Title: The neural codes underlying internally generated representations in visual working memory

Short Title: Internally generated neural codes in memory

Authors:

Qing Yu[1], Bradley R. Postle[2,3]

Affiliations:

[1]Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Shanghai, China

[2]Department of Psychiatry, University of Wisconsin–Madison, Madison, WI 53719, USA

[3]Department of Psychology, University of Wisconsin–Madison, Madison, WI 53706, USA

Corresponding author:

Qing Yu

Email: qingyu@ion.ac.cn

Conflict of Interest

The authors declare no competing financial interests.

Acknowledgements

This work was supported by NIH R01MH064498 to B.R.P.

Abstract

Humans can construct rich subjective experience even when no information is available in the external world. Here we investigated the neural representation of purely internally generated stimulus-like information during visual working memory. Participants performed delayed recall of oriented gratings embedded in noise with varying contrast during fMRI scanning. Their trialwise behavioral responses provided an estimate of their mental representation of the to-be-reported orientation. We used multivariate inverted encoding models to reconstruct the neural representations of orientation in reference to the response. We found that response orientation could be successfully reconstructed from activity in early visual cortex, even on 0% contrast trials when no orientation information was actually presented, suggesting the existence of a purely internally-generated neural code in early visual cortex. Additionally, cross-generalization and multidimensional scaling analyses demonstrated that information derived from internal sources was represented differently from typical working memory representations, which receive influences from both external and internal sources. Similar results were also observed in intraparietal sulcus (IPS), with slightly different cross-generalization patterns. These results suggest a potential mechanism for how externally-driven and internally-generated information is maintained in working memory.

Keywords

internally generated information; visual working memory; visual imagery; early visual cortex; intraparietal sulcus

Introduction

Humans have the ability to mentally retain and manipulate visual information even when the information is not in view. This ability -- visual working memory -- is fundamental to human cognition (Baddeley, 2003; Engle, Tuholski, Laughlin, & Conway, 1999; Luck & Vogel, 2013). Understanding how the brain keeps such information online is thus a critical question for cognitive neuroscience. The sensorimotor recruitment hypothesis posits that sensory cortex is an important substrate for the representation of fine-grained perceptual information in working memory (Awh & Jonides, 2001; D'Esposito & Postle, 2015; Serences, 2016); for example, early visual cortex for maintaining low-level visual information. This view is supported by evidence from multivariate analyses of functional magnetic resonance imaging (fMRI) data that stimulus-specific information can be decoded from early visual cortex during maintenance of visual feature information (Harrison & Tong, 2009; Riggall & Postle, 2012; Serences, Ester, Vogel, & Awh, 2009; Yu & Shim, 2017). With fMRI, a neural code is assessed as a systematic set of mappings between different values of a cognitive state and different patterns of fMRI activity, and a shared code is inferred if the same mapping is observed across two domains of cognition. With this logic it has been demonstrated that, in early visual cortex, visual working memory shares the same neural codes with visual perception (Harrison & Tong, 2009), attention (Yu & Shim, 2019), and imagery (Albers, Kok, Toni, Dijkerman, & de Lange, 2013), suggesting that early visual cortex may serve as a mental buffer for representing visual information across different categories of cognitive task (Roelfsema & de Lange, 2016).

Although early visual cortex recruits common neural codes for different cognitive processes, these processes can be driven by distinct sources of information. For example, visual perception is driven by external, bottom-up input received from the retina, and visual imagery is driven by internal, top-down input from higher cortical areas (Pearson, 2019). Of course, early visual cortex contains numerous reciprocal connections with higher cortical areas, and bottom-up and top-down signaling are involved in most, if not all visually mediated behaviors, including visual perception (Gilbert & Li, 2013; Muckli & Petro, 2013). Nevertheless, the fact that visual imagery shows distinct temporal dynamics, and evolves later in time, compared with visual perception (Dijkstra, Mostert, Lange, Bosch, & van Gerven, 2018), suggests at least some meaningful distinction between the processing of externally presented and internally generated information.

When considering the sensorimotor recruitment hypothesis, it is important to note that visual working memory cannot be understood as merely the prolongation of sensory processing, because many stimulus-nonspecific factors can influence representations in working memory. For example, several studies have demonstrated recall biases toward discrete color centers in visual working memory for color (Bae, Olkkonen, Allred, & Flombaum, 2015; Bae, Olkkonen, Allred, Wilson, & Flombaum, 2014; Panichello, DePasquale, Pillow, & Buschman, 2019), probably due to drift towards stable attractor states established through prior experience (Panichello et al., 2019). Information from the previous trial can also be reactivated or otherwise influence the current trial (Bae & Luck, 2019; Barbosa et al., 2020). Moreover, there is considerable physiological evidence for an important role for feedback from higher cortical areas. For example, laminar recordings indicate that delay-period input to V1 is most prominent in supra- and infragranular layers that receive feedback projections from higher areas (Lawrence et al., 2018; van Kerkoerle, Self, & Roelfsema, 2017), and delay-period local field potentials in area MT are coherent with spiking in prefrontal cortex (PFC; Mendoza-Halliday, Torres, & Martinez-Trujillo, 2014). Because typical working memory tasks, including those cited here, begin with the external presentation of to-be-remembered stimulus information, delay-period representations presumably reflect the combined influence of processes associated with externally presented and with internally generated information. How such "typical" working memory representations may differ from purely internally generated representations is the focus of the present study. Although our focus is on representations in early visual cortex, we also present results from intraparietal sulcus (IPS), because this region has also been implicated in representing working memory-related information (e.g., Bettencourt & Xu, 2016; Ester, Sprague, & Serences, 2015; Yu & Shim, 2017).

In the current study, stimulus contrast varied across trials between 0%, 10%, and 60%, but participants were instructed that a sample orientation would be presented on every trial, and that a recall response was required at the end of every trial, regardless of subjective visibility. This allowed us to use responses to infer backwards what they had represented during the delay period, including on 0% contrast ("null") trials that lacked external input. These responses could then be used to investigate internally generated representations in visual working memory maintenance. The comparison between the null and typical working memory trials (with 10% or 60% contrast) could also be used to isolate processes specific to internally generated representations.

Materials and Methods

Participants

All participants were recruited from the University of Wisconsin–Madison community. Two behavioral experiments (Experiments 1A and 1B) were performed prior to the fMRI experiment (Experiment 2) to test the visibility of the stimuli to be used in the fMRI experiment. Thirteen individuals (2 males, mean age 21.0 ± 3.3 years) participated in Experiment 1A, and 7 of these also participated in Experiment 1B, along with 9 new individuals (n = 16 in total; 3 males, mean age 19.6 ± 1.9 years). Eighteen individuals (including one who also participated in

Experiment 1B) participated in Experiment 2. One of these was excluded due to failure to comply with task instructions, resulting in 17 individuals (4 males, mean age 23.5 ± 3.8 years) as the final sample size for Experiment 2. We did not carry out power analysis a priori, but our sample size was comparable or even superior to those from recent fMRI studies that have used a similar task design (Bettencourt & Xu, 2016; Ester et al., 2015; Rademaker, Chunharas, & Serences, 2019; Yu, Teng, & Postle, 2020). All participants had normal or corrected-to-normal vision, reported no neurological or psychiatric disease, and provided written informed consent approved by the University of Wisconsin–Madison Health Sciences Institutional Review Board. All were monetarily compensated for their participation.

Stimuli and procedure

Sample stimuli were sinusoidal gratings embedded in white noise (spatial frequency = 1°/cycle, radius = 4°), presented at varying levels of Michelson contrast. In Experiment 1A, there were two types of stimuli: gratings with a *high* contrast (60%), and gratings with a 75% threshold-level contrast, determined for each subject with a thresholding task. In Experiments 1B and 2, there were three types of stimuli: gratings with a *high* contrast (contrast = 60%), gratings with a *low* contrast (contrast = 10%), and *null* stimuli (contrast = 0%). Importantly, no orientation information was visible in *null* gratings, making them equivalent to white noise patches.

All stimuli were created and presented using MATLAB (MathWorks, Natick, MA) and Psychtoolbox 3 extensions (Brainard, 1997; Pelli, 1997). In Experiments 1A and 1B, stimuli were presented on a 21.5-inch iMac screen at a viewing distance of 63 cm and behavioral responses were made with a computer mouse. In Experiment 2, stimuli were projected via a 60-Hz projector (Avotec Silent Vision 6011, Avotec, Inc., Stuart, FL), and viewed through a coilmounted mirror in the MRI scanner at a viewing distance of 69 cm, and participants' behavioral responses were made with an MR-compatible trackball response pad (Current Designs Inc., Philadelphia, PA). During the scan, eye position was monitored and recorded using the Avotec RE-5700 eye-tracking system (Avotec, Inc., Stuart, FL).

Experiment 2

We begin with a detailed description of Experiment 2, the experiment of primary theoretical interest, during which participants performed 1-item delayed recall of orientation in the fMRI scanner. On each trial, participants viewed a sample stimulus (*high*, *low*, or *null*) presented at the center of the screen for 0.5 s. After a delay of 9.5 s (or 8.5 s for two participants), an orientation dial (radius = 4º) was presented centrally, and participants rotated the dial until its needle matched the remembered orientation as precisely as possible in a 4-s response window. Critically, participants were told that an oriented grating would be presented on every trial, although its visibility would vary across trials, and they were instructed to make a best guess when they were unsure about what the orientation was. Feedback (recall error) was provided following the response period for 0.5 s, even on *null* trials, and recall error was calculated as the angular difference between sample and response orientations, regardless of whether or not the sample orientation had actually been visible (Figure 1). The sample orientation for each trial was randomly selected from 1º to 180º in steps of 1º in the orientation space. The starting position of the needle of the response dial was randomly chosen on every trial, independent of the sample.

