Restoring Latent Visual Working Memory Representations in Human Cortex

Highlights
- Visual working memory (WM) representations are impaired under increasing WM load
- No additional information for performing WM tasks can be acquired from environment
- Degraded behavioral performance and neural representations recover with a retro-cue
- Recovery requires the existence of extra “latent” information before the retro-cue

Authors
Thomas C. Sprague, Edward F. Ester, John T. Serences

Correspondence
tsprague@nyu.edu (T.C.S.), jserences@ucsd.edu (J.T.S.)

In Brief
In working memory (WM), additional information for task performance cannot be acquired from the environment. Sprague et al. use fMRI and an image-reconstruction technique to demonstrate recovery of degraded WM representations, necessitating the existence of “latent” information in neural codes.
Restoring Latent Visual Working Memory Representations in Human Cortex

Thomas C. Sprague,1,4,* Edward F. Ester,2 and John T. Serences1,2,3,*
1Neurosciences Graduate Program
2Department of Psychology
3Kavli Institute for Brain and Mind
University of California, San Diego, La Jolla, CA 92093, USA
4Present address: Department of Psychology, New York University, New York, NY 10003, USA
*Correspondence: tsprague@nyu.edu (T.C.S.), jserences@ucsd.edu (J.T.S.)
http://dx.doi.org/10.1016/j.neuron.2016.07.006

SUMMARY

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, 1998; Bays, 2015; Curtis and D’Esposito, 2003; D’Esposito and Postle, 2015; 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 (Bays and Husain, 2008; Bays, 2014, 2015; 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, 2014, 2015; Carandini and Heeger, 2011; Franconeri et al., 2013; 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, 2010, 2014; 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 retro-cue 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 blood-oxygen-level-dependent [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 (LTM) (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 one or two 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 one (remember one [R1]) or two (remember two [R2]) items from a two-item display in spatial WM as indicated by the initial color of a central fixation point over a 16-s delay period. On 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 (Figure 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 a vertical or horizontal response bar, respectively (Figure 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 (Figure 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 six invisible discs that were spaced equally along a ring 3.5° from fixation and which were rotated around fixation on every trial (Figure 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 (324–378 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 (Figure 1C; R1 versus 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; 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 versus R2-valid, p = 0.008). Performance was slightly worse on R2-valid trials compared to R1 trials (R1 versus 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; Makovsk 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 WM based on BOLD fMRI activation patterns (Brouwer and Heeger, 2009; Ester et al., 2013, 2015; Sprague and Serences, 2013; Sprague et al., 2014, 2015). We computed reconstructions in each of ten independently identified regions of interest (ROIs) we have studied previously: retinotopic occipital visual areas (V1–hV4; V3A), retinotopic 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; Srimal and Curtis, 2008). We also assayed representations encoded by the joint activation patterns of items maintained in WM, we can all be averaged by rotating and aligning all reconstructed images (Stokes and Spaak, 2016) (Figure S3E). 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 Figure 1B). We combined trials with similar relative target arrangements, and rotated reconstructions to align similar trials (Supplemental Experimental Procedures; Figure S3). On R1 trials, reconstructions computed using an early time point (4.50 s) contained representations of both targets (Figure 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 (Figure 3B): representations of items maintained in WM persisted in reconstructions.

Figure 1. Spatial WM Performance Recovers following a Retro-Cue

(A) On each trial, participants viewed two 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 one [R1]). In the remaining 67% of trials, we cued participants to remember the locations of both targets (remember two [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 two targets appeared at positions uniformly drawn from two 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 tests, Bonferroni corrected for three 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 Figure S1 for recall error histograms and Figure S2 for univariate fMRI activation for each condition.
We quantified several parameters of WM representations (amplitude, size, and baseline) by fitting a 2D surface to average coregistered reconstructions.

In the delay period (after the encoding transient subsides), compared to each of the two target representations earlier, the cued item during the late delay period appeared enhanced extended delay intervals. Furthermore, the representation of targets tracked the dynamically changing contents of WM over delay following the cue, confirming that these WM reconstructions measured in "voxel 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 (Supplemental Experimental Procedures; Figure S3D) so that targets were at known positions, enabling us to average over trials with different WM contents (Stokes and Spaak, 2016).

To quantify the strength of WM representations, we computed a "representational fidelity" metric by extracting a 1D 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 1D 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°). We quantified several parameters of WM representations (amplitude, size, and baseline) by fitting a 2D surface to average coregistered reconstructions (Figure S3D) on each of 1,000 resampling iterations (Figures 7 and 8). To assess significance, we compared distributions of best-fit parameters between conditions (Figure 7) or behavioral performance bins (Figures 8 and S8).

See also Figure S3 and Supplemental Experimental Procedures.

through the late delay period, though target representations were weaker than those in R1 trials. On R2-valid trials (Figure 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 two 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 (Figure 1B), target representations always appeared nearby and only in the position(s) corresponding to the remembered item(s) during that condition and delay period (Figure 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 two items were maintained (Figure 4A compared to Figures 4C and 4E), and during the second delay period, a valid cue restored the cued representation to a high-fidelity state (Figures 4E and 4F).

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 1D reconstruction as a function of polar angle (Figure 2C; Supplemental Experimental Procedures).

First, we plotted these rotated and aligned 1D reconstructions as a function of time (Figure 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 1D 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 1D reconstructions to a common center; Figure 2C; Supplemental Experimental Procedures, Equation 5). If this metric is reliably greater than 0 (statistically evaluated using a resampling procedure; 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-neutral trials but substantially recovered following the valid cue on R2-valid cue trials (e.g., V1; Figure 5B).

Next, we compared 1D polar angle reconstructions and representational fidelity during each delay period (Figure 6). Importantly, we found significant representational fidelity in all ROIs across both delay intervals on R1 and R2-valid cue conditions (p < 0.001; one-tailed resampling test against 0, false discovery rate [FDR] corrected; Experimental Procedures; all p values for all reported comparisons available in Tables S1–S6, S7, and S8; 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 (Figures 6A and 6B; significant ROIs all p < 0.034, maximum p value IPS0; non-significant ROIs all p ≥ 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 (Figure 6C). Representational fidelity significantly declined from delay 1 to delay 2 in V1-hV4, IPS0, IPS1, and all ROIs combined on R1 trials (p < 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 and delay 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 (Figure 4), and quantify a significant enhancement of representations on R2-valid trials following the cue (Figure 6C). This result is not contingent on this particular quantification strategy (Figure S4) or 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 (Figure 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 (Figure S6). We also pursued an exploratory analysis of prefrontal cortex WM representations (Figure 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., 2014, 2015). First, we precisely aligned all reconstructions across trials such that the target position was at a known position (Figures 7A and 7C, red dots; Figure 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 (Experimental Procedures; Figures 2D and S3E). Because fits to a reconstruction with representational fidelity indistinguishable from 0 (Figures 5 and 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 (Figure 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; Figure 7B) and posterior parietal (IPS0 and IPS1, p \( \leq 0.016 \); and IPS2 for R1 versus R2-neutral, p = 0.012; maximum p value IPS1, R1 versus 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; Experimental Procedures, “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 R2-valid trials as compared to R1 trials in V3, V3A, and hV4 and in reconstructions computed from all ROIs combined (Figure 7B; p \( \leq 0.018 \);
maximum p value V3A, R1 versus R2-neutral). In V1, V2, and IPS0, a significantly greater baseline was seen when comparing R2-valid to R1 trials (p < 0.01, maximum p value IPS0). Finally, WM representation size was significantly smaller on R2-neutral and R2-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

Figure 5. WM Representations Persist throughout the Entire Delay

We computed reconstructions over a circle surrounding fixation averaged reconstruction activation from 2.9° to 4.1° eccentricity, then rotated all reconstructions such that the probed target appeared at 0° (Figure 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 (Figure 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 < 0.05, uncorrected; shaded regions mark 95% CIs via resampling procedure.
Figure 6. WM representations Recovered after Valid Cue

(A and B) 1D polar angle reconstructions as in Figure 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 and 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 Figure S4 for an alternate means of quantifying WM representations, Figure S5 for a comparison of this effect between each pair of time points and Figure 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; Figure 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 (Supplemental Experimental Procedures; Figure 2D).

(A and C) Average reconstructions over all resampling iterations with mean (+) and size (dashed circle) of best-fit surfaces. (B and D) Best-fit parameters from surface fitting for each condition. We computed pairwise p values between all condition pairs via resampling (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 < 0.05 (uncorrected). Error bars indicate 95% CIs obtained from the distribution of best-fit parameters to resampled reconstructions.

All p values are shown in Table S4.

WM load is manipulated from one to two 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 (Figure 7B).

