Not-so-working Memory: Drift in Functional Magnetic Resonance Imaging Pattern Representations during Maintenance Predicts Errors in a Visual Working Memory Task

Phui Cheng Lim[1], Emily J. Ward[2], Timothy J. Vickery[3], and Matthew R. Johnson[1]

Abstract

■ Working memory (WM) is critical to many aspects of cognition, but it frequently fails. Much WM research has focused on capacity limits, but even for single, simple features, the fidelity of individual representations is limited. Why is this? One possibility is that, because of neural noise and interference, neural representations do not remain stable across a WM delay, nor do they simply decay, but instead, they may "drift" over time to a new, less accurate state. We tested this hypothesis in a functional magnetic resonance imaging study of a match/nonmatch WM recognition task for a single item with a single critical feature: orientation. We developed a novel pattern-based index of "representational drift" to characterize ongoing changes in brain activity patterns throughout the WM maintenance period, and we were successfully able to predict performance on the match/nonmatch recognition task using this representational drift index. Specifically, in trials where the target and probe stimuli matched, participants incorrectly reported more nonmatches when their activity patterns drifted away from the target. In trials where the target and probe did not match, participants incorrectly reported more matches when their activity patterns drifted toward the probe. On the basis of these results, we contend that neural noise does not cause WM errors merely by degrading representations and increasing random guessing; instead, one means by which noise introduces errors is by pushing WM representations away from the target and toward other meaningful (yet incorrect) configurations. Thus, we demonstrate that behaviorally meaningful drift within representation space can be indexed by neuroimaging. ■

INTRODUCTION

Working memory (WM) is critical to many aspects of cognition and behavior, yet it is far from perfect. WM is limited in capacity, both in terms of number of items (e.g., Awh, Barton, & Vogel, 2007; Xu & Chun, 2006; Vogel, Woodman, & Luck, 2001; Luck & Vogel, 1997) and in terms of the informational complexity of those items (e.g., Xu & Chun, 2006; Alvarez & Cavanagh, 2004). However, even when the number of items and their complexity are within their nominal limits, WM failures are common. How many times a day do we walk into a room and forget what we came in for? How often do we have to search for an item that we could swear we just left on the kitchen table, only to find out it was actually sitting on the couch? Despite the ubiquity of such WM failures in everyday life, its underlying mechanisms are not well understood. If we are not exceeding the design specifications of the human WM system in terms of item number or complexity, why does WM not work perfectly all the time?

Although many studies of WM failures have focused on capacity limitations (Luck & Vogel, 1997, 2013; Bays, Catalao, & Husain, 2009; Zhang & Luck, 2008; Alvarez & Cavanagh, 2004; Vogel et al., 2001) or on interference from external factors such as distractor items (Derrfuss, Ekman, Hanke, Tittgemeyer, & Fiebach, 2017; Yoon, Curtis, & D'Esposito, 2006; Kim, Kim, & Chun, 2005; Vogel, McCollough, & Machizawa, 2005), a number of other neural and psychological mechanisms have been proposed to underlie WM performance, even in cases where loads are low and no explicit distractors are present. Neuroimaging evidence has linked task performance to several different aspects of brain activity during WM. For instance, both EEG and fMRI studies have found that successful WM performance is linked to greater power and/or synchrony in certain frequency bands (Solomon et al., 2017; Balsters, Robertson, & Calhoun, 2013; Khader, Jost, Ranganath, & Rösler, 2010). In addition, above-baseline levels of brain activity during the WM maintenance period correspond with greater recall on a subsequent long-term memory test (Blumenfeld & Ranganath, 2006; Ranganath, Cohen, & Brozinsky, 2005; Brewer, Zhao, Desmond, Glover, & Gabrieli, 1998; Wagner et al., 1998). Although a number of other fMRI

[1]University of Nebraska-Lincoln, [2]University of Wisconsin-Madison, [3]University of Delaware

Journal of Cognitive Neuroscience 31:10, pp. 1520–1534
doi:10.1162/jocn_a_01427

studies have investigated the relationship between brain activity during WM maintenance and performance on the WM task itself, they found no or limited suprathreshold activation differences during maintenance between accurate and inaccurate responses (Bergmann, Daselaar, Fernández, & Kessels, 2016; Bergmann et al., 2015; Hannula & Ranganath, 2008). One potential explanation for these results comes from behavioral, animal, and modeling research suggesting that WM failures may occur as a result of "drift" in neural population activity (Schneegans & Bays, 2018; Wimmer, Nykamp, Constantinidis, & Compte, 2014), wherein mental representations become less accurate over time because of the accumulated effects of neural noise (Schneegans & Bays, 2018) or because of representational distortions (Lupyan, 2008), which do not necessarily entail a change in overall activity levels. However, the hypothesis that WM failures are because of drift has not been tested directly in humans via neuroimaging.

Aggregate activity of entire brain regions may not be sufficiently sensitive to measure fluctuations in the quality or fidelity of the information encoded by that activity, which is a necessary prerequisite for directly testing the drift hypothesis. However, WM representations may be encoded in more fine-grained brain activity patterns corresponding to specific memoranda, rather than by a region's overall activation. For example, patterns of brain activity exhibited during WM maintenance or mental imagery of a remembered stimulus reflect the patterns recorded during initial perception of that stimulus (Johnson & Johnson, 2014; Albers, Kok, Toni, Dijkerman, & de Lange, 2013; LaRocque, Lewis-Peacock, Drysdale, Oberauer, & Postle, 2013; Lee, Kravitz, & Baker, 2012; Harrison & Tong, 2009), suggesting that these mental activities transpire via the reinstantiation of perceptual activity patterns. However, activity patterns are not merely passive re-creations of perceptual stimuli. Rather, they reflect an active process wherein observers prioritize storage of relevant information, such that task-irrelevant features are encoded less strongly, if at all, compared with task-relevant features during WM maintenance (Jackson, Rich, Williams, & Woolgar, 2017; Serences, Ester, Vogel, & Awh, 2009). Given these findings, similar pattern-based analyses may provide a valuable tool for testing the hypothesis that WM failures occur because of representational drift.

Supporting the putative utility of this analytic approach, pattern-based fMRI analyses have been used to infer the fidelity of WM representations and predict how well items are later remembered. Much of this work has focused on the link between WM and long-term memory, wherein patterns that are more similar predict superior performance (Ward, Chun, & Kuhl, 2013; Kuhl, Rissman, & Wagner, 2012; Kuhl, Rissman, Chun, & Wagner, 2011; Xue et al., 2010), but there is some evidence that the fidelity of memory representations during WM maintenance affects WM performance as well (Sprague, Ester, & Serences, 2014; Ester, Anderson, Serences, & Awh, 2013). For instance, Ester and colleagues analyzed orientation-selective responses in visual cortex and created individual tuning profiles for each participant that were predictive of task performance, suggesting that the relative "quality" of each participant's WM representations was indicative of their memory acuity. However, that study only examined individual differences and did not determine whether the quality of representations also predicts WM performance within individuals.

Thus, abundant evidence supports a link between brain activity during WM maintenance and subsequent behavior, although the exact nature of the relationship between maintenance activity patterns and WM performance has not been thoroughly explored. In other words, given that brain activity patterns during WM maintenance of a particular stimulus appear to mirror those observed during visual perception of that stimulus (Johnson & Johnson, 2014; Albers et al., 2013; Harrison & Tong, 2009) and that pattern similarity during encoding and maintenance is linked to subsequent long-term memory performance (Ward et al., 2013; Kuhl et al., 2011; Xue et al., 2010), it seems plausible that fluctuations in neural activity patterns during maintenance could lead to incorrect performance during a subsequent WM probe, even with successful encoding. An example of how this drift hypothesis might unfold is illustrated in Figure 1B, which shows hypothetical brain activity patterns for two potential scenarios.

This study therefore investigates how brain activity patterns associated with specific stimuli, particularly during the maintenance period, might predict the accuracy of performance on a later WM probe. We scanned individuals using fMRI while they performed a variation of a common match/nonmatch recognition task with oriented Gabor patch stimuli (cf. Harrison & Tong, 2009) that was specifically designed to allow us to track changes in the fidelity of neural pattern representations throughout each trial. We refer to fluctuations in pattern representations as "representational drift" and relate those changes to the probability of successful versus unsuccessful WM performance in a task with nominally low demands (one item, one critical feature) on WM capacity.