<----- insert Figure 1 about here ----->

For four participants, total trial length was 22 s: for two of the four participants tested with an 8.5-s delay (S01 and S02), the inter-trial interval (ITI) was 9 s, and for the two tested with a 9.5-s delay (S03 and S04), ITI was 8 s. For all remaining participants, for whom the delay was 9.5 s and ITI was 10 s, total trial length was 24 s. To match the number of time points across participants, all analyses focused on the first 22 s of every trial.

Each run began with an 8-s fixation period, followed by 18 experimental trials, and the ratio of trial types (*high*: *low*: *null*) during each run was 3:1:2 (i.e., 9 *high* trials, 3 *low* trials, and 6 *null* trials). For one participant (S01), the experimental run in the first scan session was truncated to 12 trials (i.e., 6 *high* trials, 2 *low* trials, 4 *null* trials) due to a technical problem with scanning. Each participant completed 28 to 32 runs across two scanning sessions. In total, twelve participants completed 270 *high* trials, 90 *low* trials, and 180 *null* trials (S02, S05, S06, S08 to S12, S14 to S17); two participants completed 288 *high* trials, 96 *low* trials, and 192 *null* trials (S03 and S04); two completed 252 *high* trials, 84 *low* trials, and 168 *null* trials (S07 and S13); and one (S01) completed 231 *high* trials, 77 *low* trials, and 154 *null* trials. All participants were debriefed at the end of the study, and none of them reported awareness of the existence of null trials (i.e., all reported believing that an oriented grating was presented on every trial).

Experiments 1A and 1B

Prior to the fMRI experiment, we ran two behavioral studies to determine the contrasts of the gratings to be used in the scanner. The overarching rationale was to develop conditions that would disguise from participants the fact that a substantial proportion of samples contained no stimulus information (i.e., *null* samples). To achieve this, we sought to find two levels of contrast that were each highly discriminable, but that would create the impression for participants that subjective visibility would vary from trial to trial. The trial structure for both was similar to that from Experiment 2: one sample grating (radius = 3°) with a randomly selected orientation was presented on the screen for 0.1 s, followed by a brief delay, followed by recall with an orientation wheel. Responses were self-paced, and feedback was given after each response (0.5 s).

Experiment 1A was carried out to examine how subjects would perform at each of two levels of contrast: high and at-threshold. It began with a block of 80 trials to determine each individual's contrast threshold: After an initial 10 trials of delayed recall (delay of 0.3 s) at a fixed contrast of 12%, the sample contrast for each of the ensuing trials was adjusted using a QUEST procedure (Watson & Pelli, 1983). Responses were binarized using a cut-off criterion of 20° of recall error. Four catch trials were interleaved at randomly determined intervals, and on these catch trials the contrast was set to three times of the contrast from QUEST. The discrimination contrast threshold of the grating that generated 75% accuracy was determined at the end of the block. During the remainder of the session, participants performed 5 or 6 blocks of delayed recall of orientation, delay length was either 1 s or 7 s, and delay length and sample contrast (60%; at threshold) were fully crossed during each 60-trial block.

Experiment 1B was carried out to examine how subjects would perform at each of the three levels of contrast that would be used for the fMRI study: *high* (60%); *low* (10%), and *null* (0%). Participants performed 5 or 6 blocks with 60 trials each; again, delay length was either 1 s or 7 s, and delay length and sample contrast (*high*; *low*; *null*) were fully crossed. For both Experiments 1A and 1B, only trials with a 7-s delay were included in the behavioral analyses to better match the duration of the fMRI task.

Behavioral analyses

Behavioral performance was assessed in two ways. Within-trial recall error was calculated for *high* and *low* trials as the angular difference between the sample and response orientations, for each condition separately. Differences between conditions were evaluated by paired t-tests. Serial bias on response from the previous trial was calculated for all three conditions. This was done by calculating the difference between the current and previous response, and grouping the difference values into nine 20º-wide bins. To test whether the number of trials differed between bins, we performed a χ^2 goodness of fit test on each condition.

fMRI methods

Data acquisition

Whole-brain images were acquired with a 3 Tesla GE MR scanner (Discovery MR750; GE Healthcare, Chicago, IL) with a 32-channel head coil at the Lane Neuroimaging Laboratory at the University of Wisconsin–Madison HealthEmotions Research Institute (Department of Psychiatry). Functional images were acquired with a gradient-echo echo-planar sequence (2 s repetition time (TR), 22 ms echo time (TE), 60° flip angle) within a 64 × 64 matrix (42 axial slices, 3 mm isotropic). A high-resolution T1 image was also acquired for each session with a fast spoiled gradient-recalled-echo sequence (8.2 ms TR, 3.2 ms TE, 12° flip angle, 176 axial slices, 256 × 256 in-plane, 1.0 mm isotropic).

Preprocessing

Functional MRI data were preprocessed using AFNI (http://afni.nimh.nih.gov) (Cox, 1996). The first four volumes of each functional run were removed. The data were then registered to the first volume of the first run within each scan session, and then to the T1 volume of the same session. Data from the second session were further registered to the T1 volume of the first scanning session. The data were then motion corrected, detrended (linear, quadratic, cubic), converted to percent signal change. Data for subsequent general linear model (GLM) analyses were further spatially smoothed with a 4-mm FWHM Gaussian kernel. Data for multivariate and univariate time course analyses were zscored within each run.

Univariate analyses

Task-related changes in activity were identified with a mass-univariate GLM implemented in AFNI, with sample, delay and probe epochs of the task modeled with boxcars (0.5 s, 8.5 s or 9.5 s depending on the participant, and 4 s, respectively), each convolved with a canonical hemodynamic response function. Six nuisance regressors were also included to account for head motion artifacts in the six dimensions of rigid body motion.

Percent signal change in BOLD activity relative to baseline was calculated for each time point during the working memory task, baseline was chosen as the average BOLD activity of the first TR of each trial. BOLD signal change was averaged across trials within each condition, and across all voxels within each region of interest (ROI; see below).

Statistical significance of BOLD activity against baseline was assessed using two-tailed, one-sample t-tests against 0, and the resultant *p* values were corrected across time points using FDR (False Discovery Rate) (Benjamini & Hochberg, 1995). Statistical difference of BOLD activity between conditions at each time point was assessed using two-tailed paired t-tests, with FDR correction applied across time points and comparisons.

Region of interest (ROI) definition

We created subject-specific anatomical ROIs by warping masks from the probabilistic atlas of Wang and colleagues (2015) to each subject's structural scan in their native space. Early visual anatomical ROIs were created by merging the masks for unilateral V1, V2, and V3 within and between hemispheres. IPS anatomical ROIs were created by merging the masks for unilateral regions IPS0-5 within and between hemispheres. For the *Early Visual Cortex* functionally defined ROI, we identified the 500 voxels displaying the strongest loading on the contrast [sample - baseline], collapsing over all three conditions. For the *IPS* functionally defined ROI, we identified the 500 voxels displaying the strongest loading on the contrast [delay - baseline], collapsing over all three conditions. For completeness, an alternate "*Sample IPS* ROI" was also defined as the 500 voxels in this anatomical region displaying the strongest loading on the contrast [sample - baseline].

Multivariate inverted encoding modeling

All inverted encoding modeling (IEM) analyses were performed using custom functions in MATLAB. The IEM assumes that the responses of each voxel can be characterized by a small number of hypothesized tuning channels. Following previous work, the number of orientation tuning channels was set to nine (20º apart, equally spaced), and the idealized feature tuning curve of each channel to a specific orientation θ was defined as a half-wave-rectified sinusoid raised to the eighth power (FWHM = 0.82 rad):

$$f(\theta) = \cos(\theta - c)^8$$

Where c was the center of the channel.

We then computed the weight matrix (W, $v \times k$, v: the number of voxels; k: the number of channels) that projects the hypothesized channel responses (C_1, $k \times n$, n: the number of trials) to actual measured fMRI signals in the training dataset (B_1, $v \times n$), and extracted the estimated channel responses ($\hat{C}_2$, $k \times n$) for the test dataset (B_2, $v \times n$) using this weight matrix.

The relationship between the training dataset (*B₁*) and the channel responses (*C₁*) was characterized by:

$$B_1 = WC_1$$

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

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

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

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

Because orientations in the current study were randomly selected from the 1º - 180º orientation space (in steps of 1º), we did not pick a fixed set of channel centers, as is often done (Yu & Shim, 2017; Yu, Teng, et al., 2020). Instead, following Rademaker et al. (2019) we first picked a set of equally spaced channel centers (e.g., 0º, 20º, 40º, 60º, 80º, 100º, 120º, 140º, 160º), conducted the analysis as described above, and then shifted the channel centers by 1º and repeated the analysis. The procedure was repeated 20 times, such that all 180 orientations from 1º to 180º in 1º step served as channel centers. We then combined estimated channel responses from all iterations of these analyses to create responses of 180 orientation channels. The result, for any given orientation, can be considered a reconstruction of the model's estimate of the neural representation of that orientation. This procedure ensured that our channel estimates were not biased by any specific channel centers. All channel responses were then centered on a common center (0º on the x-axis) and averaged for visualization and for statistical comparisons.