Delay 2: Representation Amplitude Increased after Cue

During delay 2, target representations appeared weaker, though they were still identifiably present in many ROIs (Figure 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 (Figure 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 (Figure 7D; p < 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 < 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 < 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 > 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 (Figure 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-neutral 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, 2010, 2014; Sprague et al., 2014, 2015), and the latter reflect non-spatially specific increased mean activation levels in these regions following an informative cue (see also Figure S2).
DISCUSSION

Behavioral judgments about sensory stimuli rely on population-level neural representations that 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 WM can be used, as no further information can be gathered from the environment. We used an image reconstruction technique (Figure 2) to compare the fidelity of region-level WM representations across memory load conditions and replicated previous findings that behavioral performance (Figure 1) and neural representations (Figures 3, 4, 5, 6, and 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 (Figures 4, 5, 6, and 7), and the quality of this recovered representation was related to behavioral performance on the task (Figure 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, 2014, 2015). 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, 2016; 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 present in the environment, 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 that can then support improved behavioral performance.

Sources of Recovered Information

Both our behavioral (Figure 1C) and neural (Figures 4, 5, 6, 7, and 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, 2011) and beyond an initially degraded state that can then support improved behavioral performance.

Sources of Recovered Information

Both our behavioral (Figure 1C) and neural (Figures 4, 5, 6, 7, and 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, 2011) and beyond an initially degraded state that can then support improved behavioral performance.
2010, 2014; 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 (Figure 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 Tsoory, 2014). Changes in short-term synaptic efficacy, as suggested by prior theoretical and computational modeling efforts (Barak and Tsoory, 2014), the restored representation must result from neural response features inaccessible to or hidden from our BOLD activation pattern measurements before the cue.

Fidelity of Feature Representations in WM
Several previous studies cued participants to focus on a single item among multiple items maintained in WM. Lepsien and Nobre (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 retro-cued 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 (Figure 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.
Role of Long-Term Memory
Improved behavioral performance and restored representational fidelity following a valid retro-cue could also result from LTM retrieval. Recent behavioral studies have found that high-fidelity feature representations could be recalled from LTM (Brady et al., 2013; Sutterer and Awh, 2016) 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 deem this possibility unlikely. Accordingly, 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 synergetic 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 (Figures 1C, 5, 6, 7, and 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 six participants naive to the purpose of the experiment. All participants underwent three fMRI scanning sessions and one retinotopic mapping scanning session, each lasting 2 hr. Participants gave written informed consent as approved by the University of California, San Diego (UCSD) Institutional Review Board and received monetary compensation for their time ($20/hr for fMRI sessions and $10/hr for behavioral sessions).

Spatial WM Retro-Cueing Task
We presented two target stimuli (a red and a blue dot) for 500 ms 3.5° from fixation on average. The fixation point immediately changed color to red, blue, or purple. A red or blue fixation cue (one-third of trials) indicated one target should be maintained in WM over the delay interval (R1). A purple fixation cue (two-thirds 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. In one-half of R2 trials (one-third 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 one or two items over an 8-s delay interval (delay 2).

Participants also performed a spatial mapping task to estimate spatial sensitivity for each voxel (Figures S3A–S3C) and a visual localizer task to select voxels for further analysis, and each is 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 an 3-T GE MR750 scanner with a voxel size of 2.250 ms repetition time (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 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 Figures 2 and 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 (Figures 1, 6, 7, and 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 (Figures 5 and 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 Tables S1–S6, S7, and S8. All error bars and 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.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, eight figures, and eight tables and can be found with this article online at http://dx.doi.org/10.1016/j.neuron.2016.07.006.

AUTHOR CONTRIBUTIONS

T.C.S., E.F.E., and J.T.S. designed the experiment and wrote the manuscript; T.C.S. and E.F.E. acquired data; and T.C.S. analyzed data.

ACKNOWLEDGMENTS

We thank Edward Awh, Brad Postle, Sirawaj Ithipuripat, Stephanie Nelli, Roseanne Rademaker, Samantha Scudder, and V’y 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. This work was funded by National Institutes of Health (NIH) grant R01-EY025872 and a James S. McDonnell Foundation Scholar Award to J.T.S., NIH grant T32-MH20002 to T.C.S. and E.F.E., and a National Science Foundation Graduate Research Fellowship to T.C.S.

Received: December 11, 2015
Revised: May 5, 2016
Accepted: July 4, 2016
Published: August 3, 2016

REFERENCES


Supplemental Information

Restoring Latent Visual Working Memory

Representations in Human Cortex

Thomas C. Sprague, Edward F. Ester, and John T. Serences
Figure S1 Recall performance recovers when one of two items is cued (related to Figure 1)
Histograms of recall error across all trials for each participant and condition for data presented in Fig. 1C. Y axis indicates “proportion of trials”. Same participant identifiers used as in previous reports to facilitate comparison of data across experiments (Ester et al., 2015; Sprague and Serences, 2013; Sprague et al., 2014). Note that only 1 participant (AI) previously participated in a spatial WM experiment. All results presented in this report hold when excluding this participant.
Figure S2 Univariate BOLD responses from each ROI (related to Figures 1, 3)

(A) Mean BOLD activation timecourse (event-related average, time-locked to beginning of WM delay periods) averaged across all trials, all participants, and all voxels within each ROI. Replicating previous work (Emrich et al., 2013; Harrison and Tong, 2009; Riggall and Postle, 2012; Serences et al., 2009; Sprague et al., 2014), we observe no substantial activation in occipital ROIs (V1-hV4) in the univariate BOLD signal. For subsequent analyses, we identified time points primarily corresponding to the delay period before the cue (Delay 1, 6.75-9.00 s), and the delay period after the cue (Delay 2; 15.75-18.00 s).

(B) Mean delay-period activation during Delay 1 (left) and Delay 2 (right) as a function of WM condition. During Delay 1, we found trends towards increased activation with set size (R2-neutral>R1 and/or R2-valid>R1) in IPS0 and IPS1. We also observed significantly higher activation during R2-valid trials in hV4 as compared to R1, but not R2-neutral, trials. During Delay 2, we observed significant cue-related activation (R2-valid>R1 and/or R2-valid>R2-neutral) in hV4, IPS0-IPS3, and sPCS, as well as trends towards this effect in all other regions. Significant tests reflect FDR-correction for all comparisons. Trends defined as p < 0.05, uncorrected for multiple comparisons. Error bars 95% confidence intervals via resampling all trials, with replacement, 1,000 times (see Supplemental Experimental Procedures: statistical procedures). All p-values for this analysis presented in Table S1.
Figure S3 IEM procedures: mapping task, stimulus layout, and reconstruction coregistration (related to Figure 2 and Experimental Procedures)

(A) Participants performed 4 runs of a spatial mapping task during each fMRI scanning session. On each trial, we presented a single WM target stimulus (red dot) for 500 ms, followed immediately by a flickering checkerboard (1.083° radius; 6 Hz full-field flicker) overlapping the WM target location. After 3,000 ms, a probe stimulus (black dot) appeared slightly offset to either the left or right, or above or below, the remembered position (distance varied across runs to equate difficulty) for 750 ms. Simultaneously, a horizontal or vertical bar appeared at fixation, indicating the participant must make a 2AFC “left/right” or “above/below” judgment in response to the question “was the probe dot [left/above] or [right/below] [of] the remembered position?” before the end of the inter-trial interval (2-4.5 s). All stimulus features are drawn to scale. Participants performed on average 87.62% correct (target/probe separation distance adjusted across runs to maintain sufficient task difficulty).

(B) The position of the mapping stimulus varied on each trial along a hexagonal grid (black circles), both inside and outside the range of eccentricities used for the main WM task (red ring). This enabled us to reconstruct images of the contents of spatial WM across the entire visual field subtended by the projector screen inside the scanner (Fig. 2), despite only remembering items from a small range of positions in the WM task (Fig. 1). Blue dots indicate the center of spatial filters used for image reconstruction (Fig. 2).

(C) On each run of the spatial mapping task, we rotationally offset the position of the mapping stimuli by a fixed angular amount. Across sessions, we adjusted the “baseline” angle by 5° (session 1 arrangement shown here).

(D) On each trial of the primary WM task (Fig. 1), the WM targets appeared pseudo-randomly within the red dashed ring in (B). To align data across trials in “information space”, we rotated basis functions so as to zero-out the polar angle component of the WM target coordinate (1-d reconstructions & representational fidelity analyses; Figs. 5-6, S5-S7). Then, for analyses in which we precisely aligned target positions (Figs. 7-8, S8), we also shifted them horizontally to precisely align the target position to the coordinate x = 3.5°, y = 0° (see red dot, Fig 7A, C). For example, if a target appeared at 42° polar angle (up and to the right) and 3.7° eccentricity, we first rotated each basis function by 42° polar angle clockwise, then shifted all basis functions horizontally 0.2° to the left, before computing reconstructions. This means that we used a slightly different set of basis functions for computing each trial’s reconstructions (same set of basis functions used for each time point of each trial), eliminating any potential idiosyncrasies caused by the exact set of filter centers we used.

(E) Once we averaged coregistered reconstructions from all trials (on each resampling iteration, see Experimental Procedures: Statistical procedures), we fit a surface function (slice shown), which was shaped identically to each spatial filter, to the mean reconstruction. We allowed the function to vary in its size, baseline, and amplitude. Its position was constrained to be nearby the maximum pixel of the average reconstruction (see Supplemental Experimental Procedures).
Figure S4 Informative cue shifts target representations from R2- to R1-like state (related to Figures 5-6)

As an alternative visualization of the time course of WM target representations to those shown in Figures 3 and 5, we extracted the activation from each reconstruction within a 0.5° radius circular aperture centered at the exact target positions for each trial. We call this signal the “reconstruction activation”, as it reflects BOLD activation patterns transformed into visual field coordinates and extracted at the relevant visual field position. Then, we computed the difference between the activation at the probed target location and the non-probed target location (on R1 trials, the probed target was always the target in WM, on R2-neutral trials, the probed target was the one queried at the end of the trial; on R2-valid trials, the probed target was the validly-cued target in WM).