METHODS

Participants

Twenty self-reported healthy young adults (nine women, 18 right-handed; mean age = 25.1 years, SD = 4.2) participated in exchange for monetary compensation. All had normal or corrected-to-normal vision and provided informed consent. Procedures were approved by the institutional review board at the University of Nebraska-Lincoln. Seven additional participants also took part in the study, but their data sets were discarded because of

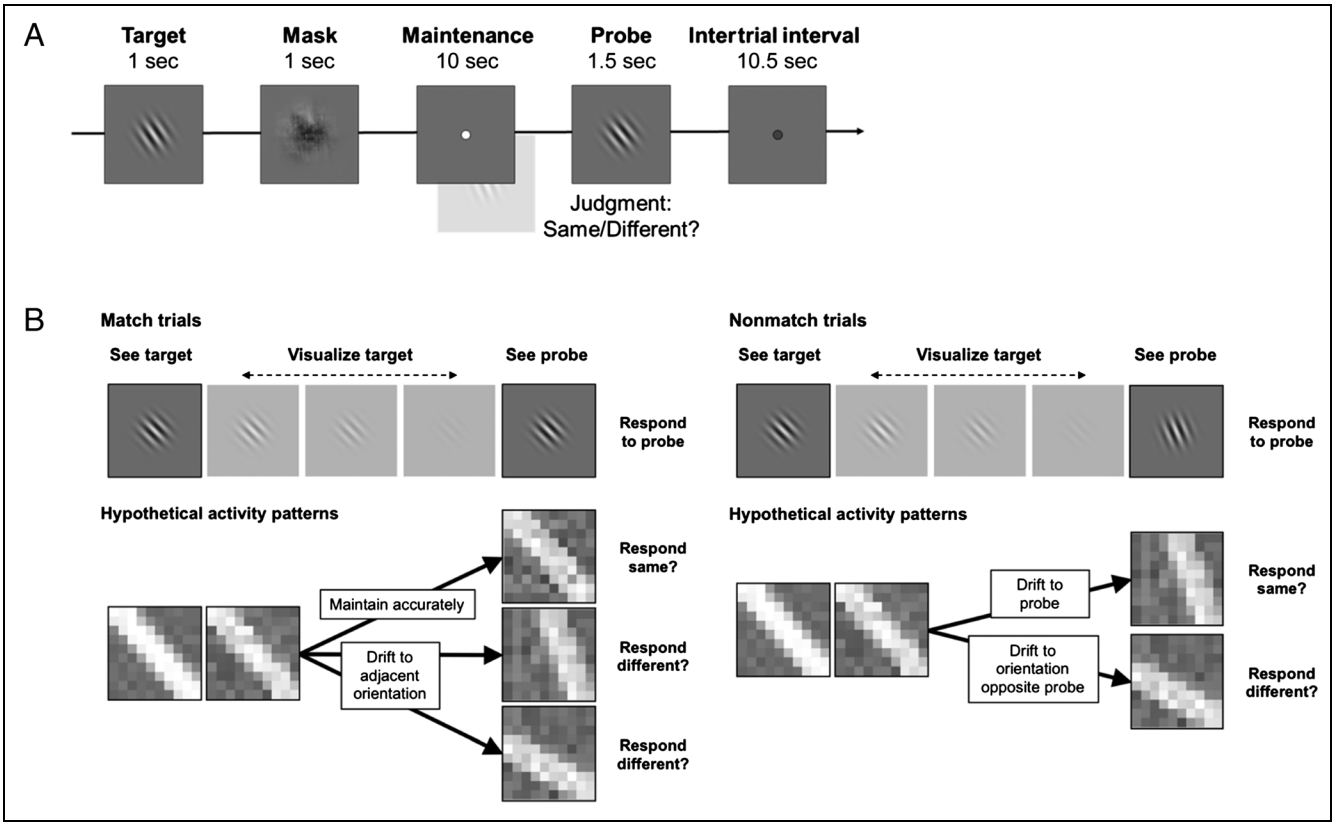

Figure 1. Task design and analysis approach. (A) Task design. Participants viewed an initial oriented Gabor patch (target) and held it in memory using a visualization strategy. A second patch (probe) then appeared, which either was the same item as the target (match trial) or had been rotated slightly (nonmatch trial). Participants pressed a button to indicate whether the target and probe were the same or different. Each participant only encountered a small number of discrete orientations with a fixed rotational distance between them; this distance was calibrated during a prescan staircasing procedure to maintain performance at approximately 75% correct. (B) Hypothetical brain activity patterns exemplifying representational drift analysis approach. For illustrative purposes, activity patterns are displayed as if they were 2-D images that resembled the stimuli/memoranda they represent. During target perception, patterns are presumed to be mostly veridical with little noise or directional bias. As maintenance begins, the representation persists but with decreasing signal to noise. The rightmost activity patterns then show hypothesized scenarios in which the patterns might drift throughout maintenance. For example, on match trials, the pattern could remain close to the true representation (top scenario; making participants more likely to correctly report a match) or drift toward an adjacent orientation (bottom two scenarios; making participants more likely to erroneously report a nonmatch). Nonmatch trials afford another comparison that is more appropriate to that condition; activity patterns could drift toward the probe orientation (top scenario; making participants more likely to erroneously report a match) or in the opposite direction (bottom scenario; making participants more likely to correctly report a nonmatch).

excessive head motion, sleepiness in the scanner, and/or low accuracy on the in-scanner behavioral task (generally <60% accuracy across all runs). All participants' data sets were considered immediately after their participation, and a decision to discard or retain their data was made based on the above criteria; thus, data sets that were rejected at this stage were never analyzed beyond motion correction. This procedure was repeated until the number of usable participants reached our target enrollment of 20.

Match/Nonmatch Recognition Task

Procedure

Participants completed seven runs of the match/nonmatch recognition task, comprising one initial prescan run to calibrate task difficulty (see Staircasing section) and six runs in the scanner. Each run had 24 trials, lasting 24 sec each (9.6 min total per run). On each trial (see Figure 1A), the target stimulus (a Gabor patch; for details, see Stimuli section) was first presented for 1 sec at the center of the display against a 50% gray background. This was immediately followed by a series of briefly presented mask images, slightly larger than the stimulus, presented for a total of 1 sec. The mask interval was followed by a fixation interval of 10 sec, during which participants maintained fixation on a small white dot at the center of the screen. Participants were instructed to remember the target using an imagery strategy, by visualizing it on the screen. At the end of the delay interval, a probe item (another Gabor patch) appeared for 1.5 sec. The probe either was identical to the target or had been rotated by an amount specific to each participant (determined earlier; see Staircasing section). Half of all trials were match trials, wherein the target and probe orientations were identical; the other half were nonmatch trials, wherein the probe was rotated either clockwise (50% of nonmatch trials) or counterclockwise

from the target orientation. Participants made a *same* or *different* response by pressing one of two buttons with the index or middle fingers, respectively, of their dominant hand. Responses were only recorded if they were made while the probe was onscreen, and a short confirmation tone was played if participants responded within this time frame, regardless of accuracy. Participants were encouraged to respond as quickly as possible while maintaining accuracy. A 10.5-sec fixation interval with a dark gray dot followed the probe. In the final 1 sec of this interval, the fixation dot changed to white to alert participants a new trial was about to begin. We chose relatively long WM delay and intertrial intervals to allow the BOLD signal to return as much as possible to baseline levels before the probe or the next trial, respectively, thus minimizing contamination of the BOLD signals by previous events' responses.

Staircasing

The first run of the match/nonmatch task took place outside the scanner and employed a staircasing procedure. During this run, task difficulty was adjusted by varying the rotation of the probe in nonmatch trials according to participants' performance, with the goal of achieving a 75% accuracy rate on the subsequent in-scanner runs. The staircasing run began with nonmatch probes rotated 10° relative to the target orientation. After each correct response, the nonmatch rotation was reduced by 1°, and after each incorrect response, it was increased by 3°. After the run, a weighted average of all probe rotation values was calculated (according to an inverse exponential function over trial number, so that later trials were weighted more heavily than earlier ones), and this rotation value was used for that participant in the in-scanner runs. Probe rotation values for the in-scanner runs were capped at a maximum of 15° and a minimum of 5°, even if staircasing performance produced higher or lower values. If, during the first two runs in the scanner, a participant's accuracy was below 60% or above 90% at the end of a run, difficulty was adjusted manually to attempt to bring performance closer to 75% on subsequent runs, and preadjustment runs were later removed from analysis. Because of this adjustment, one participant had two runs removed, and three participants had one run removed; all other participants completed all of their scanner runs with no difficulty adjustment needed. Task timing and procedure in the prescan staircasing run were largely similar to the scan runs, except that in the staircasing run, a feedback image appeared for 1 sec after the probe (smiley/frowny face for correct/incorrect), and the following intertrial interval was shortened by 1 sec accordingly, to 9.5 sec. Participants received no accuracy feedback in the scanner but did receive the confirmation tone to indicate that their response had been registered. Another difference was that, during staircasing, target orientations were randomly selected from the ranges 45° ± 35° and 135° ± 35°, whereas in the scan runs, a fixed set of target orientations was used (see below).

Stimuli

Target and probe stimuli were large, centrally presented Gabor patches (contrast 50%, phase 0), identical for all trials and participants except for their orientations. Target orientations for each in-scanner run were drawn equally from six evenly spaced orientations, wherein the spacing was determined by the earlier staircasing run. Probe orientations were one of eight possible evenly spaced orientations; six of these were the same as the target orientations, and the last two were one additional rotation step beyond the first and last target orientations. These two extreme probe orientations occurred only once per run each and only in nonmatch trials (e.g., when the most clockwise target was followed by a nonmatch probe rotated clockwise). In half of the scan runs, targets and probes were centered around a 45° orientation; in the other half of the scan runs, targets and probes were centered around a 135° orientation (with even/odd runs alternating between the 45° and 135° base orientations; starting base orientation counterbalanced across participants). For instance, if a participant's staircased difficulty was a step size of 10°, their six possible target orientations on a 45°-centered run would be oriented 20°, 30°, 40°, 50°, 60°, and 70°, and their eight possible probe orientations would be the same six target orientations with two additional orientations of 10° and 80°. Each 24-trial run was composed of a complete and balanced set of all possible target/probe configurations for the given base orientation—four trials of each possible target position, two of which were match and two of which were nonmatch, and of the two nonmatch trials, one each in which the probe was rotated clockwise/counterclockwise. Trial orders were pseudorandomized with the following constraints: (1) A maximum of three match or three nonmatch trials could occur consecutively; (2) for consecutive nonmatch trials, a maximum of two clockwise probe rotations, or two counterclockwise probe rotations, could occur consecutively; (3) the same target orientation was never presented in consecutive trials; (4) a previous trial's probe orientation could not reoccur as the next trial's target orientation (e.g., if the previous trial had used a probe orientation of 40°, the next trial's target orientation could not have been 40°); (5) for all runs centered around the same orientation for a given participant, a sequence of the same two trials was never repeated (e.g., for all 45°-centered runs, a target orientation of 40°, match trial, could have been followed by a target orientation of 60°, match trial, only once); and (6) each trial occurred in a different position order for every run (e.g., if a given participant's first trial in one run was a nonmatch trial that consisted of a target orientation at the leftmost position and probe orientation rotated clockwise, this nonmatch trial with the

same parameters would not have been presented first in any other run).

