Reconstructions of Information in Visual Spatial Working Memory Degrade with Memory Load

Thomas C. Sprague,1, * Edward F. Ester,2 and John T. Serences1,2,*
1Neurosciences Graduate Program, University of California, San Diego, La Jolla, CA 92093-0109, USA
2Department of Psychology, University of California, San Diego, La Jolla, CA 92093-0109, USA

Summary

Working memory (WM) enables the maintenance and manipulation of information relevant to behavioral goals. Variability in WM ability is strongly correlated with IQ [1], and WM function is impaired in many neurological and psychiatric disorders [2, 3], suggesting that this system is a core component of higher cognition. WM storage is thought to be mediated by patterns of activity in neural populations selective for specific properties (e.g., color, orientation, location, and motion direction) of memoranda [4–13]. Accordingly, many models propose that differences in the amplitude of these population responses should be related to differences in memory performance [14, 15]. Here, we used functional magnetic resonance imaging and an image reconstruction technique based on a spatial encoding model [16] to visualize and quantify population-level memory representations supported by multivoxel patterns of activation within regions of occipital, parietal and frontal cortex while participants precisely remembered the location(s) of zero, one, or two small stimuli. We successfully reconstructed images containing representations of the remembered—but not forgotten—locations within regions of occipital, parietal, and frontal cortex using delay-period activation patterns. Critically, the amplitude of representations of remembered locations and behavioral performance both decreased with increasing memory load. These results suggest that differences in visual WM performance between memory load conditions are mediated by changes in the fidelity of large-scale population response profiles distributed across multiple areas of human cortex.

Results

To assess the functional role that population codes in different visually responsive occipital, parietal, and frontal regions of interest (ROIs) play in spatial working memory (WM), we presented participants (n = 4, four scanning sessions each) with two target stimuli (Figure 1A) followed by a postcue instructing them to remember the location(s) of zero (R0), one (R1), or two (R2) stimuli. In behavioral testing sessions performed outside of the scanner, participants used a mouse click to indicate the exact position of the remembered target. During scanning, participants performed a two-alternative forced-choice (2AFC) discrimination task in which they compared the position of a probe stimulus to that of the corresponding remembered target stimulus (Figure 1A). We chose to test precise memory for spatial positions using either a recall task (outside the scanner) or a “same/different” task (during scanning) so that participants were required to encode exact spatial positions rather than use a verbal code or only encode a single dimension (e.g., “8 o’clock,” “far to the left”).

Behavioral performance on the analog recall task performed outside the scanner revealed lower mnemonic precision when two target locations were remembered compared to when a single target location was remembered (Figure 1C; p < 0.001, resampling test). During scanning, response accuracy did not significantly differ across set size conditions, although three out of four participants performed slightly worse with increasing set size (Figure 1D, p = 0.174, resampling test; see the Experimental Procedures). However, response times (RTs) were significantly longer when two stimuli were remembered compared to when a single stimulus was remembered (Figure 1E; p < 0.001, resampling test). Increased RTs during scanning suggest that memory representations in the R2 condition were degraded and were thus less accessible during behavioral report, consistent with previous observations of increased RTs after manipulations that impair spatial WM (e.g., [17]). Together, the behavioral data recorded inside and outside of the scanner are consistent with a degraded representation of each remembered location in the R2 condition compared to the R1 condition.

To characterize neural responses associated with WM maintenance, we first compared averaged blood-oxygenation-level-dependent (BOLD) functional magnetic resonance imaging (fMRI) responses in a set of functionally defined occipital (V1–V4 and V3A), parietal (IPS0–IPS3), and frontal (pFC; thought to be the human homolog of macaque frontal eye fields [18, 19]) ROIs as a function of memory load. We replicated previous reports that BOLD responses in frontal and parietal ROIs were larger on R2 trials compared to R1 trials [6, 20, 21] (Figure S1 available online). Interestingly, in early visual areas (V2–V3A and hV4) we observed a larger mean BOLD amplitude on R0 trials compared to R1 or R2 trials (Figure S1B, p < 0.001, resampling test). We also observed similar results using a complementary exploratory analysis in which we searched for any voxels with increased activation for larger memory loads (Figure S1C).

Next, we used a multivariate image reconstruction technique based on a spatial encoding model [16] to reconstruct remembered locations in spatial WM based on the pattern of activation across all voxels within each ROI (Figure 2). In contrast to analyses that focus solely on mean signal intensity (Figure S1), neural firing rates, or multivariate classification accuracy, this analysis uses an independently estimated model of the spatial sensitivity profile across all voxels in each ROI to transform BOLD activation patterns into an image of the remembered stimulus position(s) carried by those patterns (Figure 2; Experimental Procedures). Importantly, this analysis provides additional information compared to some other methods such as univariate population receptive field (pRF) [22] estimation or multivariate linear classification [9]: by yielding a reconstructed image of the remembered stimulus location(s), covert information held in WM can be directly visualized, quantified, and
related to behavior [16]. These reconstructions can be thought of as an image of the spatial WM contents in visual field coordinates (rather than coordinates relative to the cortical surface), and we interpret the focal bright spots found at target positions as target representations.

Spatial WM reconstructions computed based on patterns of delay-period activation from occipital (V1–V3A, hV4, IPS0, and IPS1) but became less separable in anterior parietal and frontal ROIs (IPS2–IPS3 and sPCS) The relative decrease in separability of R2 target reconstructions in these anterior parietal and frontal ROIs may reflect the rather small screen size that we used relative to the large size of spatial RFs typical of these ROIs [23, 24]. Finally, we examined the temporal structure of WM reconstructions from all ROIs across the course of the entire trial. We could readily reconstruct images of both remembered locations during target presentation when the positions were encoded into WM, but we could only reconstruct images of locations held in WM during the delay interval (Movie S1).

Next, we sought to quantify how spatial WM reconstructions differ across ROIs and under different memory loads. To do so, we rotated and shifted the reconstruction on each trial to a common reference location such that the target positions were in alignment and averaged coregistered reconstructions together (Movie S2 for coregistered reconstructions through time). Then, because the target position across all trials was now aligned, we quantified attributes of the averaged target representation by fitting a 2D surface (Figures S4A and S4B) characterized by several independent parameters (see Figures S4A–S4D for a demonstration that these parameters reflect dissociable properties of target representations). The size parameter reflects the spread (full-width half-maximum, FWHM) of the delay-period target representation: an increased fit size would reflect a less spatially precise representation of the remembered target location (note that here and elsewhere, we use “spatial” with reference to visual field space, not cortical space). The amplitude parameter reflects the height of the target representation over baseline: increased fit amplitude would correspond to a more prominent representation of the target over baseline activation not related to the target location. The baseline parameter reflects the non-spatially-selective response amplitude (i.e., a constant offset across the entire reconstructed visual field): a change in baseline reflects a change in mean signal amplitude across an entire ROI that does not carry spatial information and thus does not directly change the spatial information content of the reconstruction.

