HHS Public Access

Author manuscript

Neuron. Author manuscript; available in PMC 2017 August 03.

Published in final edited form as:

Neuron. 2016 August 3; 91(3): 694–707. doi:10.1016/j.neuron.2016.07.006.

Restoring latent visual working memory representations in human cortex

Thomas C. Sprague[*,1], **Edward F. Ester**[2], and **John T. Serences**[*,1,2,3]

[1]Neurosciences Graduate Program, University of California, San Diego, La Jolla, California 92093

[2]Department of Psychology, University of California, San Diego, La Jolla, California 92093

[3]Kavli Institute for Brain and Mind, University of California, San Diego, La Jolla, California 92093

Abstract

Working memory (WM) enables the storage and manipulation of limited amounts of information over short periods. Prominent models posit that increasing the number of remembered items decreases the spiking activity dedicated to each item via mutual inhibition, which irreparably degrades the fidelity of each item's representation. We tested these models by determining if degraded memory representations could be recovered following a post-cue indicating which of several items in spatial WM would be recalled. Using an fMRI-based image reconstruction technique, we identified impaired behavioral performance and degraded mnemonic representations with elevated memory load. However, in several cortical regions, degraded mnemonic representations recovered substantially following a post-cue, and this recovery tracked behavioral performance. These results challenge pure spike-based models of WM and suggest that remembered items are additionally encoded within latent or hidden neural codes that can help reinvigorate active WM representations.

Introduction

In many visual tasks, an observer's ability to accurately represent information declines rapidly as the complexity of the scene increases (Franconeri et al., 2013; Tsubomi et al., 2013). These processing limits are highlighted in working memory (WM) tasks, which require the maintenance and manipulation of sensory information no longer physically present in the environment (Baddeley and Hitch, 1974; Bays, 2015; Curtis and D'Esposito, 2003; D'Esposito and Postle, 2014; Gazzaley and Nobre, 2012; Luck and Vogel, 2013; Ma et al., 2014; Sreenivasan et al., 2014; Stokes, 2015). In these tasks, increasing the amount of information stored in WM leads to impaired performance when recalling visual features

[*]Corresponding authors Thomas C. Sprague, Neurosciences Graduate Program, University of California, San Diego, 9500 Gilman Dr., 92093, tsprague@nyu.edu Or John T. Serences, Department of Psychology, Neurosciences Graduate Program, Kavli Institute for Brain and Mind, University of California, San Diego, 9500 Gilman Dr., 92093, jserences@ucsd.edu.

Author contributions: TCS, EFE and JTS designed the experiment and wrote the manuscript; TCS and EFE acquired data; TCS analyzed data.

(Bays and Husain, 2008; Bays, 2015, 2014; Keshvari et al., 2013; Ma et al., 2014; Zhang and Luck, 2008).

Influential models propose that WM representations are actively maintained by sustained spiking activity in neural populations (Funahashi et al., 1989; Fuster and Alexander, 1971). Recently, WM representations have also been found in fMRI activation patterns (Harrison and Tong, 2009; Serences et al., 2009) and the pattern of EEG alpha-band potentials (Foster et al., 2015). Impaired performance with increasing WM load is accompanied by lower spike rates related to relevant memoranda in macaques, or by a diminished ability to differentiate fMRI activation patterns tied to different remembered items in humans (Buschman et al., 2011; Emrich et al., 2013; Landman et al., 2003a; Matsushima and Tanaka, 2014; Sprague et al., 2014). Importantly, the fidelity of fMRI activation patterns is tied to behavioral performance on WM tasks (Albers et al., 2013; Emrich et al., 2013; Ester et al., 2013; Reinhart et al., 2012; Sprague et al., 2014),.

According to one model, impairments in WM performance with load are due to mutually suppressive interactions between neural representations of individual items that result in degraded spiking representations for each item (Bays, 2015, 2014; Carandini and Heeger, 2012; Franconeri et al., 2013, see also Edin et al., 2009). This, in turn, results in an irreversible loss of information encoded by active spiking representations because representations are more susceptible to noise as spike rates decrease (Bays, 2014). This loss of information is permanent, as information cannot recover with any type of additional processing (Cover and Thomas, 1991; Saproo and Serences, 2014, 2010; Shannon, 1948; Sprague et al., 2015). For example, applying multiplicative gain to a noisy representation (after encoding) would amplify noise to the same extent it amplifies signal, resulting in a higher overall firing rate, but no increase in the information content of the population response.

However, the notion that increasing the number of items in WM leads to an irreversible degradation of neural representations is complicated by findings that cueing participants during the delay period with a retrospective cue (retro-cue) improves performance (Griffin and Nobre, 2003; Landman et al., 2003b; LaRocque et al., 2015; Makovski and Jiang, 2007; Matsukura et al., 2007). While these results hint that active neural WM representations may improve following retro-cues, another possibility is that a retrocue prevents representations from decaying or improves access to static representations (e.g., via attention-related mechanisms) while leaving WM representations unchanged. Without a quantitative assay of the fidelity of neural representations of remembered items, it is difficult to discriminate between these possibilities.

In the current study, we hypothesized that behavioral retro-cue benefits are observed because *active* WM representations – those that are reflected in elevated firing rates and/or sustained BOLD fMRI response patterns – can be augmented using information encoded via *latent* or activity-silent codes (Stokes, 2015; Stokes et al., 2013; Wolff et al., 2015). Such latent codes could include subthreshold membrane potential depolarization, changes in synaptic strength and/or efficacy (Briggs et al., 2013; Erickson et al., 2010), item-related fluctuations of pre-synaptic calcium concentration (Mongillo et al., 2008), changes in correlated variability

between pairs of neurons (Jeanne et al., 2013), hippocampal-dependent long term memory (Squire and Wixted, 2011), or some combination thereof.

In this framework, retro-cues improve memory performance by facilitating recovery of representations from sources of information that are each invisible to common neural measures such as spike rate or BOLD activation level. For example, a set of neurons carrying a latent WM representation in the form of elevated subthreshold membrane potential without a change in spike rate could be activated by input from other neurons, allowing the latent representation to improve the fidelity of an active (spiking) representation. While previous work has identified initial evidence for such latent representations of category-level information (LaRocque et al., 2013; Lepsien and Nobre, 2007; Lewis-Peacock et al., 2012), it remains unknown how the relative fidelity of each item's representation is updated after presentation of a retro-cue, and how those representations are related to behavioral performance on a task requiring high-precision maintenance of feature values.

This hypothesis makes several predictions. First, improvements in memory performance following a retro-cue should be accompanied by recovery of an active neural WM representation. Second, the degree to which latent information facilitates the restoration of active neural representations may co-vary with behavioral performance. However, an alternative hypothesis is that retro-cues enhance access to otherwise stable representations, which predicts no change in the fidelity of neural representations. Critically, discriminating between these requires directly evaluating the fidelity of active WM representations in a stimulus-referred feature space (Sprague et al., 2015).

We tested these predictions using a task where participants precisely maintained the spatial positions of 1 or 2 items in visual WM. On some trials, we presented a retro-cue midway through the delay validly cueing which item was relevant for behavior; on the remainder of trials we presented a non-informative neutral retro-cue. Consistent with previous results, behavioral performance and neural WM representations each degraded when more items were remembered (Emrich et al., 2013; Sprague et al., 2014). However, the retro-cue substantially improved behavioral performance and neural WM representations. Together, these results demonstrate that degraded WM representations can recover, requiring the existence of information within latent neural codes that can support improved WM performance.

Results

We tested the fidelity of WM representations using behavioral and neural measures while participants performed a retro-cued spatial recall task. Participants held 1 (Remember 1, R1) or 2 (Remember 2, R2) items from a 2-item display in spatial WM as indicated by the initial color of a central fixation point over a 16 s delay period. On some R2 trials, we changed the color of the fixation point to provide either an informative "valid" (R2-valid) or an uninformative "neutral" (R2-neutral) retro-cue at the end of the first half of the delay interval that indicated which item(s) might be cued for recall at the end of the entire delay interval (Fig. 1A). We used 100% valid retro-cues to ensure that participants utilized the cue to

optimize behavior. At the end of the delay period, participants recalled the exact horizontal or vertical position of one of the items by adjusting vertical or horizontal response bar, respectively (Fig. 1A). On R1 and R2-valid trials, only the probed item required active maintenance in WM during the second delay period, and on R2-neutral trials, we randomly selected which of the two remembered items would be queried for recall. Note that R2-neutral and R2-valid trials were identical during the first delay period, and differed only during the second delay period following the cue (Fig. 1A). The R2-valid condition allowed us to assess how a retro-cue influences behavioral performance and neural representational fidelity compared to when both items were remembered in the R2-neutral condition. We pseudo-randomly chose target positions from an array of 6 invisible discs that were spaced equally along ring 3.5° from fixation and which were rotated around fixation on every trial (Fig. 1B). We randomly positioned targets within each disc to discourage discrete or verbal encoding strategies (e.g. “8 o’clock”; Sprague et al., 2014). Each participant ($n = 6$) completed three 2-hr fMRI scan sessions (324 to 378 total trials per participant). One participant previously completed several spatial WM scan sessions as part of a previous study (Sprague et al., 2014), though all results generalized when excluding this participant from analyses (data not shown). Our behavioral measure was the distance between the response bar and the relevant target at the conclusion of a 3 s response period.

Behavioral performance improved with a retro-cue

Participants performed more poorly, as indicated by higher average recall error, on R2-neutral trials as compared to R1 trials (Fig. 1C; R1 vs. R2-neutral: $p < 0.001$, resampling test, see Experimental Procedures). This drop in recall accuracy is consistent with degraded neural representations that accompany increasing WM loads, and replicates previous findings (Sprague et al., 2014; see also: Bays and Husain, 2008; Bays, 2014; Emrich et al., 2013; Zhang and Luck, 2008). When one item was cued midway through the delay interval (R2-valid trials), behavioral performance improved as compared to R2-neutral trials (R2-neutral vs. R2-valid: $p = 0.008$). Performance was slightly worse on R2-valid trials compared to R1 trials (R1 vs. R2-valid: $p = 0.01$), suggesting substantial but imperfect recovery of WM representations with valid cues, again consistent with previous findings (Griffin and Nobre, 2003; Landman et al., 2003b; Makovski and Jiang, 2007; Matsukura et al., 2007).