Mask stimuli (presented immediately after encoding to reduce retinal afterimages) consisted of grayscale scene images that were randomly selected from a large set, phase-scrambled, passed through a circular Gaussian envelope matching that of the Gabor patch, rotated a random amount, and flashed at 15 Hz for the duration of the 1-sec mask interval.

Retinotopic Mapping Task

After the match/nonmatch task runs, participants completed four ~2.5-min runs of a standard retinotopic mapping task. Participants viewed a wedge-shaped checkerboard pattern that rotated about a central fixation dot at 2.5 cycles per minute and flickered at a rate of 10 reversals per second. The wedge rotated either clockwise or counterclockwise for the entire run, alternating between runs. Participants were instructed to maintain fixation while watching for a brief color change in the fixation dot, which occurred approximately every 10 sec on average. Whenever they detected this change, they pressed a button with the index finger on their dominant hand.

fMRI Data Acquisition

Scanning was performed on a Siemens 3-T Skyra system with a 32-channel head coil. Functional scans used a multiband EPI sequence (Feinberg et al., 2010; Moeller et al., 2010) with repetition time (TR) = 1000 msec, echo time = 30 msec, 100 × 100 in-plane resolution, 60 axial slices with a thickness of 2.2 mm and 0-mm skip, field of view = 220 mm (overall voxel size = 2.2 × 2.2 × 2.2 mm), flip angle = 60°, and interleaved acquisition with a multiband factor of 4. Scans were prescribed with slices parallel to the AC–PC line and positioned for whole-brain coverage. Participants performed six runs of the main match/nonmatch task with 580 volumes (9 min 40 sec) per run and four runs of the retinotopic mapping task with 160 volumes (2 min 40 sec) per run. In each run, to allow the fMRI signal to reach steady-state before onset of the first trial, the scanner ran for 4 sec (i.e., 4 volumes) without collecting data, and an additional 4 sec per volume of collected data were discarded from the beginning of each run. T1-weighted MPRAGE anatomical images were also collected for each participant at the beginning of each scan session (TR = 2200 msec, echo time = 3.37 msec, 256 × 256 × 192 1-mm isotropic voxels, sagittal slice prescription).

One participant completed only five match/nonmatch task runs because of technical difficulties; another completed only half of her sixth and final match/nonmatch task run before it was aborted because of physical discomfort in the scanner. (In addition, as noted above, one additional participant had two runs removed and three participants had one run removed because of manual difficulty adjustments on early scan runs.) All participants completed all four runs of the retinotopy task.

fMRI Data Preprocessing

Initial processing of fMRI data was performed using SPM8 (Wellcome Trust Centre for Neuroimaging). Data were motion corrected, and all of a participant's functional runs were coregistered to a mean image of that participant's first functional run after motion correction. Each participant's T1 anatomical image was then coregistered to the Montreal Neurological Institute (MNI) average structural template image. Participants' motion-corrected functional images were then coregistered to this anatomical image, and all functional images were resampled. Thus, all participant data were approximately aligned to MNI space, but only affine transformations were applied, keeping data in individual-participant space with no nonlinear warping and only a single resampling step at the end. For the match/nonmatch task, fMRI signal values at each time point were then z-scored across the entire volume to control for signal fluctuations over time. These z-scored versions of the functional volumes were used as the basis of all pattern analyses.

Visual Cortex ROI Definition

On the basis of the knowledge that multiple retinotopic visual areas represent information about items held in visual WM (Harrison & Tong, 2009), our analyses were based on a functionally defined ROI comprising the most responsive retinotopically mapped voxels in the brain, irrespective of anatomical location. Data from the retinotopic mapping task were used to identify the 1,000 voxels that responded most robustly to the rotating checkerboard, resulting in an ROI similar in volume to that used by Harrison and Tong. Each voxel's time course during the retinotopy task was Fourier transformed, and the amplitude at the frequency corresponding to the rotation of the checkerboard wedge was extracted. This amplitude was converted to a Pearson correlation r value, Fisher z-transformed, and averaged across the four retinotopic mapping runs. The voxels with the highest mean z-transformed correlation values across all four mapping runs were selected. Then, in the main match/nonmatch task runs, this set of 1,000 voxels was extracted from each z-scored volume of functional data and used for all subsequent pattern analyses.

fMRI Pattern Similarity and Representational Drift

We calculated pattern similarity values and representational drift indices in the ROI described above to determine how ongoing changes in brain activity patterns corresponded with performance. Separate analyses were conducted for match trials (where target and probe orientations were the same) and nonmatch trials (where

target and probe orientations were different). Statistical comparisons focused on differences in pattern similarity or drift index between accurate and inaccurate trials during time points representing the encoding, maintenance, and probe intervals of the match/nonmatch task (see Results section).

Prototypical Activity Patterns

For each participant, we first obtained prototypical activity patterns for each unique target orientation seen during the match/nonmatch task. These patterns represent the "canonical" version of the expected activity patterns corresponding to visual processing of each task-relevant orientation; pattern similarity during WM to these prototypical patterns should thus reflect successful reinstantiation of those orientations. The prototypical voxel patterns for each orientation were calculated by averaging voxel patterns from all trials in which that orientation was the target, using patterns from fMRI Volume 5 within each trial (with TR = 1000 msec, this volume represented the peak activity occurring in response to the target presentation at $t = 0$ sec of the trial, after accounting for BOLD response delay). We thus obtained prototypical activity patterns for 12 unique orientations in total (e.g., the hypothetical participant with a staircased difficulty step size of 10° would have prototypical activity patterns for the 12 orientations of 20°–70° and 110°–160°, inclusive, in steps of 10°). Because target and probe stimuli used the same orientations (with the exception of the two most extreme probe orientations), prototypical activity patterns could therefore be obtained for almost all task-relevant orientations in a scan session. Trials in which the participant did not respond (only 3.5% of all trials) were still used to generate prototypical patterns but not used in any subsequent analyses based on accuracy.

Pattern Similarity and Representational Drift Calculations

Several task-relevant orientations and their corresponding voxel patterns formed the basis of our calculations. In addition to the target orientation and the probe orientation (note that the probe orientation was the same as the target on match trials but a different, adjacent orientation on nonmatch trials), we also defined two *target-adjacent* control orientations for match trials and a *probe-opposite* control orientation for nonmatch trials. The target-adjacent orientations were the two possible stimulus orientations adjacent to the target on a given trial (one rotated one step clockwise; the other, one step counterclockwise). The probe-opposite orientation was the orientation adjacent to the target on a given trial that was not the probe (i.e., rotated one step away from the target in the opposite direction from the probe).

Raw pattern similarity values were calculated by taking the Euclidean distance between two vectors, one representing the voxel pattern in the visual cortex ROI at a specific time point and the other representing a prototypical activity pattern in that ROI for one of the task-relevant orientations. For each trial, raw pattern similarity values were calculated for every fMRI volume (1–24). To account for any differences in initial representation, values at each time point were subtracted from the value at fMRI Volume 1. Thus, all pattern similarity timelines began at 0, and the sign was inverted so that positive values indicate higher similarity (lower distance). As the TR was 1000 msec, Volume 1 was collected between $t = 0$ sec and $t = 1$ sec; Volume 2, between $t = 1$ sec and $t = 2$ sec; and so on. In all figures, fMRI volumes are represented by the average time of their collection (0.5 sec, 1.5 sec, etc.). In all analyses, pattern similarity indices were calculated separately for accurate and inaccurate trials, and all individual-trial timelines were averaged within participants before entering them into statistical analyses.

For match trials, we calculated raw pattern similarity timelines comparing ongoing activity to each of the following prototypical activity patterns: the target pattern (Figure 2A; i.e., that trial's target orientation) and the two target-adjacent control orientations defined above. We then averaged those target-adjacent pattern similarities to obtain a single combined measure for the control orientations (Figure 2B). Two participants were removed from the match trial analysis because they had very few inaccurate match trials (two and three, respectively). For nonmatch trials, we calculated raw pattern similarity timelines comparing ongoing activity to each of the following prototypical activity patterns: the target pattern (Figure 3A; same as for match trials), the probe pattern (Figure 3B; i.e., that trial's nonmatching probe orientation), and the probe-opposite control orientation defined above (Figure 3C).

However, raw pattern similarity values on their own are an insufficient means of indicating the fidelity of participants' WM representations of a specific item, as it is possible for brain activity patterns to change in similarity toward or away from multiple representations at once. For example, if a participant became distracted and stopped paying attention midtrial, their brain activity patterns would likely become less similar to all task-relevant orientations at the same time. Thus, we also calculated representational drift indices that, unlike raw pattern similarity values, are capable of conveying whether activity patterns are drifting more toward one particular representation than another.

Representational drift indices were calculated by subtracting one raw pattern similarity timeline from another, allowing a direct comparison between the two. As all raw pattern similarity timelines were baselined to begin at 0, an advantage of this approach is that the representational drift index is guaranteed to correspond to a net change in pattern similarity toward a specific representation since the beginning of the trial, with positive values indicating net drift toward one orientation and negative values indicating net drift toward the other.

For match trials, we calculated representational drift timelines comparing the target orientation with the target-adjacent control orientations (Figure 2C, which represents a subtraction of the raw pattern similarity values in Figure 2B from those in Figure 2A; i.e., $PS_{target} - PS_{control}$). Thus, positive values indicate changes in pattern similarity toward the target orientation, whereas negative values indicate changes in pattern similarity toward the control orientations. Because any general effects such as distraction should affect both raw pattern similarity timelines equally, subtracting those timelines should cancel out such nonspecific influences on brain activity patterns, and any effects seen in the representational drift index should be because of differences in particular WM representations.