(A) On R1 trials, the remembered target representation shows elevated activation relative to the non-remembered target representation throughout the entire 16 s delay interval, despite the weakening target representations as visualized in re-
constructions in Figs 3-4. 

(B) On R2-neutral trials, both target representations are equal throughout the delay period, with the queried target representation becoming stronger once the response period begins (16.0 s). 

(C) On R2-valid trials, we see a transition from R2-neutral-like target representations (both are equal, and so the difference is near zero) during the first delay period to R1-like target representations (the remaining target representations recover) during the second delay period. Black triangle at 8.0 s indicates beginning of second delay interval. Units are BOLD Z-score. Dashed lines mark 95% CI via resampling, see Experimental Procedures: Statistical procedures. 

(D-E) We also computed mean delay-period reconstruction activation separately for probed (filled bars) and non-probed (open bars) target positions for each participant individually (each symbol reflects a single participant, as in Fig. 1C; Figure S1) within Delay 1 (D) and Delay 2 (E). Asterisks in panels A-C indicate a significant change between Delay 1 and Delay 2 (two-tailed); asterisks in panels D-E indicate that the probed target representation activation is greater than the non-probed target representation activation (one-tailed). All statistics computed using a resampling procedure (see Experimental Procedures: Statistical procedures). Black asterisks indicate a significant difference after FDR-correction for multiple comparisons (q = 0.05); gray asterisks indicate a non-significant trend defined using an uncorrected threshold of α = 0.05. All p-values from this analysis available in Table S6.
Figure S5 Cue-related representation recovery does not depend on time points chosen (related to Figures 5-6)

In all figures reporting mean delay-period activation, representational fidelity, or reconstructions, we used the 3rd and 4th imaging volumes following target presentation (6.75-9.00 s) for Delay 1, and the 7th and 8th volumes for Delay 2 (15.75 and 18.00 s, approximately 8 s following Delay 1, per timing of task events). We used these time points for Delay 1 following observations from our previous report (Sprague, Ester & Serences, 2014) that sensory-related representations in image reconstructions are almost entirely absent 6.75 s following target presentation on trials in which WM maintenance was not required. Delay 2 was accordingly chosen to be ~8 s after Delay 1 concluded. However, inspection of representational fidelity time courses in Fig. 5B suggests that these time points may not reveal the maximum cue-related restoration in representational fidelity. For some regions, the Delay 1 time points average a high-fidelity and lower-fidelity representation, and the Delay 2 time points average a lower-fidelity and higher-fidelity representation. This may conservatively bias our results, and we may miss significant recovery of WM representations in some areas. We addressed this by comparing the difference in representational fidelity between each pair of time points across trials (resampled) against 0, two-tailed. We found at least one pair of significant time points suggestive of post-cue representational recovery in every ROI except sPCS, demonstrating that our general observation of cue-related representational recovery is not contingent on the time points chosen; and, if anything, our choice was conservative. Allowing for different time courses for different ROIs revealed that this phenomenon is widespread throughout the cortex.

(A) Example plot of one region and condition (V1, R2-valid). Each cell plots the signed difference between the representational fidelity for the probed WM target at each paired set of time points (T2 [row] – T1 [column]); for brevity we only plot below-diagonal cells (above-diagonal cells are the negative of their below-diagonal counterparts). Starting at the (blank) diagonal cell, moving down a column indicates the difference between that time point and each successive time point. Left columns are mostly positive, indicating that target representations emerge early in the trial and remain positive. In this R2-valid example, the positive block below the diagonal for later time points corresponds to a relative increase in fidelity between an earlier time point and later time points, consistent with a recovery in the cued WM representation.

(B) Paired time point comparisons plotted for all conditions (columns) and ROIs (rows). Black squares highlight significant cells (two-tailed, FDR-corrected across all ROIs, conditions, and time point pairs, q = 0.05). All ROIs combined corrected separately (see Experimental Procedures); gray squares highlight trends (defined as α = 0.05, uncorrected). All p-values for this analysis available in Table S8, available online as an Excel file.
Figure S6 Comparison of delay period-related changes between ROIs (related to Figure 6)
For each ROI pair, we computed the difference in the change in WM representational fidelity (Figs. 5-6) from Delay 1 to Delay 2 (computed as Delay 2 – Delay 1: positive differences reflect increased fidelity; negative differences reflect decreased fidelity) between each non-matching ROI pair. We computed this difference score on each of 1,000 resampling iterations, drawing from all trials concatenated across participants with replacement (Experimental Procedures), and compared the resampled distribution against 0 (two-tailed). Significant differences (FDR-corrected within each condition, q = 0.05) are highlighted with black boxes; trends (defined as p ≤ 0.05, uncorrected) are highlighted with gray boxes. In R1 trials, IPS1, anterior parietal (IPS2-IPS3) and frontal (sPCS) representations remained more stable throughout the entire delay period than did visual (V1-hV4) and posterior parietal (IPS0) representations (see Figs. 3-7 for visualizations of representations in each condition). All p-values for this analysis are reported in Table S7, available online as an Excel file.
Figure S7: Exploratory analysis of additional prefrontal cortex regions of interest (related to Figs. 3-6)

Previous human neuroimaging work has identified WM representations of a single remembered feature (e.g., orientation or stimulus category) in regions of visual (Albers et al., 2013; Bettencourt and Xu, 2015; Christophel et al., 2012; Ester et al., 2009; Harrison and Tong, 2009; Pratte and Tong, 2014; Riggall and Postle, 2012; Saber et al., 2015; Serences et al., 2009), parietal (Bettencourt and Xu, 2015; Christophel et al., 2012; Ester et al., 2015; Riggall and Postle, 2012; Saber et al., 2015), and frontal cortex (Ester et al., 2015; Lee et al., 2013). Interestingly, prefrontal cortex (PFC) representations seem to depend on the type of information maintained (Lee et al., 2013) and/or the analysis method used (Ester et al., 2015). As an exploratory analysis, we sought to evaluate to what extent spatial WM representations are carried by a subset of PFC regions identified using our spatial localizer task (see Supplemental Experimental Procedures). We identified several additional regions of interest in each participant’s prefrontal cortex (PFC) using their visual localizer data (used to restrict voxels analyzed for all analyses).

(A) Activation maps plotted separately for each participant on their individual inflated white/gray matter boundary surface, thresholded as indicated (chosen for each participant to maximize visibility and/or distinctness of activation clusters). Labels with arrows indicate clusters used to identify each PFC ROI (sPCS: superior precentral sulcus; iPCS: inferior precentral sulcus).
precentral sulcus; DLPFC+: dorsolateral prefrontal cortex and surrounding activation; SMA+ supplementary motor area and surrounding activation, likely including pre-SMA and human supplementary eye fields).

(B) Reconstructions from PFC activation patterns computed through time, as in Fig. 3. Cartoons above each panel indicate coregistered positions of targets, and yellow dashed circles indicate remembered positions at each point during the trial. Only iPCS appears to have an approximate representation of WM targets, and most prominently after the valid cue during R2-valid trials.

(C) Timecourse of representational fidelity (as in Fig. 5B). Filled symbols on horizontal axis indicate significant representations (one-tailed, FDR corrected across all conditions, PFC ROIs, and time points, q = 0.05); open symbols indicate trends (p ≤ 0.05, uncorrected). Shaded regions denote 95% confidence intervals, computed via resampling all trials with replacement (1,000 iterations).

(D) Comparison of representational fidelity between each delay period (as in Fig. 6C). Asterisks indicate significant differences between Delay 1 and Delay 2 (two-tailed, FDR corrected across all conditions and PFC ROIs, resampling test with 1,000 iterations). All p-values for (D) available in Table S3. Because this is an exploratory analysis, correction for multiple comparison for these p-values was conducted independently from correction for p-values used for a priori retinotopic ROIs V1-IPS3 and localizer-defined sPCS, which we have analyzed in a previous report (Sprague et al., 2014). The absence of WM representations in DLPFC+ is not necessarily surprising in the present study, as its role in the maintenance of spatial positions in humans has recently come into question (Mackey et al., 2016). Additionally, we presented WM targets at ~3.5° eccentricity in this experiment. Insofar as PFC neurons have larger receptive field sizes (often > 20° diameter, Mohler et al., 1973; Zirnsak et al., 2014), a larger stimulus display may result in more robust identification of spatial WM representations.
Figure S8 Comparison of quantified WM target representations across performance bins for each individual ROI (related to Figure 8)
Data plotted as in Figure 8, with trials sorted based on behavioral recall performance. All resampling and fitting procedures are identical to those used for Figure 8. All error bars 95% confidence intervals over resampled fitting iterations. Black asterisks indicate significant difference between low- and high-recall error trials for that WM condition, delay period, and fit parameter, FDR-corrected for multiple comparisons within each parameter (q = 0.05). Gray asterisks are trends, thresholded at α = 0.05, uncorrected for multiple comparisons. All p-values available in Table S5.