Reconstructing WM representations

To isolate and assess the information content of WM representations from alternative mechanisms, such as changes in response conflict or cue-related improvements in selection of stable representations, we implemented an inverted encoding model (IEM) of visual space to reconstruct images of the contents of spatial WM based on BOLD fMRI activation patterns (Brouwer and Heeger, 2009; Ester et al., 2015, 2013; Sprague and Serences, 2013; Sprague et al., 2015, 2014). We computed reconstructions in each of 10 independently-identified regions of interest (ROIs) we have studied previously: retinotopic occipital visual areas (V1-hV4; V3A), retinotopic areas areas along the intraparietal sulcus (IPS0-IPS3), and the superior precentral sulcus (sPCS; thought to be a human homolog to macaque frontal eye fields; the sPCS ROI was identified using an independent spatial WM localizer task; see Experimental Procedures and Srimal and Curtis, 2008). We also assayed representations

encoded by the joint activation pattern across all these regions after concatenating voxels from all areas before computing reconstructions ("All ROIs combined").

In brief, this analysis involves first estimating the spatial sensitivity profile for each voxel as a weighted sum of a discrete set of modeled information channels using activation measured during a set of 'mapping' scans reserved for this purpose (Fig. 2A; Fig. S3A). Then, once this encoding model is estimated across all voxels within a region, the model can be inverted and used to transform activation patterns measured during the WM task into images of the contents of WM, given the previously-estimated encoding model and activation pattern (Fig. 2B; Ester et al., 2015, 2013; Sprague and Serences, 2013; Sprague et al., 2015, 2014). Finally, despite each trial containing unique positions in WM, they can all be averaged by rotating and aligning all reconstructed images. We call the resulting light spots in these reconstructions "target representations" (Fig. S3D). Importantly, because the computed IEM is constant across trials and time points within a ROI, any observed differences in WM reconstructions must reflect changes in activation patterns as represented in the modeled information space.

Reconstructions track the dynamic contents of spatial WM

First, we evaluated whether WM reconstructions tracked remembered position(s). We plotted WM reconstructions computed using activation patterns from each time point during the trial (0–20.25 s) averaged over all trials with similar WM target arrangements within each WM condition (colored discs in Fig. 1B). We combined trials with similar relative target arrangements, and rotated reconstructions to align all similar trials (see Supplemental Experimental Procedures, Fig. S3).

On R1 trials, reconstructions computed using an early time point (4.50 s) contained representations of both targets (Fig. 3A). However, shortly thereafter, only the relevant target (yellow dashed circle) remained visible (6.75–18.00 s). While the target representation became less pronounced over the duration of the trial, it remained visible throughout the late delay interval. The same pattern held for R2- neutral trials (Fig. 3B): representations of items maintained in WM persisted in reconstructions through the late delay period, though target representations were weaker than those in R1 trials. On R2-valid trials (Fig. 3C), we observed a transition from two simultaneous target representations (early delay) to one target representation (late delay) following the cue, confirming that these WM reconstructions tracked the dynamically changing contents of WM over extended delay intervals. Furthermore, the representation of the cued item during the late delay period appeared enhanced compared to each of the 2 target representations earlier in the delay period (after the encoding transient subsides at ~9.00 s), consistent with enhanced WM representations following a retro-cue.

For several subsequent analyses of WM reconstructions, we focused on average reconstructions during the first delay period (Delay 1; 6.75–9.00 s) and the second delay period (Delay 2; 15.75–18.00 s). When we binned trials by the relative position of WM targets (Fig. 1B), target representations always appeared nearby and only in the position(s) corresponding to the remembered item(s) during that condition and delay period (Fig. 4). Additionally, the quality of target representations always exhibited the same pattern across

delay periods regardless of target arrangement – during the first delay period, representations degraded when 2 items were maintained (Fig. 4A compared to Fig. 4C,E), and during the second delay period, a valid cue restored the cued representation to a high-fidelity state (Fig. 4E–F).

Fidelity of WM target representations

To quantify the robustness of target representations in each ROI, we computed reconstructions over an annulus around fixation, resulting in a 1-d reconstruction as a function of polar angle (Fig. 2C; Supplemental Experimental Procedures).

First, we plotted these rotated and aligned 1-d reconstructions as a function of time (Fig. 5A). On R1 and R2-neutral trials, an initially high-fidelity representation during WM encoding subsided, but remained present in many ROIs throughout Delay 2 (e.g., V3A; IPS0). On R2-valid trials, the cued item was robust even at late time points during Delay 2, often increasing in fidelity following the cue (R2-valid, compare early and late time points, e.g. V1).

To determine the strength of a WM representation in these 1-d polar angle reconstructions, we defined a representational fidelity metric as the vector mean of a set of unit vectors pointing in each polar angle direction, weighted by the reconstruction activation for the corresponding polar angle and projected on a unit vector pointing in the WM target direction (here, always 0° polar angle, because we rotate all 1-d reconstructions to a common center; Fig. 2C; Supplemental Experimental Procedures: Eq 5). If this metric is reliably greater than 0 (statistically evaluated using a resampling procedure, see Experimental Procedures), then there is a consistently identifiable WM target representation in the corresponding reconstruction. If the reconstruction has a uniform activation profile, then this metric will be indistinguishable from 0. WM target representational fidelity gradually decreased over time on R1 and R2-netural trials, but substantially recovered following the valid cue on R2-valid cue trials (e.g., V1; Fig. 5B).

Next, we compared 1-d polar angle reconstructions and representational fidelity during each delay period (Fig. 6). Importantly, we found significant representational fidelity in all ROIs across both delay intervals on R1 and R2-valid cue conditions ($p \leq 0.001$; one-tailed resampling test against 0, FDR corrected, see Experimental Procedures; all p-values for all reported comparisons available in Supplemental Tables; maximum p-value IPS3, R2-valid, Delay 2). On R2-neutral trials we found representations in all ROIs during Delay 1 ($p < 0.001$), and all ROIs except V3A, IPS1, IPS2, and sPCS during Delay 2 (Fig. 6A–B; significant ROIs all $p \leq 0.034$, maximum p-value IPS0; non-significant ROIs all $p \geq 0.109$, minimum p-value V3A).

To ascertain the regions where WM representation fidelity changed between delay periods, we compared representational fidelity between each delay period (Fig. 6C). Representational fidelity significantly declined from Delay 1 to Delay 2 in V1-hV4, IPS0, IPS1 and All ROIs combined on R1 trials ($p \leq 0.028$; FDR-corrected, maximum p-value IPS1) and in V1–V3A, IPS0-IPS2, sPCS, and All ROIs combined on R2-neutral trials ($p < 0.001$; two-tailed resampling test of differences in representational fidelity between Delay 1 & 2 against 0). In

contrast, representational fidelity did not decline between delay periods on R2-valid trials, and in fact fidelity significantly increased after the valid cue in several occipital and parietal ROIs (V1, IPS0, IPS1, and All ROIs combined; $p \leq 0.018$, maximum p-value in IPS1).

In sum, these analyses identify reliable WM representations on R2-neutral trials, even when they are not easy to visualize in the reconstructed WM images (Fig. 4), and quantify a significant enhancement of representations on R2-valid trials following the cue (Fig. 6C). This result is not contingent on this particular quantification strategy (Fig. S4), nor the precise time points used (chosen to replicate Sprague et al., 2014); when we instead compared each pair of time points, we found evidence for representation restoration on R2-valid trials in every ROI studied, except for sPCS (Fig. S5). Furthermore, there was no strong evidence for a difference in recovery across ROIs, though visual/posterior parietal and anterior parietal/frontal cortex differed in the extent to which R1 representations decayed (Fig. S6). We also pursued an exploratory analysis of prefrontal cortex WM representations, presented in Fig. S7.

Quantifying spatial properties of target representations

Next, we sought to quantify how target representations change across WM conditions. When multiple items are remembered, representations could be weaker because they are “dimmer” above noisy background signals, as indexed by a lower amplitude over baseline, or because they are less spatially precise, as indexed by a larger size (Sprague et al., 2015, 2014). First, we precisely aligned all reconstructions across trials such that the target position was at a known position (red dots, Fig. 7A,C; Fig. S3D). Then, we fit a surface, defined by its size (i.e., spatial precision of the representation), amplitude (i.e., magnitude of the representation over spatially-global baseline in the reconstructions), and baseline (i.e., spatially-global offset in the reconstruction unrelated to WM target position), to each reconstruction using a resampling procedure (see Experimental Procedures; Fig 2D; Figure S3E). Because fits to a reconstruction with representational fidelity indistinguishable from 0 (Figs 5–6) are impossible to interpret, we only consider and report comparisons of fit parameters between pairs of conditions in which each condition has non-zero representational fidelity.

Delay 1: Representation amplitude decreased with WM load

During the first delay, averaged reconstructions qualitatively appeared higher in amplitude during R1 trials than R2 trials (Fig. 7A). Replicating previous results (Sprague et al., 2014), target representation amplitude during the first delay was higher on R1 trials as compared to both R2-neutral and R2-valid trials in visual (V1–V3A and hV4, all $p < 0.001$; Fig. 7B) and posterior parietal (IPS0 and IPS1, $p \leq 0.016$; and IPS2 for R1 vs. R2-neutral, $p = 0.012$; maximum p-value IPS1, R1 vs. R2-neutral) ROIs, as well as in reconstructions computed using all ROIs combined ($p < 0.001$; comparisons of fit parameters based on resampling test between condition pairs and FDR-corrected for multiple comparisons within fit parameter, see Statistical Procedures). No ROIs exhibited unequal representation amplitude between R2-neutral and R2-valid conditions during Delay 1 (all $p \geq 0.106$, minimum p-value in hV4), as expected given trials were identical at this point. Fit baseline was significantly greater on both R2-neutral and - valid trials as compared to R1 trials in V3, V3A, hV4, and in reconstructions computed from all ROIs combined (Fig. 7B, $p \leq 0.018$; maximum p-value

V3A, R1 vs R2-neutral). In V1, V2, and IPS0, a significantly greater baseline was seen when comparing R2-valid to R1 trials ($p \leq 0.01$, maximum p-value IPS0). Finally, WM representation size was significantly smaller on R2-neutral and -valid trials as compared to R1 trials in hV4 ($p < 0.001$). While quantitatively significant, these size changes were inconsistent across ROIs and were rarely observed compared to effects on amplitude and baseline. As such, we emphasize the consistency of observed changes in WM representation signal over noise (amplitude over a spatially-global baseline), and suggest that future datasets will help identify the extent to which changes in WM representation size are robust.

