

The Role of Location-Context Binding in Nonspatial Visual Working Memory
Context Binding in Nonspatial Working Memory

Ying Cai^{1,2}, Jacqueline M. Fulvio², Qing Yu², Andrew D. Sheldon^{3,4}, and Bradley R. Postle^{2,4*}

¹ Department of Psychology and Behavioral Science, Zhejiang University, Hangzhou, 310007, China

² Department of Psychology, University of Wisconsin–Madison, Madison, WI 53706,

³ Medical Scientist Training Program and Neuroscience Training Program, Madison, WI 53706, USA

⁴ Department of Psychiatry, University of Wisconsin–Madison, Madison, WI, 53706

4. YC, QY, and BRP Designed research, YC, JMF, ADS and Performed research, YC, JMF, and QY Analyzed data, YC, JMF, and BRP Wrote the paper.

5. Correspondence should be addressed to postle@wisc.edu

6. Number of Figures: 7

7. Number of Tables: 0

8. Number of Multimedia

9. Number of words for Abstract: 249

10. Number of words for Significance Statement: 118

11. Number of words for Introduction: 547

12. Number of words for Discussion: 760

14. Authors report no conflict of interest

15. NIH MH064498 to BRP

Abstract

Successful retrieval of an item from visual working memory (VWM) often requires an associated representation of the trial-unique context in which that item was presented. In Experiment 1, fMRI of 16 male and female humans replicated a previous dissociation of the effects of manipulating memory load in comparison to the effects of manipulating context binding, by comparing VWM for one oriented line versus for three lines individuated by their location versus for three “heterogeneous” items drawn from different categories (orientation, color, and luminance): Delay-period fMRI signal in frontal cortex and intraparietal sulcus (IPS) was sensitive to stimulus homogeneity rather than to memory load per se. Additionally, inspection of behavioral performance revealed a broad range of individual differences in the probability of responses to nontargets (a.k.a. “swap errors”), and an ad hoc comparison of *high swap-error* versus *low swap-error* groups generated several intriguing results: At recall, *high swap-error* subjects were seen to represent both the orientation and the location of the probed item less strongly, and with less differentiation from nonprobed items, and delay-period signal in IPS predicted behavioral and neural correlates of context binding at recall. In Experiment 2, which was a preregistered replication, the 27 male and female humans were grouped into *low* and *high swap-error* groups by median split, and the results were broadly consistent with Experiment 1. These results present a neural correlate of swap errors, and suggest that delay-period activity of the IPS may be more important for the operation of context binding than for representation per se of stimulus identity.

Significance Statement

Although we often think of the contents of visual working memory (VWM) as representations of the items that need to be remembered, each item’s trial-unique context is also critical for successful performance. For example, if one observes a red, then a black, then a blue car passing through an intersection, vivid memory for the colors, alone, wouldn’t allow one to execute the instruction “Follow the first of the three cars that just drove by.” Although manipulating load is commonly assumed to isolate storage functions, requiring memory for multiple items drawn from the same category also increases demands on the context binding needed to individuate these items. This experiment tracked the influence of context binding on VWM stimulus processing.

Introduction

Individual differences in human visual working memory (VWM) capacity result from several factors, including the strategic deployment of attention (Fukuda et al. 2015), retrieval-related processes (Unsworth et al. 2014) and resistance to interference (Vogel et al. 2005). In fMRI studies, delay-period activity in the intraparietal sulcus (IPS) scales with the number of items in the memory set before asymptoting at an individual's VWM capacity (Todd and Marois 2004, Todd and Marois 2005), and has often been assumed to reflect stimulus representation in VWM (Bettencourt and Xu 2015, Xu 2017). Here we test the possibility that this activity may also reflect context binding: the association of each item in the memory set with its unique episodic context. In tests that require VWM for an array of colored squares, for example, the subject must remember not only each of the colors, but also the location at which each color was presented. Intact memory for all the colors but an impaired representation of which had been presented where can lead to "swap errors" (Schneegans and Bays 2017). Indeed, some theoretical accounts hold that context binding is the factor that determines whether a stimulus can meaningfully be said to be "in" VWM (Oberauer and Lin 2017).

In one previous study (Gosseries, Yu, et al., 2018), we sought to unconfound the storage demands of a high-load condition from its context-binding demands by varying stimulus category homogeneity within the memory set: Subjects were asked to remember either the direction of motion in one random-dot kinematogram (RDK; 1M trials), the directions in three RDKs (3M), or the direction in one RDK plus the colors of two color patches (1M2C). Although delay-period activity of the IPS was elevated during 3M relative to 1M trials, it was comparable for 1M2C and 1M trials, indicating sensitivity to a factor other than memory load per se. In this experiment, the critical context was ordinal position, because stimuli on three-item trials were presented serially and all in the same location, with the item to be recalled prompted by a digit indicating "1st," "2nd," or "3rd."

In the present experiment, we tested context binding in VWM more directly. First, we changed the behavioral task such that stimulus location was the critical contextual dimension. Because the neural correlates of spatial processing are better understood than those of ordinal processing, this allowed a more direct assessment of variation in the representation of stimulus context. Second, we measured individual differences in swap error rate, a behavioral index of context binding, to relate variability in context binding efficacy with neural signals.

In a preliminary fMRI study enrolling 16 healthy young adults (Experiment 1), we replicated the earlier finding that aggregated delay-period activity in IPS is more sensitive to category homogeneity in the stimulus set than to the number of items per se, reinforcing the association of this region with functions other than stimulus representation. Additionally, post hoc categorization of subject by swap error rate revealed several intriguing differences in the neural representation of stimulus content (orientation) and context (location). These gave rise to a set of preregistered hypotheses that were tested in Experiment 2. We believed that the results from the preregistered experiment (presented here as Experiment 2) would provide novel insight into the interpretation of delay-period signals, and the computations underlying context-binding in VWM.

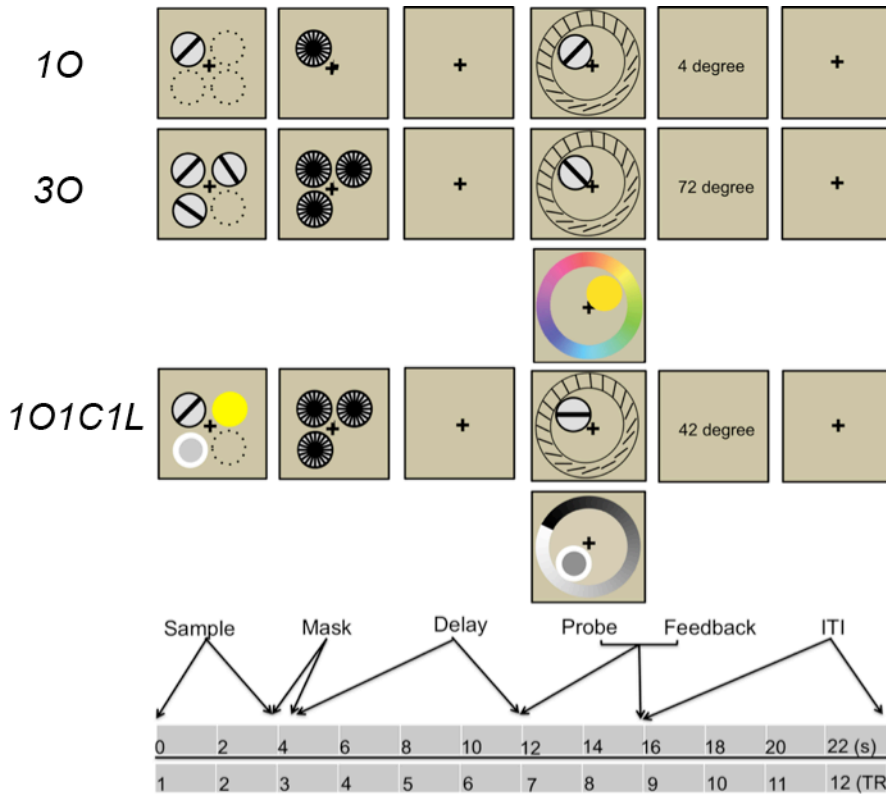


Figure 1. Schematic illustration of the three trial types from Experiment 1 (1O = one orientation; 3O = three orientations; 1O1C1L = one orientation, one color, one luminance). Dotted circles indicate the other possible stimulus presentation locations, but were not presented during experiment. Experiment 2 did not include 1O1C1L trials.