For nonmatch trials, we calculated representational drift timelines comparing the probe orientation with the probe-opposite control orientation (Figure 3D, which represents a subtraction of the values in Figure 3C from those in Figure 3B; i.e., $PS_{probe} - PS_{control}$). Thus, positive values indicate changes in pattern similarity toward the orientation that will ultimately be probed, whereas negative values indicate changes in pattern similarity toward the control orientation.

Finally, we combined representational drift timelines for both match and nonmatch trials (Figure 4). For match trials, positive values of representational drift indicate changes in pattern similarity toward the target orientation, which should bias participants toward correct behavioral responses. However, the reverse is true for nonmatch trials, where positive drift values indicate changes in pattern similarity toward the probe orientation, potentially biasing participants toward incorrect responses. Hence, we inverted the sign for nonmatch trials and averaged both drift indices, such that positive drift values for the combined index indicate changes in pattern similarity toward orientations associated with correct responses. (Note that this combined analysis included only the 17 participants used in the match trial analysis.)

As no prototypical activity patterns existed for the most extreme probe orientations (because those orientations were never seen as targets), raw pattern similarities and drift indices based on those orientations could not be calculated. Thus, we did not analyze either match or nonmatch trials where the target orientation was one of the end points of the range of possible target orientations for that run (e.g., in a target set using orientations of 20°, 30°, 40°, 50°, 60°, and 70°, the analysis would exclude trials where the target was 20° or 70°), as those calculations would have required nonexistent prototypical activity patterns for either the probe or control orientations.

RESULTS

We calculated representational drift in fMRI activity patterns during the delay period of a match/nonmatch recognition WM task to determine how ongoing changes in the quality of activity pattern representations correspond with performance. Participants viewed an initial oriented Gabor patch (the target), held it in memory for an 11-sec maintenance period, and then saw a second Gabor patch (the probe) that was either the same orientation as the target (match trial) or rotated slightly (nonmatch trial). They then pressed a button to indicate whether they thought the target/probe orientations were the same or different (see Figure 1A). Each participant only encountered a small number of discrete orientations with a fixed rotational distance between them; this distance was calibrated during a prescan staircasing procedure (for details, see Methods section) to keep performance at approximately 75% correct. Indices of representational drift were calculated based on changes in multivoxel pattern similarity during maintenance; these indicated whether brain activity patterns drifted toward (became more similar to) or away from (became less similar to) the prototypical (average) activity pattern associated with visual perception of a given orientation (e.g., the target or the probe; see Figure 1B for an illustrative example). The time courses of these representational drift indices were then compared between accurate and inaccurate trials to determine how representational drift in brain activity related to task performance.

Behavioral Performance

Each run was composed of 24 trials, 12 match trials and 12 nonmatch trials. Each trial lasted 24 sec (see Figure 1A), and thus each run was approximately 10 min long. All participants completed four to six task runs. Mean accuracy across all runs was 72.2% (*SD* = 7.2%), and mean RT was 921 msec (*SD* = 94 msec). Accuracy for match and nonmatch trials, respectively, was 81.9% (*SD* = 8.9%; excluding the two participants removed for low numbers of inaccurate trials) and 61.1% (*SD* = 9.8%). Average RT for match and nonmatch trials, respectively, was 919 msec (*SD* = 90 msec) and 939 msec (*SD* = 92 msec). Furthermore, 3.5% of the trials received no response; the average number of trials completed by each participant and included in our analyses was 132.

fMRI Analysis: Visual Cortex ROI

For each participant, we identified an omnibus visual cortex ROI comprising the top 1,000 visually responsive voxels in the brain, based on which voxels activated most in a standard retinotopic mapping task that followed the main task. For this ROI, we first report basic "raw" pattern similarity values at each point in the trial between that time point's activity pattern and the prototypical activity patterns (the average activity patterns corresponding to visual processing of that orientation; see Methods section for details) for the critical orientations used in that trial.

Temporal resolution was 1 sec. To examine effects at encoding (t = 0 sec in the trial) and probe presentation (t = 12 sec), we allowed 4–5 sec for BOLD signal delay and ran paired t tests comparing pattern similarity between accurate and inaccurate trials at fMRI Volumes 5 and 17 of the trial, respectively. To examine effects occurring during the maintenance period, we ran a linear repeated-measures analysis of variance including factors for the main effect of Accuracy and the linear and qua- dratic components of the Time × Accuracy interaction. This analysis was run on an 8-sec time window, starting from the BOLD peak of target encoding (fMRI Volume 5) and ending just before the probe appeared onscreen (Volume 12). All plots (Figures 2–4) depict the time course of an entire 24-sec trial (fMRI Volumes 1–24).

Match Trials: Target Pattern Similarity

Figure 2A shows, for match trials, pattern similarity between each time point of the trial and the prototypical activity pattern for the target orientation. A paired t test between accurate and inaccurate trials was not significant at encoding (p = .237), but was at probe, $t(17)$ = 2.14, p = .047. Specifically, at probe, pattern similarity was greater on inaccurate than accurate trials. This suggests that, counterintuitively, when there was more pattern similarity between that trial's activity pattern at probe and the prototypical target activity pattern, participants were more likely to (incorrectly) report a nonmatch. Our analysis of the maintenance period showed no significant effect of Accuracy (p = .241) and no linear (p = .194)

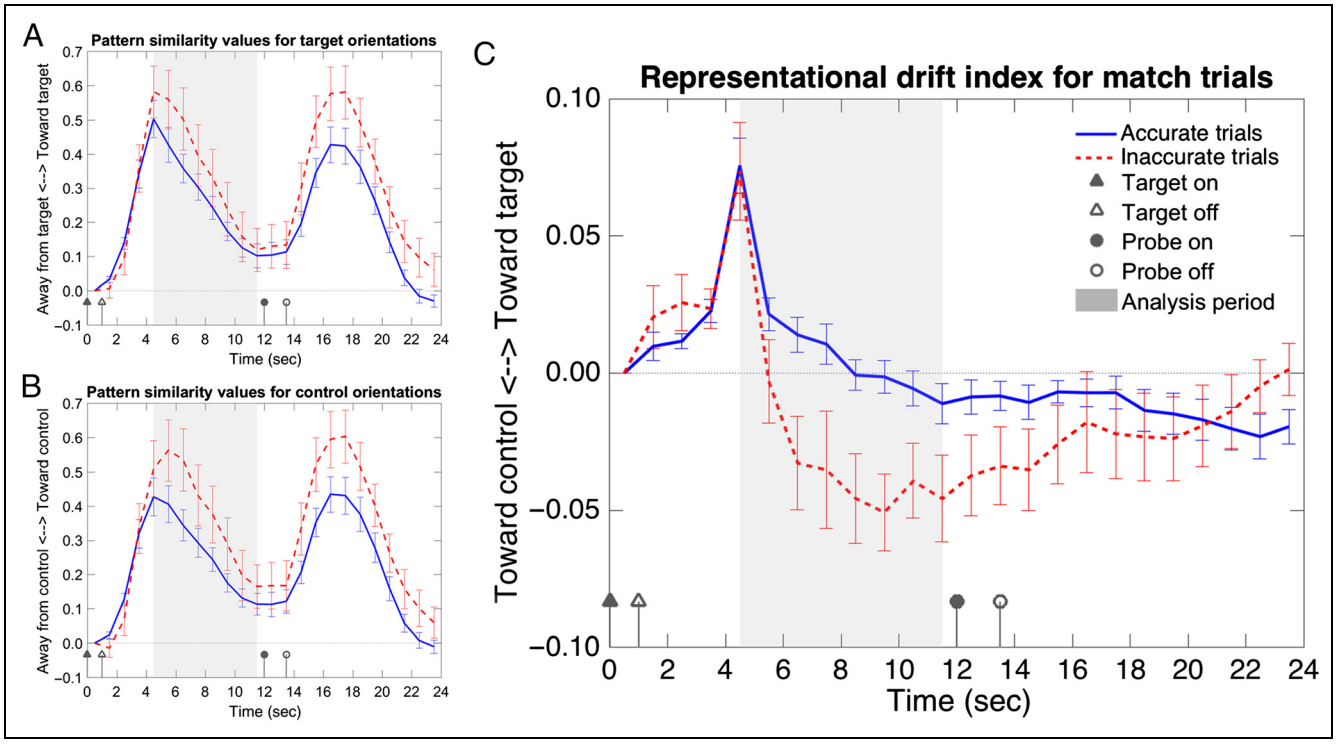

Figure 2. Match trials: Pattern similarity and representational drift. Timelines for pattern similarity and representational drift for match trials in visual cortex ROI. Accurate and inaccurate trials are plotted separately. All plots depict the time course of an entire 24-sec trial (fMRI Volumes 1–24). Data are represented as mean ± *SEM*. To examine effects occurring during the maintenance period, we analyzed an 8-sec time window (shaded in gray), starting from the BOLD peak of target encoding (fMRI Volume 5) and ending just before the probe appeared onscreen (Volume 12). The legend in C applies to all three plots. (A) Match trials: Target pattern similarity. Pattern similarity between each time point of the trial and the prototypical activity pattern for the target orientation. Positive values indicate more similar activity patterns to the prototypical target pattern for that trial. There was no difference between accurate and inaccurate trials during the maintenance period, but pattern similarity was higher for inaccurate trials at the time point representing the BOLD peak of probe presentation (fMRI Volume 17, t = 16.5 sec). (B) Match trials: Control orientation pattern similarity. Pattern similarity between each time point of the trial and the prototypical activity patterns for target-adjacent control orientations, which were the two orientations adjacent to the target. (Pattern similarity values from each control orientation were averaged to produce a single timeline.) Positive values indicate more similar activity patterns to prototypical target-adjacent orientations for that trial. Mirroring the target pattern similarity analysis in A, there was no difference between accurate and inaccurate trials during the maintenance period, but pattern similarity was higher for inaccurate trials at the time point representing the BOLD peak of probe presentation. (C) Match trials: Representational drift. Representational drift index for each time point of the trial. We subtracted the target-adjacent pattern similarities from the target pattern similarities to create a single index of representational drift, which conveys whether any changes in brain activity patterns represented a net drift in representational similarity toward either the target orientation or an adjacent (competing) orientation. Positive values indicate representational drift toward the target orientation for that trial, whereas negative values indicate representational drift toward target-adjacent orientations for that trial. We found a significant main effect of Accuracy during maintenance as well as a significant quadratic effect of Time × Accuracy; net representational drift away from the target on inaccurate trials was greater in the middle portion of the maintenance period than at the beginning or end.

or quadratic ($p = .286$) trends for the Time × Accuracy interaction. Thus, the raw pattern similarity between the prototypical target representation and participants' activity patterns during maintenance did not have a measurable effect on task accuracy for match trials.