Increasing memory load did not change the size of the best-fit surfaces to the target representations within WM reconstructions that were based on activation patterns in occipital and posterior parietal ROIs (Figure 4D; V1–IPS0; all statistics were computed via nonparametric resampling methods and Bonferroni corrected for multiple comparisons; Table S1; see the Experimental Procedures). However, fit surface size did increase with memory load in anterior parietal (IPS2–IPS3) and frontal (sPCS) ROIs. Note that in these ROIs, we did not observe strongly disjoint target representations during R2 trials (Figure 3C), so these size increases may partially reflect an inability to separately quantify the representation of each location. It is likely that a larger display and more stimulus separation would enable a more accurate reconstruction and quantification of each remembered target representation in these anterior parietal and frontal areas (like in the early visual and posterior IPS ROIs). We evaluated the possible that observed size increases may be partially an artifact of coregistering reconstructions and averaging over target positions on R2 trials, even if the “true” target representations are constant in size, by simulating reconstructions under the null assumption that target representations were equal in size.
WM Reconstructions Degrade with Memory Load

Figure 2. Inverted Spatial Encoding Model for Reconstructing the Contents of Spatial WM

(A) Each participant was scanned for three to four independent spatial mapping runs for encoding model estimation per session (see the Supplemental Experimental Procedures). Participants performed a challenging spatial WM task in which they determined whether a probe stimulus (500 ms) was in the exact same position or a slightly different position from a remembered target position (500 ms; 2AFC; see [16]). During the brief delay period (3,000 ms), a flickering checkerboard stimulus was presented near the remembered target position. This stimulus was irrelevant to the task performed by the participant but was used to drive large sensory responses to estimate a voxel-level encoding model used for computing reconstructions in the main task (see C–E). We adjusted difficulty on a run-by-run basis to maintain vigilance and equate performance across participants and sessions (73.738% ± 1.819% accuracy, mean ± SEM).

(B) We presented the mapping stimulus at each of 36 positions arrayed across a 6 × 6 square grid (one trial per position per run).

(C) To estimate spatial sensitivity profiles for each voxel, we predicted the response of each of 36 hypothetical “information channels” (spatial filters) to each stimulus used in the training runs [16]. Then, we took the measured response of each voxel and the predicted hypothetical channel responses to each stimulus position and used ordinary least-squares linear regression to estimate the contribution of each information channel to the signal observed in each voxel. This step is performed on each voxel independently (see the Supplemental Experimental Procedures, Equation 3).

(D) For each collection of voxels for which we computed reconstructions (ROIs, Figures 3 and 4), all voxels from all ROIs, Figure 4) we computed a mapping from voxel space into channel space (Supplemental Experimental Procedures, Equation 4). In contrast to “population receptive field” analyses [22], this step is multivariate and must be performed using all voxels that contribute to the image reconstruction. Using the computed linear mapping, the measured activation pattern across all voxels is transformed into “information space”—the amount each channel must have been active in order to produce the measured voxel activation pattern. A “raw” reconstruction can be computed for any single observation (e.g., one fMRI volume from area V1) by computing a sum of the spatial filters that define the information channels weighted by the estimated channel responses (right panel).

(E) When computing average reconstructions across all trials (Figures 4C and S2B), we coregistered different target positions on each trial to a common location by first rotating the spatial filters around the fixation point such that the target lies along the Cartesian x axis, then shifting the filter centers horizontally such that the target is positioned 3.25° from fixation along the x axis (white dot in reconstructions shown in Figures 4C and S2B). For R0 and R2 trials, this is done for each remembered target, and the coregistered reconstructions aligned to each target are averaged. Importantly, this coregistration procedure enables us to average the representations of spatial WM targets that appeared at different positions in the display on different trials.
across memory load conditions and performing an identical coregistration and quantification procedure as that used in Figure 4. These simulations determined that fit target representation size is artificially inflated by 8.62% on average due to the coregistration and averaging procedure. Importantly, our empirically observed size expansion in these regions (IPS2, 24.8%; IPS3, 32.7%; sPCS, 19.6%) was substantially larger than that induced by the analysis procedure itself (see Figure S4) and the Supplemental Experimental Procedures), suggesting that there are still important changes in target representation size across memory load conditions.

The amplitude of best-fit surfaces decreased with increasing memory load in striate and extrastriate occipital (V1–hV4) and posterior parietal (IPS0–IPS1) ROIs, consistent with predictions from a model in which increasing memory load results in lower gain of population-level representations of remembered stimuli [14, 15]. In contrast, fit amplitude trended toward increasing, with greater memory load in anterior parietal (IPS2–IPS3) and frontal (sPCS) ROIs (trend defined as p < 0.05, uncorrected for multiple comparisons). This latter result is consistent with previous demonstrations that average delay-period activation levels increase in frontoparietal ROIs with memory load [6, 20, 21] (Figure S1). Furthermore, simulations confirm that the fit amplitude parameter captures changes in the amplitude of the target representation and is independent of changes in baseline or size (Figure S4).

Finally, the nonspatial baseline parameter significantly increased with memory load in posterior parietal ROIs (IPS0–IPS1). The fact that nonspatial baseline levels increased only in IPS0–IPS1 with greater memory load suggests that previously documented univariate BOLD response increases in the more anterior parietal and frontal ROIs (Figure S1A; IPS2–IPS3 and sPCS) most likely correspond to a spatially focal change in target representation amplitude as opposed to spatially uninformative baseline modulations.

We observed population codes for remembered spatial positions in all of the ROIs that we examined, and the representations of remembered locations within these reconstructed images changed in different ways with increasing memory load (Figures 3 and 4). However, the activation pattern across all these ROIs may provide additional information above and beyond the activation pattern within any individual ROI, and reconstructions computed using all these across-ROI modulations may be more closely associated with behavioral memory load effects than reconstructions computed from individual ROIs alone (on the assumption that mnemonic fidelity is a function of information represented across multiple brain regions). We tested this by computing reconstructions as before (Figure 2), but using all voxels from the ten ROIs in each participant (importantly, because this is a multivariate analysis, this is not equivalent to averaging reconstructions across all ROIs; see the Supplemental Experimental Procedures). Comparison of target representations within these WM reconstructions computed using the combined ROI (Figures 4C and 4D, “all voxels combined”) across memory load conditions revealed each of the significant results found in the ROIs when analyzed individually (Figures 4C and 4D): size broadened, amplitude decreased, and baseline increased when two items were remembered compared to when one item was remembered (all p < 0.001, resampling test). As an additional exploratory analysis, we evaluated how these target representations (Figures 4C and 4D) were related to behavioral performance by computing and quantifying target representations within WM reconstructions as described above using data from each participant, ROI, and memory load individually. These results are presented and discussed in Figures S2B and S2C.