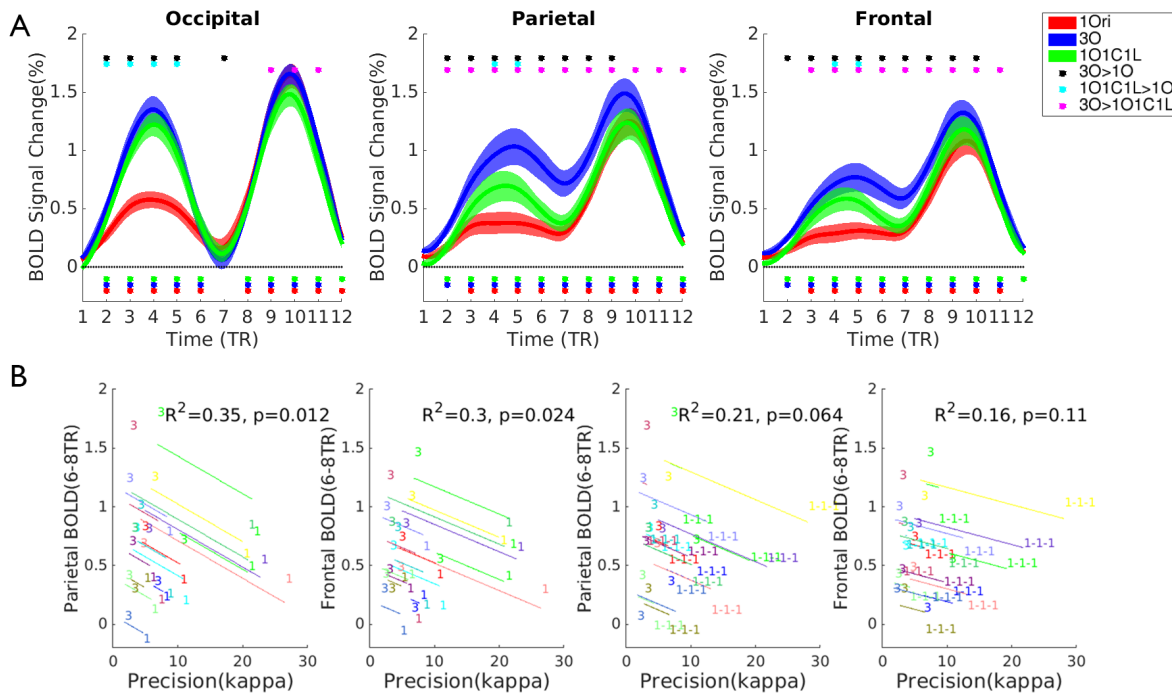


Figure 2. BOLD signal intensity results from three ROIs in Experiment 1. A. Trial-averaged BOLD signal. Dots below the x-axis indicate significance vs. baseline; dots above the plots indicate significant differences between trial types. B. Within-subject correlations (ANCOVAs) between delay-period BOLD signal intensity and behavioral precision of recall. In each plot, data from each subject are portrayed in a different color. The “1,” “3” and “1-1-1” symbols indicate individual values in the 1O, 3O, and 1O1C1L tasks, respectively. Thus, for example, the left-most plot illustrates that on 3O trials, delay-period BOLD signal in IPS is high and precision is low, whereas on 1O trials, delay-period BOLD signal in IPS is low and precision is high. The slope of the lines indicates the tendency at the group level, and the offsets between lines indicate where different individuals sit within this space. Asterisks indicate significant correlation at $p < 0.05$, triangles indicate trends ($.05 < p < 0.1$).

Material and Methods

Preregistered hypotheses

In this and all subsequent sections of this manuscript we will make the distinction between “Experiment 1” -- the preliminary dataset and results from which our predictions arose -- and “Experiment 2” for which we preregistered the a priori set of hypotheses listed in this section, and the methods that we planned to use to generate and analyze de novo data in order to test these hypotheses. Importantly, whereas two of the hypothesis generating results from Experiment 1 can be characterized as “post-hoc dichotomization of participants based on a median split on scores of some variable of interest (Yarkoni and Braver 2010),” which can be inferentially problematic, an a priori extreme groups design can be an effective way to “increase [...] power to detect effects by reducing the variance between participant groups relative to the variance within groups, thereby inflating effect sizes and making them easier to detect (Yarkoni and Braver 2010).”

(Note that for studies with preregistered hypotheses the convention is to state these hypotheses in the Introduction or the very beginning of the Methods section. As a consequence, explanation of the rationale behind specific predictions, and of the names of some of the measures in which the hypotheses are framed, appear in a subsection at the end of the Methods section, titled *Implementation of a priori hypothesis tests*.)

Hypothesis 1 (extreme-group analysis): The orientation recall specificity effect will be significantly greater, at TR 10, for the *low swap-error* group than for the *high swap-error* group.

Hypothesis 2 (extreme-group analysis): The location recall specificity effect will be significantly greater, at TR 10, for the *low swap-error* group than for the *high swap-error* group.

Hypothesis 3 (continuous analysis carried out with Spearman correlation of rank ordering): Individual differences in delay-period representation of stimulus locations, as assessed with multivariate inverted encoding modeling (IEM), will correlate with individual differences in swap-error rate (i.e., the “probability of a nontarget response,” or *pN*, parameter, to be described in the subsection on *Analysis of behavioral data*). (Note that the choice of a Spearman correlation reflects the fact that subjects in the *low swap-error* group will necessarily have *pN* values close to 0.)

Hypothesis 4 (continuous analysis carried out with Pearson correlation): Individual differences in delay-period representation of stimulus location, operationalized as the amplitude of the IEM reconstruction of the probed item, will correlate with individual differences in the location recall specificity effect.

Precise details about how each of these hypotheses was generated and tested, and how the *n* needed for adequate statistical power was calculated, are presented in the final subsection of this section, on *Implementation of a priori hypothesis tests*.

Additional analyses and “secondary hypotheses”: In addition to the theoretically important hypotheses listed above, we planned to carry out several additional analyses that can be considered of “secondary” importance for Experiment 2, because their outcomes will have little-to-no consequence for our theoretically motivated predictions. Some of these entailed replication of analyses of BOLD signal-intensity data from Experiment 1, that were, themselves, replications of previous findings. Others can be thought of as being “descriptive,” even though they may literally entail carrying out a significance test. In this latter category, for example, it is important to determine the TRs at which the IEM reconstruction of targeted stimulus orientation is statistically reliable, for both the 1O and the 3O conditions. Furthermore, it is of interest (but not critical for

Hypotheses 1-4) to know whether any reliable reconstructions for the *low swap-error* group differ quantitatively from those of the *high swap-error* group. These will be identified as “secondary hypotheses.”

Subjects

For Experiment 1, 18 right-handed volunteers (10 females, aged 18-25 years [Mean (SD) = 21.70 (1.75)]) from the [removed for blinding] community participated in a behavior-only experiment for remuneration (\$10/h). Two of the 18 elected not to participate in the subsequent fMRI study, and remaining 16 subjects (8 females, aged 18-25 years (Mean (SD) = 20.50 (1.78))) participated the fMRI experiment for additional remuneration (\$20/h). All subjects provided informed consent according to the procedures approved by the Health Sciences Institutional Review Board at the [removed for blinding]. Subjects had normal or corrected-to-normal vision, no reported history of neurological or psychiatric disease, and, for the fMRI experiment, no contraindications for MRI.

For *Experiment 2*, we carried out an extreme-group analysis in which the initial step was to recruit healthy adult volunteers within the age range of 18-25 years, from the [removed for blinding] community, for a behavioral screening session comprising 100 trials of delayed recall (a.k.a., “delayed estimation”) of orientation with a set size of three. Stimuli and procedures closely followed those from 30 trials from Experiment 1 (see *Stimuli*, below). The resultant recall data were fit with the 3-factor mixture model described in the subsection on *Behavioral Tasks* (below), and subjects with swap error rates of < 5% were selected for the fMRI sessions as part of the *low swap-error group*, and subjects with swap error rates > 12% were selected as part of the *high swap-error group*. The preregistered plan called for between 11-18 subjects for the *low swap-error group*, and between 11-18 subjects for the *high swap-error group*. The *Ns* required for each group were planned as ranges, to reflect the following factors (detailed in the subsection on *Implementation of a priori hypothesis tests*): 1) power analysis indicated that 29 subjects were required to achieve 90% power to detect the effect predicted by *Hypothesis 3*, but because this was a continuous analysis, it didn’t constrain how many of these 29 needed to belong to either extreme group; and 2) power analysis for a secondary hypothesis indicated that 11 subjects were needed to achieve 90% power to observe a significant reconstruction of stimulus orientation at TR 10, in occipital cortex, on 30 trials. Because all other a priori estimated *Ns* were lower than these numbers, our stopping point for screening was when we had successfully recruited and tested a minimum of 11 subjects with swap error rates of < 5%, a minimum of 11 subjects with swap error rates of > 12% (selected to ensure reasonable balance between the two groups), and a total of 29 subjects. (All subjects were paid at the end of the behavioral testing session and departed from the laboratory before their data were analyzed. They were told that they might be subsequently recontacted to be invited to participate in an fMRI study. Those re-contacted for the fMRI sessions of Experiment 2 were not informed of how they had been categorized with regard to *pN*, nor were they informed about the concept or definition of a swap error. It was also possible that individuals whose *pN* excluded them from Experiment 2 could be recontacted to participate in a different study being carried out by our research group.)

Note that, because it is measures of *pN* acquired during the fMRI session that we used to test our hypotheses, there were two possible developments that could have required us to recruit and behaviorally screen more individuals than the number listed above. The first was if the number of subjects whose *pN* from the fMRI scanning session satisfying the criterion for the *low swap-error group* dropped below 11, and the second was if any subject’s fMRI data were unuseable, whether due to withdrawal from the study or poor data quality.

Stimuli

In Experiment 1 there were three trial types: delayed recall (a.k.a. “delayed estimation”) of an oriented bar (1O), of one from a memory set of 3 oriented bars (3O), or of one item from a memory set of 1 oriented bar, 1 color patch, and 1 luminance patch (1O1C1L). Oriented-bar stimuli (length, 4°; width, 0.08°) were rendered as the black diameter of a white circular patch. Sample stimuli could appear in one of nine possible orientations ranging from 0–160°, in 20° increments, with a jitter of ± 1 –5° determined randomly on each trial. Color stimuli were presented on 4°-diameter circles, and drawn from a pool of 9 colors that were equidistant along a circle in CIE L*a*b* color space ($L = 70$, $a = 20$, $b = 38$, radius of 60; sample items were therefore equiluminant, varying mainly in hue and slightly in saturation), with a randomized jitter of ± 1 –5° on each trial. Luminance stimuli were comprised of a gray annulus (diameter = 2.67°) inside a white ring (RGB values [0, 0, 0]; diameter = 4°). The annulus could take on one of 9 grayscale values ranging equidistantly from light gray ([0.03, 0.03, 0.03] to darkest gray ([0.97, 0.97, 0.97])), with jitter of ± 1 –5° determined randomly on each trial.

On all trials, masks were rendered as a white circular patch bisected by 18 black 0.08° x 4° bars, all intersecting at their midpoints and each separated in orientation by 10°.

Recall displays comprised a circular stimulus patch – initially “empty” -- and a response wheel centered on fixation, with a radius to its outer edge of 9.2° and a width of 2°. Varying continuously around the response wheel were all possible values of the category being tested. For orientation, this was rendered as 20 equally spaced black bars (0.05° x 1.8°), ranging in orientation from 0–171°, in 9° increments. For color and luminance, all 180 values of that dimension were evenly distributed along the circle. The angle of rotation of the response wheel varied unpredictably from trial-to-trial, to discourage response planning during the delay. At the onset of the recall display a cursor (a conventional “mouse” arrow) was always positioned at central fixation, and the stimulus patch was rendered with a randomly determined value rendered in the format of the sample stimuli. As soon as the subject began to move the trackball of the response box (see “Behavioral tasks”) the cursor moved correspondingly, and the stimulus patch took on the value corresponding to the location on the response wheel that was nearest to the cursor. Throughout the experiment, the background screen color was gray [0.5, 0.5, 0.5] (Figure 1).