Hypothesis testing

Analysis plan. If a participant is not aware of the fact that a considerable proportion of trials will feature *null* samples that contain no orientation information, we assume that on *null* trials they will generate an orientation for response at some point prior to the onset of the response dial. Furthermore, because the initial orientation of the dial cannot be predicted prior to its onset, we assume that this response plan will not be kinematic (e.g., how many degrees they plan to turn the dial), but rather will be the representation of the orientation that the participant plans to produce at the end of trial. To validate this assumption, our first analysis would be to train an IEM using the orientation of the response on that trial (response-based IEM). Successful reconstruction of orientation with this IEM at time points preceding the response (i.e., during the delay period) would mean that participants were indeed representing the orientation of their planned response during those earlier time points (response-based neural code)."

Assuming success of this first analysis, the next step would be to determine whether a common response-based neural code was employed across conditions. This would be done by applying the response-based IEM from one trial type (e.g., *high*) to data from the other two trial types (e.g., *low* and *null*). We anticipated three possible outcomes: If reconstruction in a tested condition was significantly positive, and did not differ from that in the training condition, this would reflect "full generalization"; if reconstruction in a tested condition was significantly positive, but was also significantly lower than that in the training condition, this would reflect "partial generalization"; and if reconstruction in a tested condition was not significant, this would reflect "failed generalization". These results would be interpreted as evidence for a fully shared neural code, for a partly shared neural code, or as a failure to find evidence for a shared neural code, respectively.

Finally, because IEM relies on specific hypotheses of orientation channels, we would also perform a model-free analysis, multidimensional scaling (MDS), to see if this alternative approach would support conclusions comparable to those suggested by the IEM analyses.

Operationalizing hypothesis tests. To investigate the codes supporting the representation of orientation in the different conditions (*high*; *low*; *null*), we trained two IEMs: a response-based IEM labeled according to the orientation of the response on each trial, and a sample-based IEM labeled according to the sample orientation on each trial. Note that the response-based IEM would be the focus of our analyses, and results from the sample-based IEM in the *null* condition would not be interpretable on its own, but would serve as controls for comparing with the results from *high* and *low* conditions. IEMs were trained and tested using a leave-one-run-out cross-validation procedure, for each condition, time point (or average of time points, e.g., average of time points 8 - 10 s for delay period), and ROI separately. To compare response-based neural codes across conditions, we also used a leave-one-run-out procedure, training the response-based IEM on data from one condition, and testing the IEM on the data from all three conditions, including the training condition (which would yield the same result as the first analysis) and the two other conditions. This procedure was performed for each condition, time point (or average of time points), and ROI separately.

We also trained several complementary IEMs for testing alternative explanations for the results. First, we trained a mixed IEM using a balanced set of trials from each condition (*high*; *low*; *null*), and tested this IEM on the same balanced set of trials from each condition separately. The purpose of this IEM would be to avoid potential concerns with differences in SNR across IEMs (Liu, Cable, & Gardner, 2018; Sprague et al., 2018). Second, to examine the influence of previous-trial information on the reconstruction of current-trial information, we trained response-based IEMs using response labels from the previous trial, or trained sample-based and response-based IEMs while excluding trials with similar response to that of the previous trial.

To characterize the strength of each IEM reconstruction, we collapsed over the channel responses on both sides of the common center, averaged them, then calculated the slope of each collapsed reconstruction using linear regression (Foster, Bsales, Jaffe, & Awh, 2017; Samaha, Sprague, & Postle, 2016). A larger positive slope indicates stronger positive representation. We used a bootstrapping procedure (Ester et al., 2015; Yu, Teng, et al., 2020) to characterize the significance of the slopes. For each combination of factors (IEM, condition, time point, or ROI), seventeen orientation reconstructions were randomly sampled with replacement from the pool of seventeen participants and averaged. This procedure was repeated 10000 times, resulting in 10000 average orientation reconstructions, and correspondingly 10000 slopes. The probability of obtaining a negative slope among the 10000 slopes was counted as the one-tailed *p*-value of the slope. To characterize the difference between the slopes of two IEM reconstructions, we first calculated the difference between two bootstrapped slopes 10000 times, which generated 10000 slope differences. The significance of the slope difference was then calculated using the same one-tailed method as above. All *p*-values were corrected for multiple comparisons using the FDR method, across IEMs (sample-based, response-based), conditions (*high*, *low*, or *null*), and time points.

We also assessed evidence for differences between the slopes of delay-period response-based reconstructions with Bayes Factors (BF), which support evaluation of the amount of evidence for one hypothesis (H_1) against the null hypothesis (H_0). H_1 referred to a positive reconstruction and H_0 referred to a failed reconstruction (i.e., a slope no larger than 0). For comparison between the slopes of two reconstructions, H_1 referred to the slopes being different and H_0 referred to the absence of evidence for a difference. As an example, a BF_{10} of 3 would indicate that H_1 is three times more likely than H_0, whereas a BF_{10} of 0.33 would indicate that H_0 is three times more likely than H_1. All BF analyses were conducted using the JASP software (Love et al., 2019).

Multidimensional scaling. For each ROI and each trial epoch we categorized all response orientations into four bins (0-45º, 45-90º, 90-135º, 135-180º). Trial number for each condition was matched by subsampling data from the *high* and *null* conditions to match the number of trials in the *low* condition. The Euclidean distances between orientation bins and conditions were then computed using the covariance matrix calculated from the subsampled data. This subsampling procedure was repeated for 1000 times and averaged. Distances were averaged across participants, and multidimensional scaling was performed on the distance matrix using the "cmdscale" function in MATLAB.

Results

Behavior

Experiment 1A

Participants' 75% contrast discrimination threshold for recall of orientation against a noise background ranged between 4-6%, with a mean of 5.0% and a standard deviation of 0.6%. For delayed recall of the orientation of a sample grating, the average recall error for *high* contrast (60%) samples (9.0º ± 1.5º) was significantly lower than for the *threshold* contrast samples (17.4º ± 4.4º), $t(12) = 5.95$, $p < 0.001$.

Experiment 1B

Delayed recall of orientation did not differ between *high* contrast (60%; 10.2º ± 1.9º) and *low* contrast (10%; 10.6º ± 2.7º) conditions, $t(15) = 0.64$, $p = 0.530$ (Figure 1). The fact that average recall error did not differ between the *high* and *low* trials established the fact, critical for the logic of Experiment 2, that *low* and *high* samples were comparably visible to participants. This, plus the marked difference between performance at these two levels of contrast versus performance at 75% threshold (Experiment 1A) indicated that neither *high* nor *low* contrast trials were likely to produce trials in which the sample grating was not visible to the participant (in contrast to some of the *threshold* trials in Experiment 1A).

Experiment 2

Consistent with Experiment 1B, Recall error during scanning did not differ between the *high* (11.4º ± 4.0º) and *low* (11.7º ± 4.8º) trials, $t(16) = 0.58$, $p = 0.567$.

Although recall error could not be calculated for *null* trials, the results from several analyses suggested that participants did not treat *null* trials different from trials on which a sample grating was visible. First, angular difference between the starting position of the response needle and the recalled orientation did not differ between *high*, *low*, and *null* trials (42.3º ± 2.4º, 41.7º ± 3.9º, 41.8º ± 5.7º, respectively; all ts < 0.85, ps > 0.408), suggesting that the three conditions were comparable in terms of effort during recall. Second, although sample orientation on each trial was randomly chosen and the distribution of sample orientations was uniform (i.e., there was an equal proportion of cardinal and oblique orientations), plotting the distribution of participants' raw responses showed biased responses towards oblique orientations (relative to cardinal orientations) for all three trial types (Figure 2). This indicates that trials of all types were influenced to a similar extent by a systematic bias, perhaps from one or more stimulus-nonspecific factors such as prior knowledge (Panichello et al., 2019; Yu, Panichello, Cai, Postle,

& Buschman, 2020). In sum, *null* and *high*/*low* trials were well matched in terms of procedural details, and the only difference between conditions was the availability of external orientation information. Therefore, any orientation information observed in the *null* trials could only be internally generated.

<----- insert Figure 2 about here ----->

fMRI

Time course of BOLD activity

All analyses were carried out at the level of the *Early Visual Cortex* ROI and the *IPS* ROI. In both regions, a conventional time course of BOLD activity change was observed for all three conditions (Figure 3): sample-evoked activity reached its peak at around 4-6 s after trial onset; delay-period activity reached its trough at around 8-10 s; and response-evoked activity reached its peak at around 14-16 s. Time points 8-10 s were subsequently used to operationalize "late delay-period" activity. In early visual cortex, activity in *null* trials was slightly lower than that in *high* and *low* trials during sample and early delay epochs (2-8 s; all ps < 0.023), but not at 10 s (both ps > 0.167) nor during the response epoch (12 s and after; all ps > 0.342). In IPS, in contrast, *null* activity was lower during the sample (2-4 s; all ps < 0.005) and response epochs (12-18 s, all ps < 0.040 except for 12 s between *high* and *null*: $p = 0.073$), but not during the delay (6-10 s, all ps > 0.132).