**Discussion**

Here, we employed an image reconstruction approach implemented using a multivariate inverted encoding model [8, 16, 25–28] to reconstruct the contents of spatial WM based on activation patterns in occipital, parietal, and frontal regions of human cortex. Prior studies have used measures like
classification accuracy to correlate behavioral performance with the discriminability of neural activation patterns [6, 13]. Although these analyses have many advantages due to a relative lack of model assumptions, changes in decoding accuracy may result from many different types of neural response pattern modulation [25, 29]. In contrast, by assuming a set of spatial basis functions, our method allows us to assess whether each region encoded information about the location of a remembered stimulus (e.g., [5, 30]), as well as to visualize and quantify the characteristics of these covert representations of target locations and relate different aspects of these quantified representations to behavioral performance (e.g., [8, 16, 25, 27]). In addition, these findings reinforce the importance of measuring the effect of cognitive manipulations on population-level estimates of mnemonic representations rather than on particular properties of the underlying neural generators, as these population-level representations can be more closely associated with cognition and behavior than activity changes in single neurons or voxels [8, 16, 25–29, 31–33].

These image reconstruction and quantification analyses revealed lower amplitude and, in some anterior parietal and frontal ROIs, broader target representations with increasing memory load (Figure 4). From an information-theoretic perspective, response variability (i.e., intertrial variability in the reconstructed images) has two components: signal entropy, which is variability associated with experimental manipulations (remembered location), and noise entropy, which is variability not associated with experimental manipulations. The decrease in target representation amplitude under increased memory load should lead to less variability that is related to the remembered location(s) and thus to a decrease in the signal entropy and information about the remembered location. An increase in target representation size should also decrease signal entropy, as increased size leads to
more overlap between target representations for different locations, which would decrease the ability of the population code to discriminate between locations. In contrast, baseline shifts should not strongly influence information content as an additive shift in the entire reconstruction does not change signal entropy [14, 16, 34] (Figure S3). Thus, the observation of higher amplitude target representations corresponds to higher information content of population codes about a spatial position [14–16, 32–34] (Figure S3) and may be a consequence of changes in delay-period neural gain associated with neurons tuned to remembered locations [14]. In addition, modest increases in target representation size in anterior IPS and sPCs may reflect poorer mnemonic fidelity within particular ROIs, echoing previous results that the dispersion (analogous to size here) of reconstructed profiles of remembered features (e.g., orientation) correlates with behavioral performance [8, 25, 27]. However, future work using larger spatial stimulus arrays may help to more accurately disentangle and characterize multiple WM representations in anterior IPS and sPCs.

We were able to reconstruct the covert contents of spatial WM not only in occipital [4, 6–10, 13] and posterior parietal regions [10, 13], but also in anterior parietal and frontal cortex [5, 11]. These widespread modulations raise the possibility that distributed WM representations can be optimized to differentially contribute to complementary sensory (e.g., target localization) and motor (e.g., eye movements, reaches) behaviors. Consistent with this idea, a recent demonstration that induced alpha oscillations (which are often thought to reflect synchronized activity of large-scale cortical networks [35]) measured with scalp EEG can be used to reconstruct remembered orientations also suggests that long-range, interacting representations across much of human cortex support the maintenance of information in WM [27]. The successive representations of spatial position reported here may thus allow for a common coordinate system with which low-level stimulus features (such as spatial position and color) that are represented in occipital cortex are bound with spatial motor plans (such as eye movements and arm reaches [36]) that are more closely associated with representations in parietal and frontal cortex.

**Experimental Procedures**

**Functional Magnetic Resonance Imaging**

We scanned each participant for four sessions, each lasting 2 hr. Each session included runs of the spatial WM task (Figure 1), an independent spatial “mapping” task (Figures 2A and 2B; Supplemental Experimental Procedures) [16], and a visual locator task (5 min each).

**Encoding Model: Reconstructing Contents of Spatial WM**

We modeled the response of each voxel as a linear combination of 36 spatially selective information channels (see [16] Figure 2; Supplemental Experimental Procedures). Using a separate set of training data during which we presented a flickering checkerboard “mapping” stimulus at different locations on the screen (Figures 2A and 2B), we estimated the relative contribution of all 36 information channels to the observed signal in each voxel using ordinary least-squares regression (Figure 2C). Then, using all of these measured “channel weights” across a given ROI combined with the multivariate pattern of activation measured from that ROI during performance of the main spatial WM task (Figure 1A), we computed the channel responses that were most likely to produce the measured pattern of activation (Figure 2D). We combined these computed channel responses and the spatial filters (information channels) to produce reconstructed images of the spatial WM contents within each ROI for each measured pattern of activation (Figures 3 and 4; activation patterns measured 6.75–9 s after target onset; Movies S1 and S2, activation patterns measured at each time point during the trial).

**Quantifying Target Representations in WM Reconstructions**

We fit a surface to each reconstruction that was allowed to vary in its size, amplitude, and baseline (Figures 4A and 4B). Its center was constrained to be the position, in visual field coordinates, with the highest local reconstruction amplitude (local average within a 0.5° radius).

**Statistics**

For group-level analyses (Figures 1C–1E, 4D, and S1B), we combined data from all participants within a given ROI and memory load condition and resampled all trials with replacement and computed a mean measurement value for that resampling iteration (Figure 1C, behavioral recall error; Figure 1D, behavioral accuracy; Figure 1E, response time; Figure 4D, target representation fit parameters; Figure S1B, mean BOLD signal). We repeated this procedure 1,000 times to produce a resampled distribution of each measured value for each memory load condition. We computed p values for each ROI and each parameter as the two-tailed probability of observing an effect in the opposite direction of the mean effect observed. Comparisons are Bonferroni corrected across ROIs for each parameter (Figure 4D, ten comparisons) or across all comparisons performed (Figure S1B, 30 pairwise comparisons). All error bars are 95% confidence intervals derived from these resampled distributions unless indicated otherwise (Figure S1A).

For exploratory individual-participant analyses (Figure S2C), we performed an identical procedure but resampled only across each participant’s data when computing confidence intervals.

**Supplemental Information**

Supplemental Information includes Supplemental Experimental Procedures, four figures, one table, and two movies and can be found with this article online at http://dx.doi.org/10.1016/j.cub.2014.07.066.