Experiment 2 included only 1O and 3O trial types (see subsection on *Behavioral task during fMRI scanning* for rationale), and the stimuli differed from Experiment 1 in that sample stimuli appeared in one of six possible locations, with orientations ranging from 0–150°, in 30° increments, with a jitter of ± 1 –5° determined randomly on each trial.

Behavioral task

In Experiment 1, each trial of the 1O task began with the 4-sec presentation of a sample item equiprobably and unpredictably at one of four possible locations, each in one quadrant of the screen, and each with horizontal and vertical eccentricities from fixation of 5°. The identity of the sample, drawn equiprobably and unpredictably from a pool of 9 orientations, varied independently of location. The 8-s delay period began with a mask presented at the same location as the sample. Responses were made by moving a cursor with a trackball and “clicking” on the recalled orientation with a button press. As soon as the trackball began to move, a bar appeared within the circular patch with an orientation, updating in real-time, that matched the orientation on the wheel that was closest to the cursor. RT was computed as the latency between response-wheel movement onset and button press. Feedback, indicating the error between the recalled orientation and the sample orientation (in degrees) was presented centrally, replacing the fixation cross, appearing immediately after the response until the end of the 4s response window. ITI was 2 s for the behavior-only experiment, and 8 s for the fMRI experiment. A black fixation cross was present at the center of the screen throughout each block of trials, and subjects were instructed to fixate it throughout the block.

3O trials followed the same procedure as 1O, with the following adjustments necessitated by the greater number of items: on each trial, each of the three sample stimuli was drawn randomly from the pool of 9, without replacement; the three were displayed simultaneously, with each of the four possible configurations of location occurring equiprobably and unpredictably; each of the three stimulus locations was backward masked; and the location of the probed item occurred equiprobably and unpredictably. 1O1C1L followed the same procedure as for 3O, except that each trial featured one sample item drawn from each of the three stimulus categories, and the category of the probed item occurred equiprobably and unpredictably (Figure 1).

In Experiment 2, the task procedure was the same with that in Experiment 1, with two exceptions: only 1O and 3O trials were included, and stimuli could appear at six evenly spaced locations along the imaginary circumference of a circle 5° in diameter, centered on fixation.

Behavior-only experiment from Experiment 1

Testing was broken into two blocks of 1O and 3O trials and three blocks of 1O1C1L trials. All blocks contained 50 trials, and block order was counterbalanced across subjects by drawing the first 18 orders from a Latin square. Each block of 1O and 3O trials presented 25 of each, in a randomized sequence (thereby yielding a total of 50 1O responses and 50 3O responses per subject), and each block of 1O1C1L trials included 17 probes of two of the categories and 16 of the remaining category, randomized within block and balanced across the three blocks, thereby yielding 50 1O1C1L orientation responses, 50 1O1C1L color responses, and 50 1O1C1L luminance responses. All the experimental stimuli were controlled by the Psychophysics Toolbox (<http://psychtoolbox.org>; Brainard 1997) running in MATLAB (MathWorks), presented on a 60 Hz projector with a screen width of 32.5cm (iMac). The viewing distance was 62 cm.

Behavioral screening for Experiment 2

Procedures for the orientation WM task be identical to those from Experiment 1, with the exception that subjects only performed two 50-trial blocks of the 3O task.

Behavioral task during fMRI scanning

For Experiment 1, there were two scanning sessions, and during the first session subjects first performed 4 blocks of 3-item trials: 9 trials of 3O and 9 trials 1O1C1L in a randomly determined order during each block; 3 probes of each category on 1O1C1L trials in a randomly determined order during each block. Next, subjects completed 8 18-trial blocks of 1O trials, with each orientation appearing twice in a randomly determined order during each block. In the second fMRI scanning session, subjects performed an additional 12 18-trial blocks of 1O trials. (The larger number of 1O trials was needed to train the IEM models on which data from all trial types would be tested.)

For Experiment 2, subjects performed only 1O and 3O trials. The rationale was that the most novel effects observed in Experiment 1 were the differences in IEM reconstruction of orientation between the *low* vs. *high swap-error* groups. These differences were of greatest theoretical significance on 3O trials, which place the greatest demands on context binding. Furthermore, the results from Experiment 1 involving comparison of patterns of BOLD signal intensity in the 1O vs. 1O1C1L vs. 3O conditions, and their relation to behavior (Figure 2), were robust and adequately powered at $n = 16$, and were, themselves, replications of a previous result (Gosseries, Yu, et al., 2018). Therefore, for the Experiment 2, we maximized the sensitivity of within-subject IEM reconstructions, and of between-group comparisons, by administering 108 3O trials (an increase from the 36 that featured in Experiment 1).

For the fMRI portion of *Experiment 2*, each subject completed two 2h scanning sessions, each on a separate day. The first scanning session began with six 18-trial blocks of 3O trials, followed by seven blocks of 1O trials. The second scanning session consisted of thirteen 18-trial blocks of 1O trials. Each block of trials was performed during a 7.2-min scan. Thus, each scanning session entailed 1-h 33-min 36-sec of fMRI scanning, not including the brief pauses between scans. Over the course of all 108 3O trials, each orientation was probed at each location three times. Over the course of all 360 1O trials, each orientation was presented (and probed) at each location a total of ten times at each location.

Analysis of behavioral data

For Experiment 1, reaction time (RT) of the response-ending button press was collected and raw response error distance estimated by the distance on the response wheel between the subjects' selection and the true target value (in degrees). Trials without responses were excluded. For 1O and 1O1C1L trials, response error was fit to a two-factor mixture model that estimated the proportion of responses made to the sample (i.e., the probability of a target response (p_T), and the probability of guess responses (p_U), as well as the precision of target responses (κ) (Bays et al. 2009). For 3O trials a third factor, the probability of a response to a nontarget (p_N ; a.k.a. "swap error"), was included in the model. Parameter estimates were obtained using maximum-likelihood estimation (expectation maximization) using MATLAB routines available at <http://www.bayslab.com>. Differences across trial type in descriptive measures and in model parameters were assessed with repeated measures one-way ANOVAs, and significant effects were followed up with paired t -tests. For descriptive measures, we only focused on response error, because the RT was necessarily noisy because it included the time to adjust the response dial with a trackball positioned adjacent to the thigh of the supine subject. For model estimates, we focused on κ and p_T , because p_U was highly collinear with p_T in 1O and 1O1C1L trials (i.e., for the two-factor mixture model, $p_U + p_T = 1$). Note that p_N could only be estimated in 3O trials.

The p_N parameter estimated from 3O trials provides a measure of the efficacy of context binding, because a swap error corresponds to a trial on which the subject has forgotten the location context of the item that they are recalling. For Experiment 1, post hoc inspection of performance from the behavior-only experiment revealed a clear bimodal distribution in estimates of p_N , with 6 subjects having a p_N at or near 0 (indicating effectively no swap errors) and the remaining 10 subjects all having a p_N of 0.127 or higher. Based on this pattern in the behavior-only experiment, subjects were grouped into *low swap-error* and *high swap-error* groups for the analyses of the fMRI experiment.

For the behavioral task in the fMRI component of Experiment 1, RT and response errors were collected in the same way, but model fitting was only conducted for 1O trials due to the limited number of 3O and 1O1C1L trials. Correlations were carried out on all measures of behavioral performance on 1O trials from the two experiments, with the exception of RT, to assess the stability of subjects' performance.

For Experiment 2, procedures for analyses and model fitting of data from 1O and 3O trials were the same as those from Experiment 1.

fMRI methods

Unless specified otherwise, the fMRI methods for Experiment 2 followed those from Experiment 1, as described in this section.

General procedure and behavioral tasks

For Experiment 1, the fMRI experiment comprised two scanning sessions, each lasting about 1.5hr. The first of the two scanning sessions followed the behavior-only task by 6-21 days, and the

second scanning session followed the first by 2-28 days. Scanning of 3-item trials preceded scanning of 10 trials to minimize the likelihood that subjects would process orientation stimuli different from color and luminance stimuli on 101C1L trials. All the experimental stimuli were controlled by the Psychophysics Toolbox (<http://psychtoolbox.org>; Brainard 1997) running in MATLAB (MathWorks), presented via a 60 Hz projector (Silent Vision 6011; Avotec) backprojecting onto a screen mounted inside the bore of the scanner, and viewed through a coil-mounted mirror. The viewing distance was 68.58 cm and screen width was 33.02 cm.

Data acquisition

Whole-brain images were acquired with a 3 Tesla scanner (Discovery MR750; GE Healthcare) at the [removed for blinding]. For all subjects, a high-resolution T1-weighted image was acquired with a fast spoiled gradient-recalled-echo sequence (TR = 8.2 ms, TE = 3.2 ms, Flip angle = 12°, 160 axial slices, 256 x 256 in-plane, 1 mm isotropic). A T2*-weighted gradient echo pulse sequence was used to acquire data sensitive to the BOLD signal while subjects performed the VSTM task (TR=2000 ms, TE=25 ms, Flip angle = 60°, within a 64 x 64 matrix, 42 sagittal slices, 3 mm isotropic). Each of the twenty fMRI scanning runs generated 213 volumes (excluding disaqs).

Preprocessing

fMRI data were preprocessed using the Analysis of Functional Neuroimages (AFNI) software package (<http://afni.nimh.nih.gov>; Cox 1996). All volumes were spatially aligned to the first volume of the first run using rigid-body realignment, then aligned to the T1 volume. Volumes were corrected for slice-time acquisition, and linear, quadratic, and cubic trends were removed from each run to reduce the influence of scanner drift. For univariate analyses, data were spatially smoothed with a 4 mm FWHM Gaussian, and z-scored separately within run for each voxel. For IEM and MVPA analyses (see below), data were z-scored separately within run for each voxel, but were not smoothed. All analyses were carried out in each subject's native space.

Data analysis

Univariate analyses and ROI creation. For Experiment 1, a modified general linear model (GLM) was fit to data from all 30 and 101C1L trials and from 36 randomly selected 10 trials. It included regressors modeling the sample presentation, delay, and recall periods with boxcars of 4 s, 8 s, and 4 s, respectively, each convolved with the canonical hemodynamic response function supplied with AFNI.