<----- insert Figure 3 about here ----->

Inverted encoding modeling

Early visual cortex. To assess the time course of neural representations of orientation, for each trial type, we applied a sample-based IEM (i.e., trained on the sample label) and a response-based IEM (trained on the response label) to every time point of the trial. For *high* and *low* trials, reconstruction with the sample-based IEM was significant beginning at 4 s after sample onset and sustained for the remainder of the trial (all ps < 0.001). Similarly, reconstruction with the response-based IEM were significant for the duration of trial, beginning at 4 s for *high* trials and at 2 s for *low* trials (all ps < 0.040). Sample-based and response-based reconstructions did not differ at any time point, for either of these two conditions (all ps > 0.157). These results validated the approach of using participants' responses as an estimate of the orientation that they represented earlier in the trial, prior to the response.

Turning next to *null* trials, reconstruction with sample-based IEMs did not achieve statistical significance except for two isolated time points: 2 s ($p = 0.017$) and 16 s ($p = 0.036$), probably due to statistical noise. Note that these null results amounted to confirmation of a sanity check, because the labels used to train the sample-based IEM did not correspond to what subjects were presented on these trials. Reconstructions with response-based IEMs, were significant beginning with 6 s and for the duration of trial (all ps < 0.020; Figure 4A). Critically, these response-based reconstructions were significantly different from the sample-based reconstructions for 6-8 s and from 12 s onwards (green asterisks; all ps < 0.012), suggesting that robust orientation representations specific to the response started from 6 s after trial onset. This indicates that, beginning relatively early in the trial, participants generated and maintained a representation with exclusively internally derived information.

Intraparietal sulcus. In IPS, results were generally comparable to those from early visual cortex, albeit weaker in magnitude. When focusing on the late-delay period (Figure 4D), sample and response reconstructions were significant in all conditions (all ps < 0.037), except for the sample reconstruction in the null condition (*p* = 0.259). Time point-by-time point reconstructions were also qualitatively similar to early visual cortex (Figure 4C): on *high* trials sample and response reconstructions emerged during the sample period and were sustained throughout the trial, as were sample reconstructions on *low* trials (all *p*s < 0.041). Response reconstructions were smaller in slope on *low* trials, and, with the exception of a single time point (6 s after trial onset, *p* = 0.007), did not survive correction for multiple comparisons during the delay. Note that the lower number of trials for the *low* condition might have been responsible for the lack of significance here. Indeed, robust reconstruction of orientation was observed for *low* trials when averaging across time points in the delay period (Figure 4D).

Turning to *null* trials, reconstructions with sample-based IEMs only achieved statistical significance at 2 s, a result probably be due to statistical noise. Reconstruction with response-trained IEMs, however, was significant for all time points beginning with 4 s (all *p*s < 0.028), with the exception of 10 s of the delay period (*p* = 0.076).

We also carried out these analyses in the *Sample IPS* ROI (IPS ROI defined using sample-period activity), and the results (not shown) were qualitatively similar to those in the *Delay IPS* ROI.

<----- insert Figure 4 about here ----->

One possible concern about the finding of principal theoretical interest from these analyses -- the reconstruction of response-related stimulus information from the delay period of *null* trials (Figure 4) -- is that this might reflect “spillover” of information processed during the previous trial, rather than evidence for genuinely internally generated stimulus representations. Additional analyses carried out to assess this alternative possibility ruled it out as a major concern, and these are presented at the end of the *Results* section (see *Secondary analyses to assess the influence of the previous trial on response-based IEMs*).

Comparison of neural codes across high, low, and null trials

Having established robust measurements for internally generated neural representations of orientation, we next sought to examine the nature of these representations. Specifically, because representations on *null* trials were purely internally generated, whereas representations on *high* and *low* trials reflected influences from both external and internal sources, we tested whether the representations maintained during these different trial types recruited a common neural code, in keeping with previous demonstrations of a shared neural code between working memory and perception (Harrison & Tong, 2009), between working memory and attention (Yu & Shim, 2019), and between working memory and imagery (Albers et al., 2013). To this end we trained IEMs on one condition, and tested it on all three conditions (see *Methods*). Note that only response-based IEMs were recruited for this purpose. For these analyses, we emphasized the results from the late delay period (8-10 s after trial onset; also see Figure 5 for results for the full time courses). Here we also employed Bayes Factors (BF) to assess the amount of evidence in generalization. A BF of larger than 3 or smaller than 1/3 can be considered substantial evidence supporting or rejecting the hypothesis.

<----- insert Figure 5 about here ----->

In early visual cortex, we successfully reconstructed orientation from the late-delay period of *low* trials with IEMs trained on *high* trials ($p < 0.001$, $BF_{10} = 280.3$), and of *high* trials with IEMs trained on *low* trials ($p < 0.001$, $BF_{10} = 79.5$). Furthermore, these results demonstrated full generalization: reconstructions for *high* and *low* trials with the *high*-trained IEM did not differ from each other (p = 0.552, BF_{10} = 0.5); nor did reconstructions for *high* and *low* trials with the *low*-trained IEM (p = 0.477, BF_{10} = 0.8; Figure 6A and 6B). When comparing each of these visible trial types with *null* trials, in contrast, cross-condition generalization was asymmetric: For *high* trials, although the IEM trained on *high* trials failed to generalize to *null* trials (p = 0.135, BF_{10} = 0.7; Figure 6A), the IEM trained on *null* trials did successfully reconstruct orientation on *high* trials (p = 0.0014, BF_{10} = 6.0), and reconstructions for *high* and *null* trials with the *null*-trained IEM did not differ from each other (p = 0.552, BF_{10} = 0.2; Figure 6C). For *low* trials, on one hand, the IEM trained on *null* trials successfully reconstructed orientation on *low* trials (p = 0.0013, BF_{10} = 8.8), and reconstructions for *low* and *null* trials with the *null*-trained IEM did not differ from each other (p = 0.552, BF_{10} = 0.2; Figure 6C); On the other hand, the IEM trained on *low* trials did generalize to *null* trials (p = 0.0052, BF_{10} = 5.3), although the slope of this reconstruction was lower than that on *low* trials with the *low*-trained IEM (p = 0.030, BF_{10} = 28.4; Figure 6B), suggesting only partial generalization from *low* to *null* trials.

In IPS, although response-based neural codes were also fully generalizable between *high* and *low* trials (train *high*-test *low*, p = 0.007, BF_{10} = 6.2; train *low*-test *high*, p = 0.006, BF_{10} = 6.9; train *high*-test *high* vs. train *high*-test *low*, p = 0.732, BF_{10} = 0.2; train *low*-test *low* vs. train *low*-test *high*, p = 0.497, BF_{10} = 0.3; Figure 6D and 6E), there was no evidence for cross-generalization from *null* trials to *high* or *low* trials (train *null*-test *high*, p = 0.215, BF_{10} = 0.5; train *null*-test *low*, p = 0.061, BF_{10} = 1.6; Figure 6F), nor from *high* or *low* trials to *null* trials (train *high*-test *null*, p = 0.252, BF_{10} = 0.4; and train *low*-test *null*, p = 0.187, BF_{10} = 0.6; Figure 6D and 6E).

<----- insert Figure 6 about here ----->

Although cross-generalization is a common approach for assessing commonality of neural codes (Albers et al., 2013; Rademaker et al., 2019; Yu & Shim, 2019), interpreting failures to generalize can be complicated by technical considerations arising from training the IEM on the same versus on different datasets (Liu, Cable, & Gardner, 2018; Sprague et al., 2018). Therefore, we repeated these analyses but with a single IEM trained on a balanced set of trials drawn in equal number from *high*, *low*, and *null* trials. Results with this mixed IEM were complementary to the cross-generalization analyses: in both early visual cortex and IPS, the mixed IEM generated successful reconstructions of orientation from *high* and *low* trials (4-10 s: all *p*s < 0.006), but failed on *null* trials (4-10 s: all *p*s > 0.140; Figure 7).

<----- insert Figure 7 about here ----->

Model-free analyses

Lastly, to determine whether a difference between *null* and *high*/*low* trials would be observed when no model-based approach was applied to the data, we compared the representational distances between conditions using MDS. MDS analyses were performed for the sample (4-6 s after trial onset), delay (8-10 s after trial onset), and response (14-16 s after trial onset) epochs of the working memory task, separately for early visual cortex and for IPS. For visualization purposes, response orientations were grouped into four 45º-wide bins.

In early visual cortex, during the sample epoch, the three conditions were discriminable along Dimension 1, confirming that differences in stimulus contrast influenced sensory processing (Figure 8A). During the delay period the distance between the *high* and *low* conditions decreased, such that the two now overlapped along Dimension 1, while the *null* condition remained separated from the other two. This suggested that, as stimulus-driven influences diminished, trials that relied exclusively on internally derived information remained distinct. This discriminative element carried on into the response period, along Dimensions 2 and 3, despite the fact that participants performed the same type of motor response on every trial. In IPS, a similar discriminative pattern was also observed between conditions (Figure 8B). Thus, in both brain areas, patterns of activity on *null* trials were distinct from those on *high*/*low* trials in multidimensional representational space. The fact that this was true for all epochs of the trial suggests that this separability was not simply a result of perceptual differences between memory samples.