Match Trials: Control Orientation Pattern Similarity

We then calculated pattern similarity between each time point of the trial and the prototypical patterns for two control orientations, which were the two orientations adjacent to the target. This allowed us to determine to what extent any pattern similarity effects seen in Figure 2A were unique to the target orientation or, conversely, whether they were because of less specific phenomena (e.g., general inattention on some trials, leading to inaccurate responses). Figure 2B shows, for match trials, pattern similarity between each time point of the trial and the prototypical activity patterns for the control orientations. Similar to the analysis of target pattern similarity, a paired t test between accurate and inaccurate trials was not significant at encoding ($p = .237$), but was at probe, $t(17) = 2.27$, $p = .037$. Pattern similarity at probe was again greater on inaccurate than accurate trials, suggesting that greater pattern similarity to any task-relevant orientation (target or target-adjacent) at retrieval is associated with (incorrect) reporting of a nonmatch. Our analysis of the maintenance period, as for target orientations, showed no significant effect of Accuracy ($p = .093$) and no linear ($p = .340$) or quadratic ($p = .096$) trends for the Time × Accuracy interaction. Thus, the raw pattern similarity during maintenance did not appear to have a measurable effect on match trial accuracy for either the target orientation or our target-adjacent control orientations.

Match Trials: Representational Drift

The pattern similarity analyses above did not strongly support any effects of raw pattern similarity during maintenance on participants' accuracy, for either the target orientation or the target-adjacent control orientations. Furthermore, the similarity between the timelines shown in Figure 2A and B suggests that raw pattern similarity alone may not be a good indicator of the quality of participants' WM representations for the target orientation, specifically. However, we hypothesized that the difference in raw pattern similarities between the target and control orientations ($PS_{target} - PS_{control}$) might better reflect the quality of the target's WM representation. Because any general effects such as distraction should affect both the target and control pattern similarity timelines equally, subtracting those timelines should cancel out such nonspecific influences on brain activity patterns. Thus, we subtracted the target-adjacent pattern similarities from the target pattern similarities to create a single index of representational drift. This index conveys, independent of any more general factors (e.g., waxing and waning attention), a true shift in pattern space toward either the target orientation or the control orientations, that is, whether any changes in brain activity patterns represented a net drift in representational similarity toward either the target orientation or an adjacent (competing) orientation.

Figure 2C shows, for match trials, the representational drift index for each time point of the trial. A paired t test between accurate and inaccurate trials was not significant at either encoding ($p = .866$) or probe ($p = .622$). However, our analysis of the maintenance period showed a significant main effect of Accuracy, $F(1, 17) = 5.84$, $p = .027$, with more net representational drift away from the target orientation (toward target-adjacent orientations) on inaccurate trials. The Time × Accuracy interaction showed no significant linear trend ($p = .232$) but did show a significant quadratic trend, $F(1, 17) = 5.52$, $p = .031$, where net representational drift away from the target on inaccurate trials was greater in the middle portion of the maintenance period than at the beginning or end. This suggests that participants were more likely to incorrectly report a nonmatch when their activity patterns drifted away from the target orientation and toward target-adjacent orientations; furthermore, this effect was largest in the middle portion of the maintenance period.

Nonmatch Trials: Target Pattern Similarity

Figure 3A shows, for nonmatch trials, pattern similarity between each time point of the trial and the prototypical activity pattern for the target orientation. A paired t test between accurate and inaccurate trials was not significant at either encoding ($p = .173$) or probe ($p = .132$). Our analysis of the maintenance period showed no significant effect of Accuracy ($p = .281$) and no quadratic trend ($p = .863$) for the Time × Accuracy interaction. However, there was a near-significant linear trend for the interaction, $F(1, 19) = 4.17$, $p = .055$, where pattern similarity was initially numerically higher for accurate than inaccurate trials, but the difference between accurate and inaccurate trials disappeared by the end of the maintenance period. This suggests that, when there was more initial pattern similarity between that trial's activity pattern and the prototypical target activity pattern, participants may have been more likely to (correctly) report a nonmatch, even though such starting differences were negated later in the maintenance period.

Nonmatch Trials: Probe Pattern Similarity

Figure 3B shows, for nonmatch trials, pattern similarity between each time point of the trial and the prototypical activity pattern for the probe orientation. Similar to the target pattern similarity analysis above, a paired t test between accurate and inaccurate trials was not significant at either encoding ($p = .103$) or probe ($p = .199$). Also