A different set of anatomically constrained, functionally defined regions of interest (ROI) was created for each of the three categories of analysis: BOLD signal intensity, IEM of stimulus orientation, and MVPA of stimulus location. For all three, anatomical regions were generated from the standard anatomical masks for occipital, parietal, and frontal cortex from the MNI152_T1_1mm template and warping them to each subject's native space. For BOLD signal intensity analyses, an *Occipital Sample* ROI was generated for each subject by selecting the 400 voxels with the highest *t*-values for the contrast [Sample₃₀ – Sample₁₀] within anatomically defined occipital cortex, and *Parietal Delay* and *Frontal Delay* ROIs were generated by selecting the 400 voxels with the highest *t*-values for the contrast [Delay₃₀ – Delay₁₀] within each of these anatomically defined regions. Within each of these ROIs, trial-averaged time series for each of the three trial types were generated and converted to mean percentage signal change from baseline (first TR of the trial), and contrasts versus baseline and between conditions carried out with *t* tests (all *p* values FDR-corrected across TRs, ROIs and comparisons). For IEM analyses, four sets of “sample location-specific” ROIs were generated with the top 400 voxels within each of the three anatomical regions responding to the contrasts [Sample_{upper left} – baseline], [Sample_{upper right} – baseline], [Sample_{lower left} – baseline], and [Sample_{lower right} – baseline]. Finally, for MVPA analyses, “*location-general*” ROIs were created with the top 400 voxels within each of the three anatomical regions identified with the

contrast [(Sample_{upper left} + Sample_{upper right} + Sample_{lower left} + Sample_{lower right}) – baseline]. (Analyses of fMRI data from 1O trials, unrelated to those described here, are presented in (Cai et al. 2019).)

For Experiment 2, ROI creation differed only in that it used data from all 3O trials and from 108 randomly selected 1O trials.

Task-related patterns of covariation. For Experiment 1, we used ANCOVA to evaluate evidence for correlated sensitivity across trial types (specifically, across 1O vs. 3O and across 1O1C1L vs. 3O) of two of dependent variables – BOLD signal intensity and behavioral precision – seen to covary in previous studies (Emrich et al. 2013, Grosser et al. 2018, Yu et al. 2018). Unlike simple correlations, ANCOVA accommodates the fact that each subject contributes a value for each level of the factor of trial type. It removes between-subject differences and assesses evidence for “within-subject correlation” – the extent to which variation in one dependent variable can be explained by variation in a second (Bland and Altman 1995). For analyses including delay-period fMRI activity, the BOLD signal was averaged across TRs 6 and 7, those least likely to be contaminated by Sample-related signal.

For the Experiment 2, comparable ANCOVAs, carried out with data from the 3O and 1O trials, were considered *secondary hypotheses*.

Multivariate inverted encoding modeling. In Experiment 1, we estimated population-level neural representations of the orientation of oriented-bar stimuli with multivariate inverted encoding modeling (IEM; Serences and Saproo 2012, Sprague et al. 2018). To optimize estimation from 1O trials, four IEMs were trained within each of the three anatomical ROIs, one for each location at which a sample could appear on the screen. The four resultant location-specific reconstructions were averaged prior to assessment of the results of IEM training and testing.

To build our IEMs we assumed that the responses of each voxel can be characterized by activity in 9 hypothesized tuning channels, one corresponding to each of the 9 possible sample orientations. Following previous work (Ester et al. 2015, Yu and Shim 2017), the idealized feature tuning curve of each channel was defined as a half-wave-rectified and squared sinusoid raised to the seventh power. Before feeding the preprocessed data into the IEM, a baseline from each voxel’s response was removed in each run using the following equation from (Brouwer and Heeger 2011):

$$B = B - m(m^T B)$$

in which B represented the data matrix from each run with size $v \times c$ (v : the 400 location-specific voxels; c : the 9 orientations) and m represented the mean response across all stimulus conditions of length v . Next, for trials corresponding to each of the four sample locations, we randomly divided the data into a training set (81 trials) and a test set (9 trials; these numbers were selected to match the total number of 1O1C1L and 3O trials). We computed the weight matrix (W) that projects the hypothesized channel responses (C_1) to actual measured fMRI signals in the training dataset (B_1), and extracted the estimated channel responses (\hat{C}_2) for the test dataset (B_2) using this weight matrix. The relationship between the training dataset (B_1 , $v \times n$, n : the number of repeated measurements) and the channel responses (C_1 , $k \times n$) was characterized by:

$$B_1 = W C_1$$

Where W was the weight matrix ($v \times k$).

Next, the least-squared estimate of the weight matrix (\hat{W}) was calculated using linear regression:

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

The channel responses (\hat{C}_2) for the test dataset (B_2) were then estimated using the weight matrix (\hat{W}):

$$\hat{C}_2 = (\hat{W}^T \hat{W})^{-1} \hat{W}^T B_2$$

Having thus estimated the weight matrix mapping each voxel's response to each orientation channel from the training dataset, we inverted this matrix to estimate channel responses on each test trial. The average response output for each channel across trials was obtained by circularly shifting each response to a common center of 0°. To generate smooth, 180-point channel tuning functions (CTF, also referred to as “reconstructions”) we repeated the encoding model analysis 180 times and shifted the centers of the orientation channels by 1° on each iteration (Brouwer and Heeger 2009). The CTFs were averaged across permutations and averaged across the four location-specific models. The same weight matrix trained in this fashion on data from 1O trials was used to reconstruct the neural representation of orientation in 1O, 3O, and 1O1C1L trials. More specifically, within each ROI, separate IEMs were trained on 1O data from each time point in the trial and then tested (i.e., reconstructions attempted) at same time point with data from 1O, 3O, and 1O1C1L trials separately. For 3O trials, the location-specific IEM to be used for IEM testing was assigned according to the orientation to be tested for recall. For 1O1C1L trials, testing was carried out with the location-specific IEM congruent with the location occupied by oriented-bar in the sample array.

To quantify the results, the CTF in each ROI for each subject was fit with an exponentiated cosine function of the form:

$$f(x) = \alpha (e^{\kappa[\cos(\mu - x) - 1]}) + \beta.$$

Here, α and β control the vertical scaling (i.e., signal over baseline) and baseline of the function, respectively, and κ and μ control the concentration (the inverse of dispersion) and center of the function, respectively. No biases in reconstruction centers were expected or observed, so we fixed μ at 0. Fitting was performed by combining a GLM with a grid search procedure. We first defined a range of plausible κ values (from 1 to 30 in 0.1 increments). For each possible value of κ , we generated a response function using the fitting equation after setting α to 1 and β to 0. Next, we generated a design matrix containing the predicted response function and a constant term (i.e., a vector of 1s) and used ordinary least-squares regression to obtain estimates of α and β (defined by the regression coefficient for the response function and constant term, respectively). We then selected the combination of κ , α and β that minimized the sum of squared errors between the observed and predicted reconstructions.

In Experiment 2, the same IEM procedure was used to estimate neural representations of stimulus orientation and stimulus location, except that orientation models characterized each voxel in terms of the activity in 6 hypothesized orientation tuning channels (one corresponding to each possible orientation). And six separated IEM were trained (one for each location) and these location-specific reconstructions were averaged prior to assessment of the results of IEM training and testing.

Implementation of a priori hypothesis tests

Hypothesis 1 and *Hypothesis 2* of Experiment 2 corresponded to analyses from Experiment 1 carried out in the occipital ROI and focusing on TR 10, one of the TRs at which CTFs in multi-item trials were reliable at the group level ($N = 16$). To carry out tests of *Hypotheses 1* and *2* for Experiment 2, we first needed to ensure sufficient statistical power to achieve the *secondary hypothesis* that, for the *low swap-error* group, the amplitude of the reconstruction of stimulus orientation at TR 10, on 3O trials, would be statistically greater than baseline. To do this, we followed our procedure from Experiment 1 by assessing the significance of CTF amplitude using a bootstrapping procedure in which, for each ROI, we randomly select, with replacement, the number CTFs corresponding to the number of subjects in the group, and averaged them. This step was repeated 2500 times, yielding 2500 unique stimulus reconstructions. We then estimated the amplitude of each reconstruction and a p value was computed as the proportion of permutations

for which amplitude estimates ≤ 0 were obtained (Ester et al. 2015, Ester et al. 2016). For this and all subsequent bootstrapping analyses, all the p values were one-tailed and FDR-corrected across ROIs, TRs, and tasks. Whereas in Experiment 1 Cohen's d for the reconstruction of stimulus orientation at TR 10 from 3O trials was .729, simulations indicated that we could expect effect sizes for reconstruction of stimuli from 3O trials for Experiment 2 to double relative to Experiment 1, due to the increased number of 3O trials. Therefore, to calculate the number of subjects needed in the *low swap-error* group to achieve 90% power for this *secondary hypothesis*, we used an estimate of Cohen's d of $[\text{.729} * 2 = 1.458]$. To achieve 90% power to detect this effect with $\alpha = .025$, we needed an N of 11. (Note that, because none of the primary hypotheses tested in Experiment 2 required significant reconstruction of orientation in the occipital ROI at TR 10 in the *high swap-error* group, we did not carry out power analyses for this effect.)

Additional *secondary hypotheses* were whether, at any TR of the task, orientation reconstructions differed between the *low swap-error* and *high swap-error* groups. These were assessed by first carrying out the bootstrapping procedure separately for each group, then comparing the CTF estimates from each permutation with each group by subtracting the amplitude from the *high swap-error* group from the corresponding value from the *low swap-error* group. A one-tailed was conducted with p referring to the proportion of the 2500 subtractions with a value ≤ 0 . Finally, we planned to assess the test-retest reliability of pN as estimated from the behavioral screening session vs. pN as estimated from the fMRI session, by Pearson correlation.

Hypothesis 1 of Experiment 2 formalized the intuition that one would expect that individual differences in pN , a behavioral measure, would be reflected in neural evidence for inappropriate activation of the identity of non-probed items. To assess this prediction, as we did in Experiment 1, we operationalized "orientation recall specificity" as the difference (on 3O trials) between probe-epoch CTFs of the orientation of the probed versus of a non-probed item, then compared this measure between swap-error groups. To compute orientation recall specificity, for each trial one of the two non-probed sample items was selected at random and the amplitude of the reconstruction of the orientation of that non-probed item was subtracted from the amplitude of the reconstruction of the orientation of the probed item. Significance of all comparisons was assessed with bootstrapping. Using the results from Experiment 1 as a starting point, and taking into account that fact that the modified design of Experiment 2 increased effect sizes for reconstruction of stimuli from 3O trials, we determined that we required an N of 10 subjects to achieve 90% power to detect an orientation recall-specificity effect with $\alpha = .025$ (two-tailed t test).