**Author Contributions**

T.C.S., E.F.E., and J.T.S. developed the experiment protocol and wrote the manuscript. T.C.S. and E.F.E. collected data. T.C.S. analyzed data. J.T.S. supervised the project.

**Acknowledgments**

We thank Miranda Scolari and Mary Smith for assistance in developing parietal cortex mapping protocols, Anna Byers for assistance with data collection, Sirawaj Itthipuripat, Vy Vo, and Alexander Heitman for discussion, and Sirawaj Itthipuripat, Vy Vo, and Stephanie Nell for comments on the manuscript. This work was supported by a NSF Graduate Research Fellowship to T.C.S., NIH T32-MH020002-12 to E.F.E., and NIH R01 MH-092345 to J.T.S. Data from each cortical region of interest and scripts required to produce results as presented in this manuscript are available upon request.

Received: April 30, 2014

Revised: June 13, 2014

Accepted: July 25, 2014

Published: September 4, 2014

**References**


**Supplemental Data**

**A**

![Graph A](image)

**B**

![Graph B](image)

**C**

![Image C](image)

**Figure S1** Mean BOLD signal depends on WM maintenance, related to Figures 1 and 3.

For each participant and each ROI we extracted the target-locked hemodynamic response function (HRF; event-related average) for each memory load condition. Then, we averaged HRFs across participants. Error bars ±1 SEM (across participants, n = 4, sessions 1 & 2). Time courses qualitatively replicate previous findings [S1] in which early IPS visual field maps (e.g., IPS0) show more transient activation, while later IPS visual field maps (e.g., IPS3) show more sustained activation during delay periods. (B) Average delay-period activation is significantly reduced during WM maintenance in occipital ROIs V2-V3A and hV4 (R1 < R0 and R2 < R0, p < 0.001, resampling test, see Supplemental Experimental Procedures). In contrast, delay-period activation is higher during WM maintenance in parietal and frontal ROIs (IPS1-sPCS, R1 > R0 and R2 > R0, p < 0.024). In IPS2 and IPS3, we observed a trend (defined at uncorrected α = 0.05) towards greater delay-period activation during R2 than during R1 trials (p = 0.012 and 0.03, respectively). Despite these decreases in mean BOLD response with WM performance, we can still reconstruct robust spatial representations of 1 or 2 remembered stimuli (Fig. 3B-C). All significant p-values pass a Bonferroni-corrected threshold for 30 comparisons of α = 0.0017 (10 ROIs, 3 comparisons for each ROI: R2 vs. R0, R1 vs. R0, R2 vs. R1) and are indicated by asterisks in B. Trends, defined at uncorrected α = 0.05, are indicated by †. P-values presented in Table S1. Error bars 95% CI over resampled distribution pooled across participants. (C) For each participant we analyzed neuroimaging data using a
univariate general linear model (GLM) with 3 predictors corresponding to the 3 memory load conditions (Remember 0, Remember 1, and Remember 2). Here, we show significant voxels for the contrast Remember 2 > Remember 1, corrected within each participant for multiple comparisons using the false discovery rate as implemented in BrainVoyager 2.6.1 ($q < 0.05$). Within 2 of 4 individual observers, we observe data similar to previous reports that increasing the number of remembered items results in increased BOLD responses in parietal and frontal cortex [S2–S4]. Note that even though there are few significant voxels in the remaining two participants, patterns of activation for both Remember 1 and Remember 2 conditions can be used to reconstruct spatial WM contents (albeit with lower amplitude, see Fig. S2). Symbols match those used in Figure 1C-E and Figure S2.
Figure S2 Individual-participant WM reconstructions and fit surface parameters compared to behavioral recall error, related to Figures 3-4 (continued on next page)
Figure S2 (continued) Individual-participant WM reconstructions and fit surface parameters compared to behavioral recall error, related to Figures 3-4

(A) Data as in Figure 3 displayed for each participant individually. Symbols match those used in Figure 1C-E and Figure S1. Color scale is the same for all participants, memory load conditions, and ROIs within this panel. (B) Coregistered spatial WM reconstructions for each participant for each memory load condition (R1: Remember 1, R2: Remember 2) for each ROI reported in Figures 3 and 4 (coregistered as in Figure 4C). Black dot indicates the target position; black + is mean centroid of best-fitting surface; black dashed circle is drawn around the mean centroid at the mean full-width half-maximum (FWHM) of best-fitting surface. Color scale is the same for all participants, memory load conditions, and ROIs within this panel. (C) Best-fit size, amplitude, and non-spatial baseline surface parameters to target representations for each participant and each memory load condition. Error bars 95% CI computed via resampling all trials per condition within each participant and ROI. Though we do not have adequate statistical power to identify whether an effect is present, for any given ROI, these scatterplots suggest that target representation amplitude, more so than size or baseline, is best correlated with behavioral recall performance (especially IPS0 and All voxels combined) such that high representation amplitude is associated with better behavioral recall performance, both within and across participants. These analyses imply that the amplitude of representations across the visual hierarchy provides the primary constraint on behavioral performance in our spatial WM task, suggesting that the amplitude of population-level representations of remembered locations are more closely related to their fidelity as indexed by corresponding measures of behavioral performance than are other parameters, like their size [S5].
Population-level codes for a remembered spatial position can change in several ways, each of which may have different consequences on the information content of the population code for the remembered position. For a given remembered target position (A), a brain region carries a representation of the remembered position as a bivariate Gaussian-like representation (B). This representation could be modulated in several ways. If the size of the target representation decreased (C, F), this would reflect a more precise representation of spatial position, and could lead to more accurate localization of a remembered target. However, depending on the noise amplitude relative to the amplitude of the target representation, this type of modulation could be more susceptible to noise across trials, resulting in poor localization. Instead, an increase in the amplitude of the target representation (D, G) would increase the signal-to-noise ratio (SNR) of the population code even if the spatial precision (size) of the representation remained fixed. Because such an amplitude modulation would increase the SNR, this type of representational modulation would be more robust to high levels of cross-trial neural noise in the population code [S6, S7]. Alternatively, the baseline response level in the WM reconstruction could increase (E, H), which would not change the information content of the target representation given most reasonable noise models. For example, the absolute effect of a baseline shift on the information content of the population code depends on the across-trial noise distribution. If noise scales with the mean (e.g. Poisson noise) then a pure baseline shift would increase the noise level without a corresponding change in the amplitude of the target representation over baseline, which would decrease the information content of the neural code (i.e. noise would go up, but the dynamic range of stimulus-locked signals would remain the same). Under conditions of independent and identically distributed (IID) noise across positions and trials, a baseline shift would have no effect on the information content of the population code, as the dynamic range of the target representation amplitude over reconstruction baseline compared to the noise level would not change. However, note that under an unlikely scenario where noise decreases with the signal amplitude, a baseline shift would be beneficial, as increasing signal amplitude would decrease the corrupting influence of noise on the population-level target representation. All figures are modeled using the fit surface function (Equation 5) and adjustments either to the size (C, F), amplitude (D, G), or baseline (E, H) parameters. (F-H) slices through the center of
the representation. Note that combinations of these modulations are possible and these are meant only as illustrative examples.
Figure S4 Fits to simulated surfaces demonstrate sensitivity and specificity of fitting approach, related to Figure 4.