(A) Remember 1 trials, Delay 1. IPS0 baseline was significantly higher on low recall error trials (p < 0.001).
(B) Remember 1 trials, Delay 2.
(C) Remember 2-neutral trials, Delay 1. V2 baseline trended to be lower for low recall error trials (p = 0.02).
(D) Remember 2-neutral trials, Delay 2. In hV4, baseline trended to be smaller for low recall error trials (p = 0.01).
(E) Remember 2-valid trials, Delay 1.
(F) Remember 2-valid trials, Delay 2. In V3 and hV4, amplitude trended towards being larger on low recall error trials (p = 0.002 and p = 0.004, respectively), and size trended towards being larger on low recall error trials in hV4 (p = 0.048).
Table S1 (related to Figure S2)
P-values for comparisons of mean delay-period activation over all voxels within each ROI between WM conditions (two-tailed). All p-values reflect pair-wise comparisons between conditions (R1 vs R2-neutral, R2-neutral vs. R2-valid, and R1 vs. R2-valid). P-value of 0 indicates $p < 0.001$, the minimum p-value achievable per our resampling procedure with 1,000 iterations. Bold numbers indicate significant differences after FDR correction for all comparisons ($q = 0.05$, all comparisons and all individual ROIs). Italicized numbers indicate trends, defined using $\alpha = 0.05$, uncorrected. Significant comparisons and trends are shown graphically in Figure S2. FDR threshold for V1-sPCS is $p \leq 0.002$.
<table>
<thead>
<tr>
<th>Representational fidelity (Fig. 6A-B and Fig. S7)</th>
<th>Delay 1 (one-tailed)</th>
<th>Delay 2 (one-tailed)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R1</td>
<td>R2-neutral</td>
</tr>
<tr>
<td>V1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>V2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>V3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>V3A</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>hV4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>IPS0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>IPS1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>IPS2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>IPS3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>sPCS</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>All ROIs combined</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>iPCS</td>
<td>0.017</td>
<td>0.028</td>
</tr>
<tr>
<td>DLPCF+</td>
<td>0.52</td>
<td>0.541</td>
</tr>
<tr>
<td>SMA+</td>
<td>0.23</td>
<td>0.085</td>
</tr>
</tbody>
</table>

**Table S2 (related to Figure 6 and Figure S7)**

P-values for comparisons between representational fidelity computed separately within each WM delay (one-tailed, against the null hypothesis that representational fidelity ≤ 0). P-value of 0 indicates p < 0.001, the minimum p-value achievable per our resampling procedure with 1,000 iterations. Bold numbers indicate significant differences after FDR correction for all comparisons (q = 0.05, all conditions and all individual *a priori* ROIs [V1-sPCS], and separately for “All ROIs combined” and the exploratory PFC ROIs [Figure S7; iPCS-SMA+] across all conditions, see Experimental Procedures). Italicized numbers indicate trends, defined using α = 0.05, uncorrected. Significant comparisons and trends are shown graphically in Figure 6A-B. FDR threshold for V1-sPCS is p ≤ 0.034, for All ROIs combined is p < 0.001, and for PFC ROIs is p ≤ 0.002.
**Table S3 (related to Figure 6 and Figure S7)**

P-values for comparisons of representational fidelity between Delay 1 and Delay 2 (two-tailed). P-value of 0 indicates $p < 0.001$, the minimum p-value achievable per our resampling procedure with 1,000 iterations. Bold numbers indicate significant differences after FDR correction for all comparisons ($q = 0.05$, all conditions and all individual *a priori* ROIs [V1-sPCS], separately for “All ROIs combined”, and separately for exploratory PFC ROIs [iPCS-SMA+] across all conditions, see Experimental Procedures). Italicized numbers indicate trends, defined using $\alpha = 0.05$, uncorrected. Significant comparisons and trends are shown in Figure 6C and Figure 7D. FDR threshold for V1-sPCS is $p \leq 0.028$, for All ROIs combined is $p \leq 0.002$, and for PFC ROIs is $p \leq 0.008$.

<table>
<thead>
<tr>
<th>Representational fidelity (Fig. 6C; Fig. S7D)</th>
<th>Delay 1 vs. Delay 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R1</td>
</tr>
<tr>
<td>V1</td>
<td>0</td>
</tr>
<tr>
<td>V2</td>
<td>0</td>
</tr>
<tr>
<td>V3</td>
<td>0</td>
</tr>
<tr>
<td>V3A</td>
<td>0</td>
</tr>
<tr>
<td>hV4</td>
<td>0</td>
</tr>
<tr>
<td>IPS0</td>
<td>0</td>
</tr>
<tr>
<td>IPS1</td>
<td>0.028</td>
</tr>
<tr>
<td>IPS2</td>
<td>0.848</td>
</tr>
<tr>
<td>IPS3</td>
<td>0.362</td>
</tr>
<tr>
<td>sPCS</td>
<td>0.538</td>
</tr>
<tr>
<td>All ROIs combined</td>
<td>0</td>
</tr>
<tr>
<td>iPCS</td>
<td>0.84</td>
</tr>
<tr>
<td>DLPFC+</td>
<td>0.918</td>
</tr>
<tr>
<td>SMA+</td>
<td>0.934</td>
</tr>
</tbody>
</table>
**Table S4 (related to Figure 7)**

P-values for comparisons of parameters (size, amplitude, baseline) for best-fit surfaces to coregistered reconstructions between conditions for each delay period individually (two-tailed). All p-values reflect pair-wise comparisons between conditions (R1 vs R2-neutral, R2-neutral vs R2-valid, and R1 vs. R2-valid). P-value of 0 indicates $p < 0.001$, the minimum p-value achievable per our resampling procedure with 1,000 iterations. Bold numbers indicate significant differences after FDR correction for all comparisons ($q = 0.05$, all conditions and all individual ROIs, and separately for “All ROIs combined” across all conditions, see Experimental Procedures). Italicized numbers indicate trends, defined using $\alpha = 0.05$, uncorrected. Significant comparisons and trends are shown graphically in Figure 7.

<table>
<thead>
<tr>
<th>Delay</th>
<th>ROI</th>
<th>Size R1 vs R2-neutral to R2-valid</th>
<th>Size R2-neutral vs R2-valid</th>
<th>Amplitude R1 vs R2-neutral to R2-valid</th>
<th>Amplitude R2-neutral vs R2-valid</th>
<th>Baseline R1 vs R2-neutral to R2-valid</th>
<th>Baseline R2-neutral vs R2-valid</th>
<th>FDR thresh</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>V1</td>
<td>0.27</td>
<td>0.278</td>
<td>0</td>
<td>0.328</td>
<td>0</td>
<td>0.144</td>
<td>0.022</td>
</tr>
<tr>
<td>1</td>
<td>V2</td>
<td>0.506</td>
<td>0.092</td>
<td>0.182</td>
<td>0</td>
<td>0.24</td>
<td>0.138</td>
<td>0.07</td>
</tr>
<tr>
<td>1</td>
<td>V3</td>
<td>0.032</td>
<td>0.878</td>
<td>0.02</td>
<td>0</td>
<td>0.94</td>
<td>0.002</td>
<td>0.422</td>
</tr>
<tr>
<td>1</td>
<td>V3A</td>
<td>0.3</td>
<td>0.31</td>
<td>0.04</td>
<td>0</td>
<td>0.214</td>
<td>0.018</td>
<td>0.512</td>
</tr>
<tr>
<td>1</td>
<td>hV4</td>
<td>0</td>
<td>0.594</td>
<td>0</td>
<td>0</td>
<td>0.106</td>
<td>0.016</td>
<td>0.446</td>
</tr>
<tr>
<td>1</td>
<td>IPS0</td>
<td>0.685</td>
<td>0.852</td>
<td>0.702</td>
<td>0</td>
<td>0.368</td>
<td>0.052</td>
<td>0.664</td>
</tr>
<tr>
<td>1</td>
<td>IPS1</td>
<td>0.304</td>
<td>0.852</td>
<td>0.464</td>
<td>0.016</td>
<td>0.5</td>
<td>0.004</td>
<td>0.408</td>
</tr>
<tr>
<td>1</td>
<td>IPS2</td>
<td>0.706</td>
<td>0.702</td>
<td>0.3</td>
<td>0.012</td>
<td>0.556</td>
<td>0.042</td>
<td>0.396</td>
</tr>
<tr>
<td>1</td>
<td>IPS3</td>
<td>0.15</td>
<td>0.576</td>
<td>0.354</td>
<td>0.938</td>
<td>0.394</td>
<td>0.242</td>
<td>0.21</td>
</tr>
<tr>
<td>1</td>
<td>sPCS</td>
<td>0.788</td>
<td>0.488</td>
<td>0.478</td>
<td>0.422</td>
<td>0.872</td>
<td>0.474</td>
<td>0.418</td>
</tr>
<tr>
<td>1</td>
<td>All ROIs</td>
<td>0.822</td>
<td>0.256</td>
<td>0.108</td>
<td>0</td>
<td>0.818</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>V1</td>
<td>0.572</td>
<td>0.964</td>
<td>0.022</td>
<td>0.072</td>
<td>0</td>
<td>0</td>
<td>0.89</td>
</tr>
<tr>
<td>2</td>
<td>V2</td>
<td>0.626</td>
<td>0.842</td>
<td>0.152</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.804</td>
</tr>
<tr>
<td>2</td>
<td>V3</td>
<td>0.498</td>
<td>0.728</td>
<td>0.41</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.104</td>
</tr>
<tr>
<td>2</td>
<td>V3A</td>
<td>0.016</td>
<td>0.096</td>
<td>0.292</td>
<td>0.004</td>
<td>0</td>
<td>0</td>
<td>0.35</td>
</tr>
<tr>
<td>2</td>
<td>hV4</td>
<td>0.282</td>
<td>0.066</td>
<td>0.244</td>
<td>0.002</td>
<td>0</td>
<td>0</td>
<td>0.048</td>
</tr>
<tr>
<td>2</td>
<td>IPS0</td>
<td>0.06</td>
<td>0.1</td>
<td>0.492</td>
<td>0</td>
<td>0</td>
<td>0.002</td>
<td>0.012</td>
</tr>
<tr>
<td>2</td>
<td>IPS1</td>
<td>0.002</td>
<td>0.03</td>
<td>0.112</td>
<td>0</td>
<td>0</td>
<td>0.118</td>
<td>0.004</td>
</tr>
<tr>
<td>2</td>
<td>IPS2</td>
<td>0.046</td>
<td>0.116</td>
<td>0.92</td>
<td>0.074</td>
<td>0.042</td>
<td>0.626</td>
<td>0.224</td>
</tr>
<tr>
<td>2</td>
<td>IPS3</td>
<td>0.398</td>
<td>0.024</td>
<td>0.072</td>
<td>0.084</td>
<td>0.016</td>
<td>0.366</td>
<td>0.044</td>
</tr>
<tr>
<td>2</td>
<td>sPCS</td>
<td>0.584</td>
<td>0.016</td>
<td>0.114</td>
<td>0.122</td>
<td>0</td>
<td>0.016</td>
<td>0.07</td>
</tr>
<tr>
<td>2</td>
<td>All ROIs</td>
<td>0.17</td>
<td>0.23</td>
<td>0.578</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>FDR thresh:</td>
<td>V1-sPCS</td>
<td>0.002</td>
<td>V1-sPCS</td>
<td>0.016</td>
<td>V1-sPCS</td>
<td>0.018</td>
<td></td>
<td></td>
</tr>
<tr>
<td>All ROIs</td>
<td>&lt; 0.001</td>
<td>All ROIs</td>
<td>&lt; 0.001</td>
<td>All ROIs</td>
<td>0.018</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**Table S5 (related to Figure 8 and Figure S8)**