Hypothesis 2 of Experiment 2 was analogous to *Hypothesis 1*, with the difference being that it operationalized the specificity of recall of the location of the probed item. Thus, we operationalized "location recall specificity" as the difference between probe-epoch CTFs of the location of the probed versus of a non-probed item, then compared this measure between swap-error groups. To compute location recall specificity, for each trial one of the two non-probed sample items was selected at random and the amplitude of the reconstruction of the location of this non-probed item was subtracted from the amplitude of the reconstruction of the location of the probed item. Significance of all comparisons were assessed with bootstrapping. Because Experiment 1 did not use IEM to assess the neural representations of stimulus location, we used a different dataset from our lab to estimate power, and these data suggested that data from only three subjects were needed to achieve 90% power to detect a reliable location recall-specificity effect with $\alpha = .025$ (two-tailed t test).

Results from Experiment 1 suggested that delay-period representation of sample locations in parietal cortex was stronger in the *low swap-error* group. Specifically, MVPA decoding of stimulus location was superior for the *low swap-error* group at TRs 6 and 7. Follow-up correlational analyses indicated that individual differences in MVPA of sample location information in parietal delay-period activity correlated with pN and with the location recall specificity effect. *Hypothesis 3*

was therefore intended to assess the prediction that individual differences in the strength of delay-period representation of the sample location in parietal cortex, as operationalized by the amplitude of the CTF of the location of the to-be-probed item at TRs 6 and 7, would predict individual differences in pN . We planned to carry this out with Spearman correlation of rank ordering of these two variables, because subjects in the *low swap-error* group would necessarily have pN values close to 0. The coefficient of determination (R^2) for this effect in Experiment 1, with an N of 16, was .282. Because *Hypothesis 4* would also use the same measure of delay-period representation of the sample location in parietal cortex, we set $\alpha = .025$ and calculated that, to achieve 90% power to find this effect, we would need to enroll an N of 29.

Hypothesis 4 assessed the prediction that individual differences in the strength of delay-period representation of sample locations in parietal cortex, as operationalized by the amplitude of the CTF of the location of the to-be-probed item, would predict individual differences in location recall specificity. We planned to carry this out with a Pearson correlation, because we expected, based on Experiment 1, that both variables would be normally distributed. (However, if either distribution turned out to be skewed (as estimated via Shapiro-Wilk analysis) we would instead use Spearman correlation.) Based on the fact that the R^2 for this effect in Experiment 1, with an N of 16, was .303, we estimated that we would need $N = 25$ to achieve 90% power to detect this effect at $\alpha = .025$.

Results

Experiment 1

The only results from Experiment 1 that are reported are those that were not used to generate the preregistered hypotheses, which were tested in Experiment 2.

Behavioral Session. Here, we report results from only the 16 subjects who also participated in the fMRI experiment. An average of 2.286 trials ($SD = 2.920$); 4.857 trials ($SD = 2.824$), and 1.857 trials ($SD = 2.824$) per subject were excluded in the 1O, 3O, and 1O1C1L condition, respectively. Descriptive statistics suggested that task difficulty increased from 1O to 1O1C1L to 3O, as reflected in the mean response error ($F(2,30) = 41.830$, $p < 0.0001$; paired t -test, $t_s > 3.960$, $p_s < 0.002$; Table 1). Results from mixture modeling mirrored this pattern, with pT highest for 1O, followed by 1O1C1L and 3O ($F(2,30) = 16.791$, $p < 0.0001$), paired t -test $t_s > 2.644$, $p_s < 0.018$; Table 1). Recall precision (κ) also differed across trials types ($F(2,30) = 16.458$, $p < 0.001$), being significantly different between 1O and 3O ($t(15) = 5.253$, $p < 0.0001$) and between 3O and 1O1C1L ($t(15) = 4.325$, $p < 0.001$) trials, although not differing between 1O and 1O1C1L trials ($t(15) = 1.387$, $p = 0.186$). Finally, although the group mean pN (swap errors) on 3O trials was 0.120 ($SD = 0.116$), 6 subjects had a pN at or near 0 (all $pNs < 0.006$), indicating that these subjects made effectively no swap errors, whereas the remaining 10 subjects all had a pN of 0.127 or higher, corresponding to an average of nearly 20% swap errors on 3O trials for these ten subjects.

Table 1: Behavioral results from Experiment 1 ($n = 16$ for both sessions)

Trial type	Descriptive Data		Three-factor mixture model (parameter estimates)			
	RT(s)	Response error (degree)	pT	pN	pU	$\kappa(\text{rad}^{-1})$
Behavior-only session						
1O	2.825(0.247)	7.395(2.360)	0.987(0.019)	N/A	0.013(0.019)	12.899(7.706)
3O	2.941(0.231)	16.214(5.889)	0.828(0.120)	0.120(0.116)	0.052(0.064)	4.166(2.230)
1O1C1L	2.687(0.336)	10.764(5.076)	0.895(0.114)	N/A	0.105(0.114)	10.718(7.004)
fMRI session						
1O	2.912(0.149)	7.765(4.219)	0.946(0.001)	N/A	0.054(0.001)	13.748(6.881)

3O	3.029(0.223)	16.510(7.308)	--	--	--	--
1O1C1L	3.009(0.248)	12.284(4.028)	--	--	--	--

fMRI sessions

Behavior.

An average of 4.62 (SD = 2.07) 1O trials was excluded, no trials for any subject for either of the other two trial types were excluded. Performance on 1O trials was highly correlated with pre-scan testing across the 16 subjects who participated in both ($r_s > 0.841$, $p_s < 0.001$; Table 1).

fMRI.

Occipital ROI. Trial-averaged signal on all three trial types showed the expected sample-related increase, return to baseline by the end of the delay period, and probe/recall-related increase. BOLD signal intensity did not differ between 3O and 1O1C1L trials during TRs corresponding to the encoding and delay epochs (from TR2-TR8, $t_s < 1.604$, $p_s > 0.130$), but was higher on 3O than 1O1C1L trials during the peak response to probe/recall (TR9-11, $t_s > 2.524$, $p_s < 0.048$). BOLD signal from all three trial types did not differ from each other or from baseline at TR7 corresponding to late delay ($t_s < 0.806$, $p_s > 0.433$; Figure 2A). The ANCOVA relating delay-period BOLD signal intensity to behavioral precision revealed no significant within-subject correlations between either 1O and 3O trials or 1O and 1O1C1L trials (Figure 2B).

Parietal and Frontal ROIs. In both regions, trial-averaged signal on all three trial types remained elevated across the duration of the trial, with sample- and delay-related activity greater for 3O trials than for the other two trial types (from TR3 to TR8, $t_s > 3.041$, $p_s < 0.019$), and sample-related activity for 1O1C1L trials greater than for 1O trials during the encoding epoch (TRs 3-4, $t_s > 2.514$, $p_s < 0.048$). Beginning with TR6 in the delay period, however, BOLD signal no longer differed between 1O1C1L and 1O trials ($t_s < 1.848$, $p_s > 0.138$). That is, delay-period activity in parietal and frontal cortex was not sensitive to memory load (one item vs. three items) per se, but was sensitive to stimulus category homogeneity (operationalized as 1O1C1L vs. 3O; Figure 2A). ANCOVAs relating delay-period BOLD signal intensity to behavioral precision revealed significant within-subject correlations between 1O and 3O trials in both regions ($r_s > 0.539$; $p_s < .05$), and a nonsignificant trend in this direction between 1O1C1L and 3O trials in parietal cortex ($r = 0.458$; $p = .06$. Figure 2B).

Experiment 2

Behavioral.

Behavioral Session. 39 of 72 subjects who met a behavioral screening criterion of performing with a “low” swap error rate ($< .05$) or “high” swap error rate ($> .12$) were invited to participate two subsequent fMRI sessions. From these, 29 subjects completed the two fMRI scanning sessions (*low swap-error* group, mean(SD): $pT = 0.923(0.085)$, $pN = 0.015(0.015)$, $pU = 0.062(0.080)$, $\kappa = 3.844(2.402)$; *high swap-error* group, $pT = 0.674(0.133)$, $pN = 0.254(0.085)$, $pU = 0.071(0.075)$, $\kappa = 3.294(1.815)$).

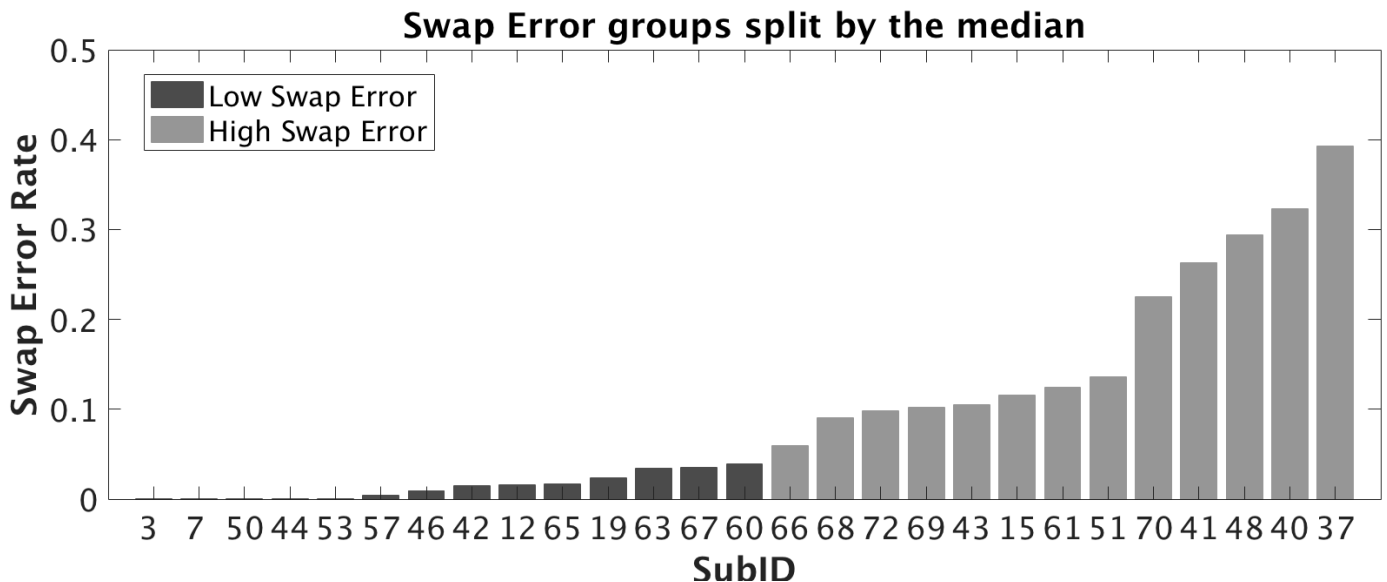


Figure 3. Swap error rates for each subject in the fMRI session of Experiment 2. Low swap-error subjects (dark gray) and high swap-error subjects (light gray) were defined via median split.

