock JA, Kessler Y, Oberauer K (2018) The removal of information from working memory. Annals of the New York Academy of Sciences 1424:33-44.
Lewis-Peacock JA, Drysdale AT, Oberauer K, Postle BR (2012) Neural evidence for a distinction between short-term memory and the focus of attention. J Cog Neuroscience 24:61-79.
Liu T, Cable D, Gardner JL (2018) Inverted encoding models of human population response conflate noise and neural tuning width. J Neurosci 38:398–408.
Lorenc ES, Vandenbroucke ARE, Nee DE, de Lange FP, D'Esposito M (2020) Dissociable neural mechanisms underlie currently-relevant, future-relevant, and discarded working memory representations. Scientific Reports 10:11195.
Lundqvist M, Miller EK, Nordmark J, Liljefors J, Herman P (2024) Beta: bursts of cognition. Trends in Cognitive Sciences 28:662-676.
Luo C, Ester EF (2024) Traveling waves link human visual and frontal cortex during working memory-guided behavior. bioRxiv:2024.2003.2012.584543.
Ma WJ, Husain M, Bays PM (2014) Changing concepts of working memory. Nature neuroscience 17:347.
Marr D, Poggio T (1976) From understanding computation to understanding neural circuitry. Artificial Intelligence Laboratory AI Memo Massachusetts Institute of Technology:hdl:1721.1721/5782. AIM-1357.
Mohan UR, Zhang H, Ermentrout B, Jacobs J (2024) The direction of theta and alpha travelling waves modulates human memory processing. Nature Human Behaviour 8:1124-1135.
Monsell S (1978) Recency, immediate recognition memory, and reaction time. Cognitive Psychology 10:465-501.
Muller L, Chavane F, Reynolds J, Sejnowski TJ (2018) Cortical travelling waves: mechanisms and computational principles. Nature Reviews Neuroscience 19:255-268.
Norman KA, Newman EL, Detre G (2007) A neural network model of retrieval-induced forgetting. Psychological Review 114:887–953.
Oberauer K, Lin H-Y (2017) An interference model of visual working memory. Psychological review 124:21.
Oostenveld R, Fries P, Maris E, Schoffelen J-M (2011) FieldTrip: Open Source Software for Advanced Analysis of MEG, EEG, and Invasive Electrophysiological Data. Computational Intelligence and Neuroscience 2011:156869.
Pascucci D, Mancuso G, Santandrea E, Della Libera C, Plomp G, Chelazzi L (2019) Laws of concatenated perception: Vision goes for novelty, decisions for perseverance. PLOS Biology 17:e3000144.
Pelli DG (1997) The VideoToolbox software for visual psychophysics: transforming numbers into movies. Spatial Vision 10:437-442.
Pietrelli M, Samaha J, Postle BR (2022) Spectral Distribution Dynamics across Different Attentional Priority States. The Journal of Neuroscience 42:4026.

Pion-Tonachini L, Kreutz-Delgado K, Makeig S (2019) ICLabel: An automated electroencephalographic independent component classifier, dataset, and website. NeuroImage 198:181-197.
Postle BR, Berger JS, Goldstein JH, Curtis CE, D'Esposito M (2001) Behavioral and neurophysiological correlates of episodic coding, proactive interference, and list length effects in a running span verbal working memory task. Cognitive, Affective, & Behavioral Neuroscience 1:10-21.
Rose NS, LaRocque JJ, Riggall AC, Gosseries O, Starrett MJ, Meyering EE, Postle BR (2016) Reactivation of latent working memories with transcranial magnetic stimulation. Science 354:1136-1139.
Sahan MI, Sheldon AD, Postle BR (2020) The neural consequences of attentional prioritization of internal representations in visual working memory. Journal of Cognitive Neuroscience 32:917-944.
Samaha J, Switzky M, Postle BR (2019) Confidence boosts serial dependence in orientation estimation. Journal of vision 19:25-25.
Schurgin MW, Wixted JT, Brady TF (2020) Psychophysical scaling reveals a unified theory of visual memory strength. Nature Human Behaviour 4:1156-1172.