P-values for comparisons of parameters (size, amplitude, baseline) for best-fit surfaces to coregistered reconstructions between low recall error and high recall error trials (two-tailed, always equal number of trials in each bin per participant and session). P-value of 0 indicates \( p < 0.001 \), the minimum p-value achievable per our resampling procedure with 1,000 iterations. Bold numbers indicate significant differences after FDR correction for all comparisons within each parameter (\( q = 0.05 \), all conditions and all individual ROIs, and separately for “All ROIs combined” across all conditions, see Experimental Procedures; FDR thresholds indicated at bottom of table). Italicized numbers indicate trends, defined using \( \alpha = 0.05 \), uncorrected. Significant comparisons and trends are shown graphically in Figure 8 and Figure S8. For the All ROIs combined comparisons, use of a threshold derived with Bonferroni’s method produces identical significant comparisons.
Table S6A (related to Figure S4A-C)
P-values for comparisons of target activation differences (probed target – non-probed target) between Delay 1 and Delay 2 (two-tailed). P-value of 0 indicates \( p < 0.001 \), the minimum p-value achievable per our resampling procedure with 1,000 iterations. Bold numbers indicate significant differences after FDR correction for all comparisons (\( q = 0.05, \) all conditions and all individual ROIs, and separately for “All ROIs combined” across all conditions, see Experimental Procedures). Italicized numbers indicate trends, defined using \( \alpha = 0.05 \), uncorrected. Significant comparisons and trends are shown in Figure S4A-C. FDR thresholds for V1-sPCS and for All ROIs combined are \( p < 0.001 \). Identical comparisons remain significant when correcting with Bonferroni’s method.

<table>
<thead>
<tr>
<th>Figure: S4A-C</th>
<th>V1</th>
<th>V2</th>
<th>V3</th>
<th>V3A</th>
<th>hV4</th>
<th>IPS0</th>
<th>IPS1</th>
<th>IPS2</th>
<th>IPS3</th>
<th>sPCS</th>
<th>All ROIs combined</th>
</tr>
</thead>
<tbody>
<tr>
<td>Remember 1</td>
<td>0.522</td>
<td>0.904</td>
<td>0.566</td>
<td>0.226</td>
<td>0.246</td>
<td>0.262</td>
<td>0.31</td>
<td>0.126</td>
<td>0.35</td>
<td>0.756</td>
<td>0.932</td>
</tr>
<tr>
<td>Remember 2 - neutral</td>
<td>0.764</td>
<td>0.3</td>
<td>0.556</td>
<td>0.788</td>
<td>0.57</td>
<td>0.944</td>
<td>0.382</td>
<td>0.2</td>
<td>0.682</td>
<td>0.234</td>
<td>0.862</td>
</tr>
<tr>
<td>Remember 2 - valid</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.028</td>
<td>0.38</td>
<td>0</td>
</tr>
</tbody>
</table>

Table S6B (related to Figure S4D)
P-values for comparisons between probed target (PT) activation and non-probed target (NPT) activation computed separately within each WM delay (one-tailed, against the null hypothesis that PT \( \leq \) NPT). P-value of 0 indicates \( p < 0.001 \), the minimum p-value achievable per our resampling procedure with 1,000 iterations. Bold numbers indicate significant differences after FDR correction for all comparisons (\( q = 0.05, \) all conditions and all individual ROIs, and separately for “All ROIs combined” across all conditions, see Experimental Procedures). Italicized numbers indicate trends, defined using \( \alpha = 0.05 \), uncorrected. Significant comparisons and trends are shown in Figure S4D. FDR threshold for V1-sPCS is \( p \leq 0.013 \) and for All ROIs combined is \( p < 0.001 \).

<table>
<thead>
<tr>
<th>Condition:</th>
<th>Remember 1</th>
<th>Remember 2 - neutral</th>
<th>Remember 2 - valid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Delay 1</td>
<td>Delay 2</td>
<td>Delay 1</td>
</tr>
<tr>
<td>V1</td>
<td>0</td>
<td>0</td>
<td>0.919</td>
</tr>
<tr>
<td>V2</td>
<td>0</td>
<td>0</td>
<td>0.959</td>
</tr>
<tr>
<td>V3</td>
<td>0</td>
<td>0</td>
<td>0.626</td>
</tr>
<tr>
<td>V3A</td>
<td>0</td>
<td>0</td>
<td>0.974</td>
</tr>
<tr>
<td>hV4</td>
<td>0</td>
<td>0</td>
<td>0.588</td>
</tr>
<tr>
<td>IPS0</td>
<td>0</td>
<td>0</td>
<td>0.937</td>
</tr>
<tr>
<td>IPS1</td>
<td>0.006</td>
<td>0</td>
<td>0.727</td>
</tr>
<tr>
<td>IPS2</td>
<td>0.012</td>
<td>0</td>
<td>0.173</td>
</tr>
<tr>
<td>IPS3</td>
<td>0.013</td>
<td>0</td>
<td>0.191</td>
</tr>
<tr>
<td>sPCS</td>
<td>0.127</td>
<td>0.068</td>
<td>0.153</td>
</tr>
<tr>
<td>All ROIs combined</td>
<td>0</td>
<td>0</td>
<td>0.925</td>
</tr>
</tbody>
</table>
**Table S7 (related to Figure S5)**  
(Available online as Excel spreadsheet)  
P-values for comparisons of restoration in representational fidelity (Delay 2 – Delay 1) between each pair of ROIs. All tests are two-tailed against the null hypothesis of no difference in restoration effect between each pair of ROIs. Italicized numbers indicate trends, defined using $\alpha = 0.05$, uncorrected. Bold indicates significant differences, FDR-corrected within each condition ($q = 0.05$; FDR thresholds for R1: $p \leq 0.014$, R2-neutral: $p < 0.001$, R2-valid: $p < 0.001$).