<----- insert Figure 8 about here ----->

Secondary analyses to assess the influence of the previous trial on response-based IEMs

Recent perceptual history can bias behavior on the current trials (Fischer & Whitney, 2014), including during working memory tasks (Barbosa et al., 2020; Samaha, Switzky, & Postle, 2019), and it has been shown that the no-longer-relevant content of the previous trial can be decoded from electroencephalography (EEG) signals recorded during a visual working memory task (Bae & Luck, 2019). Consequently, we carried out a series of analyses to assess whether the response-related reconstructions from *null* trials (Figure 4), rather than reflecting internally generated stimulus representations, might instead be due to “spillover” of information processed during the previous trial. We tested this possibility with two approaches. First, we examined if the response of the previous trial could be reconstructed from patterns of activity of the current trial in the current data. In early visual cortex we found that the response of the previous trial could indeed be reconstructed in all conditions, especially during the earlier portion of the trial, (all *p*s < 0.031; Figure 9). However, because above-baseline-level reconstruction of the previous-trial response was present at the very beginning of the trial (i.e., 0 s), and reconstruction of the current-trial response did not emerge until 6 s after trial onset, we believe it was unlikely that these two sets of results reflected the same piece of information. Furthermore, in IPS, reconstruction of the previous trial's response was almost absent, with the exception of three isolated time points across all three conditions (all *p*s < 0.03). This effect alone again cannot explain the sustained reconstructions of the response orientation on *null* trials.

<----- insert Figure 9 about here ----->

A second approach to assess the possible influence of information from previous trials on response-related reconstructions from *null* trials was to redo the analyses after removing the trials for which the response was most similar to the response on the previous trial. We did this by first calculating the difference between each trial's response and the response on the previous trial, for all three conditions, and grouping the trials by difference values into nine 20º-wide bins. For *high* and *null* trials, the distribution of the differences was not uniform ($\chi^2(8) = 19.5$ and 81.9, $p = 0.012$ and $p < 0.001$, respectively; Figure 10), suggesting a potential influence from previous trials on the performance of the current trial. Next, for *null* trials, we removed the influence from the responses that were closest to the previous response (difference < 10º; bins highlighted in red in Figure 10) by excluding trials that belonged to this bin and repeating the IEM analyses on the remaining trials. Significant response reconstructions were still observed in this subset of *null* condition trials (Figure 11), increasing our confidence that the representation of response-related orientation information on *null* trials cannot be simply explained as reactivation of perceptual history from the previous trial.

<----- insert Figures 10 and 11 about here ----->

Finally, we examined whether the potency of the spillover effect varied with sample type, by sorting every *high* trial as a function of whether it was preceded by a *high*, *low*, or *null* trial. Results (not shown) indicated that the spillover effect was comparable for each trial type, and that the time course of each mimicked the pattern seen in Figure 9.

Discussion

The human brain processes massive amounts of information every day, from both external and internal sources. To explore how internally-generated information is represented in the brain during working memory, we incorporated a null-sample condition into a delayed-recall task. First, we demonstrated that, after the presentation of a null sample, participants generated a neural representation that corresponded to the response that they would make at the end of the trial, confirming that our procedure was successful at producing internally generated working-memory representations. Next, we assessed cross-generalization of the neural representation of orientations between conditions, and observed an asymmetric pattern in early visual cortex: IEMs trained on data from *null* trials generalized fully to data from visible-sample trials, but the converse was not true. This suggested some difference in the processing of internally generated representations versus conventional working memory representations that receive influences from both external and internal sources. This difference in neural codes was also evident when the data were projected into multidimensional representational space: the patterns of activity for *high* and *low* trials were clearly segregated from *null* trials in both early visual cortex and IPS. Therefore, stimulus information that is derived from an external source is represented differently than stimulus information that is generated internally.

Our findings might seem inconsistent with previous work that has demonstrated a shared neural code between visual working memory and visual imagery in early visual cortex (Albers et al., 2013). However, because visual imagery tasks often involve elements such as mental rotations (Albers et al., 2013) or retrocueing manipulations (Dijkstra et al., 2018), they typically refer overtly to previously presented (i.e., externally originated) information, and this may explain why similar neural codes are recruited by these two classes of task. It had thus remained unclear whether "purely" internally-derived representations also share the same neural code as "conventional" working memory representations. The present results -- indicating that the representation of orientation in early visual cortex fully generalizes from the *null* to the *high* and *low* conditions, but not in the other direction – suggest that all three conditions share the same purely internally generated neural codes, and that conventional working memory representations contain one or more additional dimensions that are lacking from "purely" internally originated visual representations. The additional dimension(s) are likely related to processes that are involved in the initial processing of externally presented information.

The differences between working memory and internally originated imagery were also preserved in IPS, where we found that *null* and *high*/*low* trials did not generalize in either direction, although the effects were generally weaker compared with those in early visual cortex. These results are in line with previous work demonstrating failures to find evidence -- in higher-order parietal and/or frontal cortex -- for generalization of neural codes between working memory and visual perception (Rademaker et al., 2019), attention (Yu & Shim, 2019), and imagery (Albers et al., 2013).

What is the nature of the internally generated representations observed in the delay period of the *null* condition in the current study? One possibility is a preparatory motor code, similar to what has been demonstrated for visual working memory for orientation on a task that allowed for concurrent selection of visual and motor responses (van Ede, Chekroud, Stokes, & Nobre, 2019). If so this would need to be a highly abstract code, akin to an intention, because the starting position of the probe in our experiment was randomized from trial-to-trial, and so participants would not have been able to plan their specific motor response prior to the onset of the response wheel. Another possibility is that they reflected internally generated representations of participants' best guess of the orientation of the sample. This would be consistent with the fact that the time course of the representation of orientation developed later in time in the *null* condition relative to the *high* and *low* conditions, especially in early visual cortex. Similarly, it has been observed that representations of visual imagery develop later in time than do representations associated with visual perception (Dijkstra et al., 2018) and with visual working memory (Albers et al., 2013). It is likely that internally generated representations are influenced by many stimulus-nonspecific factors, such as prior knowledge (Bae et al., 2015; Bae et al., 2014; Panichello et al., 2019; Yu, Panichello, et al., 2020) and recent history (Bae & Luck, 2019; Fischer & Whitney, 2014), and that these stimulus-nonspecific factors may serve, in part, as differentiating factors in the coding of externally driven versus internally generated information. Indeed, we did observe influences from the previous trial in the current experiment, although spillover from the previous trial alone cannot explain the sustained, robust representations of the response in the *null* condition. It should be possible to use the "null-sample" paradigm in combination with other visual tasks to better understand the nature of internally generated visual representations. For example, it would be interesting to compare internally generated representations directly with the codes that support visual perception. It would also be interesting to include confidence ratings in future tasks to better understand the subjective experience of the null task. Finally, by combining the paradigm with ultra-high field fMRI, one would be able to investigate whether there exist layer-specific representations for purely internally generated representations.

Our results, together with previous work (Albers et al., 2013; Harrison & Tong, 2009; Rademaker et al., 2019; Yu & Shim, 2019), suggest a potential mechanism for how the brain processes information originating from different sources. Early visual cortex represents stimulus properties with a common neural code that is insensitive to behavioral/cognitive context, such that the same neural code is shared between visual perception, attention, and working memory, consistent with the sensorimotor recruitment hypothesis. However, early visual cortex also registers the source of origination of this information, such that externally originated and internally originated representations can be differentiated. This distinction between externally and internally originated representations was also observed in a higher-order cortical area, IPS, although perhaps with a slightly different pattern. These signals may underlie the neural basis for how the brain differentiates and maintains signals from different sources.

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. *Curr Biol, 23*(15), 1427-1431. doi:10.1016/j.cub.2013.05.065

Awh, E., & Jonides, J. (2001). Overlapping mechanisms of attention and spatial working memory. *Trends Cogn Sci, 5*(3), 119-126.

Baddeley, A. (2003). Working memory: looking back and looking forward. *Nat Rev Neurosci, 4*(10), 829-839. doi:10.1038/nrn1201

Bae, G. Y., & Luck, S. J. (2019). Reactivation of Previous Experiences in a Working Memory Task. *Psychol Sci, 30*(4), 587-595. doi:10.1177/0956797619830398

Bae, G. Y., Olkkonen, M., Allred, S. R., & Flombaum, J. I. (2015). Why some colors appear more memorable than others: A model combining categories and particulars in color working memory. *J Exp Psychol Gen, 144*(4), 744-763. doi:10.1037/xge0000076

Bae, G. Y., Olkkonen, M., Allred, S. R., Wilson, C., & Flombaum, J. I. (2014). Stimulus-specific variability in color working memory with delayed estimation. *Journal of Vision, 14*(4), 7-7. doi:10.1167/14.4.7

Barbosa, J., Stein, H., Martinez, R. L., Galan-Gadea, A., Li, S., Dalmau, J., . . . Compte, A. (2020). Interplay between persistent activity and activity-silent dynamics in the prefrontal cortex underlies serial biases in working memory. *Nat Neurosci*. doi:10.1038/s41593-020-0644-4

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B-Statistical Methodology, 57*(1), 289-300.

Bettencourt, K. C., & Xu, Y. (2016). Decoding the content of visual short-term memory under distraction in occipital and parietal areas. *Nat Neurosci, 19*(1), 150-157. doi:10.1038/nn.4174

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spat Vis, 10*(4), 433-436.

Cox, R. W. (1996). AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Comput Biomed Res, 29*(3), 162-173.

D'Esposito, M., & Postle, B. R. (2015). The cognitive neuroscience of working memory. *Annu Rev Psychol, 66*, 115-142. doi:10.1146/annurev-psych-010814-015031

Dijkstra, N., Mostert, P., Lange, F. P., Bosch, S., & van Gerven, M. A. (2018). Differential temporal dynamics during visual imagery and perception. *Elife, 7*. doi:10.7554/eLife.33904

Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. (1999). Working memory, short-term memory, and general fluid intelligence: a latent-variable approach. *J Exp Psychol Gen, 128*(3), 309-331.