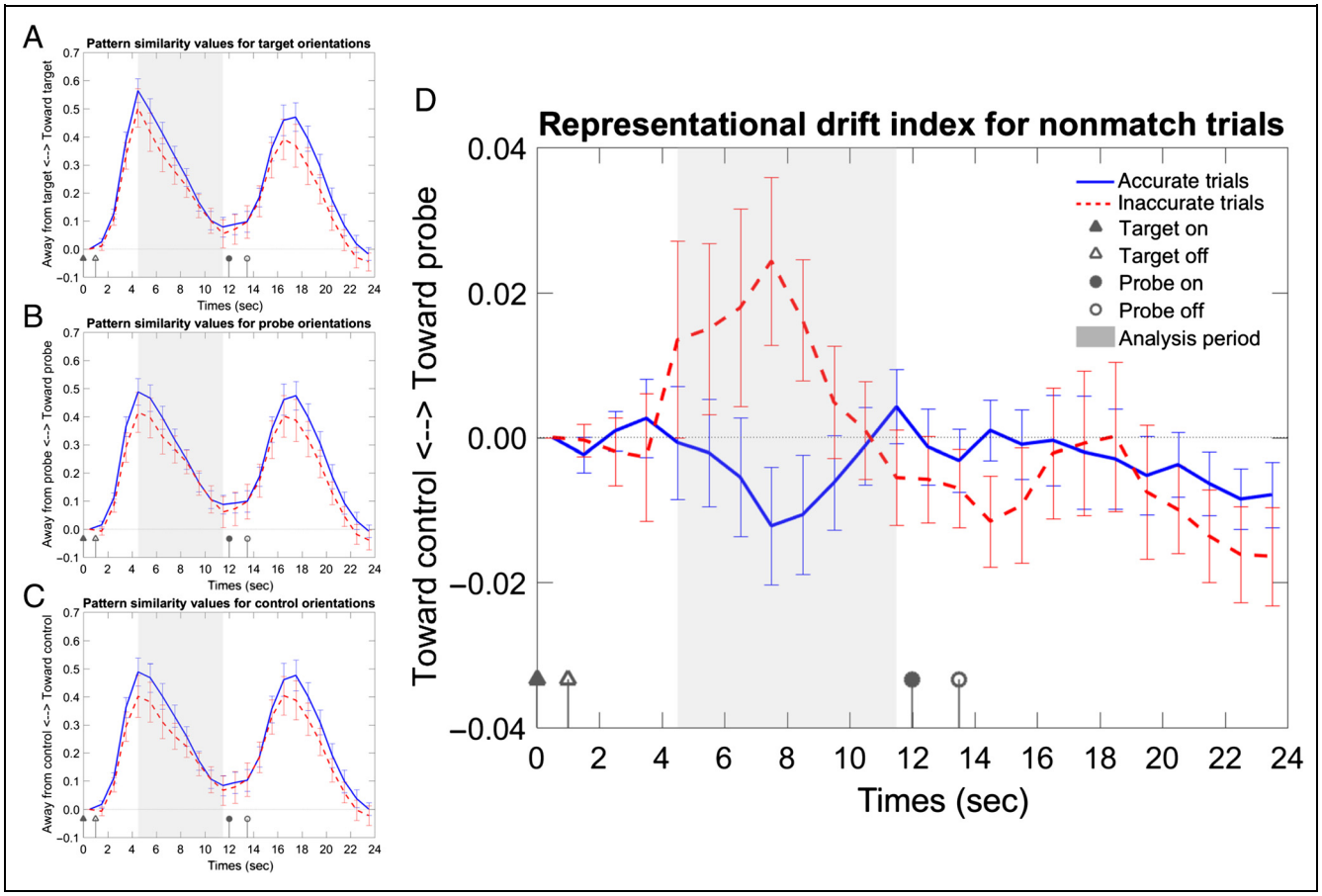

Figure 3. Nonmatch trials: Pattern similarity and representational drift. Timelines for pattern similarity and representational drift for nonmatch trials in visual cortex ROI. Accurate and inaccurate trials are plotted separately. All plots depict the time course of an entire 24-sec trial. Data are represented as mean ± *SEM*. As in match trials, we analyzed an 8-sec time window (shaded in gray), starting from the BOLD peak of target encoding (fMRI Volume 5) and ending just before the probe appeared onscreen (Volume 12). The legend in D applies to all four plots. (A) Nonmatch trials: Target pattern similarity. Pattern similarity between each time point of the trial and the prototypical activity pattern for the target orientation. Positive values indicate more similar activity patterns to the prototypical target pattern for that trial. Pattern similarity was initially numerically higher for accurate than inaccurate trials, but the difference disappeared by the end of the maintenance period. (B) Nonmatch trials: Probe pattern similarity. Pattern similarity between each time point of the trial and the prototypical activity pattern for the probe orientation. Positive values indicate more similar activity patterns to the prototypical probe pattern for that trial. Similar to target pattern similarity in A, probe pattern similarity was initially numerically higher for accurate than inaccurate trials, but the difference disappeared by the end of the maintenance period. (C) Nonmatch trials: Control pattern similarity. Pattern similarity between each time point of the trial and the prototypical activity pattern for the probe-opposite control orientation, which was the orientation that, like the probe, was adjacent to the target orientation, but was rotated in the opposite direction. Positive values indicate more similar activity patterns to the prototypical control orientation pattern for that trial. Similar to target (A) and probe (B) pattern similarity, control orientation pattern similarity was initially numerically higher for accurate than inaccurate trials, but the difference disappeared by the end of the maintenance period. (D) Nonmatch trials: Representational drift. Representational drift index for each time point of the trial. We subtracted the control pattern similarities from the probe pattern similarities to create a single index of representational drift, which conveys whether any changes in brain activity patterns represented a net drift in representational similarity toward either the probe orientation or the probe-opposite control orientation. Positive values indicate representational drift toward the probe orientation for that trial, whereas negative values indicate representational drift toward the probe-opposite orientation for that trial. We found a significant quadratic effect of Time × Accuracy; inaccurate trials showed net representational drift toward the probe in the middle portion of the maintenance period, but not at the beginning or end.

closely mirroring the analysis of target pattern similarity, our analysis of the maintenance period showed no significant effect of Accuracy ($p = .362$) and no quadratic trend ($p = .329$) for the Time × Accuracy interaction, but there was a significant linear trend for the Time × Accuracy interaction, $F(1, 19) = 4.82$, $p = .041$. As above, pattern similarity during maintenance was initially numerically higher for accurate than inaccurate trials, but that difference disappeared by the end of the maintenance period.

Nonmatch Trials: Control Orientation Pattern Similarity

We also calculated pattern similarity on nonmatch trials for a control orientation that, like the probe, was adjacent to the target orientation, but was rotated in the opposite direction (the probe-opposite orientation). As with the control orientation analysis for match trials, this allowed us to determine to what extent any pattern similarity effects seen in Figure 3A and B were unique to the

target/probe orientations or, conversely, whether they were because of less specific phenomena. Figure 3C shows pattern similarity between each time point of the trial and the prototypical activity pattern for the probe-opposite control orientation. A paired t test between accurate and inaccurate trials trended toward significance at encoding, $t(19) = 2.01$, $p = .059$, but not at probe ($p = .172$). Specifically, at encoding, pattern similarity was greater on accurate than inaccurate trials. Similar to the analyses of target and probe pattern similarity, our analysis of the maintenance period showed no significant main effect of Accuracy ($p = .159$) and no quadratic trend ($p = .922$) for the Time × Accuracy interaction but a significant linear trend for the interaction, $F(1, 19) = 7.70$, $p = .012$, where pattern similarity was initially numerically higher for accurate than inaccurate trials, but the difference disappeared by the end of the maintenance period. This suggests that greater pattern similarity to any task-relevant orientation (target, probe, or control) at encoding is associated with (correct) reporting of a nonmatch. However, none of these analyses of raw pattern similarity suggested that the fidelity of participants' WM representations for specific orientations during the maintenance period had an effect on accuracy.

Nonmatch Trials: Representational Drift

Given that the raw pattern similarity analyses above did not strongly support any orientation-specific effects on participants' accuracy, as well as the general similarity between the timelines shown in Figure 3A–C, it appeared that raw pattern similarity alone (as for match trials) may not be a good indicator of the quality of participants' WM representations for specific orientations. However, similar to our representational drift analysis for match trials, we hypothesized that the difference in raw pattern similarities between the probe and control orientations ($PS_{probe} - PS_{control}$) might better convey consequential changes in participants' WM representations during maintenance. Thus, we subtracted the control pattern similarities from the probe pattern similarities to create a single index of representational drift. Critically, this index of representational drift during nonmatch trials, unlike the representational drift index for match trials above, is capable of reflecting a directional effect of representational drift on accuracy; in other words, this index can indicate whether participants are more likely to (incorrectly) report a match when their WM representations drift toward the nonmatching probe orientation than when their WM representations drift away from the probe orientation and toward the probe-opposite control orientation.

Figure 3D shows, for nonmatch trials, the representational drift index for each time point of the trial. A paired t test between accurate and inaccurate trials was not significant at either encoding ($p = .307$) or probe ($p = .896$). Our analysis of the maintenance period showed no significant main effect of Accuracy ($p = .206$) and no linear trend ($p = .191$) for the Time × Accuracy interaction. However, there was a significant quadratic trend for the interaction, $F(1, 19) = 6.66$, $p = .018$, with inaccurate trials having net representational drift toward the probe in the middle portion of the maintenance period, but not at the beginning or end. This suggests that participants were more likely to incorrectly report a match when their activity patterns drifted toward the probe orientation (or, conversely, more likely to correctly report a nonmatch when their activity patterns drifted toward the probe-opposite orientation), with the representational drift effect being maximal in the middle portion of the maintenance period.

Generalized Representational Drift

We combined drift results from match and nonmatch trials to create a single index of generalized representational drift, shown in Figure 4, where positive values of representational drift indicate pattern similarity changes

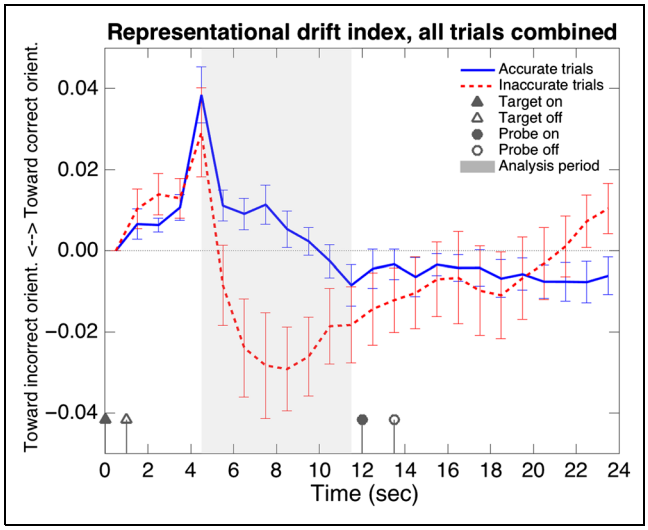

Figure 4. Generalized representational drift (match and nonmatch trials combined). Timeline for representational drift for match and nonmatch trials, combined, in visual cortex ROI. Accurate and inaccurate trials are plotted separately. The plot depicts the time course of an entire 24-sec trial. Data are represented as mean ± *SEM*. As in previous analyses, we analyzed an 8-sec time window (shaded in gray), starting from the BOLD peak of target encoding (fMRI Volume 5) and ending just before the probe appeared onscreen (Volume 12). We averaged results from match and nonmatch trials to create a single index of generalized representational drift, which conveys whether any changes in brain activity patterns represented a net drift in representational similarity toward orientations (orient.) associated with correct or incorrect behavioral responses. Positive values indicate representational drift toward orientations associated with correct responses for that trial, whereas negative values indicate representational drift toward orientations associated with incorrect responses for that trial. We found a significant main effect of Accuracy during maintenance as well as a significant quadratic effect of Time × Accuracy; net representational drift away from correct orientations on inaccurate trials was greater in the middle portion of the maintenance period than at the beginning or end.

toward orientations associated with correct behavioral responses (which we will refer to as "correct orientations") and negative values indicate pattern similarity changes toward orientations associated with incorrect behavioral responses. A paired t test between accurate and inaccurate trials was not significant at either encoding (p = .351) or probe (p = .857). Our analysis of the maintenance period showed a significant main effect of Accuracy, $F(1, 17) = 5.25, p = .035$, with more net representational drift away from correct orientations on inaccurate trials. The Time × Accuracy interaction showed no significant linear trend (p = .841) but did show a significant quadratic trend, $F(1, 17) = 10.17, p = .005$, where net representational drift away from correct orientations on inaccurate trials was greater in the middle portion of the maintenance period than at the beginning or end. This suggests, as the separate match and nonmatch analyses indicated, that participants were more likely to respond incorrectly when their activity patterns drifted away from the orientations associated with correct behavioral responses; furthermore, this effect was largest in the middle portion of the maintenance period.

DISCUSSION

We calculated representational drift in brain activity patterns during the delay period of a match/nonmatch recognition task to determine how ongoing changes in brain activity corresponded with WM performance. Analyses revealed that, overall, participants were more likely to respond incorrectly when their brain activity patterns drifted away from the orientations associated with correct behavioral responses in the recognition task. Separate analyses were also conducted for match trials (where target and probe orientations were the same) and nonmatch trials (where target and probe orientations were different), with similar results. In match trials, participants were more likely to incorrectly report that orientations did not match when their activity patterns drifted away from the target orientation and toward target-adjacent orientations. In nonmatch trials, participants were more likely to incorrectly report that orientations matched when their activity patterns drifted toward the probe orientation and away from a control orientation rotated, relative to the target, in the opposite direction of the probe.