**Table S8 (related to Figure S6)**  
(Available online as Excel spreadsheet)  
P-values for comparisons of representational fidelity between each pair of time points (Time 2 – Time 1) for each condition and ROI. All tests are two-tailed against the null hypothesis of no difference in representational fidelity between each pair of time points. FDR threshold for V1-sPCS: $p \leq 0.028$, All ROIs combined: $p \leq 0.028$ (indicated with black squares in Figure S6).
SUPPLEMENTAL EXPERIMENTAL PROCEDURES

Participants
We recruited \( n = 6 \) participants (5 female; aged 22-29 yrs) naïve to the purpose of the experiment from the UC San Diego community. We used a small sample size, but acquired substantial data from each participant to maximize sensitivity to subtle WM representations, similar to our previous report (Sprague et al., 2014). Participant identifiers are identical to those used in previous reports to facilitate comparison of data across experiments (Ester et al., 2015; Sprague and Serences, 2013; Sprague et al., 2014). Participants AI and AL participated in the experiments reported in Sprague & Serences (2013). Participant AI participated in the experiments reported in Sprague et al (2014). Participants AI, AL and AP participated in Ester et al (2015). Participants gave written informed consent as approved by the UCSD Institutional Review Board and received monetary compensation for their time ($20/hr for fMRI sessions, $10/hr for behavioral sessions).

Spatial WM retro-cueing task
All participants underwent 3 fMRI scanning sessions and 1 retinotopic mapping scanning session, each lasting 2 hrs. Participants also completed 2-4 behavioral sessions, each lasting 1-1.5 hrs. The size of the stimulus display was fixed across all behavioral and scanning sessions. However, the size of the screen, which constantly contained a gray background, differed (inside scanner: 18.18° × 13.64°, aspect ratio 4:3; outside scanner: 44.71° × 25.15°, aspect ratio 16:9).

We adapted a spatial WM task reported previously (Sprague et al., 2014). On each trial, we presented 2 target stimuli (a red and a blue dot, 0.15° diameter) for 500 ms at pseudorandom locations 3.5° from fixation on average. Following target presentation, the fixation point (square, 0.2°/side) immediately changed color to either red, blue, or purple. A red or blue fixation cue (1/3 of trials) indicated the target to be maintained in WM over the delay interval (Remember 1). A purple fixation cue (2/3 of trials) indicated both targets should be maintained in WM (Remember 2). After an 8,000 ms delay interval (Delay 1), the fixation cue changed color once again. On Remember 1 trials, the cue always changed to black, indicating participants should maintain the encoded target in WM over the subsequent second delay interval. On ½ of Remember 2 trials (1/3 of trials overall), the fixation cue turned black, providing a neutral cue as to which target was relevant for behavior (Remember 2-neutral condition). On the remaining ½ of Remember 2 trials (1/3 of trials overall), the fixation cue changed from purple to either red or blue, cueing the participants with 100% validity to remember only one of the targets (Remember 2-valid condition). Following this cue change, participants continued to maintain 1 or 2 items in WM over an additional 8,000 ms delay interval (Delay 2).

At the end of each trial after both delay intervals, participants recalled the exact horizontal or vertical coordinate of the item cued by the color of the fixation point. The response coordinate was randomly chosen on every trial so that participants could not implement a uni-dimensional encoding scheme (i.e., encode only \( x \) or \( y \) coordinate). Participants responded by adjusting the position of a gray horizontal bar up or down (for \( y \) coordinate trials) or a vertical bar left or right (for \( x \) coordinate trials) using either a computer keyboard or an MR-compatible button box (bar thickness: 0.02°). We took the adjusted bar coordinate at the end of a 3,000 ms response window as the participant’s response.

Target locations were drawn from an isoeccentric ring 3.5° from fixation at 60° polar angle intervals along the ring, where the starting angle was jittered by up to ±15° on each trial. The position of the second target relative to the first target was always offset from the first target by 60°, 120°, or 180° in either direction (clockwise or counterclockwise, see colored discs in Fig. 1B). This resulted in a minimum target separation distance of 2.3° and a maximum separation distance of 8.2°. By using random target positions on each trial, we ensured that participants maintained precise spatial locations in WM rather than using alternative coding strategies, like verbally labeling the location(s). Additionally, constraining relative target positions within one of several discs allowed for comparison of data from trials with similar target arrangements (Figs. 3-4).

We counterbalanced trials for target position (1 of 6 discs), relative target position (1 of 3 relative angular separation distances, Fig. 1B), and memory condition (Remember 1, Remember 2-neutral, or Remember 2-valid), resulting in 54 trials per full counterbalanced repetition. Each full set of trials (or “super-run”) was broken up into 3 runs, each with 18 trials, each 19.5 s long. Trials were separated by a random inter-trial interval chosen from a uniform distribution from 3 to 6 s.

Spatial mapping task
Inside the scanner, participants completed 4 runs per session of a spatial mapping task used to estimate voxel-level encoding models for reconstruction analyses described below. On each trial, participants remembered the exact position of
a single target stimulus over a 3,000 ms delay interval during which a flickering checkerboard disc (6 Hz, full-field flicker, 1.083° radius, 1.474 cycles/°; Figure S3A) appeared nearby the memorized location. Following checkerboard presentation, participants indicated whether a probe stimulus (black dot) was either to the left or right or above or below the remembered stimulus position, as cued by an oriented bar at fixation (horizontal bar: respond left vs. right; vertical bar: respond above vs. below; probe and response bar presented for 750 ms). Participants could respond until the beginning of the next trial (after 2,000 – 4,500 ms inter-trial interval, uniform distribution). We maintained performance at ~75% correct by adjusting the target-probe separation distance between runs, but due to a programming error, accuracy was computed incorrectly during scanning (“null” trials were counted as incorrect responses, so accuracy on task trials was actually ~89%, not ~75% as computed within the stimulus presentation script, Figure S3 caption). To ensure participants did not just encode one target coordinate dimension (x or y), we jittered the irrelevant coordinate on each trial by a small amount, preventing a scenario in which the presentation of the probe stimulus added certainty to the position maintained in WM. Each run included 6 null trials (no target/mapping stimulus/probe presented) during which participants passively fixated until the subsequent trial began.

During each run of this spatial mapping task the checkerboard stimuli were presented at each of 36 positions arrayed along a hexagonal grid (see Figure S3B-C) and the target position was randomly chosen from a uniform disc centered at the checkerboard position with radius 0.542°. On each run, we rotated the angular orientation of the entire hexagonal grid by 15° polar angle (Figure S3C). Across sessions, we rotated the “baseline” angular orientation of the grid by 5° polar angle. This resulted in 4 × 3 × 36 = 432 unique stimulus positions across all scanning sessions. We used different grid orientations (and thus stimulus positions) on each scanning run to maximize the number of unique stimulus positions so that we could estimate as robust a spatial encoding model as possible (see below), as well as to ensure our model was not identifying peculiarities specific to a given set of mapping stimulus positions.

**Localizer task**

To focus our neuroimaging analyses to voxels responsive during spatial WM maintenance over the area subtended by our display setup, we scanned each participant on 6-8 runs (AI: 6, AL: 7, AS: 8, AR: 7, AP: 7, BC: 8) of a visual spatial WM localizer task similar to one we have described before (Sprague et al., 2014). On each trial we presented a flickering radial checkerboard annulus in one visual hemifield extending from 0.8° to 6.0° from fixation (1.25 cycles/° from fixation, 12° per polar angle cycle, 6 Hz contrast-reversing) for 10 s. During the stimulus interval, we presented 2 spatial WM trials in which participants remembered the precise position of 1 red dot over a 3 s delay interval. At the end of each delay interval, participants responded whether a green probe stimulus was to the left or to the right, or above or below, the remembered target position as indicated by a horizontal or vertical bar at fixation, respectively. WM targets could only appear within the stimulated hemifield. We maintained performance at ~75% by adjusting the task difficulty (target/probe separation distance) across trials. Stimulus epochs were separated by 3 – 5 s ITIs (uniform distribution). Each run contained 4 null trials that were the same duration as normal trials but did not contain checkerboard stimuli.

**Behavioral analysis**

For the main WM task, we defined behavioral recall error as the absolute distance along the relevant coordinate dimension (x or y) between the position of the response bar at the conclusion of the response window and the actual coordinate of the recalled target. We averaged all recall errors across all trials from scanning sessions within each participant.

In fMRI analyses in which we split trials based on behavioral performance, we computed the median recall error within each WM condition (R1, R2-neutral, R2-valid) within each scanning session. Trials with recall error greater than or equal to the median value were labeled “high recall error” and trials with recall error less than the median value were labeled “low recall error” (Figure 8 and Figure S8, Table S5).