Ester, E. F., Sprague, T. C., & Serences, J. T. (2015). Parietal and Frontal Cortex Encode Stimulus-Specific Mnemonic Representations during Visual Working Memory. *Neuron, 87*(4), 893-905. doi:10.1016/j.neuron.2015.07.013

Fischer, J., & Whitney, D. (2014). Serial dependence in visual perception. *Nat Neurosci, 17*(5), 738-743. doi:10.1038/nn.3689

Foster, J. J., Bsales, E. M., Jaffe, R. J., & Awh, E. (2017). Alpha-Band Activity Reveals Spontaneous Representations of Spatial Position in Visual Working Memory. *Curr Biol, 27*(20), 3216-3223 e3216. doi:10.1016/j.cub.2017.09.031

Gilbert, C. D., & Li, W. (2013). Top-down influences on visual processing. *Nat Rev Neurosci, 14*(5), 350-363. doi:10.1038/nrn3476

Harrison, S. A., & Tong, F. (2009). Decoding reveals the contents of visual working memory in early visual areas. *Nature, 458*(7238), 632-635. doi:10.1038/nature07832

Lawrence, S. J. D., van Mourik, T., Kok, P., Koopmans, P. J., Norris, D. G., & de Lange, F. P. (2018). Laminar Organization of Working Memory Signals in Human Visual Cortex. *Curr Biol, 28*(21), 3435-3440 e3434. doi:10.1016/j.cub.2018.08.043

Liu, T., Cable, D., & Gardner, J. L. (2018). Inverted Encoding Models of Human Population Response Conflate Noise and Neural Tuning Width. *J Neurosci, 38*(2), 398-408. doi:10.1523/JNEUROSCI.2453-17.2017

Love, J., Selker, R., Marsman, M., Jamil, T., Dropmann, D., Verhagen, J., . . . Wagenmakers, E. J. (2019). JASP - graphical statistical software for common statistical designs. *Journal of Statistical Software, 88*.

Luck, S. J., & Vogel, E. K. (2013). Visual working memory capacity: from psychophysics and neurobiology to individual differences. *Trends Cogn Sci, 17*(8), 391-400. doi:10.1016/j.tics.2013.06.006

Mendoza-Halliday, D., Torres, S., & Martinez-Trujillo, J. C. (2014). Sharp emergence of feature-selective sustained activity along the dorsal visual pathway. *Nat Neurosci, 17*(9), 1255-1262. doi:10.1038/nn.3785

Muckli, L., & Petro, L. S. (2013). Network interactions: non-geniculate input to V1. *Curr Opin Neurobiol, 23*(2), 195-201. doi:10.1016/j.conb.2013.01.020

Panichello, M. F., DePasquale, B., Pillow, J. W., & Buschman, T. J. (2019). Error-correcting dynamics in visual working memory. *Nat Commun, 10*(1), 3366. doi:10.1038/s41467-019-11298-3

Pearson, J. (2019). The human imagination: the cognitive neuroscience of visual mental imagery. *Nat Rev Neurosci, 20*(10), 624-634. doi:10.1038/s41583-019-0202-9

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spat Vis, 10*(4), 437-442.

Rademaker, R. L., Chunharas, C., & Serences, J. T. (2019). Coexisting representations of sensory and mnemonic information in human visual cortex. *Nat Neurosci, 22*(8), 1336-1344. doi:10.1038/s41593-019-0428-x

Riggall, A. C., & Postle, B. R. (2012). The Relationship between Working Memory Storage and Elevated Activity as Measured with Functional Magnetic Resonance Imaging. *Journal of Neuroscience, 32*(38), 12990-12998. doi:10.1523/Jneurosci.1892-12.2012

Roelfsema, P. R., & de Lange, F. P. (2016). Early Visual Cortex as a Multiscale Cognitive Blackboard. *Annu Rev Vis Sci, 2*, 131-151. doi:10.1146/annurev-vision-111815-114443

Samaha, J., Sprague, T. C., & Postle, B. R. (2016). Decoding and Reconstructing the Focus of Spatial Attention from the Topography of Alpha-band Oscillations. *J Cogn Neurosci, 28*(8), 1090-1097. doi:10.1162/jocn_a_00955

Samaha, J., Switzky, M., & Postle, B. R. (2019). Confidence boosts serial dependence in orientation estimation. *J Vis, 19*(4), 25. doi:10.1167/19.4.25

Serences, J. T. (2016). Neural mechanisms of information storage in visual short-term memory. *Vision Res, 128*, 53-67. doi:10.1016/j.visres.2016.09.010

Serences, J. T., Ester, E. F., Vogel, E. K., & Awh, E. (2009). Stimulus-specific delay activity in human primary visual cortex. *Psychol Sci, 20*(2), 207-214. doi:10.1111/j.1467-9280.2009.02276.x

Sprague, T. C., Adam, K. C. S., Foster, J. J., Rahmati, M., Sutterer, D. W., & Vo, V. A. (2018). Inverted Encoding Models Assay Population-Level Stimulus Representations, Not Single-Unit Neural Tuning. *eNeuro, 5*(3). doi:10.1523/ENEURO.0098-18.2018

van Ede, F., Chekroud, S. R., Stokes, M. G., & Nobre, A. C. (2019). Concurrent visual and motor selection during visual working memory guided action. *Nat Neurosci, 22*(3), 477-483. doi:10.1038/s41593-018-0335-6

van Kerkoerle, T., Self, M. W., & Roelfsema, P. R. (2017). Layer-specificity in the effects of attention and working memory on activity in primary visual cortex. *Nat Commun, 8*, 13804. doi:10.1038/ncomms13804

Wang, L., Mruczek, R. E., Arcaro, M. J., & Kastner, S. (2015). Probabilistic Maps of Visual Topography in Human Cortex. *Cereb Cortex, 25*(10), 3911-3931. doi:10.1093/cercor/bhu277

Watson, A. B., & Pelli, D. G. (1983). QUEST: a Bayesian adaptive psychometric method. *Percept Psychophys, 33*(2), 113-120. doi:10.3758/bf03202828

Yu, Q., Panichello, M. F., Cai, Y., Postle, B. R., & Buschman, T. J. (2020). Delay-period activity in frontal, parietal, and occipital cortex tracks noise and biases in visual working memory. *PLoS Biol, 18*(9), e3000854. doi:10.1371/journal.pbio.3000854

Yu, Q., & Shim, W. M. (2017). Occipital, parietal, and frontal cortices selectively maintain task-relevant features of multi-feature objects in visual working memory. *Neuroimage, 157*, 97-107. doi:10.1016/j.neuroimage.2017.05.055

Yu, Q., & Shim, W. M. (2019). Temporal-Order-Based Attentional Priority Modulates Mnemonic Representations in Parietal and Frontal Cortices. *Cereb Cortex, 29*(7), 3182-3192. doi:10.1093/cercor/bhy184

Yu, Q., Teng, C., & Postle, B. R. (2020). Different states of priority recruit different neural representations in visual working memory. *PLoS Biol, 18*(6), e3000769. doi:10.1371/journal.pbio.3000769

Figures and legends

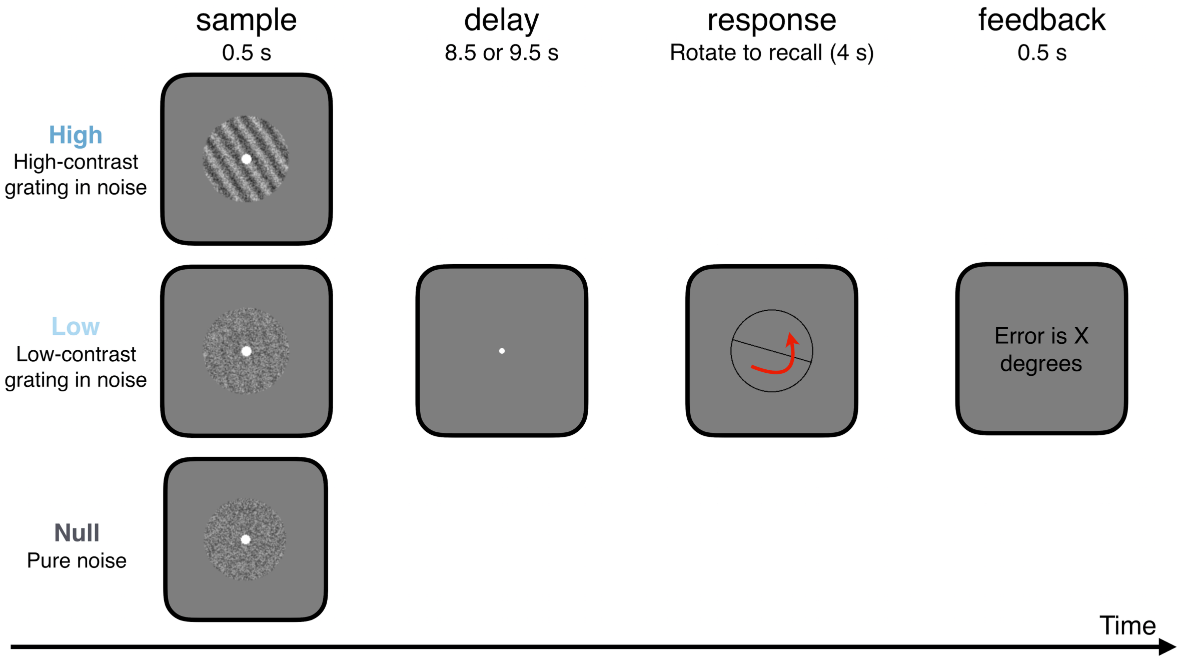

Figure 1. Trial sequence of the fMRI task

Participants performed a 1-item delayed recall task on oriented gratings. On different trials, they viewed a high-contrast (60%) grating embedded in noise, a low-contrast (10%) grating embedded in noise, or a null-contrast (0%) grating embedded in noise (i.e., pure noise patch). Participants were told orientation information was always presented, despite that the visibility of the gratings might differ. After a prolonged delay (8.5 s for two participants and 9.5 s for fifteen participants), they recalled the remembered orientation on an orientation wheel. Feedback was provided at the end of every response.