These results suggest that WM failures can be at least partially explained by representational drift during maintenance. Neural drift effects analogous to those observed here have been theorized, and effects consistent with neural population activity drift have been observed in behavioral, animal, and modeling research (Rademaker, Park, Sack, & Tong, 2018; Schneegans & Bays, 2018; Wimmer et al., 2014; Burak & Fiete, 2012); however, this study represents, to our knowledge, the first human neuroimaging study to directly demonstrate the consequences of representational drift in brain activity patterns for WM performance.

Consequences of Representational Drift in Match and Nonmatch Trials

Representational drift for match trials assessed whether participants' WM representations drifted toward the target orientation or target-adjacent orientations. Drift for accurate and inaccurate trials was similar at encoding but then quickly diverged; activity patterns drifted closer to target-adjacent orientations for inaccurate than accurate trials, suggesting that participants were more likely to incorrectly report that orientations did not match when their WM representations were more similar to target-adjacent orientations. Interestingly, representational drift showed a quadratic trend, with maximal differentiation between correct and incorrect trials in the early-to-mid maintenance period; drift indices were not significantly different between accurate and inaccurate trials at encoding or probe. (Note that effects were typically maximal at the middle of the period we analyzed, but the analysis period terminated at probe presentation; thus, after accounting for BOLD lag, the center of the analysis period represented brain activity corresponding to ~4 sec into the 11-sec maintenance period.) This suggests that, even with successful encoding, disruption of WM patterns during the maintenance period could lead to an incorrect response on the subsequent probe. Furthermore, it did not appear necessary for this disruption to persist into the probe period to have an effect on behavior.

Representational drift for nonmatch trials measured whether participants' ongoing WM representation was relatively more similar to the probe orientation or a control orientation also adjacent to the target, but in the opposite direction from the probe. Generally, representational drift for accurate and inaccurate trials was similar throughout the trial except for the maintenance period, where representations drifted toward the probe for inaccurate trials but toward the control orientation for accurate trials. This suggests that, when participants' WM representation of the target was more similar to the probe orientation, they were more likely to incorrectly report that probe and target orientations matched. As in match trials, representational drift diverged between accurate and inaccurate trials primarily during the early-to-mid maintenance period; there was no significant difference in representational drift by accuracy at either encoding or probe.

Both these results suggest that the early-to-mid maintenance period is critical to WM accuracy (in line with long-term memory findings; Bergmann, Kiemeneij, Fernández, & Kessels, 2013; Ranganath et al., 2005) and that pattern similarity at encoding and probe may not necessarily guarantee accuracy if WM patterns are disrupted during maintenance. Our findings may indicate that participants' WM representations are most labile in the first few seconds after encoding, as early drift activity predicted WM accuracy, but drift activity later in the

maintenance period did not. Past research has documented a form of "activity-silent" WM, wherein neural activity for an unattended item drops to baseline during the maintenance period, even when the item is later successfully remembered at probe (Rose et al., 2016; Sprague, Ester, & Serences, 2016; Stokes, 2015). It is possible, then, that participants in our study actively maintained representations in early-to-mid maintenance and then allowed those representations to become dormant, which could account for the lack of differentiation in WM drift between accurate and inaccurate trials at probe. However, it appears that, even if participants' representations became relatively activity-silent by probe time, their behavioral decisions may have been based on their WM representations from earlier in the maintenance period. In turn, this suggests that those first few seconds of maintenance may comprise a critical consolidation period, after which the representations become more crystallized. This reflects findings in long-term memory research, where memories for events are labile during only a limited period when the memory is active (Lee, 2009; Nader, Schafe, & Le Doux, 2000).

Although "activity-silent" WM seems plausible, there are several other possibilities. Past studies have reported a strong dynamic component in WM coding (Murray et al., 2017; Wolff, Jochim, Akyürek, & Stokes, 2017), suggesting that items may be represented by a sequence of neural activity rather than a static pattern. If this is the case, early and not late drift activity in our study may have predicted task accuracy not because WM representations became dormant, but because their coding scheme changed more drastically than our drift index could account for. Furthermore, because of the slow nature of the BOLD response, it is difficult to infer the time scale of neural events very precisely, leaving open the possibility that drift on individual trials may have a shorter and/or more temporally variable time course than implied by these averaged results (a caveat that holds, of course, for any fMRI study that presents averaged timelines of BOLD data). It is also possible that behaviorally relevant representational drift may have persisted into the probe period at a neural level but was no longer reliably observable with fMRI because of the overall lower signal (and thus lower pattern similarity) at the end of the delay period.

In addition, given the challenging nature of the task, some participants may have adopted their own strategies for remembering the target stimulus, rather than using visualization as they were specifically instructed to do. For instance, participants may have learned, either implicitly or explicitly, that they were shown a discrete set of orientations and remembered them categorically, or they may have estimated and verbally encoded target orientations. Although no participants spontaneously reported these strategies, we did not formally survey them about this post-task, so it is unknown whether such alternative strategies may have affected our results. Still, if anything, such aberrant strategies should only serve to weaken our results but could not produce them spuriously, and thus it is worth noting that we observed results consistent with the drift hypothesis in visually responsive voxels regardless.

Accuracy-based Differences in "Raw" Pattern Similarity to Various Task-relevant Orientations

Although our primary hypotheses required calculating the novel representational drift indices described above to capture effects associated with specific WM representations (e.g., drift toward the probe orientation), we also observed differences by accuracy in the raw pattern similarities (a measure more commonly used in previous studies) used to compute those indices. In match trials, raw pattern similarity values between ongoing brain activity and any task-relevant orientation (i.e., the target orientation for that trial and the two target-adjacent orientations used as controls; see Figure 2A and B) were generally greater for inaccurate trials; although this varied over the time course of the trial (for instance, the difference was statistically significant at probe but not at encoding or maintenance), pattern similarity was numerically greater at every time point throughout the trial. In other words, when pattern similarity was greater for any orientation, participants tended to report a nonmatch. Conversely, in nonmatch trials, raw pattern similarity values to any task-relevant orientation (i.e., the target orientation, the nonmatching probe, and the probe-opposite control orientation; see Figure 3A–C) were generally greater for accurate trials. Again, this varied over the time course of the trial, but pattern similarity was numerically greater at most time points throughout the trial; this means that, as with match trials, when pattern similarity was greater for any orientation, participants tended to report a nonmatch. One likely explanation for this pattern of results is that higher pattern similarity to task-relevant orientations, in a manner that is not particularly specific to any one orientation or portion of the trial, reflects general alertness or task-focused states of mind; in other words, a participant who is staying focused on performing the WM task is likely to present brain activity patterns more similar to any task-relevant orientation than a participant who is distracted or otherwise inattentive.[1]

Our pattern of results thus suggests an overall task strategy wherein participants tend to report nonmatches more often when their attention is more focused and their representations of the WM target are clearer. Conversely, when participants' attention was less focused and thus they had representations of the target that were less clear, they may have been more likely to report a match. Put another way, it seems likely that participants adopted a violation-detection strategy in which they tended to report a nonmatch when their WM representations were sufficiently clear to establish confidence in their decision, whereas a match response could occur

either because they were actually confident of a match or because their WM representations were not clear enough to be confident of detecting a nonmatch. If most low-confidence responses defaulted to "match" rather than "nonmatch," this would also explain the greater number of "match" responses (assuming high-confidence responses are equally distributed between "match" and "nonmatch"; future follow-up studies could confirm this hypothesis by including an explicit confidence judgment in the probe). Thus, these findings based on raw pattern similarities may offer some insights into participants' strategies for performing the task, although they were not particularly effective for isolating representation-specific effects and instead appeared primarily to reflect overall attention or task focus. Rather, for representation-specific effects, the difference between pattern similarity timelines provided the critical measure, namely, the representational drift index described in the section above.

Conclusions

What are the causes of WM failure? In summary, our results constitute neural evidence that representational drift is among the factors that underlie such failure. When an item is held in WM, its representation is subject to random fluctuations. If those fluctuations bring the representation closer to those of nontarget items that may also appear in the environment, WM errors can occur.

Acknowledgments

We thank Lauren Bandel, Aaron Halvorsen, Rafay Khan, Joanne Murray, and Kerry Hartz for assistance with data collection and scanning. This work was supported by the National Science Foundation's Established Program to Stimulate Competitive Research (EPSCoR) award (#1632849) to M. R. J., T. J. V., and colleagues.

Reprint requests should be sent to Phui Cheng Lim, Department of Psychology, University of Nebraska-Lincoln, 238 Burnett, Lincoln, NE 68588, or via e-mail: cheng.lim@unl.edu.

Note

1. During peer review, the question was raised as to whether these non-orientation-specific effects might be eliminated by using a different similarity/distance metric than our chosen Euclidean distance. Thus, we reran our analyses using Pearson correlation and cosine similarity, as these alternative similarity measures are less sensitive to such global changes in brain activity (although they are not perfectly insensitive, as global factors such as activation amplitude are correlated with signal to noise, which would still affect those metrics). As expected, the differences between accurate and inaccurate trials shown in Figures 2A and 2B and 3A–C disappeared, but the drift effects shown in Figures 2C and 3D remained, with no meaningful changes in statistical significance and negligible numeric differences in the statistical analyses overall. Thus, in our study, Euclidean distance but not Pearson correlation or cosine similarity indexed these more global activity changes that we believe were because of greater general alertness or task focus on some trials than others. In this article, we present only the results using the Euclidean distance metric, but we note that, for future researchers who wish to ignore these more global effects, Pearson correlation and cosine similarity are also viable metrics that could help simplify data interpretation. Our thanks to the anonymous reviewer for the suggestion.

REFERENCES

Albers, A. M., Kok, P., Toni, I., Dijkerman, H. C., & de Lange, F. P. (2013). Shared representations for working memory and mental imagery in early visual cortex. *Current Biology*, *23*, 1427–1431.

Alvarez, G. A., & Cavanagh, P. (2004). The capacity of visual short-term memory is set both by visual information load and by number of objects. *Psychological Science*, *15*, 106–111.