**fMRI acquisition**

We scanned all participants on a 3 T research-dedicated GE MR750 scanner located at the UCSD Keck Center for Functional Magnetic Resonance Imaging with a 32 channel send/receive head coil (Nova Medical, Wilmington, MA) using identical sequences to those we have reported previously (Sprague and Serences, 2013; Sprague et al., 2014). We acquired functional data using a gradient echo planar imaging (EPI) pulse sequence (19.2 × 19.2 cm field of view, 96 × 96 matrix size, 31 3-mm-thick slices with 0-mm gap, obliquely-oriented through occipital, parietal & dorsal frontal cortex, TR = 2,250 ms, TE = 30 ms, flip angle = 90°, voxel size 2x2x3 mm, xyz).
To anatomically coregister images across sessions, and within each session, we also acquired a high resolution anatomical scan during each scanning session (FSPGR T1-weighted sequence, TR/TE = 11/3.3 ms, T1 = 1,100 ms, 172 slices, flip angle = 18°, 1 mm³ resolution). For all sessions but one, anatomical scans were acquired with ASSET acceleration. For the remaining session, we used an 8 channel send/receive head coil and no ASSET acceleration to acquire anatomical images with minimal signal inhomogeneity near the coil surface, which enabled improved segmentation of the gray-white matter boundary. We transformed these anatomical images to Talairach space and then reconstructed the gray/white matter surface boundary in BrainVoyager 2.6.1 (BrainInnovations) which we used for identifying ROIs.

fMRI preprocessing
We preprocessed fMRI data similarly to our previous report (Sprague et al., 2014). We coregistered functional images to a common anatomical scan across sessions (used to identify gray/white matter surface boundary as described above) by first aligning all functional images within a session to that session’s anatomical scan, then aligning that session’s scan to the common anatomical scan. We performed all preprocessing using FSL (Oxford, UK) and BrainVoyager 2.6.1 (BrainInnovations). Preprocessing included unwarping the EPI images using routines provided by FSL, then slice-time correction, three-dimensional motion correction (six-parameter affine transform), temporal high-pass filtering (to remove first-, second- and third-order drift), transformation to Talairach space (resampling to 2×2×2 mm resolution) in BrainVoyager, and finally normalization of signal amplitudes by converting to Z-scores separately for each run using custom MATLAB scripts. We did not perform any spatial smoothing beyond the smoothing introduced by resampling during the co-registration of the functional images, motion correction and transformation to Talairach space. All subsequent analyses were computed using custom code written in MATLAB (release 2014b, The Mathworks, Inc).

One participant (AS) changed positions inside the scanner substantially during one session. As a result, the field inhomogeneities estimated with the field map scan used for unwarping were only accurate for half of the runs during this session and could not be used to un warp the other half of scans. To mitigate this problem with the raw data, we did not perform unwarping on any session for this participant in order to maintain consistency in the analysis procedure across sessions for this participant. This did not appear to affect any aspect of their results.

Identifying regions of interest (ROIs)
Based on our previous work, we identified 10 a priori ROIs using independent scanning runs from those used for all analyses reported in the text. For retinotopic ROIs (V1-V3, hV4, V3A, IPS0-IPS3), we utilized a combination of retinotopic mapping techniques. Each participant completed several scans of meridian mapping in which we alternately presented flickering checkerboard “bowties” along the horizontal and vertical meridians. Additionally, each participant completed several runs of an attention-demanding polar angle mapping task in which they detected brief contrast changes of a slowly-rotating checkerboard wedge (described in detail in Sprague and Serences, 2013). We used a combination of maps of visual field meridians and polar angle preference for each voxel to identify retinotopic ROIs (Engel et al., 1994; Swisher et al., 2007). Polar angle maps computed using the attention-demanding mapping task for most participants are available in previous publications (AI: Sprague & Serences, 2013; AL and AP: Ester et al., 2015). We combined left- and right-hemispheres for all ROIs, as well as dorsal and ventral aspects of V2 and V3 for all analyses by concatenating voxels.

We defined superior precentral sulcus (sPCS) by plotting voxels active during either the left or right conditions of the localizer task described above (FDR corrected, q = 0.05) on the reconstructed gray/white matter boundary of each participant’s brain and manually identifying clusters appearing near the superior portion of the precentral sulcus, following previous reports (Srimal and Curtis, 2008). Additionally, for an exploratory post-hoc analysis of prefrontal cortex ROIs, we used activation maps from this localizer to identify significant voxels nearby the inferior aspect of the precentral sulcus (iPCS), dorsolateral prefrontal cortex (DLPFC+), and a medial region comprising the supplementary and pre-supplementary motor areas (SMA+). Because these ROIs were not always observable at the rigorous FDR-corrected threshold used to identify sPCS (an a priori chosen ROI), in some participants we adjusted the statistical threshold to maximize visibility and/or discriminability of the activation patches (see Fig. S7A).

The “All ROIs combined” region reported throughout the text consists of all voxels from all 10 individual a priori ROIs (V1, V2, V3, V3A, hV4, IPS0, IPS1, IPS2, IPS3, sPCS) concatenated together, and so all multivariate analyses involving this ROI reflect the net information content of the entire set of regions studied (reported also in Sprague et al., 2014).
**fMRI analysis: univariate**

For all ROI analyses, we used data from the localizer scans to identify voxels significantly active during checkerboard stimulus presentation and WM maintenance (FDR corrected, $q = 0.05$) for inclusion in further analyses. All analyses include only those voxels.

We computed BOLD time series by extracting signal at each time point averaged over all voxels within an ROI on each trial from 0 to 24.75 s (0 to 11 TRs) after the beginning of the first delay (rounded to the nearest TR), then averaging time series over all trials. We extracted mean activation levels for each delay period by averaging the TRs 6.75-9.00 s after probe onset for Delay 1 and 15.75-18.00 s after probe onset for Delay 2.

**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 (sensitivity profile over the relevant feature dimension(s) as parameterized by a small number of modeled information channels) 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 a “test set” into activation patterns in a modeled set of information channels.

We built an encoding model for spatial position based on a linear combination of spatial filters (Sprague and Serences, 2013; Sprague et al., 2015, 2014). Each voxel’s response was modeled as a weighted sum of 37 identical spatial filters arrayed in a hexagonal grid (Fig. 2A). Centers were spaced by 2.293° and each filter was a Gaussian-like function with full-width half-maximum of 2.523°:

$$f(r) = \left(0.5 + 0.5 \cos \frac{2\pi r}{s}\right)^7$$ for $r < s$; 0 otherwise

Where $r$ is the distance from the filter center and $s$ is a “size constant” reflecting the distance from the center of each spatial filter at which the filter returns to 0. Values greater than this are set to 0, resulting in a single smooth round filter at each position along the hexagonal grid ($s = 6.349°$; see Fig. 2A, Figure S3E for illustration of filter layout and shape; see also Sprague and Serences, 2013; Sprague et al., 2014).

This hexagonal grid of filters forms the set of information channels for our analysis. Each mapping task stimulus is converted from a contrast mask (1’s for each pixel subtended by the stimulus, 0’s elsewhere) to a set of filter activation levels by taking the dot product of the vectorized stimulus mask and the sensitivity profile of each filter. This results in each mapping stimulus being described by 37 filter activation levels rather than $1,024 \times 768 = 786,432$ pixel values. Once all filter activation levels are estimated, we normalize so that the maximum filter activation is 1.

We model the response in each voxel as a weighted sum of filter responses (which can loosely be considered as hypothetical discrete neural populations, each with spatial RFs centered at the corresponding filter position).

**Equation 2:**

$$B_1 = C_1 W$$

Where $B_1$ ($n$ trials $\times$ $m$ voxels) is the observed BOLD activation level of each voxel during the spatial mapping task (averaged over 6.75 – 9.00 s after mapping stimulus onset; Figure S3A), $C_1$ ($n$ trials $\times$ $k$ channels) is the modeled response of each spatial filter, or information channel, on each trial of the mapping task (normalized from 0 to 1), and $W$ is a weight matrix ($k$ channels $\times$ $m$ voxels) quantifying the contribution of each information channel to each voxel. Because we have more stimulus positions than modeled information channels, we can solve for $W$ using ordinary least-squares linear regression:

**Equation 3:**

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

This step is univariate and can be computed for each voxel in a region independently. Next, we used all estimated voxel encoding models within a ROI ($\hat{W}$) and a novel pattern of activation from the WM task (each TR from each trial, in turn) to compute an estimate of the activation of each channel ($\hat{C}_2$, $n$ trials $\times$ $k$ channels) which gave rise to that observed activation pattern across all voxels within that ROI ($B_2$, $n$ trials $\times$ $m$ voxels):

**Equation 4:**

$$\hat{C}_2 = B_2 \hat{W}^T (\hat{W} \hat{W}^T)^{-1}$$
The Moore-Penrose pseudoinverse of the estimated weight matrix from the training set \((\hat{\mathbf{W}})\) is the inverted part of the IEM: all encoding models across all voxels are used, and this step is multivariate. This analysis is only feasible when more voxels are measured than information channels are modeled. The Moore-Penrose pseudoinverse acts as a linear mapping from data measured in voxel space \((\mathbf{B}_d)\) into channel space \((\mathbf{C}_2)\), and accordingly stretches, scales and skews voxel activation patterns during this transformation, but importantly does not result in any nonlinear transformations. This analysis can be considered a directed form of dimensionality reduction in which activation patterns are transformed from an idiosyncratic activation pattern across voxels (unique to each individual participant and ROI, and thus difficult to directly compare) to a common information space, common across ROIs and participants, which allows for direct manipulation, quantification, and comparison of activation patterns in an intuitive and stimulus-referred coordinate space.

Once channel activation patterns are computed (Equation 4), we compute spatial reconstructions by weighting each filter’s spatial profile by the corresponding channel’s reconstructed activation level and summing all weighted filters together. This step aids in visualization, quantification, and coregistration of trials across WM target positions, but does not confer additional information.