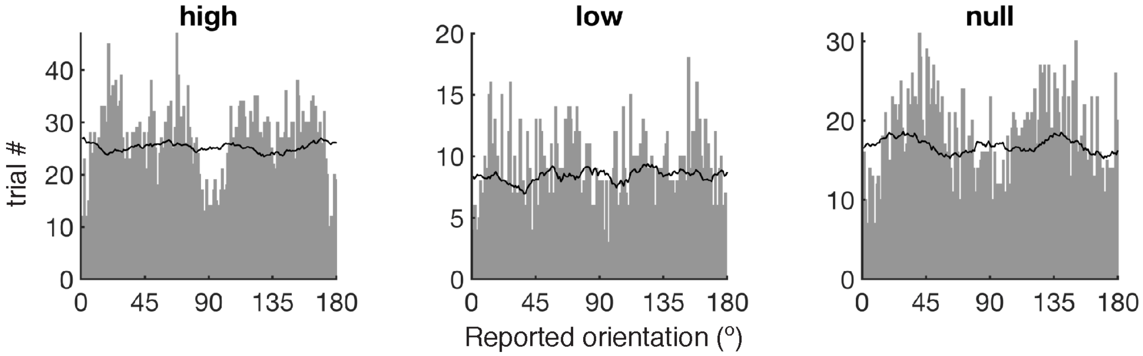

Figure 2. Raw response distribution

The raw response distribution of *high*, *low*, and *null* conditions, indicated by the gray histograms. The black lines indicate the envelope of sample distribution.

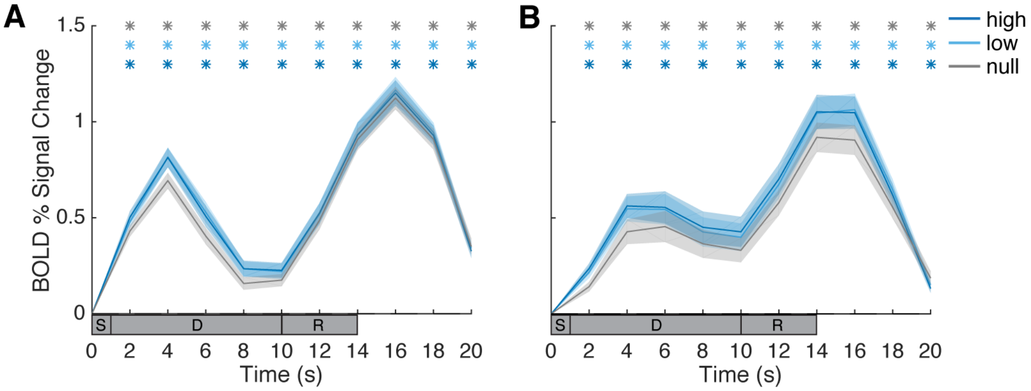

Figure 3. Time course of BOLD activity in early visual cortex and IPS

A. Trial-averaged BOLD activity in the *Early Visual Cortex* ROI. **B**. Time course of BOLD activity in the *IPS* ROI. Dark blue, light blue, and gray lines correspond to the *high*, *low*, and *null* conditions, respectively. Data from the two subjects with 8.5-sec delay periods were included in the averaged results, but event labels below the x-axis represent the trial sequence for subjects with 9.5-sec delay periods only for illustration purposes in this and subsequent figures, with S, D, R representing the Sample, Delay, and Response periods, respectively. Shaded areas indicate ± 1 SEM.

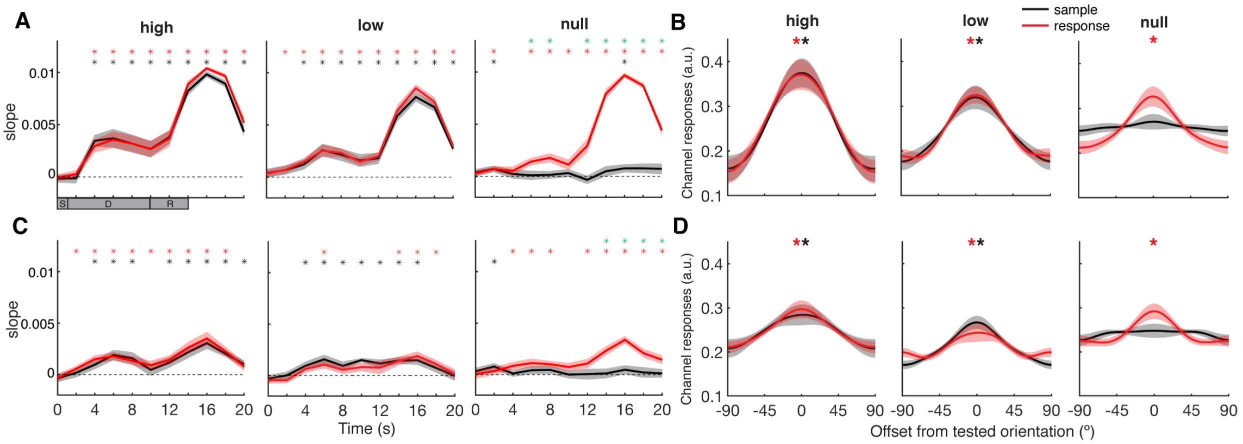

Figure 4. Orientation reconstructions in early visual cortex and IPS in the *high*, *low*, and *null* conditions.

A. Time course of orientation reconstructions (indexed by the slope of reconstruction) in early visual cortex, from left to right: in *high*, *low*, and *null* condition. **B**. Orientation reconstruction in the late delay period (8-10 s after trial onset) in early visual cortex. **C**. Time course of orientation reconstructions (indexed by the slope of reconstruction) in IPS, from left to right: in *high*, *low*, and *null* condition. **D**. Orientation reconstruction in the late delay period in IPS. Black and red lines correspond to sample and response, respectively. Black and red asterisks at the top of each plot denote significance of reconstructions, for sample and response, respectively. Green asterisks denote significance of difference between response and sample reconstructions. Event labels below the x-axis represent the Sample (S), Delay (D), and Response (R) periods, respectively. All *p*-values were corrected with False Discovery Rate (FDR) across conditions and time points. Shaded areas denote ± 1 SEM. *$p < 0.05$.

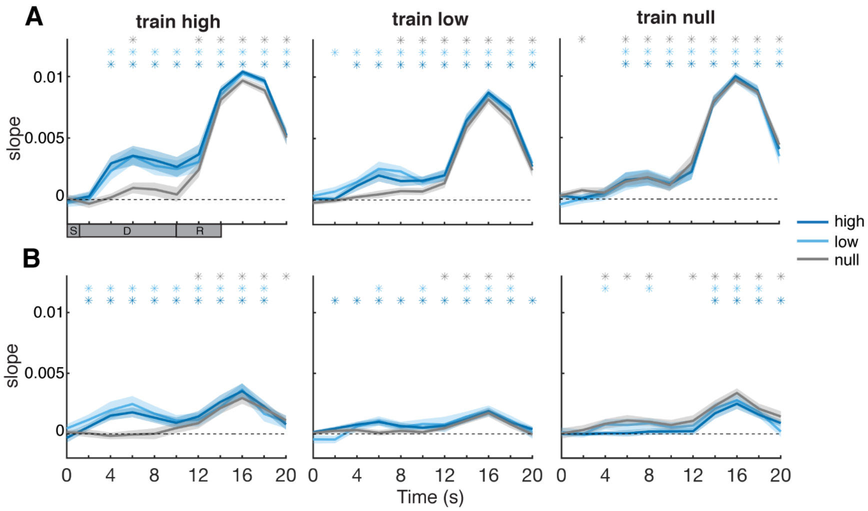

Figure 5. Time course of generalization of neural codes between the *high*, *low* and *null* conditions

A. Time course of orientation reconstructions for *high*, *low*, and *null* conditions in early visual cortex, from left to right: results from *high*-, *low*-, and *null*-trained IEMs. **B**. Time course of orientation reconstructions for *high*, *low*, and *null* conditions in IPS, from left to right: results from *high*-, *low*-, and *null*-trained IEMs. Dark blue, light blue, and gray lines correspond to the *high*, *low*, and *null* conditions, respectively. Dark blue, light blue, and gray asterisks at the top of each plot denote significance of reconstruction at each time point relative to baseline, for *high*, *low*, and *null* conditions, respectively. Event labels below the x-axis represent the Sample (S), Delay (D), and Response (R) periods, respectively. All *p*-values were corrected with False Discovery Rate (FDR) across conditions and time points. Shaded areas indicate ± 1 SEM. $*p < 0.05$.