It may be the case that our fitting procedure is unable to assess changes in a given fit surface parameter independent of changes in another parameter. For example, changes in amplitude of the “true” target representation may be incorrectly attributed to changes in baseline. To address this concern, we generated simulated target representations using known parameters (vertical dashed lines in A-C, size: 5°; amplitude: 0.2, baseline: 0, approximately the values from R1 condition in the ‘All voxels combined’ analysis, Fig. 4D), then varied one parameter at a time while keeping all others constant. After generating simulated representations, we performed an identical surface fitting procedure to that performed using experimental data (see Supplemental Experimental Procedures: Simulating and fitting representations with known parameters; Fig. 4). We plot each fit parameter as a function of the target representation parameter that was varied. (A) Size was varied; (B) amplitude was varied; (C) baseline was varied. For each of these manipulations, only the fit parameter matching that which was manipulated changes as a function of parameter value; all others remain constant at the value indicated by the vertical dashed lines in their respective panels. An additional concern is that our narrow field of
view (FOV; 9.30°) over which we reconstruct WM contents and fit a surface to target representations may result in erroneous estimates of size, amplitude or baseline. To address this issue, we kept the simulated representation constant, but expanded the FOV over which we simulated target representations and fit surfaces (D). Best-fit size, amplitude, and baseline are largely constant across a wide range of FOVs. Importantly, we choose to maintain our original FOV, as that is the window over which we “trained” the spatial encoding model during the spatial mapping runs (Fig. 2A-C; Supplemental Experimental Procedures: Spatial encoding model). Allowing reconstructions to evaluate to 0 at high-eccentricity positions (as would be the case with an artificially-enlarged FOV) may not capture properties of the actual target representations, and so we intentionally avoid this. Additionally, it might be that the coregistration and averaging procedure (Fig. 2E) may artificially inflate estimates of fit size on R2 trials relative to R1 trials, even if there were no “true” change in target representation size. If this were the case, our observations of size increases in anterior parietal and frontal ROIs (IPS2-3, sPCS) might be partially due to the analysis procedure. Importantly, this may disproportionately mask results for ROIs in which target representations are not readily separable during R2 trials (Fig. 3C). To evaluate this possibility, we simulated reconstructions from both R1 and R2 trials under the null assumption that representations of the remembered target(s) do not change as a function of memory load. The representation parameters used to generate R2 reconstructions from a given ROI were taken from the best-fit parameter values reported in Figure 4D. Results from this analysis are presented in E. Using actual target positions remembered by participants and an identical analysis procedure, fit size indeed increased by an average of 8.62% across all ROIs for R2 compared to R1, even when the “true” simulated representations were constant in size. However, this effect is small when compared to the actual (significant) size increases observed in IPS2 (simulated: 10.62%, measured: 24.93% [18.85% 31.59%], mean [95% CI]), IPS3 (simulated: 11.73%, measured: 32.7% [23.28% 44.58%]), and sPCS (simulated: 8.25%, measured: 19.6% [13.50% 26.53%]). Because the size increases from simulations seeded with R1 representation parameters in these ROIs lie outside the 95% CIs of the observed size increases, the observed size changes are above and beyond those that would be expected given the analysis procedure alone. Finally, we also simulated the effect of increasing spatial overlap (via simulating reconstructions of 2 targets with increasing sizes) in order to evaluate whether increased overlap alone can lead to overestimates of fit size. We observed a decrease in the size overestimation once we exceeded simulated representation sizes of 5.8°. Though the size of R1 target representations in IPS2 (5.591°), IPS3 (5.889°), and sPCS (5.696°) hover near this maximum bias point, the size increase we observe in those regions for R2 remains substantially larger than that introduced by the analysis procedure alone. In all simulations performed using realistic parameter values, we never observed a decrease in representation amplitude as a result of the coregistration and fitting procedure.
Table S1 P values for resampled tests reported in Figure 4D and Figure S1B.

P-values derived from resampling tests described in Supplemental Experimental Procedures: Statistical methods. A value of 0 indicates \( p < 0.001 \) (due to 1,000 sampling iterations). Bold indicates significant test after correcting for multiple comparisons within parameter using Bonferroni’s method (Figure 4D, 10 comparisons, \( \alpha = 0.005 \); Figure 4D, All voxels combined, 1 comparison, \( \alpha = 0.05 \); Figure S1B, 30 comparisons, \( \alpha = 0.0017 \)). Italics indicate trends at \( p < 0.05 \), uncorrected for multiple comparisons.

<table>
<thead>
<tr>
<th></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 voxels combined</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 4D</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Size</td>
<td>0.824</td>
<td>0.682</td>
<td>0.362</td>
<td>0.73</td>
<td>0.244</td>
<td>0.022</td>
<td><strong>0.004</strong></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Amplitude</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td><strong>0.002</strong></td>
<td>0.02</td>
<td>0.02</td>
<td>0.03</td>
<td>0</td>
</tr>
<tr>
<td>Baseline</td>
<td>0.064</td>
<td>0.196</td>
<td>0.02</td>
<td>0.234</td>
<td>0.08</td>
<td>0</td>
<td>0</td>
<td>0.582</td>
<td><strong>0.008</strong></td>
<td>0.402</td>
<td>0</td>
</tr>
<tr>
<td>Figure S1B</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R0 vs R1</td>
<td><strong>0.004</strong></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.148</td>
<td><strong>0.024</strong></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>R0 vs R2</td>
<td><strong>0.002</strong></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.344</td>
<td><strong>0.004</strong></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>R1 vs R2</td>
<td>0.720</td>
<td>0.428</td>
<td>0.682</td>
<td>0.628</td>
<td>0.938</td>
<td>0.650</td>
<td>0.298</td>
<td><strong>0.012</strong></td>
<td>0.030</td>
<td>0.584</td>
<td></td>
</tr>
</tbody>
</table>
Supplemental Experimental Procedures

Participants

We used 4 healthy participants from the UCSD community (aged 24-31 yrs, 2 female, all right-handed). All participants were experienced psychophysical observers, and data from these participants has been reported previously using the same participant identifiers [S5]. One participant (AA) was an author (TCS). All participants gave informed consent as approved by the UCSD Institutional Review Board and were compensated for their time ($20/hr fMRI sessions, $10/hr behavioral sessions).