These Delay 1 results closely replicate our previous report in which we characterized how WM representations change as WM load is manipulated from 1 to 2 items (Sprague et al., 2014). In that report, we found extensive evidence for decreases in WM representation amplitude with increasing set size across visual and posterior parietal cortex, which we fully reproduced here (Fig. 7B).

Delay 2: Representation amplitude increased after cue

During Delay 2, target representations appeared weaker, though they were still identifiably present in many ROIs (Fig. 7C). Because our fitting procedure did not restrict the best-fit position of surfaces to be near the actual target position, the identification of WM representations nearby the true target position suggests a WM target representation was present (see also Fig. 6A).

WM representation amplitude was significantly higher during R2-valid trials than R1 trials in V1, V2, V3, V3A, hV4, IPS0, sPCS, and All ROIs combined (Fig. 7D, $p \leq 0.016$, maximum p-value sPCS), and was higher than representation amplitude in R2-neutral trials in all individual ROIs with WM representations during these conditions ($p \leq 0.016$, maximum p-value IPS3). Additionally, several ROIs showed a similar WM load effect for amplitude during Delay 2 as during Delay 1, such that R1 amplitude was significantly greater than R2-neutral amplitude (V2, V3, hV4, IPS0, and All ROIs combined, $p \leq 0.002$; maximum p- value hV4). Importantly, WM representation size during Delay 2 was always similar between R1 and R2- valid conditions, during which participants are maintaining the same number of items in WM (all p's ≥ 0.022, minimum p-value in V1, does not survive FDR correction). Finally, fit baseline was higher during R2-neutral and R2-valid conditions than R1 in several ROIs (R2-neutral > R1: IPS0 and All ROIs Combined; R2-valid > R1: V3A, IPS0, IPS1, IPS2, IPS3, sPCS, All ROIs Combined, maximum p-value 0.018 for All ROIs Combined, R2-neutral > R1), as well as higher on R2 trials with a valid cue than with a neutral cue (Fig. 7D, R2-valid > R2-neutral: IPS0, IPS3, and All ROIs Combined, all $p < 0.001$).

Improvements in WM representations of the cued item during Delay 2 of R2-valid trials were primarily found in their amplitude, with additional increases in spatially-global reconstruction baseline levels. The former amplitude increases reflect increased information content about the cued target position over a noisy baseline (Saproo and Serences, 2014, 2010; Sprague et al., 2015, 2014), and the latter reflect non-spatially- specific increased mean activation levels in these regions following an informative cue (see also Fig S2).

Cued WM representation amplitude tracks behavior

Finally, we tested whether any properties of the cued WM representation on R2-valid trials were related to participants' behavioral performance on each trial. Since performance is likely related to the fidelity of WM representations over many brain regions, we focused on the All ROIs combined ROI.

We separated trials into low- and high-recall error groups based on the median recall error within each condition, session, and participant. During Delay 2, cued WM target representations were qualitatively clearer and quantitatively they were significantly higher in amplitude on low recall error trials compared to high recall error trials (Fig. 8, $p < 0.001$; see Fig. S8 for results from each ROI individually). This observation suggests that participant performance is related to the signal-to-noise ratio (i.e. amplitude over baseline) of the validly-cued WM representation.

Discussion

Behavioral judgments about sensory stimuli rely on population-level neural representations which decrease in fidelity as the amount of relevant information increases (Drew et al., 2012, 2011; Tsubomi et al., 2013). When performing a demanding task in which stimuli that are used to guide behavior are no longer present in the display, only sustained internal representations held in working memory (WM) can be used, as no further information can be gathered from the environment. We used an image reconstruction technique (Fig. 2) to compare the fidelity of region-level WM representations across memory load conditions and replicated previous findings that behavioral performance (Fig. 1) and neural representations (Figs. 3–7) degrade with increasing load (Buschman et al., 2011; Emrich et al., 2013; Landman et al., 2003a; Sprague et al., 2014). However, upon presentation of an informative cue indicating which WM representation was relevant for behavior, the fidelity of a degraded representation substantially recovered (Figs. 4–7), and the quality of this recovered representation was related to behavioral performance on the task (Fig. 8). These data challenge the notion that WM representations rely primarily on active codes (e.g., spiking activity), for which degraded representations resulting from mutual suppression are permanently impaired (Bays, 2015, 2014). Instead, these data suggest that WM is supported by additional 'spike-silent' information that is manifest in a latent state invisible to typical measurement techniques (single unit firing rates or fMRI activation), but can be reinvigorated to an accessible, active state when task demands require an updated representation.

Our demonstration that a valid retro-cue enhanced the fidelity of WM representations primarily via an increase in their amplitude bears a striking similarity to the effects of spatial attention on visual representations as measured using neuroimaging and behavior (Gazzaley and Nobre, 2012; Itthipuripat and Serences, 2015; Lepsien and Nobre, 2007; Nobre et al., 2004; Saproo and Serences, 2014, 2010; Sprague and Serences, 2013; Sprague et al., 2015). However, in these experiments information used to improve neural representations and performance on the task is directly accessible in the sensory input to the visual system. As such, it is not possible to make strong inferences about the ability of neural codes to store latent information that can augment degraded representations, as information is still

available in the environment during the performance of the task. By using a visual WM task, in which all task-relevant information is necessarily represented in the nervous system, we could demonstrate directly that latent information sources must be present in the brain to bolster neural representations above and beyond an initially degraded state which can then support improved behavioral performance.

Sources of recovered information

Both our behavioral (Fig. 1C) and neural (Figs. 4–8) results suggest that the fidelity of neural representations can improve following the presentation of an informative retro-cue. What was the format of this information before the cue appeared? In information theory, the data processing inequality theorem provides the strong constraint that the total information about one variable given the observed state of another variable (i.e. mutual information) can never increase with additional processing; it can at best remain constant (Cover and Thomas, 1991; Quian Quiroga and Panzeri, 2009; Saproo and Serences, 2014, 2010; Shannon, 1948; Sprague et al., 2015). Accordingly, we can conclude that the information used to complete the behavioral recall task more accurately following the presentation of a retro-cue must be, in some way, present in the system before the cue appears. However, because WM item representations in fMRI-based image reconstructions were already degraded by the time the retro-cue appeared (Fig. 6; Sprague et al., 2014), the restored representation must result from neural response features inaccessible to or hidden from our BOLD activation pattern measurements before the cue.

One potential source of the restored representational fidelity is WM-related patterns of sub-threshold membrane potential and/or changes in short-term synaptic efficacy, as suggested by prior theoretical and computational modeling efforts (Barak and Tsodyks, 2014; Edin et al., 2009; Mongillo et al., 2008; Stokes, 2015; Stokes et al., 2013). Both of these mechanisms render a circuit dynamically sensitive to input as a function of WM contents, and both processes may be difficult to detect with typical electrophysiological or neuroimaging techniques in animals or humans. Consistent with this view, a recent study found that motion-sensitive visual area MT did not carry information about the memorized stimulus over a brief delay interval via changes in spike counts (Mendoza-Halliday et al., 2014). However, changes in local field potentials (LFP) did carry information about the contents of WM. Such LFP changes may reflect aggregate changes in the membrane potentials of nearby neural populations, which could enable more robust WM coding following re-allocation of attention (Griffin and Nobre, 2003; Landman et al., 2003b; Lepsien et al., 2011; Makovski and Jiang, 2007) or a sweep of non-specific activity (Mongillo et al., 2008; Stokes, 2015; Stokes et al., 2013; Wolff et al., 2015). In fact, Mendoza- Halliday et al. found evidence for such top-down control of LFP representations by identifying spike-field coherence between prefrontal spikes and MT LFP beta band activity (Mendoza-Halliday et al., 2014), and a recent study that decoded WM representations from EEG scalp potentials found evidence that nonspecific visual input can reveal such hidden states in visual WM (Wolff et al., 2015). Future experiments measuring membrane potentials of single cells while animals perform demanding WM tasks under varying load conditions (Buschman et al., 2011; Kornblith et al., 2015; Landman et al., 2003a; Lara and Wallis, 2014, 2012) may reveal how such non-spiking sources of neural

information can augment neural population codes that are typically described solely in terms of spiking activity (Bays, 2014; Ma et al., 2006; Tan et al., 2014).

These potential physiological mechanisms could be part of a neural normalization process (Bays, 2015, 2014; Carandini and Heeger, 2012; Edin et al., 2009; Franconeri et al., 2013; Ma et al., 2014) whereby each of several simultaneously-held representations mutually suppresses the spiking output of (but not the synaptic input to) all other WM representations, resulting in degraded behavioral performance and degraded representations as measured with fMRI. This could allow for latent information encoded via short-term synaptic plasticity of inputs or subthreshold membrane potentials to exert an influence on spiking activity of cells after the presentation of an informative cue (e.g. the mid-delay retro-cue on R2- valid trials in the present study), perhaps by removing the suppressive influence of the irrelevant item on other representations. Then, synaptic input, which is “latent” in this case because it does not cause spiking while both representations are present, would now enable reinvigoration of active neural representations as measured by spike rates due to reduced inhibitory inputs from the neural representation of the non-cued item. A similar normalization account has also been used to predict attentional modulations as a function of the spatial extent of items attended (Reynolds and Heeger, 2009), which is supported experimentally by EEG and behavioral measurements of representational fidelity (Herrmann et al., 2010; Itthipuripat et al., 2014). Normalization of simultaneous representations may reflect a general neural constraint on representing information for the guidance of behavior (Carandini and Heeger, 2012).