Awh, E., Barton, B., & Vogel, E. K. (2007). Visual working memory represents a fixed number of items regardless of complexity. *Psychological Science*, *18*, 622–628.

Balsters, J. H., Robertson, I. H., & Calhoun, V. D. (2013). BOLD frequency power indexes working memory performance. *Frontiers in Human Neuroscience*, *7*, 207.

Bays, P. M., Catalao, R. F. G., & Husain, M. (2009). The precision of visual working memory is set by allocation of a shared resource. *Journal of Vision*, *9*, 7.

Bergmann, H. C., Daselaar, S. M., Beul, S. F., Rijpkema, M., Fernández, G., & Kessels, R. P. C. (2015). Brain activation during associative short-term memory maintenance is not predictive for subsequent retrieval. *Frontiers in Human Neuroscience*, *9*, 479.

Bergmann, H. C., Daselaar, S. M., Fernández, G., & Kessels, R. P. C. (2016). Neural substrates of successful working memory and long-term memory formation in a relational spatial memory task. *Cognitive Processing*, *17*, 377–387.

Bergmann, H. C., Kiemeneij, A., Fernández, G., & Kessels, R. P. C. (2013). Early and late stages of working-memory maintenance contribute differentially to long-term memory formation. *Acta Psychologica*, *143*, 181–190.

Blumenfeld, R. S., & Ranganath, C. (2006). Dorsolateral prefrontal cortex promotes long-term memory formation through its role in working memory organization. *Journal of Neuroscience*, *26*, 916–925.

Brewer, J. B., Zhao, Z., Desmond, J. E., Glover, G. H., & Gabrieli, J. D. E. (1998). Making memories: Brain activity that predicts how well visual experience will be remembered. *Science*, *281*, 1185–1187.

Burak, Y., & Fiete, I. R. (2012). Fundamental limits on persistent activity in networks of noisy neurons. *Proceedings of the National Academy of Sciences, U.S.A.*, *109*, 17645–17650.

Derrfuss, J., Ekman, M., Hanke, M., Tittgemeyer, M., & Fiebach, C. J. (2017). Distractor-resistant short-term memory is supported by transient changes in neural stimulus representations. *Journal of Cognitive Neuroscience*, *29*, 1547–1565.

Ester, E. F., Anderson, D. E., Serences, J. T., & Awh, E. (2013). A neural measure of precision in visual working memory. *Journal of Cognitive Neuroscience*, *25*, 754–761.

Feinberg, D. A., Moeller, S., Smith, S. M., Auerbach, E., Ramanna, S., Glasser, M. F., et al. (2010). Multiplexed echo planar imaging for sub-second whole brain fMRI and fast diffusion imaging. *PLoS One*, *5*, e15710.

Hannula, D. E., & Ranganath, C. (2008). Medial temporal lobe activity predicts successful relational memory binding. *Journal of Neuroscience*, *28*, 116–124.

Harrison, S. A., & Tong, F. (2009). Decoding reveals the contents of visual working memory in early visual areas. *Nature*, *458*, 632–635.

Jackson, J., Rich, A. N., Williams, M. A., & Woolgar, A. (2017). Feature-selective attention in frontoparietal cortex: Multivoxel codes adjust to prioritize task-relevant information. *Journal of Cognitive Neuroscience*, *29*, 310–321.

Johnson, M. R., & Johnson, M. K. (2014). Decoding individual natural scene representations during perception and imagery. *Frontiers in Human Neuroscience*, *8*, 59.

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

Kim, S.-Y., Kim, M.-S., & Chun, M. M. (2005). Concurrent working memory load can reduce distraction. *Proceedings of the National Academy of Sciences, U.S.A.*, *102*, 16524–16529.

Kuhl, B. A., Rissman, J., Chun, M. M., & Wagner, A. D. (2011). Fidelity of neural reactivation reveals competition between memories. *Proceedings of the National Academy of Sciences, U.S.A.*, *108*, 5903–5908.

Kuhl, B. A., Rissman, J., & Wagner, A. D. (2012). Multi-voxel patterns of visual category representation during episodic encoding are predictive of subsequent memory. *Neuropsychologia*, *50*, 458–469.

LaRocque, J. J., Lewis-Peacock, J. A., Drysdale, A. T., Oberauer, K., & Postle, B. R. (2013). Decoding attended information in short-term memory: An EEG study. *Journal of Cognitive Neuroscience*, *25*, 127–142.

Lee, J. L. C. (2009). Reconsolidation: Maintaining memory relevance. *Trends in Neurosciences*, *32*, 413–420.

Lee, S.-H., Kravitz, D. J., & Baker, C. I. (2012). Disentangling visual imagery and perception of real-world objects. *Neuroimage*, *59*, 4064–4073.

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*, 279–281.

Luck, S. J., & Vogel, E. K. (2013). Visual working memory capacity: From psychophysics and neurobiology to individual differences. *Trends in Cognitive Sciences*, *17*, 391–400.

Lupyan, G. (2008). From chair to "chair": A representational shift account of object labeling effects on memory. *Journal of Experimental Psychology: General*, *137*, 348–369.

Moeller, S., Yacoub, E., Olman, C. A., Auerbach, E., Strupp, J., Harel, N., et al. (2010). Multiband multislice GE-EPI at 7 tesla, with 16-fold acceleration using partial parallel imaging with application to high spatial and temporal whole-brain fMRI. *Magnetic Resonance in Medicine*, *63*, 1144–1153.

Murray, J. D., Bernacchia, A., Roy, N. A., Constantinidis, C., Romo, R., & Wang, X.-J. (2017). Stable population coding for working memory coexists with heterogeneous neural dynamics in prefrontal cortex. *Proceedings of the National Academy of Sciences, U.S.A.*, *114*, 394–399.

Nader, K., Schafe, G. E., & Le Doux, J. E. (2000). Fear memories require protein synthesis in the amygdala for reconsolidation after retrieval. *Nature*, *406*, 722–726.

Rademaker, R. L., Park, Y. E., Sack, A. T., & Tong, F. (2018). Evidence of gradual loss of precision for simple features and complex objects in visual working memory. *Journal of Experimental Psychology: Human Perception and Performance*, *44*, 925–940.

Ranganath, C., Cohen, M. X., & Brozinsky, C. J. (2005). Working memory maintenance contributes to long-term memory formation: Neural and behavioral evidence. *Journal of Cognitive Neuroscience*, *17*, 994–1010.

Rose, N. S., LaRocque, J. J., Riggall, A. C., Gosseries, O., Starrett, M. J., Meyering, E. E., et al. (2016). Reactivation of latent working memories with transcranial magnetic stimulation. *Science*, *354*, 1136–1139.

Schneegans, S., & Bays, P. M. (2018). Drift in neural population activity causes working memory to deteriorate over time. *Journal of Neuroscience*, *38*, 4859–4869.

Serences, J. T., Ester, E. F., Vogel, E. K., & Awh, E. (2009). Stimulus-specific delay activity in human primary visual cortex. *Psychological Science*, *20*, 207–214.

Solomon, E. A., Kragel, J. E., Sperling, M. R., Sharan, A., Worrell, G., Kucewicz, M., et al. (2017). Widespread theta synchrony and high-frequency desynchronization underlies enhanced cognition. *Nature Communications*, *8*, 1704.

Sprague, T. C., Ester, E. F., & Serences, J. T. (2014). Reconstructions of information in visual spatial working memory degrade with memory load. *Current Biology*, *24*, 2174–2180.

Sprague, T. C., Ester, E. F., & Serences, J. T. (2016). Restoring latent visual working memory representations in human cortex. *Neuron*, *91*, 694–707.

Stokes, M. G. (2015). 'Activity-silent' working memory in prefrontal cortex: A dynamic coding framework. *Trends in Cognitive Sciences*, *19*, 394–405.

Vogel, E. K., McCollough, A. W., & Machizawa, M. G. (2005). Neural measures reveal individual differences in controlling access to working memory. *Nature*, *438*, 500–503.

Vogel, E. K., Woodman, G. F., & Luck, S. J. (2001). Storage of features, conjunctions, and objects in visual working memory. *Journal of Experimental Psychology: Human Perception and Performance*, *27*, 92–114.

Wagner, A. D., Schacter, D. L., Rotte, M., Koutstaal, W., Maril, A., Dale, A. M., et al. (1998). Building memories: Remembering and forgetting of verbal experiences as predicted by brain activity. *Science*, *281*, 1188–1191.

Ward, E. J., Chun, M. M., & Kuhl, B. A. (2013). Repetition suppression and multi-voxel pattern similarity differentially track implicit and explicit visual memory. *Journal of Neuroscience*, *33*, 14749–14757.

Wimmer, K., Nykamp, D. Q., Constantinidis, C., & Compte, A. (2014). Bump attractor dynamics in prefrontal cortex explains behavioral precision in spatial working memory. *Nature Neuroscience*, *17*, 431–439.

Wolff, M. J., Jochim, J., Akyürek, E. G., & Stokes, M. G. (2017). Dynamic hidden states underlying working-memory-guided behavior. *Nature Neuroscience*, *20*, 864–871.

Xu, Y., & Chun, M. M. (2006). Dissociable neural mechanisms supporting visual short-term memory for objects. *Nature*, *440*, 91–95.

Xue, G., Dong, Q., Chen, C., Lu, Z., Mumford, J. A., & Poldrack, R. A. (2010). Greater neural pattern similarity across repetitions is associated with better memory. *Science*, *330*, 97–101.

Yoon, J. H., Curtis, C. E., & D'Esposito, M. (2006). Differential effects of distraction during working memory on delay-period activity in the prefrontal cortex and the visual association cortex. *Neuroimage*, *29*, 1117–1126.

Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature*, *453*, 233–235.