fMRI Session. Two subjects were excluded from the fMRI analyses -- one because responses were not registered on more than 40% trials, and one because mixture modeling indicated that pU was greater than .75 (outside 3 SD from the mean value) – and so the data from 27 are reported here. The behavioral performance of many of these 27 subjects, specifically their pN , differed considerably during the scanning sessions in comparison to the behavioral screening session, with 8 subjects initially classified as “low swap error” having pN values $> .05$ and one subject initially classified as “high swap error” having a pN value $< .12$. A consequence of this that the same sample of subjects who had been selected to form a bimodal distribution of “extreme” values of pN yielded a continuous distribution of values from the fMRI session, with 15 subjects with pN values $< .05$, 6 with pN values $> .05$ and $< .12$, and 8 with pN values $> .12$ (Figure 2). As a consequence, with too few *high swap-error* subjects, we were unable to test our hypotheses with the preregistered analyses. We reasoned, however, that the original intent of the preregistered hypothesis tests would be met if we deviated from the preregistered criteria for classifying subjects and, instead, carried out the planned analyses on two groups determined by a median split of pN values. From the perspective of scientific publishing, this deviation from the preregistered plan would qualify these analyses as “exploratory” were this a formal Registered Report. From the perspective of our experimental design, it might be expected a priori to decrease the power of the design, because several subjects in the *high swap-error* group had with pN values lower than the lower bound of $pN > .12$. that had been used to calculate power.

Analyses of data from fMRI session grouped by median split. The median split assigned 14 subjects, with pNs ranging from 0 – 0.039, to the *low swap-error* group (mean(SD): $pT = 0.896(0.127)$, $pN = 0.012(0.013)$, $pU = 0.091(0.125)$, and $\kappa = 3.092(1.498)$), and 13 subjects, with pNs ranging from 0.059 – 0.393, to the *high swap-error* group (mean(SD): $pT = 0.700(0.170)$, $pN = 0.169(0.109)$, $pU = 0.131(0.177)$, and $\kappa = 3.652(1.881)$; Figure 3). Independent t -tests between the groups revealed no significant difference in pU or κ ($ts < 0.851$, $ps > 0.403$). Test-retest stability between the behavioral-screening and fMRI sessions was assessed with correlation analyses. Pearson correlations revealed significant test-retest correlations for pT and for κ ($R^2 > 0.477$, $ps < 0.001$). Because pN from the behavioral screening session was selected to be dichotomous, a Spearman correlation was carried out on this measure, and it failed to provide evidence for a correlation of performance across the two sessions ($R^2 = 0.054$, $p = 0.247$).

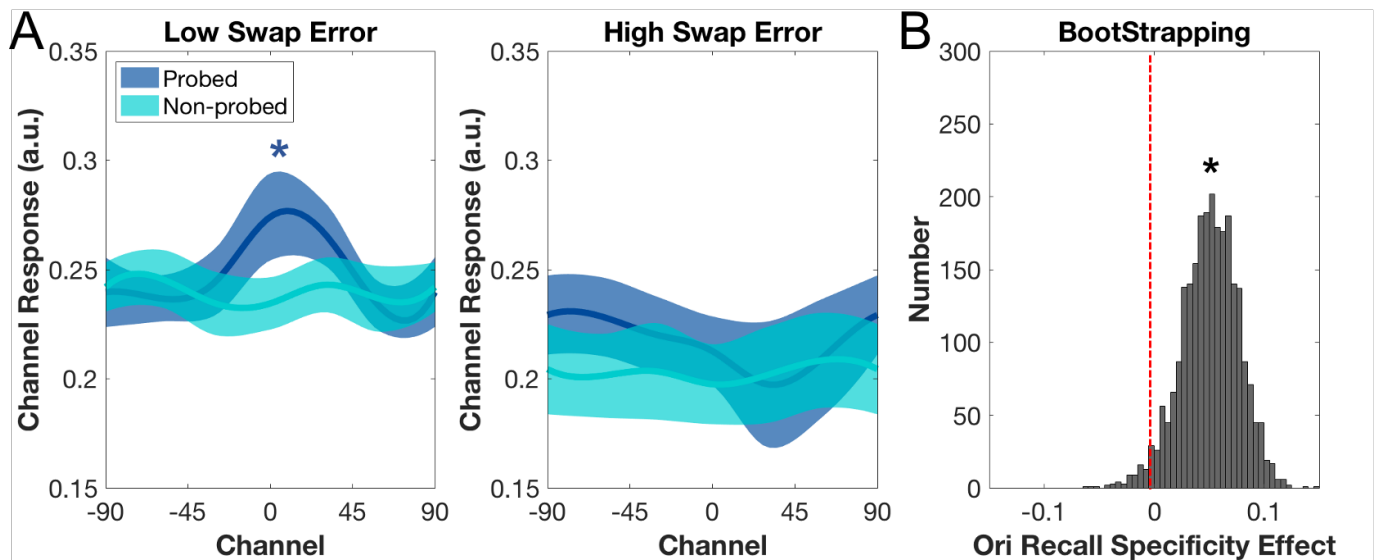


Figure 5. Orientation recall specificity effect in the occipital ROI, at TR10 on 30 trials (Experiment 2). A. Neural reconstructions for probed and non-probed orientations in the two groups. Asterisks indicate significant reconstructions at $p < 0.05$. B. Distributions of bootstrapping estimates of orientation recall-specificity effects. The recall specificity effects were identified by the recall specificity difference between low and high swap error group. Asterisk indicates significant difference between the two distributions ($p < 0.05$).

Hypothesis tests.

Hypothesis 1. In the occipital ROI, at TR10, for the *low swap-error* group, IEM reconstruction of stimulus orientation on 30 trials was significantly larger than baseline ($p = 0.047$; Figure 4.A), thereby meeting the prerequisite for being able to test Hypothesis 1. (Neither the comparable result for the *high swap-error* group, nor the reconstructions of the orientation of a non-probed item for either group, approached significance ($ps > 0.671$; Figure 4.A)). Comparison of orientation recall specificity between the two groups indicated that this measure was higher for the *low swap-error* group than for the *high swap-error* ($p = 0.049$; Figure 4.B), and thus supported Hypothesis 1.

Hypothesis 2. In the occipital ROI, at TR10, IEM reconstruction of the probed location on 30 trials was significantly larger than baseline for both the *low swap-error* group ($p < 0.001$) and the *high swap-error* group ($p = 0.021$; Figure 5.A). Comparable reconstructions for non-probed locations were nonsignificant for both groups (low swap error, $p = 0.059$ [trending toward negative]; high swap error $p = 0.770$, Figure 5.A). Comparison of location recall specificity between the two groups indicated that although this measure was numerically higher for the *low swap-error* group than for the *high swap-error* group (Figure 5.B), this difference approached, but did not achieve, the threshold for significance ($p = 0.067$).

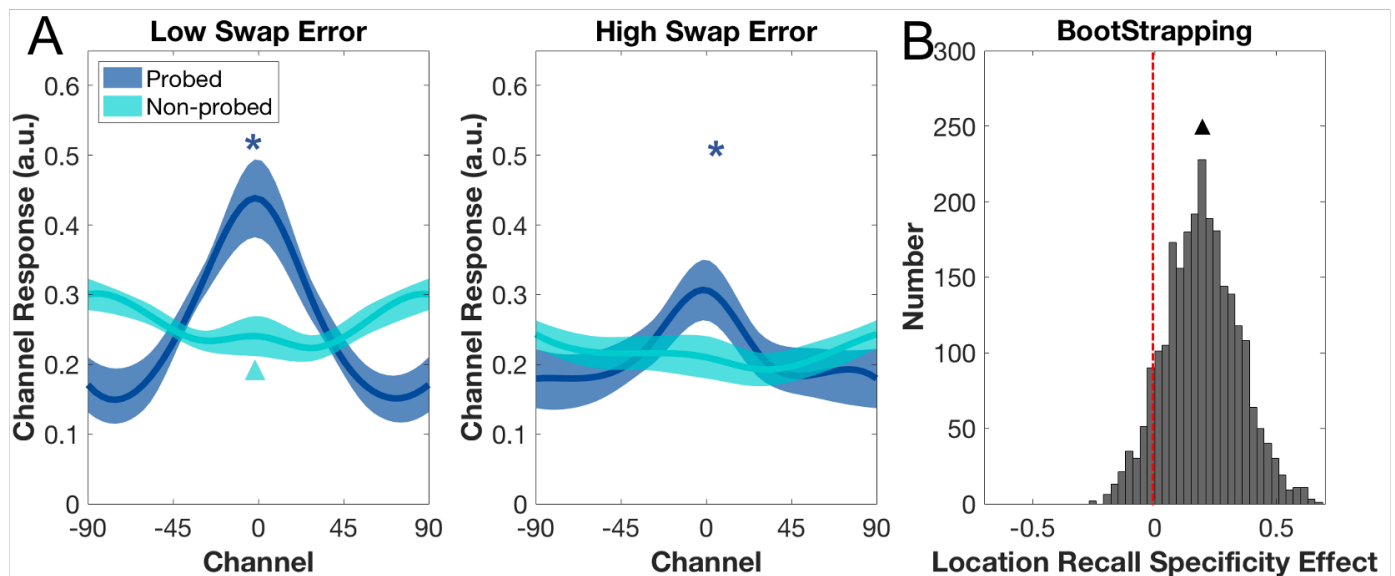


Figure 5. Location recall specificity effect in the occipital ROI, at TR 10 on 30 trials (Experiment 2). A. Neural reconstructions for probed and non-probed locations in the two groups. Asterisks indicate significant reconstructions ($p < 0.05$), and triangle indicates trend-level evidence for a significantly negative-going reconstruction ($.05 < p < 0.1$). B. Distributions of bootstrapping estimates of location recall-specificity effects in the two groups. Triangle indicates trend-level evidence for a difference between the two distributions ($.05 < p < 0.1$).