Shan J, Postle BR (2022) The Influence of Active Removal from Working Memory on Serial Dependence. Journal of Cognition.
Sheehan TC, Serences JT (2022) Attractive serial dependence overcomes repulsive neuronal adaptation. PLOS Biology 20:e3001711.
Sprague TC, Ester EF, Serences JT (2016) Restoring latent visual working memory representations in human cortex. Neuron 91:694-707.
Sprague TC, Boynton GM, Serences JT (2019) The importance of considering model choices when interpreting results in computational modeling. eNeuro 6:ENEURO.0196-0119.2019.
Sprague TC, Adam KCS, Foster JJ, Rahmati M, Sutterer DW, Vo VA (2018) Inverted encoding models assay population-level stimulus representations, not single-unit neural tuning. eNeuro 5.
Suchow JW, Brady TF, Fougnie D, Alvarez GA (2013) Modeling visual working memory with the MemToolbox. Journal of Vision 13:9-9.
Teng C, Fulvio JM, Jiang J, Postle BR (2022) Flexible top-down control in the interaction between working memory and perception. Journal of Vision 22:3-3.
Teng C, Fulvio JM, Pietrelli M, Jiang J, Postle BR (2023) Temporal dynamics and representational consequences of the control of processing conflict between visual working memory and visual perception. bioRxiv:2023.2012.2007.570647.
Thut G, Nietzel A, Brandt SA, Pascual-Leone A (2006) Alpha-band electroencephalographic activity over occipital cortex indexes visuospatial attention bias and predicts visual target detection. The Journal of neuroscience : the official journal of the Society for Neuroscience 26:9494-9502.
Trapp S, Pascucci D, Chelazzi L (2021) Predictive brain: Addressing the level of representation by reviewing perceptual hysteresis. Cortex 141:535-540.
Tsubomi H, Fukuda K, Kikumoto A, Mayr U, Vogel EK (2024) Task Termination Triggers Spontaneous Removal of Information From Visual Working Memory. Psychological Science 35:995-1009.

Wan Q, Menendez JA, Postle BR (2022) Priority-based transformations of stimulus representation in visual working memory. PLOS Computational Biology 18:e1009062.
Wan Q, Cai Y, Samaha J, Postle BR (2020) Tracking stimulus representation across a 2-back visual working memory task. Royal Society Open Science 7:190228.
Wang L, Mruczek REB, Arcaro MJ, Kastner S (2015) Probabilistic maps of visual topography in human cortex. Cerebral Cortex 25:3911-3931.
Wang TH, Placek K, Lewis-Peacock JA (2019) More Is less: Increased processing of unwanted memories facilitates forgetting. The Journal of Neuroscience 39:3551–3560.
Wessel JR, Anderson MC (2024) Neural mechanisms of domain-general inhibitory control. Trends in Cognitive Sciences 28:124-143.
Wolff MJ, Ding J, Myers NE, Stokes MG (2015) Revealing hidden states in visual working memory using electroencephalography. Frontiers in systems neuroscience 9.
Wolff MJ, Jochim J, Akyürek EG, Stokes MG (2017) Dynamic hidden states underlying working-memory-guided behavior. Nature Neuroscience.
Xu Y, Long X, Feng J, Gong P (2023) Interacting spiral wave patterns underlie complex brain dynamics and are related to cognitive processing. Nature Human Behaviour 7:1196-1215.
Yu Q, Teng C, Postle BR (2020) Different states of priority recruit different neural representations in visual working memory. PLOS Biology 18:e3000769.
Zeng Y, Sauseng P, Alamia A (2024) Alpha Traveling Waves during Working Memory: Disentangling Bottom-Up Gating and Top-Down Gain Control. The Journal of Neuroscience 44:e0532242024.
Zhang Z, Lewis-Peacock J (2023) Prioritization Sharpens Working Memories but Does Not Protect Them From Distraction. Journal of Experimental Psychology: General 152.
Zhang Z, Lewis-Peacock JA (2024) Reactivation of prior responses drives serial dependence and stabilizes working memory representations. The Journal of Neuroscience:e2399232024.