Fidelity of feature representations in WM

Several previous studies cued participants to focus on a single item among multiple items maintained in WM. Lepsien et al (2007) post-cued participants to remember either a face or scene after both types of stimuli were encoded, and Lewis-Peacock, LaRocque and colleagues cued participants during the delay period to focus on one from among two different stimulus categories (LaRocque et al., 2013; Lewis- Peacock et al., 2012) . These studies found evidence for enhanced representations of the cued item category by either comparing mean signal amplitude in different category-selective ROIs (Lepsien and Nobre, 2007) or comparing multivariate classifier evidence for each item category during the delay interval before and after the post-cue (LaRocque et al., 2013; Lewis-Peacock et al., 2012). These studies suggest that cueing one of several items in WM can effectively trigger a switch in the focus of attention to internal category-level representations, resulting in increased activation levels (or classifier evidence) associated with that category (LaRocque et al., 2013; Lepsien and Nobre, 2007; Lewis-Peacock et al., 2012). Such results are broadly consistent with our framework that information about items in WM can additionally be maintained via latent or unobserved neural signals. However, because these studies did not evaluate the fidelity of WM representations of the category members themselves (i.e., are the retrocued face representations in FFA more informative about which face is in WM?), they cannot rule out a competing account whereby the retro-cue triggers a shift in rehearsal strategy and/or category-specific prospective attention to the probe stimulus, but no change in the representations themselves. Moreover, they do not establish a relationship between

behavioral retro-cue benefits and improvements in representational fidelity of precise feature information in WM.

In contrast, we show here that latent information can be revealed by (1) cueing participants to one of several items of the same category (spatial positions) and (2) quantitatively evaluating the feature-specific information content of WM representations carried by fMRI activation patterns throughout the trial. Our results thus conceptually replicate the general finding that the contents of visual WM are dynamic and can be modulated by delay-period cues (Fig. 3). However, we show here that such cues can directly enhance the fidelity with which an individual cued item is represented via the use of latent information (Figs. 6–7) in a manner related to behavioral performance (Fig. 8).

Role of long-term memory

Improved behavioral performance and restored representational fidelity following a valid retro-cue could also result from long-term memory (LTM) retrieval. Recent behavioral studies have found that high-fidelity feature representations could be recalled from LTM (Brady et al., 2013; Sutterer and Awh, 2015) in tasks in which participants recalled precise features (e.g., color) associated with images or drawings of distinguishable objects. Performance on these tasks was nearly as robust as when maintaining an item in WM, though recall from LTM was poorer than WM for a single item (Brady et al., 2013). Thus, while there is a possibility that participants transfer spatial positions to LTM during the long delay intervals of our task and then recall those positions when given a valid cue, it would likely result in a degraded representation.

Information in measurements as a lower bound

In this study, we examined markers of WM representations using fMRI activation patterns. Consequently, all conclusions we draw about changes in neural information are inferred based on changes in information in our measured signals (BOLD activation patterns). While it could theoretically be the case that such changes do not relate in any meaningful way to neural activity occurring below the spatial, temporal, and physiological resolution of our measurements, we interpret our findings as placing a lower bound on the information content of the true neural codes. That is, the observation of information in a measurement is sufficient to infer information in the underlying cause of that measurement. But the observation of information with a measurement (e.g., a BOLD activation pattern) is not necessary given information in the underlying cause (neural activity state). Similar constraints hold when measuring neural spiking extracellularly: the observation of spikes is sufficient to conclude a change in the membrane potential of a cell, but changes in membrane potentials can occur without spikes. Similar arguments hold for all techniques in use in modern neuroscience, including additional information that can be available in synergistic codes across multiple neurons (Schneidman et al., 2003), wherein information could be missed by focusing on single neurons in isolation. Accordingly, the absence of evidence for information in a given technique should not be used to argue that information is absent (Dubois et al., 2015; Ester et al., 2016). This is well-illustrated in our observations that representations degrade in fidelity early in the delay period, but can recover with a valid cue (Fig. 1C, Figs. 5–8). The poorer fidelity of WM representations identified in measured

signals underestimated the actual information accessible within the brain, which was revealed upon cue presentation. The existence of the information after the cue is sufficient to conclude that information must have been available in the brain before the cue appeared.

Conclusions

We show that post-cuing an item accessible only in WM can enhance the fidelity of its item-specific representation.. Information theoretic constraints preclude spike-based models from accounting for these post-cue effects because spike-based models predict that a loss of spiking integrity should be irreversible. Thus, these data suggest the maintenance of additional information about the cued item in a latent, high-fidelity state that can restore degraded active representations in response to changing behavioral demands. Finally, representations of information in neural activity patterns may more broadly rely on such sub-threshold components that are not typically assayed in neuroimaging or electrophysiological experiments.

Experimental Procedures

Participants—We recruited 6 participants naive to the purpose of the experiment. All participants underwent 3 fMRI scanning sessions and 1 retinotopic mapping scanning session, each lasting 2 hrs.

Spatial WM retro-cueing task—We presented 2 target stimuli (a red and a blue dot) for 500 ms 3.5° from fixation on average. The fixation point immediately changed color to either red, blue, or purple. A red or blue fixation cue (1/3 of trials) indicated one target should be maintained in WM over the delay interval (R1). A purple fixation cue (2/3 of trials) indicated both targets should be maintained in WM (R2). After an 8 s delay interval (Delay 1), the fixation cue changed color once again. On ½ of R2 trials (1/3 of trials overall), the fixation cue changed from purple to either red or blue, cueing the participants to remember only the cued target (R2-valid condition). During all other trials, the fixation point became black, acting as a neutral cue. Following this cue, participants maintained 1 or 2 items over an 8 s delay interval (Delay 2).

Participants also performed a spatial mapping task to estimate spatial sensitivity for each voxel (Fig. S3A–C) and a visual localizer task to select voxels for further analysis, each described in Supplemental Experimental Procedures.

Behavioral analysis—We defined behavioral recall error as the absolute distance between the position of the response bar and the actual coordinate of the recalled target. In fMRI analyses in which we split trials based on behavioral performance, we computed the median recall error within each WM condition within each scanning session and split trials based on performance above or below this median.

fMRI acquisition and preprocessing—We scanned on a 3 T GE MR750 scanner with a voxel size of 2×2×3 mm and 2,250 ms TR. Preprocessing included coregistration of scans across sessions, unwarping, slice time correction, motion correction, temporal high-pass filtering, transformation to Talairach space, and Z-score normalization.

fMRI analysis: inverted encoding model—To reconstruct images of spatial WM contents, we implemented an inverted encoding model (IEM) for spatial position. This analysis involves first estimating an encoding model for each voxel in a region using a training set of data reserved for this purpose. Then, the encoding models across all voxels within a region are inverted to estimate a mapping used to transform novel activation patterns from the WM task into activation patterns in a modeled set of information channels. Details of the IEM analysis and quantification strategies are presented in detail in Figs. 2 & S3 and Supplemental Experimental Procedures.

Statistical procedures—All statistical inferences are based on resampling tests whereby a variable was computed over 1,000 iterations. During each, all single-trial variables from a given condition were resampled with replacement and averaged, resulting in 1,000 resampled averages. For analyses comparing conditions (Figs. 1, 6–8), we computed the distribution of differences between one resampled distribution (e.g., R1) and another (e.g., R2), yielding 1,000 difference values. We tested whether these difference distributions significantly differed from 0 in either direction by comparing against 0 (p = % of values > or < 0; null hypothesis that difference = 0) and doubling the smaller p value. For comparisons of representational fidelity against 0 (Figs. 5–6), we used the % of values < 0, one-tailed.

Unless otherwise stated, we corrected all repeated tests within an analysis using the false discovery rate (Benjamini and Yekutieli, 2001), $q = 0.05$. All p-values for all tests are reported in Supplemental Tables. All error bars/intervals reflect 95% confidence intervals via this resampling procedure. The All ROIs Combined ROI was not independent of the other ROIs, so we independently corrected for multiple comparisons within that ROI alone.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Edward Awh, Brad Postle, Sirawaj Itthipuripat, Stephanie Nelli, Roseanne Rademaker, Samantha Scudder, and Vy Vo for helpful discussions and/or comments on a draft of this manuscript, and Haider Al-Hakeem for helpful discussions and assistance with data collection. Funded by NIH R01-EY025872 and a James S. McDonnell Foundation Scholar Award to J.T.S., NIH T32-MH20002 to T.C.S. and E.F.E., and an NSF Graduate Research Fellowship to T.C.S.

References

Albers AM, Kok P, Toni I, Dijkerman HC, de Lange FP. Shared representations for working memory and mental imagery in early visual cortex. Curr Biol. 2013; 23:1427–31. DOI: 10.1016/j.cub.2013.05.065 [PubMed: 23871239]

Baddeley AD, Hitch G. Working memory. Psychol Learn Motiv. 1974; 8:47–89. DOI: 10.1016/S0079-7421(08)60452-1

Barak O, Tsodyks M. Working models of working memory. Curr Opin Neurobiol. 2014; 25:20–4. DOI: 10.1016/j.conb.2013.10.008 [PubMed: 24709596]

Bays PM. Spikes not slots: noise in neural populations limits working memory. Trends Cogn Sci. 2015; 19:431–8. DOI: 10.1016/j.tics.2015.06.004 [PubMed: 26160026]

Bays PM. Noise in neural populations accounts for errors in working memory. J Neurosci. 2014; 34:3632–45. DOI: 10.1523/JNEUROSCI.3204-13.2014 [PubMed: 24599462]

Bays PM, Husain M. Dynamic shifts of limited working memory resources in human vision. Science (80-). 2008; 321:851–4. DOI: 10.1126/science.1158023

Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. Ann Stat. 2001; 29:1165–1188.

Brady TF, Konkle T, Gill J, Oliva A, Alvarez GA. Visual long-term memory has the same limit on fidelity as visual working memory. Psychol Sci. 2013; 24:981–90. DOI: 10.1177/0956797612465439 [PubMed: 23630219]

Briggs F, Mangun GR, Usrey WM. Attention enhances synaptic efficacy and the signal-to-noise ratio in neural circuits. Nature. 2013; 499:476–80. DOI: 10.1038/nature12276 [PubMed: 23803766]

Brouwer G, Heeger D. Decoding and Reconstructing Color from Responses in Human Visual Cortex. J Neurosci. 2009; 29:13992–14003. [PubMed: 19890009]

Buschman TJ, Siegel M, Roy JE, Miller EK. Neural substrates of cognitive capacity limitations. Proc Natl Acad Sci U S A. 2011; 108:11252–5. DOI: 10.1073/pnas.1104666108 [PubMed: 21690375]

Carandini M, Heeger DJ. Normalization as a canonical neural computation. Nat Rev Neurosci. 2012; 13:51–62. DOI: 10.1038/nrn3136 [PubMed: 22108672]

Cover, T.; Thomas, J. Elements of information theory. Wiley; New York: 1991.