Hypothesis 3. IEM did not yield reliable reconstruction of the to-be-probed location in the parietal ROI during TRs 6 and 7 of the delay period for either group ($ps > 0.213$), and so a direct test of Hypothesis 3, as it was preregistered, could not be carried out. (Note that the exclusion of 2 subjects was not factor here, because the planned N of 29 was needed for the planned correlation, not for the reliable reconstruction of the to-be-probed location, which would have been expected to require considerably fewer subjects). When we nevertheless carried out the planned Spearman correlation with data from TRs 6 and 7, it was not significant ($p > .23$). However, because the reconstruction of stimulus location was significant for the *low swap-error* group at adjacent TR 8 ($p = 0.041$), we applied the procedures for testing Hypothesis 3 to this adjacent time point, as an exploratory analysis. Location representation-specificity was reliable at TR 8 for the *low swap-error* group ($p = 0.041$) but not for the *high swap-error* group ($p = 0.566$). Across all subjects, Spearman correlation indicated that higher values of location representation-specificity in IPS were associated with lower values of pN ($R^2 = 0.269$, $p = 0.01$).

Hypothesis 4. For the same reason as Hypothesis 3, a direct test of Hypothesis 4, as it was preregistered, could not be carried out. Thus, analogous to Hypothesis 3, we also applied the procedures for testing Hypothesis 4 to TR 8, as an exploratory analysis. Across all subjects, Spearman correlation indicated that higher values of location representation-specificity in IPS were associated with higher values of location recall specificity in occipital cortex, at TR 10 ($R^2 = 0.357$, $p = 0.001$).

Secondary Hypotheses.

BOLD signal intensity. In all three ROIs, for 10 and 30 trials, the time course of fMRI activity was comparable to what was observed in Experiment 1: In the occipital ROI, at TR 7, signal in the occipital ROI did not differ between 30 and 10 trials ($t_{(26)} = 1.617$, $p = 0.118$); whereas it was significantly higher for 30 than for 10 trials in the parietal and frontal ROIs ($ts > 3.572$, $ps < 0.001$). Furthermore, the ANCOVAs relating load-related changes in signal at TR 7 to load-related changes in behavioral recall precision was not significant in the occipital ROI, but the analogous ANCOVAs were significant in both the parietal ($R^2 = .48$) and frontal ($R^2 = .4$; $ps < 0.001$) ROIs.

The time course of IEM reconstructions of orientation during 3O trials. In the occipital ROI, the reconstruction of the to-be-probed orientation was successful during the sample-presentation and recall epochs (*low swap-error* group: TR4 and TR 10; *high swap-error* group: TR 4 and TR 8; *p* values for these IEM reconstructions not FDR-corrected), but not during the delay period. Orientation could not be reconstructed in the parietal and frontal ROIs (Figure 6.A). (Note that it has previously been reported that, on 1O trials, when collapsing across all subjects, orientation can be reconstructed during the entirety of the trial in each of these three regions (Cai et al. 2019).)

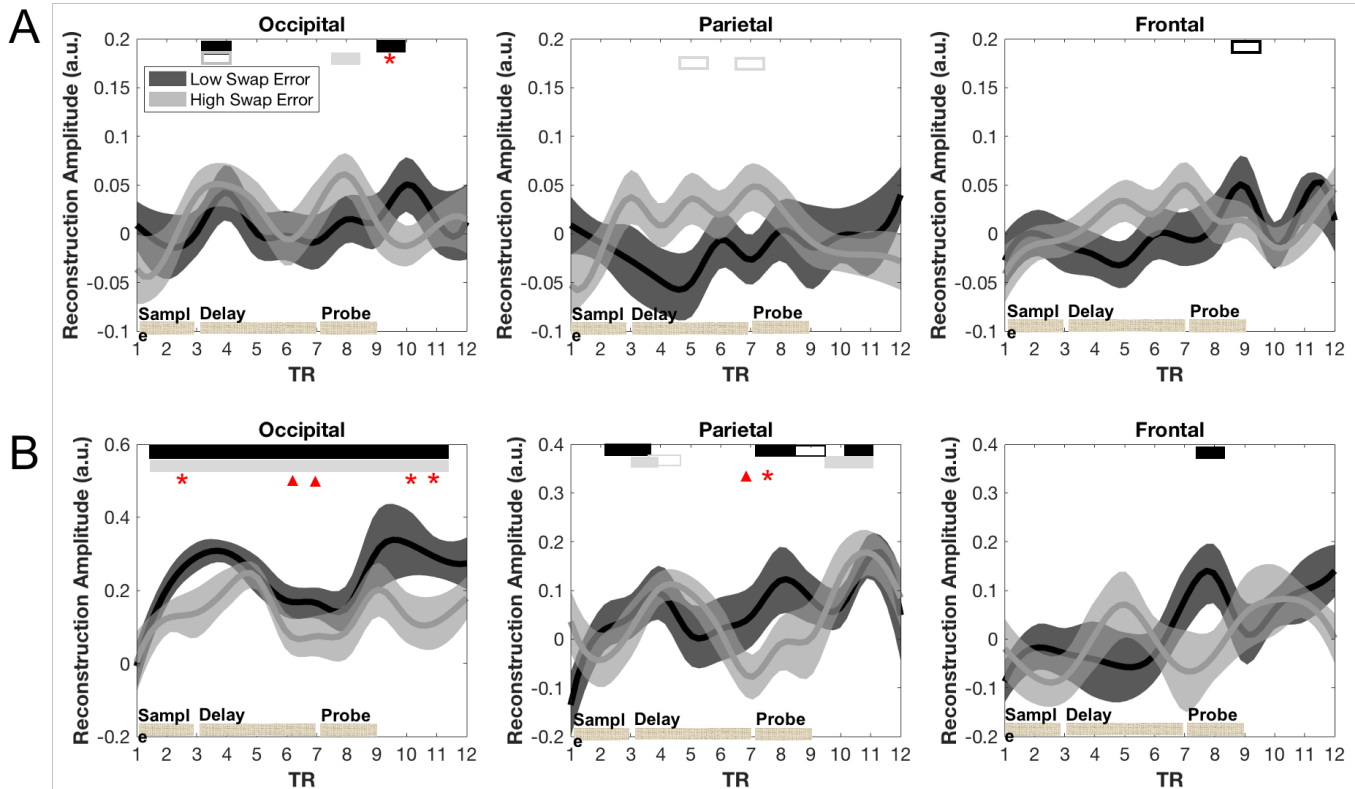


Figure 6. Time course of IEM reconstructions of unshifted data from 3O trials.

A. Reconstructions of the orientation of the probed stimulus broken out by swap-error group, in the three ROIs. B. Reconstruction of the location of the probed stimulus broken out by swap-error group, in the three ROIs. Bars along the top indicate values statistically different from baseline, with color corresponding to swap-error group. * indicates statistical difference between swap-error group at that TR (not FDR-corrected). Bars along the bottom indicate the times when stimuli are on the screen.

Turning to the reconstruction of the location of the probed item, in the occipital ROI location could be reconstructed across the entirety of the trial in both groups, with reconstruction amplitude significantly greater for the *low swap-error* group during portions of sample presentation (TR 3) and recall (TRs 10 and 11) epochs, and trending in this direction during the delay period (TRs 6 and 7; Figure 6.B). In the parietal ROI, location could be reconstructed during the sample presentation (TRs 3 and 4) and during recall (TRs 8-11) in both groups, with a trend toward greater amplitude in the *low swap-error* group at TRs 7 and 8 (Figure 7.B). In the frontal ROI location reconstruction was only successful at one time point during recall (TR 8) in the *low swap-error* group (Figure 6.B).

Discussion

Success on visual working memory tasks that are nominally nonspatial can often nonetheless require memory for the location at which each item was presented, and thus require context binding (c.f., Olson and Marshuetz 2005, Jiang et al. 2009). The results presented here are consistent with the proposition that delay-period activity of the IPS may correspond more closely to

the operation of context binding than to the representation of stimulus identity per se. Experiment 1 varied the demands on context binding by using two types of 3-item trials, 3O trials that could only be solved if the subject remembered which item had been presented at the location of the recall dial, and 1O1C1L trials in which memory for item-location association was not needed, because the nonspatial features of the recall dial uniquely indicated which item was to be recalled. The delay-period BOLD response in IPS (and in frontal cortex) was sensitive to demands on context binding (3O vs. 1O1C1L), but not to the number of items that needed to be remembered (1O vs. 1O1C1L; Fig. 2). This finding replicates and extends a previous finding (Gosseries, Yu, et al., 2018), and has at least two implications for our understanding of visual working memory. First, it suggests that the interpretation of results from many previous reports of load-sensitive activity in IPS (e.g., Todd and Marois 2004, Todd and Marois 2005, Xu and Chun 2006) may need to be revisited, due to the possibility that such studies have often confounded the factors of load and of context binding. Second, it provides an explicit, mechanistic account of the “benefit of memory set heterogeneity” (e.g., Olson and Jiang 2002, Delvenne and Bruyer 2004), and may provide support for models that explicitly incorporate context binding (e.g., Swan and Wyble 2014, Oberauer and Lin 2017, Bouchacourt and Buschman 2019) as well as a challenge for models that do not (including slot and resource models, c.f., Oberauer et al. 2016).