Stimulus & Task

All participants took part in 4 fMRI testing sessions and 2-4 behavioral testing sessions. The size of the stimulus display was fixed across sessions. During each trial, two target stimuli were presented at pseudorandomly chosen positions on the screen for 500 ms, followed by a post-cue (fixation point changing color) and 8,000 ms delay interval. The color of the post-cue during this delay interval instructed participants to precisely remember the position of one target (matching the color of the fixation point; Remember 1), remember the position of both targets (purple fixation point; Remember 2), or to passively fixate and wait for the next trial to begin (only in some fMRI sessions; Fig. 1A; Remember 0).

Each target was presented within one of 8 discs with 0.6° radius evenly spaced around a 3.25° circle, offset from the horizontal and vertical meridians. Both targets were presented in different discs. On every trial, the exact target position within the area of the disc was randomly chosen with uniform density to discourage alternative coding strategies (e.g., verbally labeling the location, such as “up and to the left”; “8 o’clock”, etc.).

During fMRI scanning sessions, a probe stimulus appeared at the end of the delay period (500 ms). The probe stimulus was always presented near to and in the color of the remembered target during Remember 1 trials, and with equal probability in either color/location during Remember 2 trials (matching the color of the corresponding target). The probe stimulus was in exactly the same position as the target stimulus on 50% of trials, and an offset position in a random direction on the other 50% of trials. The magnitude of the offset was uniformly chosen from the range 0.1° - 1.5° at 0.2° steps. Participants performed a two-alternative forced-choice task at the end of each Remember 1 or Remember 2 trial, comparing whether the probe stimulus was in exactly the same or in a slightly different position relative to the corresponding target position.

In the behavioral recall task performed outside the scanner, we instructed participants to use the mouse to click as accurately as possible at the remembered target position on the screen. For Remember 2 trials, the response was cued by a change in fixation color from purple to red or blue.

During the first 2 fMRI sessions acquired for each participant we included Remember 0 trials, but these were dropped from the final 2 sessions in an effort to maximize data acquired during memory maintenance. Remember 0 trials were omitted from the behavioral recall task.

In order to independently estimate a spatial encoding model for each voxel, participants also performed a “spatial mapping” task (Fig. 2A-B), which is similar to the Spatial WM condition in our previous work [S5]. Briefly, they remembered the exact position of a target dot (500 ms) over a brief delay interval (3,000 ms) and made a same/different 2AFC judgment on a probe dot presented after the delay (500 ms). The principal difference between this task and the main spatial WM task described above (Fig. 1A) is the presentation of a high-contrast flickering checkerboard (1.163° radius) around the remembered target position (Fig. 2A). Also, these memory targets and interstitial delay-period checkerboards were presented on a 6 × 6 grid (spaced by 1.163°, Fig. 2B). This stimulus and spatial arrangement allowed us to drive a strong BOLD response and efficiently estimate the spatial sensitivity of each voxel across the entire visual field. Task
difficulty was adjusted between runs to ensure sufficient task engagement (mean ± standard error of accuracy: 73.738 ± 1.819 %).

Finally, observers also performed a spatial WM task while a large full-hemifield flickering checkerboard was presented. Data from this task was used to independently identify which voxels were significantly modulated by visual input from our stimulus display for inclusion in further analyses (see ROI definition).

**Behavioral analysis**

For each participant, we combined data across all behavioral recall sessions. We discarded trials in which errant clicks were made (responses > 7.0° eccentricity or within 1.0° of fixation; these typically were a result of accidentally pressing the mouse before moving the cursor or errant mouse movements; 11 of 3,264 trials discarded across all participants). Behavioral recall error was defined as the mean Euclidean distance between the response position and the correct target position across all trials for each condition within a participant.

**fMRI scanning**

We scanned all 4 participants for 4 sessions, with each session lasting 2 hrs. Each session included 3 types of runs: spatial mapping runs (Fig. 2A-B, 5 min each, 3-4 per session), used for encoding model estimation (Fig 2C), spatial WM main task runs (Fig. 1A; 2-3 per session, each subdivided over 4 shorter runs lasting 4-6 min each depending on whether Remember 0 trials were included), and visual localization runs for identifying visually-responsive voxels involved in WM maintenance (6 min each, 1-3 per session). These participants were additionally scanned for another 1.5-2 hr session in order to map retinotopically organized IPS subregions IPS0-IPS3 using methods described previously (Methods and IPS retinotopic maps for 3 of 4 participants can be found in [S5]).

We scanned all participants on a 3 T research-dedicated GE MR750 scanner at the UCSD Keck Center for Functional Magnetic Resonance Imaging using a 32 channel send/receive head coil (Nova Medical, Wilmington, MA). We acquired functional data using a gradient echo planar imaging (EPI) pulse sequence [S5] (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 2 × 2 × 3 mm, xyz).

During each session we also acquired a high resolution anatomical scan (FSPGR T1-weighted sequence, TR/TE = 11/3.3 ms, TI = 1,100 ms, 172 slices, flip angle = 18°, 1 mm³ resolution). Functional images were coregistered to a separate anatomical scan collected during a different session by aligning each session’s functional images to the respective session’s anatomical scan, and then aligning the anatomical scan to the target anatomical scan. Images were preprocessed as described previously [S5] using FSL (Oxford, UK) and BrainVoyager 2.3 (BrainInnovations). Preprocessing included unwarping the EPI images using routines provided by FSL, 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 and normalization of signal amplitudes by converting to Z-scores separately for each run. 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 (version 2012a, The Mathworks, Inc).

**ROI definition**

All reported ROIs except sPCS were defined using standard retinotopic mapping procedures [S5, S8, S9]. For analysis, voxels from ROIs in the left and right hemispheres which showed a significant response to either hemifield during the visual localizer runs (across sessions, corrected using the false discovery rate (FDR, [S10]) across all measured voxels within each participant, q < 0.05) were concatenated to produce bilateral ROIs. We concatenated voxels from the dorsal
and ventral aspects of V2 and V3. The sPCS ROI was defined using voxels with significant activation localized in the posterior part of the superior precentral sulcus during the visual localizer runs, FDR-corrected for multiple comparisons. The “All voxels combined” ROI presented in Figure 4C was defined by concatenating all voxels from all 10 ROIs.