Curtis CE, D'Esposito M. Persistent activity in the prefrontal cortex during working memory. Trends Cogn Sci. 2003; 7:415–423. [PubMed: 12963473]

D'Esposito M, Postle BR. The Cognitive Neuroscience of Working Memory. Annu Rev Psychol. 2014; 66:115–42. DOI: 10.1146/annurev-psych-010814-015031 [PubMed: 25251486]

Drew T, Horowitz TS, Wolfe JM, Vogel EK. Neural measures of dynamic changes in attentive tracking load. J Cogn Neurosci. 2012; 24:440–50. DOI: 10.1162/jocn_a_00107 [PubMed: 21812558]

Drew T, Horowitz TS, Wolfe JM, Vogel EK. Delineating the neural signatures of tracking spatial position and working memory during attentive tracking. J Neurosci. 2011; 31:659–68. DOI: 10.1523/JNEUROSCI.1339-10.2011 [PubMed: 21228175]

Dubois J, de Berker AO, Tsao DY. Single-unit recordings in the macaque face patch system reveal limitations of fMRI MVPA. J Neurosci. 2015; 35:2791–802. DOI: 10.1523/JNEUROSCI.4037-14.2015 [PubMed: 25673866]

Edin F, Klingberg T, Johansson P, McNab F, Tegner J, Compte A. Mechanism for top-down control of working memory capacity. Proc Natl Acad Sci U S A. 2009; 106:6802–7. DOI: 10.1073/pnas.0901894106 [PubMed: 19339493]

Emrich SM, Riggall AC, Larocque JJ, Postle BR. Distributed patterns of activity in sensory cortex reflect the precision of multiple items maintained in visual short-term memory. J Neurosci. 2013; 33:6516–23. DOI: 10.1523/JNEUROSCI.5732-12.2013 [PubMed: 23575849]

Erickson MA, Maramara LA, Lisman J. A single brief burst induces GluR1-dependent associative short-term potentiation: a potential mechanism for short-term memory. J Cogn Neurosci. 2010; 22:2530–40. DOI: 10.1162/jocn.2009.21375 [PubMed: 19925206]

Ester EF, Anderson DE, Serences JT, Awh E. 2013; A Neural Measure of Precision in Visual Working Memory. J Cogn Neurosci. :754–761. DOI: 10.1162/jocn_a_00357 [PubMed: 23469889]

Ester EF, Rademaker RL, Sprague TC. How do visual and parietal cortex contribute to visual short-term memory? eNeuro. 2016; ENEURO.0041–16. doi: 10.1523/ENEURO.0041-16.2016

Ester EF, Sprague TC, Serences JT. Parietal and Frontal Cortex Encode Stimulus-Specific Mnemonic Representations during Visual Working Memory. Neuron. 2015; 87:893–905. DOI: 10.1016/j.neuron.2015.07.013 [PubMed: 26257053]

Foster JJ, Sutterer DW, Serences JT, Vogel EK, Awh E. The topography of alpha-band activity tracks the content of spatial working memory. J Neurophysiol. 2015; jn.00860.2015. doi: 10.1152/jn.00860.2015

Franconeri SL, Alvarez GA, Cavanagh P. Flexible cognitive resources: competitive content maps for attention and memory. Trends Cogn Sci. 2013; 17:134–41. DOI: 10.1016/j.tics.2013.01.010 [PubMed: 23428935]

Funahashi S, Bruce CJ, Goldman-Rakic PS. Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. J Neurophysiol. 1989; 61:331–49. [PubMed: 2918358]

Fuster JM, Alexander GE. Neuron activity related to short-term memory. Science. 1971; 173:652–4. [PubMed: 4998337]

Gazzaley A, Nobre AC. Top-down modulation: bridging selective attention and working memory. Trends Cogn Sci. 2012; 16:129–35. DOI: 10.1016/j.tics.2011.11.014 [PubMed: 22209601]

Griffin IC, Nobre AC. Orienting attention to locations in internal representations. J Cogn Neurosci. 2003; 15:1176–94. DOI: 10.1162/089892903322598139 [PubMed: 14709235]

Harrison SA, Tong F. Decoding reveals the contents of visual working memory in early visual areas. Nature. 2009; 458:632–635. DOI: 10.1038/nature07832 [PubMed: 19225460]

Herrmann K, Montaser-Kouhsari L, Carrasco M, Heeger DJ. When size matters: attention affects performance by contrast or response gain. Nat Neurosci. 2010; 13:1554–1559. [PubMed: 21057509]

Itthipuripat S, Garcia JO, Rungratsameetaweemana N, Sprague TC, Serences JT. Changing the spatial scope of attention alters patterns of neural gain in human cortex. J Neurosci. 2014; 34:112– 23. DOI: 10.1523/JNEUROSCI.3943-13.2014 [PubMed: 24381272]

Itthipuripat S, Serences JT. Integrating Levels of Analysis in Systems and Cognitive Neurosciences: Selective Attention as a Case Study. Neuroscientist. 2015; doi: 10.1177/1073858415603312

Jeanne JM, Sharpee TO, Gentner TQ. Associative learning enhances population coding by inverting interneuronal correlation patterns. Neuron. 2013; 78:352–63. DOI: 10.1016/j.neuron.2013.02.023 [PubMed: 23622067]

Keshvari S, van den Berg R, Ma WJ. No evidence for an item limit in change detection. PLoS Comput Biol. 2013; 9:e1002927.doi: 10.1371/journal.pcbi.1002927 [PubMed: 23468613]

Kornblith S, Buschman TJ, Miller EK. Stimulus Load and Oscillatory Activity in Higher Cortex. Cereb Cortex. 2015; doi: 10.1093/cercor/bhv182

Landman R, Spekreijse H, Lamme VAF. Set size effects in the macaque striate cortex. J Cogn Neurosci. 2003a; 15:873–82. DOI: 10.1162/089892903322370799 [PubMed: 14511540]

Landman R, Spekreijse H, Lamme VAF. Large capacity storage of integrated objects before change blindness. Vision Res. 2003b; 43:149–64. [PubMed: 12536137]

Lara AH, Wallis JD. Executive control processes underlying multi-item working memory. Nat Neurosci. 2014; 17:876–83. DOI: 10.1038/nn.3702 [PubMed: 24747574]

Lara AH, Wallis JD. Capacity and precision in an animal model of visual short-term memory. J Vis. 2012; :12.doi: 10.1167/12.3.13

LaRocque J, Lewis-Peacock J, Drysdale A, Oberauer K, Postle BR. Decoding attended information in short-term memory: An eeg study. J Cogn Neurosci. 2013; 25:127–142. [PubMed: 23198894]

LaRocque JJ, Eichenbaum AS, Starrett MJ, Rose NS, Emrich SM, Postle BR. The shortand long-term fates of memory items retained outside the focus of attention. Mem Cognit. 2015; 43:453–68. DOI: 10.3758/s13421-014-0486-y

Lepsien J, Nobre AC. Attentional modulation of object representations in working memory. Cereb Cortex. 2007; 17:2072–83. DOI: 10.1093/cercor/bhl116 [PubMed: 17099066]

Lepsien J, Thornton I, Nobre AC. Modulation of working-memory maintenance by directed attention. Neuropsychologia. 2011; 49:1569–77. DOI: 10.1016/j.neuropsychologia.2011.03.011 [PubMed: 21420421]

Lewis-Peacock JA, Drysdale AT, Oberauer K, Postle BR. Neural evidence for a distinction between short-term memory and the focus of attention. J Cogn Neurosci. 2012; 24:61–79. DOI: 10.1162/jocn_a_00140 [PubMed: 21955164]

Luck SJ, Vogel EK. Visual working memory capacity: from psychophysics and neurobiology to individual differences. Trends Cogn Sci. 2013; 17:391–400. DOI: 10.1016/j.tics.2013.06.006 [PubMed: 23850263]

Ma WJ, Beck JM, Latham PE, Pouget A. Bayesian inference with probabilistic population codes. Nat Neurosci. 2006; 9:1432–8. DOI: 10.1038/nn1790 [PubMed: 17057707]

Ma WJ, Husain M, Bays PM. Changing concepts of working memory. Nat Neurosci. 2014; 17:347–56. DOI: 10.1038/nn.3655 [PubMed: 24569831]

Makovski T, Jiang YV. Distributing versus focusing attention in visual short-term memory. Psychon Bull Rev. 2007; 14:1072–8. [PubMed: 18229477]

Matsukura M, Luck SJ, Vecera SP. Attention effects during visual short-term memory maintenance: protection or prioritization? Percept Psychophys. 2007; 69:1422–34. [PubMed: 18078232]

Matsushima A, Tanaka M. Different neuronal computations of spatial working memory for multiple locations within versus across visual hemifields. J Neurosci. 2014; 34:5621–6. DOI: 10.1523/JNEUROSCI.0295-14.2014 [PubMed: 24741052]

Mendoza-Halliday D, Torres S, Martinez-Trujillo JC. Sharp emergence of feature-selective sustained activity along the dorsal visual pathway. Nat Neurosci. 2014; 17:1255–62. DOI: 10.1038/nn.3785 [PubMed: 25108910]

Mongillo G, Barak O, Tsodyks M. Synaptic theory of working memory. Science. 2008; 319:1543–6. DOI: 10.1126/science.1150769 [PubMed: 18339943]

Nobre AC, Coull JT, Maquet P, Frith CD, Vandenberghe R, Mesulam MM. Orienting attention to locations in perceptual versus mental representations. J Cogn Neurosci. 2004; 16:363–73. DOI: 10.1162/089892904322926700 [PubMed: 15072672]

Quian Quiroga R, Panzeri S. Extracting information from neuronal populations: information theory and decoding approaches. Nat Rev Neurosci. 2009; 10:173–85. DOI: 10.1038/nrn2578 [PubMed: 19229240]

Reinhart RMG, Heitz RP, Purcell BA, Weigand PK, Schall JD, Woodman GF. Homologous mechanisms of visuospatial working memory maintenance in macaque and human: properties and sources. J Neurosci. 2012; 32:7711–22. DOI: 10.1523/JNEUROSCI.0215-12.2012 [PubMed: 22649249]