We analyzed all data within each session: we used the 4 mapping task runs for a given session to estimate the encoding model for each voxel, then inverted that encoding model to reconstruct WM representations during all main WM task runs within that same session. Then, we averaged reconstructions over sessions within each participant.

Because WM target positions were unique on each and every trial, direct comparison of WM reconstructions on each trial is not possible without coregistration of reconstructions so that WM targets appeared at a common position across trials. To accomplish this, we adjusted the center position of the spatial filters on each trial such that we could rotate (and sometimes translate) the resulting reconstruction. For Figures 3-4, we rotated each trial such that one target (the target not queried at the end of each trial) was on average centered at \(x = 3.5^\circ\) and \(y = 0^\circ\) and the other target was in the upper visual hemifield (which required flipping ½ of reconstructions across the horizontal meridian). For Figures 7 and 8 and Figure S8, we coregistered each trial so that the queried target position was always centered at exactly \(x = 3.5^\circ\) and \(y = 0^\circ\) by first rotating the reconstruction so that the target was aligned along the positive x Cartesian axis, then horizontally translating it so that its x coordinate was exactly 3.5° (Figure S3D).

Because we carefully designed our task such that we presented an equal number of trials for each target separation condition (+60°, +120°, +180°, -60°, -120°, and -180° polar angle) in order to minimize the potential for participants to discover geometric regularities in the target arrangements, there was an overabundance of trials at ±180° polar angle separation distance, which led to a non-uniform distribution of positions for the non-coregistered target (that is, there were double the number of trials with non-coregistered targets at 180° polar angle from the coregistered target as there were for +60°, -60°, +120° and -120°). As a result, we excluded the second half of 180° separation condition trials from each super-run from all reconstruction-based analyses. When the other half of these trials is included, there is often a noticeable “bump” along the negative x axis corresponding to the greater number of trials in which a non-coregistered target appeared near that position, which renders quantification of target representations via curvefitting methods (see below) suboptimal.

**Quantifying WM representations: fidelity**

We took three approaches to WM representation quantification. First, we defined a “representational fidelity” metric that quantifies the extent to which a target representation reliably appeared within a reconstruction. To accomplish this, we first reduced the reconstruction from a 2-d image to a 1-d line plot by averaging over each of 220 evenly-spaced polar angle arms subtending 2.9-4.1° eccentricity (subset illustrated in Fig. 2C). The resulting 1-dimensional reconstruction reflects the average profile along an annulus around fixation. A target representation in these reconstructions would be a “bump” near 0° after the reconstructions have been rotated to a common center (where 0° corresponds to the actual target polar angle). To reduce these 1-d reconstructions to a single number which could be used to quantify the presence of target representations \((F)\), we computed a vector mean of the 1-d reconstruction \((r(\theta))\), where \(\theta\) is the polar angle of each point and \(r(\theta)\) is the reconstruction activation) when plotted as a polar plot, as projected along the x axis (because the reconstructions were rotated such that the target was presented at 0°; Fig. 2C):

**Equation 5:** \[F = \text{mean}(r(\theta) \cos \theta)\]

If \(F\) is reliably greater than zero, over a resampling procedure (see Statistical Procedures), this quantitatively demonstrates that the net activation over the entire reconstruction carries information above chance about the target position. This
measure is independent of baseline activation level in the reconstruction, as the mean of $r(\theta)$ is removed by averaging over the full circle. We computed timecourses of representational fidelity (Fig. 5B), as well as representational fidelity for each delay period (Fig. 6). To determine whether the cue on Remember 2-valid trials restores representations, we compared $F$ between Delay 2 and Delay 1 for each ROI, and between each pair of time points individually (Fig. S5). To evaluate whether ROIs differed in the extent to which their WM representations changed over the long delay interval, we compared the difference in fidelity between Delay 2 and Delay 1 between each pair of ROIs (Fig. S6).

**Quantifying WM representations: fit surfaces**

Additionally, we sought to evaluate the size, amplitude, and baseline of the WM target representation(s) from each WM condition and WM delay interval to establish how the information content of the population code changed across conditions. We followed procedures developed previously (Sprague et al., 2014) whereby we resampled all trials with replacement concatenated across all sessions from all participants from a condition 1,000 times and computed a single mean coregistered reconstruction (Figs. 7-8, Figure S8) on each resampling iteration. Then, we fit the mean reconstruction with a round Gaussian-like function parameterized by its center position, size, amplitude, and baseline:

$$f(r) = b + a \left(0.5 + 0.5 \cos \frac{2\pi r}{s}\right)^7 \text{ for } r < s; \text{ 0 otherwise}$$

Where $r$ is the distance from the center of the surface, $s$ is the size constant (as in Eq. 1), and $a$ and $b$ are the amplitude and baseline, respectively. Because there are many free parameters and some reconstructions are noisy, we adopted several heuristics to constrain our optimization problem. First, we found the maximum point on the entire reconstruction and used this as the center position (Sprague et al., 2014). Then, we performed a search through different sizes of fit surface function (FWHM: 0.099° to 9.934° in 0.099° steps). At each search iteration, we used ordinary least squares linear regression to find the amplitude and baseline which minimized residual errors between the reconstruction and the fit function. Finally, we used the best-fit amplitude, baseline, and size parameters from this search procedure and the global maximum position on the reconstruction as seed values for a constrained nonlinear optimization fitting algorithm (Matlab’s fmincon function) subject to several constraints: position could not deviate more than one reconstruction “pixel size” (0.235° x 0.235°) from the global maximum position; size could not surpass the range used in the grid search procedure (0.099° to 9.934°), and amplitude/baseline could each not go below -5 or above 10 (BOLD Z-score units). This entire curvefitting procedure was repeated on each resampling iteration, for each condition described in the text (R1, R2-neutral, R2-valid broken down by Delay 1 and Delay 2 for Fig. 7, each of those broken down by High and Low recall error for Fig. 8 and Figure S8), resulting in 1,000 resampled estimates of each fit parameter on each condition for each ROI. Average resampled reconstructions over all resampling iterations are shown in Figure 7A,C, Figure 8 and Figure S8.

**Quantifying WM representations: target activation**

As a third means of quantifying the integrity of WM representations, we evaluated the relative strengths of each target representation at each time point of the trial by extracting the average reconstruction activation within a 0.5° radius circle centered at each target position. Then, we took the difference between the reconstructed target representation activation of the target probed at the end of each trial and that of the target which was not probed at the end of each trial (on R1 trials, the probed target was always the remembered target; on R2-neutral trials, the probed target was the target queried at the end of the trial; on R2-valid trials, the probed target was always the remaining target following the valid retro-cue; Figure S4). This allowed us to directly compare the strength of the representation through time for each target in a manner which did not require fitting a surface with many free parameters.

**Statistical procedures**

All statistical statements reported in the text are based on resampling procedures in which a variable of interest is computed over 1,000 iterations. In each iteration, all single-trial variables from a given condition are resampled with replacement and averaged, resulting in 1,000 resampled averages for a given condition. We then subjected these distributions of resampled averages to pairwise comparisons by computing the distribution of differences between one resampled distribution (e.g., R1) and another resampled distribution (e.g., R2), yielding a new distribution of 1,000 difference values. We tested whether these difference distributions significantly differed from 0 in either direction by performing two one-tailed tests ($p = \text{proportion of values greater than or less than 0; null hypothesis that difference between conditions } = 0$) and doubling the smaller $p$ value. For the supplemental analysis in which we compared the change in fidelity between the two delay periods between each pair of ROIs, we compared the distribution of differences of delay period differences against 0, two-tailed ((ROI1: Delay
2 – ROI1: Delay 1) – (ROI2: Delay 2 – ROI2: Delay 1)). For tests in which we compared whether representations were present in 1-d reconstructions using the representational fidelity measure, we performed one-tailed tests (null hypothesis that $F \leq 0$).

Because we performed 1,000 iterations of these analyses, we cannot identify $p$ values less than 0.001, so all comparisons in which resampled difference distributions were all greater than or less than 0 are reported as $p < 0.001$. Because we performed many pairwise comparisons, we corrected all repeated tests within an analysis using the false discovery rate (Benjamini and Yekutieli, 2001) and a threshold of $q = 0.05$ (except for tests of behavioral performance, which we corrected using Bonferroni’s method due to the small number of comparisons performed). All $p$-values for all tests are reported in Supplementary Tables. All error bars/intervals reflect 95% confidence intervals as estimated using this resampling procedure. Because PFC ROIs were examined in an exploratory manner, we corrected all tests for multiple comparisons independently for the a priori ROIs (V1–sPCS), and PFC ROIs (iPCS, SMA+, DLPFC+). The All ROIs Combined ROI (consisting of concatenated voxels across the a priori ROIs) was not independent of the constituent ROIs, which required us to independently correct for multiple comparisons within that ROI alone.

Code and data availability
In an effort to improve reproducibility, all data and code required to perform the analyses described here and to generate figures appearing in the text and supplement, as well as task scripts, are freely available in the Open Science Framework (http://osf.io/s5r6g). Additionally, tutorial and stimulus presentation scripts for implementing inverted encoding model (IEM)-based image reconstruction analyses are freely available at bit.ly/IEM_tutorial. Any questions regarding code or data can be addressed to author TCS.

Supplemental References