**fMRI analysis**

For each trial of the spatial mapping runs, we extracted the Z-scored BOLD signal from each voxel averaged over the two TRs occurring 6.75 – 9.00 s after target onset as the observed signal in that voxel for that trial. For each trial of the main spatial WM task runs, we extracted signal from each TR following the target onset and computed spatial representations for each TR independently.

**Spatial encoding model**

We implemented an inverted spatial encoding model [S5] to compute spatial WM reconstructions using the pattern of activation over subsets of voxels during each time point (TR) of the WM delay period. This method assumes (1) the BOLD signal reflects an approximately linear combination of neural responses within each voxel, (2) at least some voxels within each ROI have non-uniform responses to stimuli presented at different positions on the screen during the “training” phase of the analysis (Fig. 2A-C), and (3) a voxel’s estimated encoding model during the training phase is maintained during the main task runs.

For this method we first estimated the spatial sensitivity profile of each voxel (Fig. 2A-C). Then, with novel data, we used the pattern of activation across all voxels of interest and the independently estimated spatial sensitivity profile estimated for each of those voxels to compute the reconstruction carried by the activity across all voxels in a ROI (Fig. 2D). In this implementation of the inverted encoding model we used entirely different datasets for estimating the encoding model and for reconstructing spatial WM contents (i.e., there is no need for a “leave-one-out” procedure). We trained and tested the encoding model using data from each session independently, then combined resulting reconstructions across sessions. Thus, we used 3-4 runs of “training” data from spatial mapping runs at a time for encoding model estimation, and 8-12 runs of “testing” data for computing reconstructions.

We modeled the response of each voxel as a linear combination of a grid of 36 spatial filters (or information channels; Fig. 2C) where $B_1$ is the observed signal in each voxel on each trial ($m$ voxels x $n$ trials), $C_1$ is the predicted responses for each channel on every trial ($k$ channels x $n$ trials; see below), and $W$ is a matrix describing the mapping between “channel space” and “voxel space” ($m$ voxels x $k$ channels):

$$B_1 = WC_1 \text{ (Equation 1)}$$

During the training phase, we estimated the spatial sensitivity profile of each voxel by first filtering all training stimuli (Fig. 2A-B) by a “basis set” of 36 spatial filters, each of the form:

$$f(r) = 0.5 + 0.5 \cos(\pi r/s) \text{ for } r < s, \text{0 elsewhere} \text{ (Equation 2)}$$

Where $r$ is the distance from the center of the basis function, and $s$ is a size constant which corresponds to the distance from the basis function center at which the function reaches zero (same function used in [S5]). Spatial filter centers are spaced by 1.86° and have a full-width at half-maximum (FWHM) of 2.31° (the corresponding size constant, $s$, is 5.82°).

The resulting filtered predicted channel responses ($C_1$, $k$ channels x $n$ trials) enter into an ordinary least-squares regression, along with the observed signal in each voxel ($B_1$) to estimate the contribution of each spatial information channel to the observed response in each voxel across all training trials ($\hat{W}$, $m$ voxels x $k$ channels):

$$\hat{W} = B_1 C_1^T (C_1 C_1^T)^{-1} \text{ (Equation 3)}$$
This step amounts to a simple general linear model of the form commonly implemented in fMRI analyses, and is performed on each voxel individually (i.e., it is a univariate analysis).

Once the spatial sensitivity profile (“channel weights”, $\mathbf{W}$) was estimated for each voxel using all spatial mapping runs within a session, we computed the pseudoinverse of the channel weights to obtain a mapping from the observed pattern of activation across all voxels within an ROI (“voxel space”) into estimated channel responses ($\mathbf{C}_2$, $k \times n$, “information space”):

$$\mathbf{C}_2 = (\mathbf{W}^T \mathbf{W})^{-1} \mathbf{W}^T \mathbf{B}_2$$ (Equation 4)

This mapping is computed using all voxels assigned to a given ROI (e.g., all V1 voxels, Figs. 3-4, or all voxels across all ROIs, Fig. 4), and is different for any given combination of voxels (i.e., it is a multivariate operation).

The resulting estimated channel responses reflect the response of each information channel that is most likely to have given rise to the observed pattern of activation across all voxels within an ROI, given our linear model of observed BOLD responses as a function of information channel activation.

To compute spatial WM reconstructions from each vector of channel responses (one 36-dimensional channel response vector for each fMRI data volume for each ROI) we computed a sum of the basis functions, each weighted by the corresponding estimated channel response for each channel response vector (Fig. 2D). For Figures 2-4 and Figure S2 we averaged reconstructions from 6.75 – 9.00 s (2 volumes during the delay period). For all Supplemental Movies, we did not average reconstructions across time. Each frame of the movies corresponds to a single fMRI data volume (averaged across all trials and participants).

Critically, because we used data from an independent task (Fig. 2A-B) to estimate the encoding model which was then used to generate reconstructions during the memory task (Fig. 1A), we can be confident that stimulus-specific idiosyncrasies in the data (i.e., overfitting noise in a leave-one-out cross-validation design) were not responsible for our ability to reconstruct the contents of spatial WM.

Coregistering reconstructions

Because the representations of remembered targets within WM reconstructions are rather weak (especially compared to stimulus representations of flickering checkerboard stimuli, [S5]) we implemented several different coregistration procedures in order to visualize different aspects of the WM reconstructions. Because the pattern of BOLD activation was mapped into an information space, we can manipulate the functions describing information space in a manner that allows us to average target representations within the WM reconstructions from trials in which remembered visual stimuli appeared in different positions by effectively “rotating” the WM reconstructions in order to match target positions. That is, we can combine the target representations when the remembered stimulus was on the top left of the screen with the target representation when the remembered stimulus was at the top right of the screen. This removes any potential inhomogeneities in representations that are dependent upon particular screen positions and allows us to ascertain how target representations generally change across the visual system and across task demands independent of the exact position of a remembered stimulus. This is a unique ability afforded by this analysis method – it is otherwise very challenging to determine how one would average responses to stimuli at different parts of the screen in “voxel space”.

First, we sought to qualitatively evaluate WM reconstructions for different stimulus arrangements across the 3 memory load conditions. We collapsed all trials in which the 2 target stimuli were an equal average angular distance apart. To do this, we rotated all reconstructions clockwise such that the non-probed target (for Remember 1, this was the non-cued target; for Remember 2, this was the target which was not probed at the end of the trial) appeared along the positive $x$ axis,
and flipped reconstructions in which the probed stimulus was below the x axis across the horizontal midpoint such that there are four possible target arrangements (the four columns in Fig. 3, Fig. S2A, Movies S1A-C).