Reynolds JH, Heeger DJ. The normalization model of attention. Neuron. 2009; 61:168–185. [PubMed: 19186161]

Saproo S, Serences JT. Attention Improves Transfer of Motion Information between V1 and MT. J Neurosci. 2014; 34:3586–3596. DOI: 10.1523/JNEUROSCI.3484-13.2014 [PubMed: 24599458]

Saproo S, Serences JT. Spatial Attention Improves the Quality of Population Codes in Human Visual Cortex. J Neurophysiol. 2010; 104:885–895. DOI: 10.1152/jn.00369.2010 [PubMed: 20484525]

Schneidman E, Bialek W, Berry MJ. Synergy, redundancy, and independence in population codes. J Neurosci. 2003; 23:11539–53. [PubMed: 14684857]

Serences JT, Ester EF, Vogel EK, Awh E. Stimulus-Specific Delay Activity in Human Primary Visual Cortex. Psychol Sci. 2009; 20:207–214. DOI: 10.1111/j.1467-9280.2009.02276.x [PubMed: 19170936]

Shannon CE. A Mathematical Theory of Communication. Bell Syst Tech J. 1948; 27:379–423. DOI: 10.1145/584091.584093

Sprague TC, Ester EF, Serences JT. Reconstructions of Information in Visual Spatial Working Memory Degrade with Memory Load. Curr Biol. 2014; doi: 10.1016/j.cub.2014.07.066

Sprague TC, Saproo S, Serences JT. Visual attention mitigates information loss in small- and large-scale neural codes. Trends Cogn Sci. 2015; 19:215–26. DOI: 10.1016/j.tics.2015.02.005 [PubMed: 25769502]

Sprague TC, Serences JT. Attention modulates spatial priority maps in the human occipital, parietal and frontal cortices. Nat Neurosci. 2013; 16:1879–87. DOI: 10.1038/nn.3574 [PubMed: 24212672]

Squire LR, Wixted JT. The cognitive neuroscience of human memory since H.M. Annu Rev Neurosci. 2011; 34:259–88. DOI: 10.1146/annurev-neuro-061010-113720 [PubMed: 21456960]

Sreenivasan KK, Curtis CE, D'Esposito M. Revisiting the role of persistent neural activity during working memory. Trends Cogn Sci. 2014; 18:82–9. DOI: 10.1016/j.tics.2013.12.001 [PubMed: 24439529]

Srimal R, Curtis CE. Persistent neural activity during the maintenance of spatial position in working memory. Neuroimage. 2008; 39:455–468. [PubMed: 17920934]

Stokes MG. "Activity-silent" working memory in prefrontal cortex: a dynamic coding framework. Trends Cogn Sci. 2015; 19:394–405. DOI: 10.1016/j.tics.2015.05.004 [PubMed: 26051384]

Stokes MG, Kusunoki M, Sigala N, Nili H, Gaffan D, Duncan J. Dynamic coding for cognitive control in prefrontal cortex. Neuron. 2013; 78:364–75. DOI: 10.1016/j.neuron.2013.01.039 [PubMed: 23562541]

Sutterer DW, Awh E. Retrieval practice enhances the accessibility but not the quality of memory. Psychon Bull Rev. 2015; doi: 10.3758/s13423-015-0937-x

Tan AYY, Chen Y, Scholl B, Seidemann E, Priebe NJ. Sensory stimulation shifts visual cortex from synchronous to asynchronous states. Nature. 2014; 509:226–9. DOI: 10.1038/nature13159 [PubMed: 24695217]

Tsubomi H, Fukuda K, Watanabe K, Vogel EK. Neural limits to representing objects still within view. J Neurosci. 2013; 33:8257–63. DOI: 10.1523/JNEUROSCI.5348-12.2013 [PubMed: 23658165]

Wolff MJ, Ding J, Myers NE, Stokes MG. Revealing hidden states in visual working memory using electroencephalography. Front Syst Neurosci. 2015; 9:123.doi: 10.3389/fnsys.2015.00123 [PubMed: 26388748]

Zhang W, Luck SJ. Discrete fixed-resolution representations in visual working memory. Nature. 2008; 453:233–235. DOI: 10.1038/nature06860 [PubMed: 18385672]

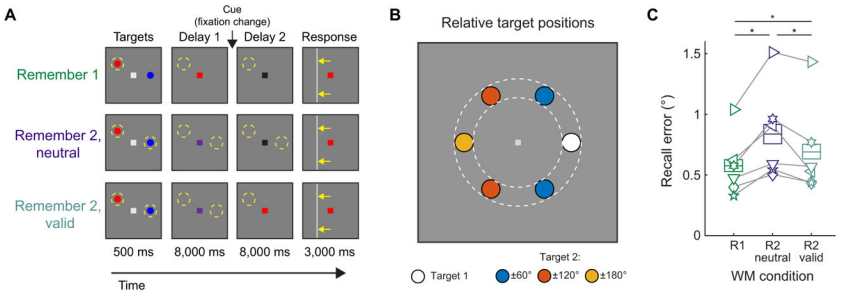

Figure 1. Spatial WM performance recovers following a retro-cue

(A) On each trial, participants viewed 2 target stimuli (red and blue dots). A subsequent change in the color of the fixation point to red, blue, or purple cued participants to remember the location of the red target, the blue target, or both targets (respectively). On 33% of trials, we cued participants to remember the location of one target over the entire delay interval (fixation became red or blue, Remember 1; R1). On the remaining 67% of trials we cued participants to remember the locations of both targets (Remember 2; R2). This set of trials was further divided in half: on R2-neutral trials, we gave no information about which item was relevant (fixation point became black after an 8 s delay); on R2- valid trials the fixation point became red or blue, indicating which target would be probed at the end of the trial. After the 16 s delay, participants adjusted a horizontal or vertical bar to match the position of the remembered target. Dashed yellow circles illustrate remembered locations.

(B) The 2 targets appeared at positions uniformly drawn from 2 discs (0.6° radius centered 3.5° from fixation; colored circles within dashed annulus). Targets never appeared within the same disc; they appeared ± 60°, ± 120°, or ± 180° polar angle apart on each trial. We randomly rotated the entire target arrangement on each trial.

(C) Memory performance was lower (i.e., higher recall error) during R2-neutral trials than R1 trials in all $n = 6$ participants. However, a valid cue (R2-valid) improved performance relative to R2-neutral trials, though performance remained lower than R1 trials. Asterisks indicate significance determined by pairwise resampling test, Bonferroni corrected for 3 comparisons. Boxes with horizontal lines indicate 95% confidence intervals (CIs) computed via resampling and mean over resampling iterations, respectively. Each symbol in (C) is a single participant. See Fig. S1 for recall error histograms, and Fig. S2 for univariate fMRI activation for each condition.

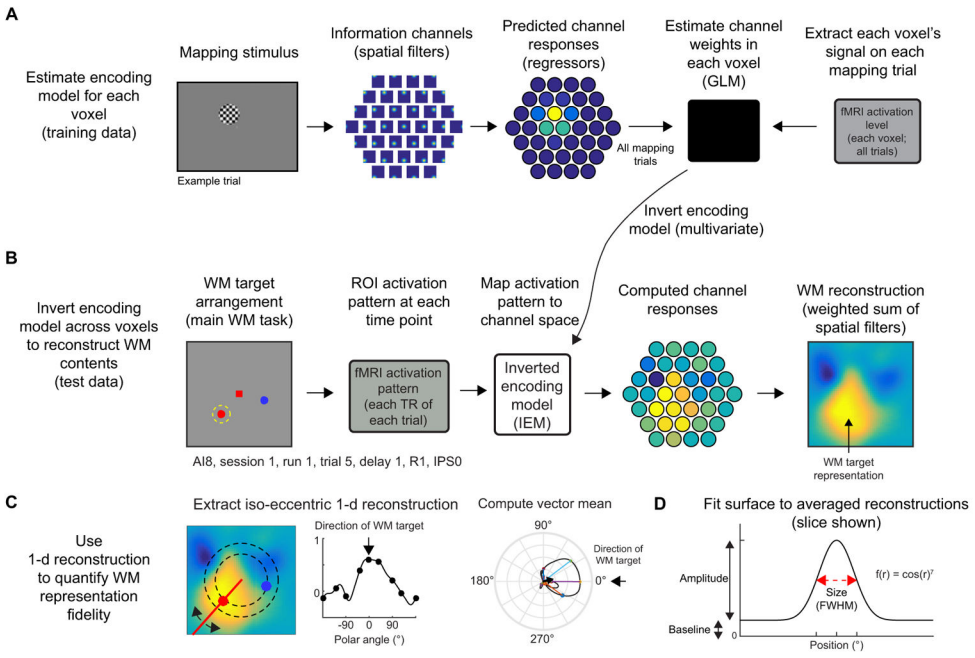

Figure 2. Inverted encoding model (IEM) for reconstructing and quantifying WM representations

To evaluate whether fMRI-based measurements of spatial WM representations are modulated by task demands, we implemented an inverted encoding model (IEM) which enabled reconstruction of spatial WM representations in retinotopically organized visual, parietal, and frontal regions.

(A) To estimate voxel-level encoding models, we modeled the response of each voxel as a weighted sum of 37 information channels, each defined as a round smooth spatial filter, spanning a hexagonal spatial grid. After predicting the activation of each channel on each trial of a separate mapping task (Fig. S3A–C), we used measured activation levels from all trials to estimate the contribution of each channel to each voxel. This results in 37 weights for each voxel, describing the contribution of the 37 channels to the observed signal in that voxel. Example trial shown; we used all mapping trials within a session.

(B) After estimating encoding models for all voxels within an ROI, we used the pattern of encoding models (37-dimensional weight vector for every voxel) across all voxels in an ROI to compute an IEM. We used the resulting inverted matrix to map WM delay activation patterns measured in “voxel space” into the “information space” defined by the 37 modeled channels of our encoding model (A). Next, we summed the spatial filters weighted by their estimated channel activation, resulting in a reconstructed image of the visual field. On this example trial, the bright region in the reconstruction (right) matches the position held in WM (left, dashed circle), and we call these “target representations”. We reconstructed images at each imaging volume (TR) in the trial and aligned all reconstructions across trials (see Supplemental Experimental Procedures, Fig. S3D) so that targets were at known positions, enabling us to average over trials with different WM contents.