The results from Experiment 2 broadly replicated the intriguing preliminary findings from Experiment 1 that individual differences in the rate of swap errors were associated with neural measures of the strength of the representation of the (nonspatial) orientation and the location of stimuli. Low swap-error subjects showed higher orientation recall specificity in occipital cortex (Hypothesis 1), and a trend in the same direction for location recall specificity (Hypothesis 2). These results provide a demonstration, at the neural level, for a phenomenon predicted by a theoretical and computational model of visual working memory: stronger context binding should result in a lower level of competition between the probed item and nonprobed items at the time of retrieval (Oberauer and Lin 2017).

For IPS, exploratory results from Experiment 2 showed a pattern comparable to what was seen Experiment 1, but at a later time point: the strength of delay-period representation of stimulus location (i.e., of context) in IPS predicted lower swap-error rates (Hypothesis 3), and higher location recall specificity in occipital cortex (Hypothesis 4). These are consistent with the proposal from Gosseries, Yu et al. (2018) that delay-period activity in IPS may be particularly important for the maintenance of the bindings of context to content. We note that whereas the present results are consistent with the proposed functions of a spatial priority map (e.g., Bisley and Goldberg 2010, Jerde et al. 2012, Bisley and Mirpour 2019), the task-critical context in Gosseries, Yu et al. (2018) was ordinal position. Thus, an important question for future research is determining whether the processing of context in IPS is carried out in a domain-specific or a domain-general manner.

Interpretation of some of the results from Experiment 2 is complicated by a pattern in the results of this multisession experiment that was unanticipated: Subject’s swap-error rates (pN) were less stable between the behavioral screening session and the fMRI sessions than were other indices of performance, including precision (κ) and the probability of a target response (pT). As a result, several subjects selected for the *low swap-error* group based on performance at behavioral screening perform the task with a higher rate of swap errors during the fMRI sessions. The resultant decrease in sensitivity relative to what had been expected for the preregistered extreme-groups design might explain why the between-group difference in location recall specificity did not achieve statistical significance. There are at least two possible explanations for the greater volatility of estimates of pN than those of κ and of pT . The less interesting possibility is a technical one: Model estimates of pN may simply be less robust than those of other parameters, particularly at the relatively small number of trials that features in the fMRI sessions from Experiment 2. (Note, however, that we did try two other models that we only became aware of after preregistering Experiment 2 (Bays 2016, Oberauer et al. 2017), but these did not produce appreciably different

results.) The potentially more interesting possibility is that the hypothesized mechanism that is indexed by pN , context binding, may be more sensitive to challenging testing conditions (i.e., the MRI scanner session vs. the behavioral lab) than are the mechanisms underlying κ and pT .

References

- Bays, P. (2016). "Evaluating and excluding swap errors in analogue tests of working memory." *Scientific Reports* **6**: 19203.
- Bays, P. M., R. F. G. Catalao and M. Husain (2009). "The precision of visual working memory is set by allocation of a shared resource." *Journal of vision* **9**(10): 7.1-711.
- Bettencourt, K. C. and Y. Xu (2015). "Decoding the content of visual short-term memory under distraction in occipital and parietal areas." *Nature Neuroscience* **19**: 150.
- Bisley, J. W. and M. E. Goldberg (2010). "Attention, intention, and priority in the parietal lobe." *Annual Review of Neuroscience* **33**: 1–21.
- Bisley, J. W. and K. Mirpour (2019). "The neural instantiation of a priority map." *Current Opinion in Psychology* **29**: 108-112.
- Bland, J. M. and D. G. Altman (1995). "Comparing methods of measurement: why plotting difference against standard method is misleading." *The Lancet* **346**(8982): 1085-1087.
- Bouchacourt, F. and T. J. Buschman (2019). "A flexible model of working memory." *Neuron* **103**: 147-160.
- Brainard, D. H. (1997). "The Psychophysics Toolbox." *Spatial Vision* **10**(4): 433-436.
- Brouwer, G. J. and D. J. Heeger (2009). "Decoding and reconstructing color from responses in human visual cortex." *The Journal of neuroscience : the official journal of the Society for Neuroscience* **29**(44): 13992-14003.
- Brouwer, G. J. and D. J. Heeger (2011). "Cross-orientation suppression in human visual cortex." *J Neurophysiol* **106**(5): 2108-2119.
- Cai, Y., A. D. Sheldon, Q. Yu and B. R. Postle (2019). "Overlapping and distinct contributions of stimulus location and of spatial context to nonspatial visual short-term memory." *Journal of Neurophysiology* **121**: 1222-1231.
- Cox, R. W. (1996). "AFNI: software for analysis and visualization of functional magnetic resonance neuroimages." (0010-4809 (Print)).
- Delvenne, J.-B. and R. Bruyer (2004). "Does visual short-term memory store bound features?" *Visual Cognition* **11**: 1-27.
- Emrich, S. M., A. C. Riggall, J. J. Larocque and B. R. Postle (2013). "Distributed patterns of activity in sensory cortex reflect the precision of multiple items maintained in visual short-term memory." *J Neurosci* **33**(15): 6516-6523.
- Ester, E. F., T. C. Sprague and J. T. Serences (2015). "Parietal and Frontal Cortex Encode Stimulus-Specific Mnemonic Representations during Visual Working Memory." *Neuron* **87**(4): 893-905.
- Ester, E. F., D. W. Sutterer, J. T. Serences and E. Awh (2016). "Feature-Selective Attentional Modulations in Human Frontoparietal Cortex." *The Journal of Neuroscience* **36**(31): 8188-8199.
- Fukuda, K., G. Woodman and E. Vogel (2015). *Individual Differences in Visual Working Memory Capacity: Contributions of Attentional Control to Storage*, In P. Jolicoeur, B. Lefebvre & J. Martinez-Trujillo (Eds.), *Mechanisms of Sensory Working Memory: Attention and Performance XXV* (1 ed.): Academic Press. .
- Gosseries, O., Q. Yu, J. J. LaRocque, M. J. Starrett, N. Rose, N. Cowan and B. R. Postle (2018). "Parieto-occipital interactions underlying control- and representation-related processes in working memory for nonspatial visual features." *Journal of Neuroscience* **38**: 4357-4366.
- Gosseries, O., Q. Yu, J. J. LaRocque, M. J. Starrett, N. S. Rose, N. Cowan and B. R. Postle (2018). "Parietal-occipital interactions underlying control- and representation-related processes in working memory for nonspatial visual features." *The Journal of Neuroscience*.
- Jerde, T., E. P. Merriam, A. C. Riggall, J. H. Hedges and C. E. Curtis (2012). "Prioritized Maps of Space in Human Frontoparietal Cortex." *The Journal of Neuroscience* **32**: 17382-17390.
- Jiang, Y. V., T. Makovski and W. M. Shim (2009). Visual memory for features, conjunctions, objects, and locations. *The visual world in memory*. J. R. Brockmole, Psychology Press.: 33-65.
- Oberauer, K., S. Farrell, C. Jarrold and S. Lewandowsky (2016). "What limits working memory capacity?" *Psychological Bulletin* **142**: 758-799.
- Oberauer, K. and H.-Y. Lin (2017). "An interference model of visual working memory." *Psychological Review* **124**(1): 21-59.

- Oberauer, K. and H.-Y. Lin (2017). "An interference model of visual working memory." Psychological Review **124**: 21-59.
- Oberauer, K., C. Stoneking, D. Wabersich and H.-Y. Lin (2017). "Hierarchical Bayesian measurement models for continuous reproduction of visual features from working memory." Journal of Vision **17**.
- Olson, I. R. and Y. Jiang (2002). "Is visual short-term memory object based? Rejection of the "strong-object" hypothesis." Perception & Psychophysics **64**: 1055-1067.
- Olson, I. R. and C. Marshuetz (2005). "Remembering "what" brings along "where" in visual working memory." Perception & Psychophysics **67**: 185-194.
- Schneegans, S. and P. M. Bays (2017). "Neural Architecture for Feature Binding in Visual Working Memory." The Journal of Neuroscience **37**(14): 3913-3925.
- Serences, J. T. and S. Saproo (2012). "Computational advances towards linking BOLD and behavior." Neuropsychologia **50**(4): 435-446.
- Sprague, T. C., K. C. S. Adam, J. J. Foster, M. Rahmati, D. W. Sutterer and V. A. Vo (2018). "Inverted Encoding Models Assay Population-Level Stimulus Representations, Not Single-Unit Neural Tuning." eneuro.
- Swan, G. and B. Wyble (2014). "The binding pool: A model of shared resources for distinct items in visual working memory." Attention, Perception, and Psychophysics **76**: 2136-2157.
- Todd, J. J. and R. Marois (2004). "Capacity limit of visual short-term memory in human posterior parietal cortex." Nature **428**: 751-754.
- Todd, J. J. and R. Marois (2004). "Capacity limit of visual short-term memory in human posterior parietal cortex." Nature **428**: 751.
- Todd, J. J. and R. Marois (2005). "Posterior parietal cortex activity predicts individual differences in visual short-term memory capacity." Cognitive, Affective, & Behavioral Neuroscience **5**: 144-155.
- Todd, J. J. and R. Marois (2005). "Posterior parietal cortex activity predicts individual differences in visual short-term memory capacity." Cognitive, Affective, & Behavioral Neuroscience **5**(2): 144-155.
- Unsworth, N., K. Fukuda, E. Awh and E. K. Vogel (2014). "Working Memory and Fluid Intelligence: Capacity, Attention Control, and Secondary Memory Retrieval." Cognitive psychology **71**: 1-26.
- Vogel, E. K., A. W. McCollough and M. G. Machizawa (2005). "Neural measures reveal individual differences in controlling access to working memory." Nature **438**: 500.
- Xu, Y. (2017). "Reevaluating the Sensory Account of Visual Working Memory Storage." Trends in Cognitive Sciences **21**(10): 794-815.
- Xu, Y. and M. M. Chun (2006). "Dissociable neural mechanisms supporting visual short-term memory for objects." Nature **440**: 91-95.
- Yarkoni, T. and T. S. Braver (2010). Cognitive neuroscience approaches to individual differences in working memory and executive control: Conceptual and methodological issues. Handbook of individual differences in cognition: Attention, memory, and executive control. New York, NY, US, Springer Science + Business Media: 87-107.
- Yu, Q. and W. M. Shim (2017). "Occipital, parietal, and frontal cortices selectively maintain task-relevant features of multi-feature objects in visual working memory." Neuroimage **157**: 97-107.