Second, we sought to coregister reconstructions so that the remembered target was always centered at exactly the same position (along the x axis, 3.25° from fixation). We accomplished this by rotating each trial’s reconstruction by the angular distance between the target and the horizontal axis, then horizontally shifting the reconstruction to remove any remaining jitter (Fig. 2E). For Remember 0 and Remember 2 trials, we did this for each target in turn and averaged the resulting reconstructions (Fig. 4; Fig. S2B; Movie S2). Note that this procedure was identical across all ROIs, so any changes in WM reconstructions resulting directly from the coregistration procedure would be similar across all ROIs. Because the effects of memory load on target representations within WM reconstructions we observed are different across different ROIs, those effects must be due to changes in fidelity of the target representation rather than artifacts of the analysis procedure (see Fig. S4).

**Surface fitting**

For each exactly-coregistered WM reconstruction (aligned to optimally reveal the target representation) we fit a surface (Fig. 4A-B) parameterized by its position, size ($s$, distance from center at which surface reaches zero), amplitude ($a$) and baseline ($b$, non-spatially-selective) shift:

$$f(r) = b + a(0.5 + 0.5 \cos(r\pi/s))^7$$ for $r < s, \ 0$ elsewhere \hspace{1cm} (Equation 5)

To ensure robust fits, we restricted the position of the best-fit function to be the point on the reconstruction with the largest local average. The local average was computed across a disc 0.5° in radius (via convolution of each reconstruction with a circular disc). All other parameters ($s, a, b$) were allowed to freely vary.

We computed fits using fminsearch in MATLAB (R2012a, The Mathworks, Inc), which implements the Nelder-Mead algorithm. For every fit, we began from 10 different randomly selected initial values and chose the fit that minimized the sum of squared errors between the surface function and the coregistered WM reconstruction.

**Statistical procedures**

To evaluate behavioral effects of the memory load manipulation (Fig. 1) we performed a resampling test in which we resampled all valid trials across all participants (with a response, for RT/accuracy in the scanner, Fig. 1C-D; within a reasonable spatial response window, behavioral recall task, Fig. 1E) with replacement for each memory load condition 1,000 times and computed the distribution of differences between R2 and R1 for inside-scanner accuracy, RT and outside-scanner recall error. We defined $p$-values as the percentage of resampling iterations in which an effect was observed in the opposite direction of the mean effect and multiplied by two, as we did not make any a priori predictions about the direction of the effect. Each of these comparisons were tested against a threshold of $\alpha = 0.05$.

When comparing best-fit parameters to target representations with WM reconstructions across memory load conditions (Fig. 4), we implemented an across-participant resampling procedure. We combined data across all 4 participants into one large pool of 1,248 trials per condition. Within each condition, we resampled across all trials, with replacement, 1,000 times and computed an averaged WM reconstruction. We then implemented the surface fitting procedure described above to quantify the target representation on each resampling iteration, resulting in a resampled distribution of 1,000 best fit values for each parameter (size, amplitude, and baseline). Error bars on figures are 95% confidence intervals derived from this resampled fit parameter distribution. $P$-values are computed as the percentage of resampled iterations in which a difference opposite to the mean difference was observed (e.g., if Remember 2 had a larger value on average than Remember 1, $p$ would correspond to the percentage of resampled iterations in which Remember 2 had a smaller value than Remember 1). All $p$-values were doubled (because we made no a priori hypothesis about the direction of the effect).
and compared against an alpha threshold corrected for multiple comparisons across 10 ROIs for each parameter using Bonferroni’s method (corrected $\alpha = 0.005$), and trends are defined by $p$-values below an uncorrected threshold $\alpha = 0.05$ and are indicated using gray asterisks in Figure 4D (see also Table S1 for $p$-values for all comparisons in Fig. 4D).

Because the results from the “All voxels combined” ROI include data from each of the other 10 ROIs, we performed statistics separately for the 10 constituent ROIs and for the combined ROI (that is, we corrected for 10 comparisons within each parameter when evaluating the statistical significance of each ROI separately, then performed no corrections when evaluating the statistical significance of the combined ROI). However, as can be seen in Table S1, correcting for an additional comparison would not change which tests are found to be significant. Since we computed 1,000 iterations in this resampling procedure, we only report $p$-values as $< 0.001$ if we do not observe an effect opposite the mean in the resampled distribution. It is possible that with more iterations we could see a result, and so it is inappropriate to report $p = 0$.

When evaluating statistical significance of mean delay-period BOLD responses (Fig. S1B), we resampled the average BOLD signal during the delay period (6.75 – 9.00 s after target onset, as in Figs. 3-4, S2) on each trial for each set size condition pooled across all participants (1,000 resampling iterations, resampling all trials of a given set size condition with replacement). Because we only included Remember 0 trials during Sessions 1 and 2, we only included those sessions in this analysis. Then, we compiled 3 resampled distributions corresponding to the difference between each pair of set size conditions (R0 vs. R1, R0 vs. R2, R1 vs. R2) and computed $p$-values as the percentage of resampled iterations in which a difference opposite to the mean difference was observed, doubled (as described above). To evaluate significance for this exploratory analysis, we corrected for multiple comparisons using Bonferroni’s method across all 30 computed $p$-values (3 pairwise comparisons for each of 10 ROIs; $\alpha = 0.0017$, Table S1).

**Simulating and fitting target representations with known parameters**

To evaluate the accuracy and sensitivity of our fitting procedure (Fig. 4A-B), we simulated WM reconstructions containing a target representation (or multiple target representations) with known parameters, then implemented the fitting procedure in the same way as used on the actual data (see Surface Fitting, above). We simulated target representations as a single surface (Equation 5) with fixed $s$, $a$, and $b$ parameters, centered at $(3.25^\circ, 0^\circ)$ and computed over a square field of view (FOV) from $-4.65^\circ$ to $4.65^\circ$ along the $x$ and $y$ axes, sampled at a grid of $151 \times 151$ points (along each axis). Then, we varied a single parameter while keeping the others constant and plotted best-fit parameters as a function of the value of the single parameter that we varied (Fig. S4A-C). Additionally, we performed this same procedure while allowing the FOV to vary (Fig. S4D), but keeping all target representation parameters constant, in order to evaluate how allowing the representation to artificially return to “baseline” at distant edges of the reconstruction might influence the fitting routine.

Finally, to quantify how the coregistration procedure (Fig. 2E) might result in changes in best-fit parameter estimates without any “true” changes in parameter values of the underlying target representations, we generated a simulated dataset in which we centered representations at the actual target positions participants remembered during scanning, performed an identical coregistration procedure as to that in the main analysis, and fit the resulting coregistered and averaged representations. We did this using the actual best-fit parameters for the R1 condition in each ROI plotted in Figure 4D (Fig. S4E).
Supplemental References