(C) To quantify the strength of WM representations, we computed a “representational fidelity” metric by extracting a 1–d reconstruction as a function of polar angle by computing the mean reconstruction activation from 2.9–4.1° eccentricity (dashed black lines). Then, we

used this 1–d reconstruction to compute a vector mean of a circular set of unit vectors, each weighted by its corresponding activation. Finally, we projected this vector mean onto a unit vector pointing in the polar angle direction of the WM target (subset of unit vectors shown as colored lines; vector mean shown as black arrow; reconstruction rotated so that target at 0°).

(**D**) We quantified several parameters of WM representations (amplitude, size, and baseline) by fitting a 2-d surface to average coregistered reconstructions (Fig. S3D) on each of 1,000 resampling iterations (Figs. 7–8). To assess significance, we compared distributions of best-fit parameters between conditions (Fig. 7) or behavioral performance bins (Fig. 8 and Fig. S8). See also Fig. S3 and Supplemental Experimental Procedures.

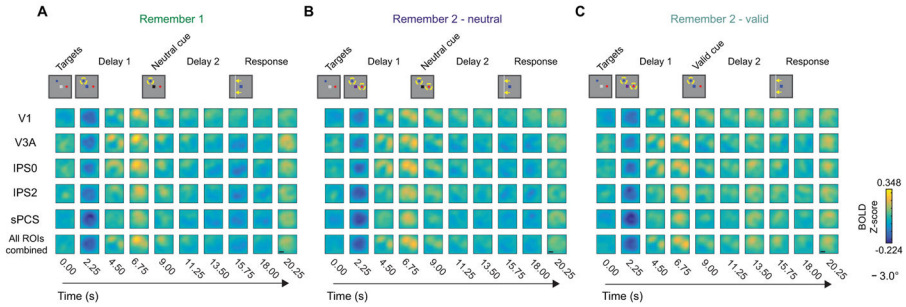

Figure 3. Reconstructions show dynamic contents of WM

We reconstructed the contents of spatial WM at each TR during the trial using activation patterns from several ROIs defined using independent localizers (subset shown for brevity). Here we show reconstructions from an example target arrangement condition where WM targets were separated by an average of 120° polar angle (top row). We rotated all single-trial reconstructions to match the cartoons and averaged over trials and participants (n = 6, 3 2-hr scanning sessions each). Yellow circles indicate the position(s) held in WM at each TR. Each image shows a a reconstruction generated using activation patterns measured at a specific time point (columns) and ROI that we examined (rows). All images show a 12° × 12° square visual field aperture and are plotted on the same color scale. See also Fig. S7 for exploratory prefrontal cortex ROIs.

(A) During R1 trials, stable WM representations emerged ~6–9 s following the first delay cue (delayed onset reflects hemodynamic lag) and remained present throughout the entire 16 s delay interval.

(B) During R2-neutral trials, stable WM representations were preserved over the entire 16 s delay interval, although they remained substantially weaker than those on R1 trials.

(C) On R2-valid trials, 2 representations appeared initially, then a single representation appeared after the valid cue, tracking the manipulated contents of WM.

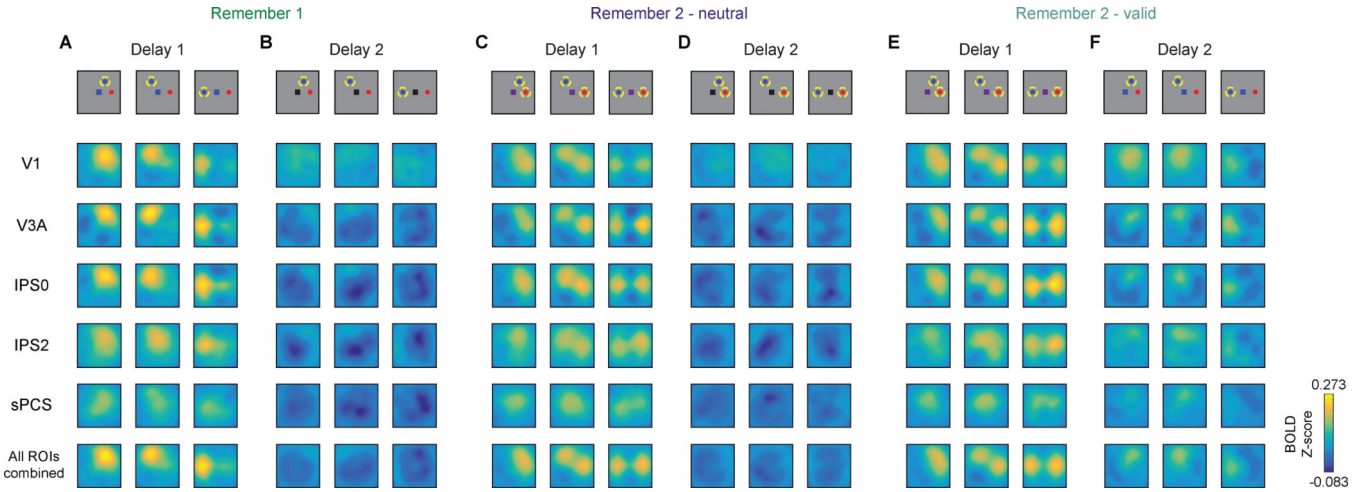

Figure 4. WM reconstructions track target positions

WM reconstructions computed and plotted as in Fig. 3, for each target arrangement condition and averaged over 2 TRs during each delay (Delay 1: 6.75 and 9.00 s; Delay 2: 15.75 and 18.00 s). A subset of ROIs is shown for brevity. In all cases, target representations track remembered positions. All reconstructions plotted on same color scale.

(A–B) During R1 trials, only the relevant item was represented over both delays, ruling out contributions from sensory transients (see also Sprague et al., 2014). Calibri,Bold

(C–D) During R2-neutral trials, both items were represented, though more weakly than during R1 trialsCalibri,Bold

(E–F) During R2-valid trials, both items were represented during Delay 1, then only the cued item was represented during Delay 2. The cued representation during Delay 2 appeared qualitatively stronger than each of the representations before the cue.

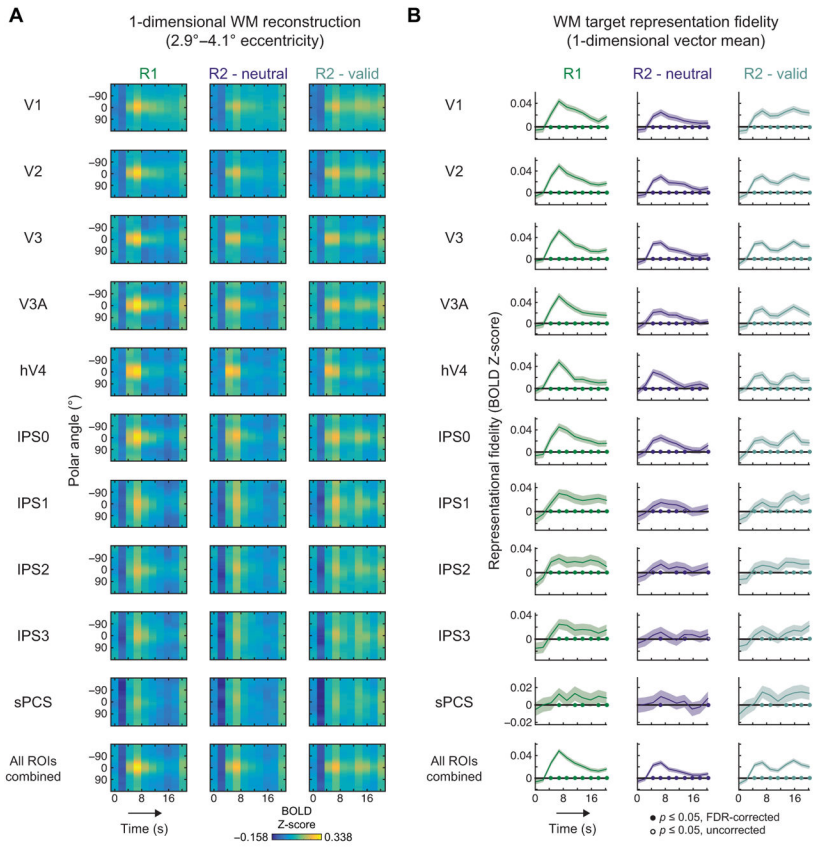

Figure 5. WM representations persist throughout entire delay

We computed reconstructions along radial vectors spanning the full circle and averaged reconstruction activation from 2.9–4.1° eccentricity, then rotated all reconstructions such that the probed target appeared at 0° (Fig. 2C).

(A) Reconstructed target representations for each ROI and WM condition throughout the trial, averaged over all participants. A bright streak appears at 0° on many plots, indicating that a WM representation was present throughout the delay interval.

(B) WM representational fidelity (Fig. 2C) computed for each time point. Although representational fidelity weakened later in the trial on R1 and R2-neutral trials, representations could still be identified. On R2-valid trials, representational fidelity increased following the informative cue, indicating that the cue enabled the remaining representation to be bolstered. Filled symbols at $y = 0$ indicate significant WM representations, FDR corrected ($q = 0.05$; across all ROIs, WM conditions and time points); open symbols indicate non-significant trends at $p \leq 0.05$, uncorrected; shaded regions mark 95% CIs via resampling procedure.

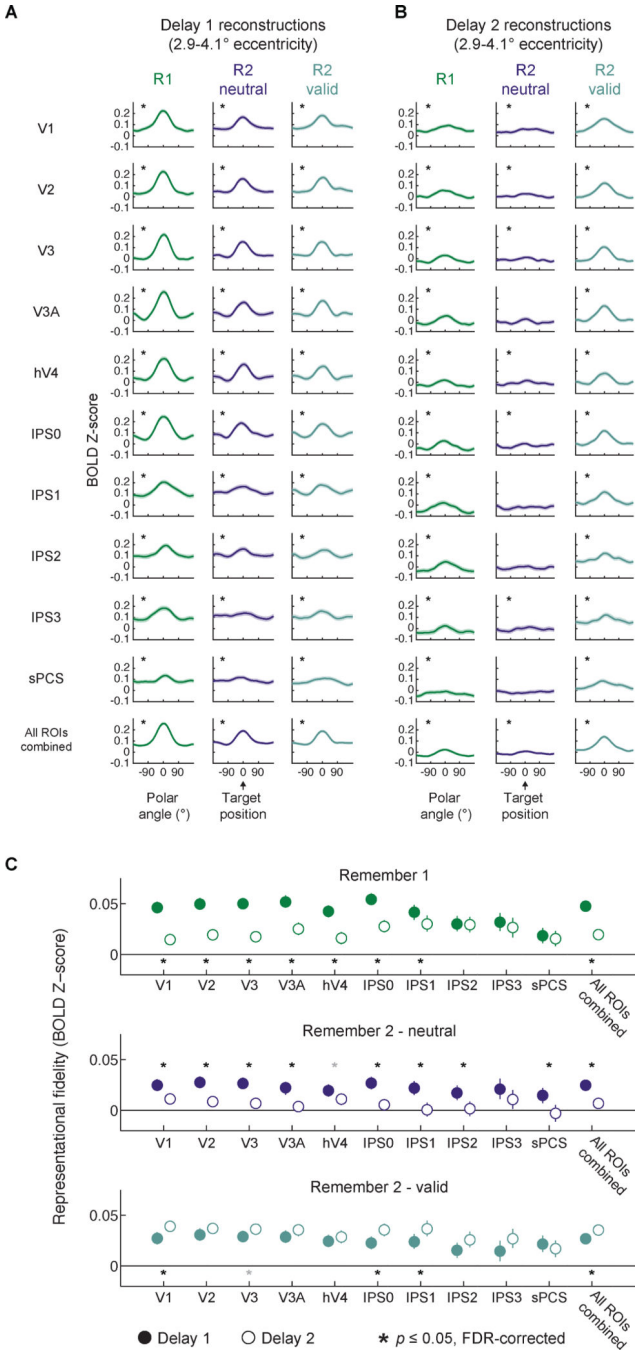

Figure 6. WM representations recovered after valid cue

(A–B) 1-d polar angle reconstructions as in Fig. 5A, averaged over each delay. Black asterisks indicate significant WM representations (FDR-corrected); gray asterisks indicate non-significant trends ($p \leq 0.05$; uncorrected; see Table S2 for all p-values from this analysis); shaded regions mark 95% confidence intervals via resampling procedure.

(C) Direct comparison of WM representations between delay periods. After a neutral cue (R1 and R2- neutral), the fidelity of representations decreased in many ROIs. In contrast, a valid cue significantly enhanced WM representations in V1, IPS0, IPS1, and All ROIs

combined. Black asterisks indicate significant differences between delay periods, two-tailed, FDR-corrected ($q = 0.05$); gray asterisks indicate non-significant trends defined as $p \leq 0.05$, uncorrected. Error bars mark 95% CIs via resampling procedure. See Table S3 for all p-values from this analysis. See Fig. S4 for an alternate means of quantifying WM representations, Fig. S5 for a comparison of this effect between each pair of time points, and Fig. S6 for a comparison of this effect between each pair of ROIs.

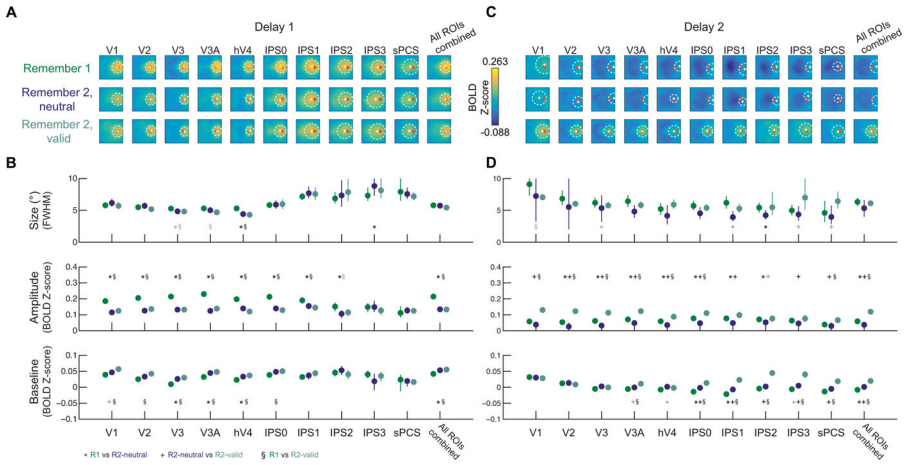

Figure 7. WM load and retro-cue altered WM representation amplitude

To quantify WM target representations, we coregistered reconstructions from each trial so that all targets appeared at the same position (red circle in A; see Fig. S3D). We resampled all trials within each condition, with replacement, 1,000 times, computed an average reconstruction from the resampled trials, and fit a surface allowed to vary in its size (full-width half-maximum; FWHM), amplitude, and baseline constrained to the position with maximum reconstruction activation for that resampling iteration (see Supplemental Experimental Procedures; Fig. 2D).

(A,C) Average reconstructions over all resampling iterations with mean (+) and size (dashed circle) of best-fit surfaces.

(B,D) Best-fit parameters from surface fitting for each condition. We computed pair-wise *p*-values between all condition pairs via resampling (see Experimental Procedures). Black symbols indicate significant pairwise differences after FDR correction within each fit parameter ($q = 0.05$). Gray symbols indicate trends, defined as $p \leq 0.05$, uncorrected. Error bars indicate 95% CIs obtained from the distribution of best-fit parameters to resampled reconstructions. All *p*-values shown in Table S4.

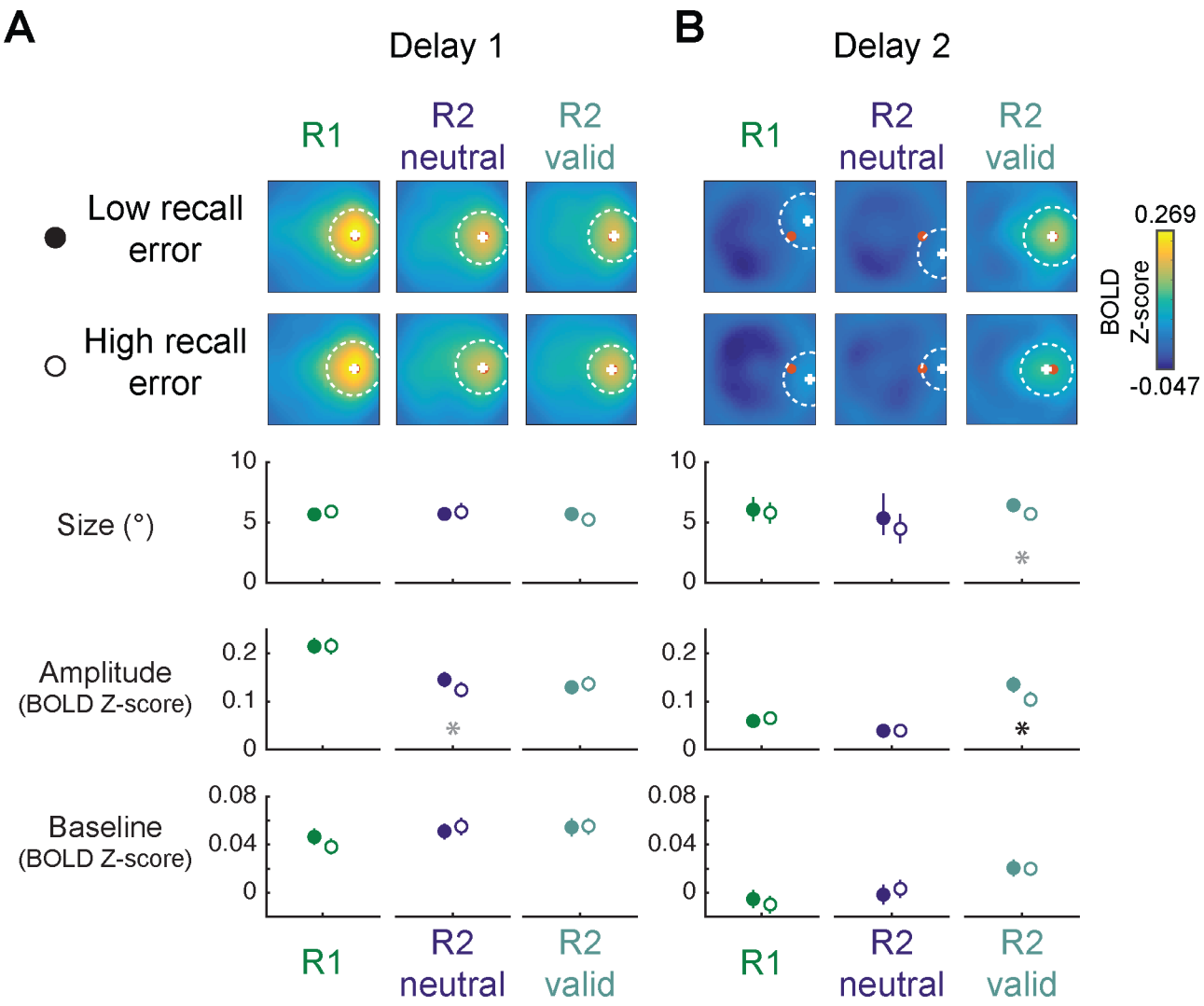

Figure 8. Recovered WM representation amplitude on R2-valid trials tracked behavioral performance

Within each participant, session, and WM condition, we split trials based on median recall error, then quantified low- and high-error reconstructions separately via a resampling procedure. All data shown are from reconstructions computed from All ROIs Combined (see Fig. S8 for each individual ROI). Plotted as in Fig. 7.

(A) During Delay 1, reconstructions were similar across high and low recall error trials.

(B) During Delay 2, the cued representation on R2-valid trials was visibly more robust on low- compared to high-error trials. This recovered WM representation was related to behavioral performance selectively via amplitude: on trials when participants performed more accurately, cued representation amplitude was higher. All other WM conditions and parameters showed no differences across behavioral performance bins. Error bars mark 95% CIs of fit parameters to resampled reconstructions. Black asterisk indicates significant differences after FDR correction ($q = 0.05$); gray asterisk indicates trends defined as $p \leq 0.05$, uncorrected. See Table S5 for all p-values.