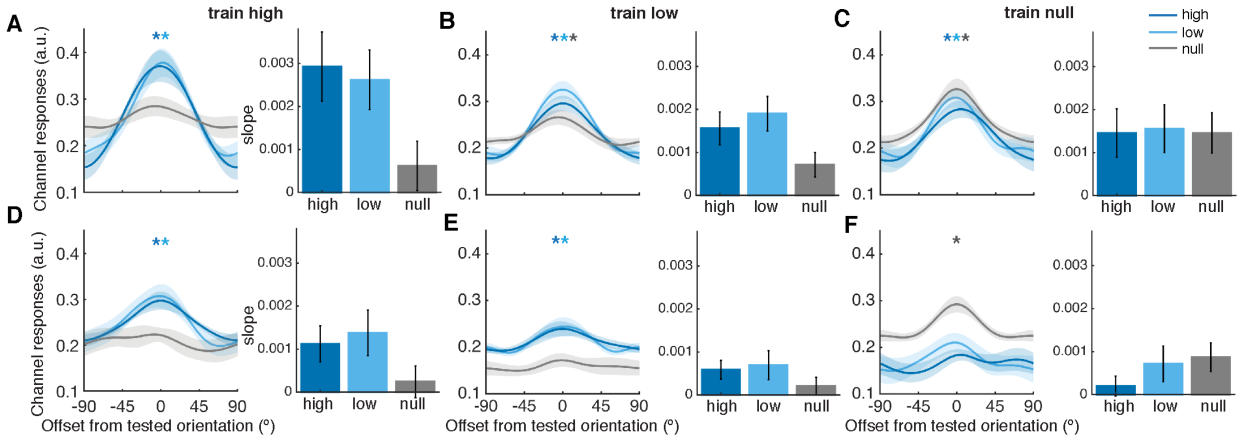

Figure 6. Generalization of neural codes between the *high*, *low* and *null* conditions

A. Left panel: orientation reconstructions for *high*, *low*, and *null* conditions from the *high*-trained IEM, in the late delay period (8-10 s after trial onset) in early visual cortex. Right panel: slopes of these reconstructions. **B**. Orientation reconstructions and slopes for *high*, *low*, and *null* conditions from the *low*-trained IEM, in the late delay period (8-10 s after trial onset) in early visual cortex. **C**. Orientation reconstructions and slopes for *high*, *low*, and *null* conditions from the *null*-trained IEM, in the late delay period (8-10 s after trial onset) in early visual cortex. **D**. Orientation reconstructions and slopes for *high*, *low*, and *null* conditions from the *high*-trained IEM, in the late delay period (8-10 s after trial onset) in IPS. **E**. Orientation reconstructions and slopes for *high*, *low*, and *null* conditions from the *low*-trained IEM, in the late delay period (8-10 s after trial onset) in IPS. **F**. Orientation reconstructions and slopes for *high*, *low*, and *null* conditions from the *null*-trained IEM, in the late delay period (8-10 s after trial onset) in IPS. Dark blue, light blue, and gray colors correspond to the *high*, *low*, and *null* conditions, respectively. Dark blue, light blue, and gray asterisks at the top of each plot denote significance of reconstruction, for *high*, *low*, and *null* conditions, respectively. All p-values were corrected with False Discovery Rate (FDR) across conditions. Shaded areas and error bars indicate ± 1 SEM. *$p < 0.05$.

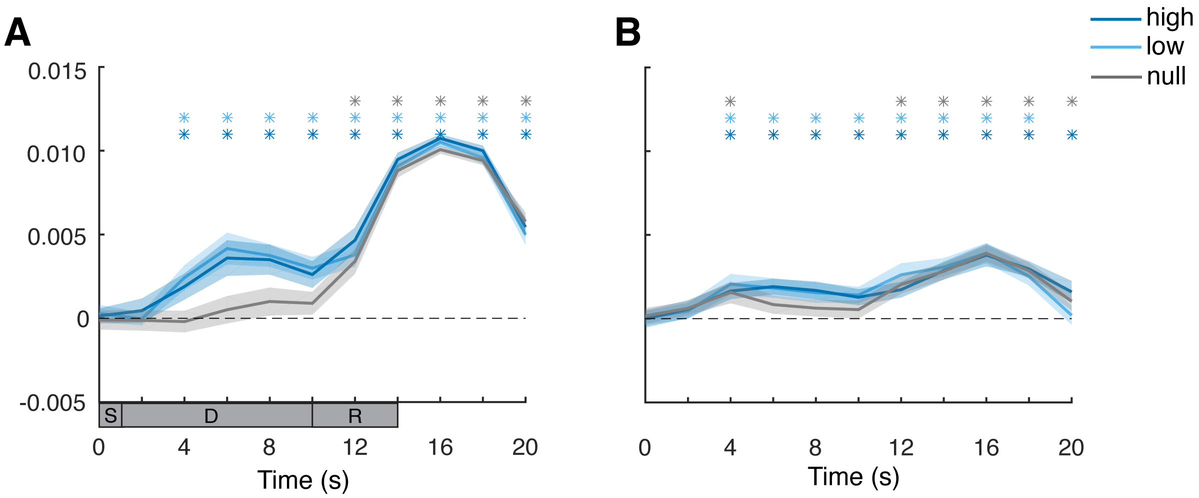

Figure 7. Time course of generalization of neural codes between the *high*, *low* and *null* conditions, using a mixed IEM

A. Time course, in early visual cortex, of orientation reconstructions for *high*, *low*, and *null* conditions using a mixed IEM. **B**. Time course, in IPS, of orientation reconstructions for *high*, *low*, and *null* conditions using a mixed IEM. Dark blue, light blue, and gray lines correspond to the *high*, *low*, and *null* conditions, respectively. Dark blue, light blue, and gray asterisks at the top of each plot denote significance of reconstruction at each time point relative to baseline, for *high*, *low*, and *null* conditions, respectively. Event labels below the x-axis represent the Sample (S), Delay (D), and Response (R) periods, respectively. All *p*-values were corrected with False Discovery Rate (FDR) across conditions and time points. Shaded areas indicate ± 1 SEM. $*p < 0.05$.

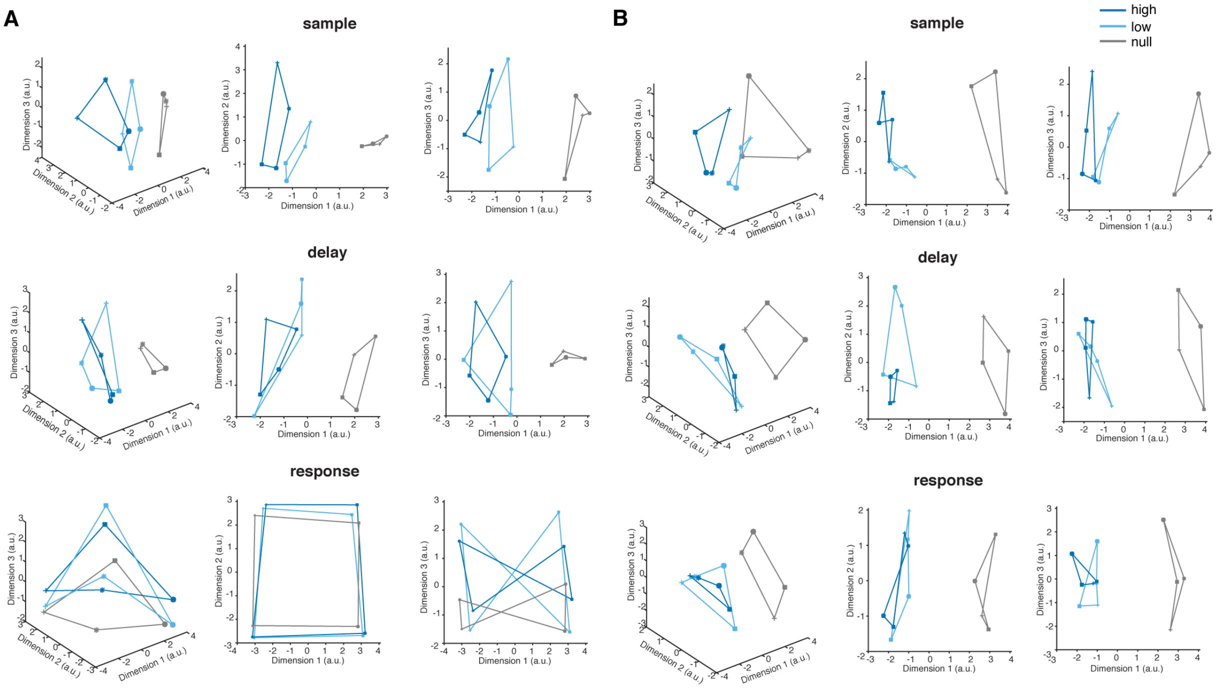

Figure 8. Visualization of distances between conditions in multidimensional representational space

A. Visualization of representational distances between *high*, *low*, and *null* conditions, in early visual cortex. The top panel shows the same MDS projection of data from the sample epoch from three perspectives: 3D; a 2D view of dimension 1 vs. dimension 2; and a 2D view of dimension 1 vs. dimension. The middle panel shows three comparable views of the MDS projection of data from the late delay period, and the bottom panel three comparable views of the MDS projection of data from the response epoch. **B**. MDS analyses of data from IPS, using the same display conventions as A. Each marker represents one of the orientation bins (0-45º, 45-90º, 90-135º, 135-180º). Dark blue, light blue, and gray colors correspond to the *high*, *low*, and *null* conditions, respectively. a.u.: arbitrary units.

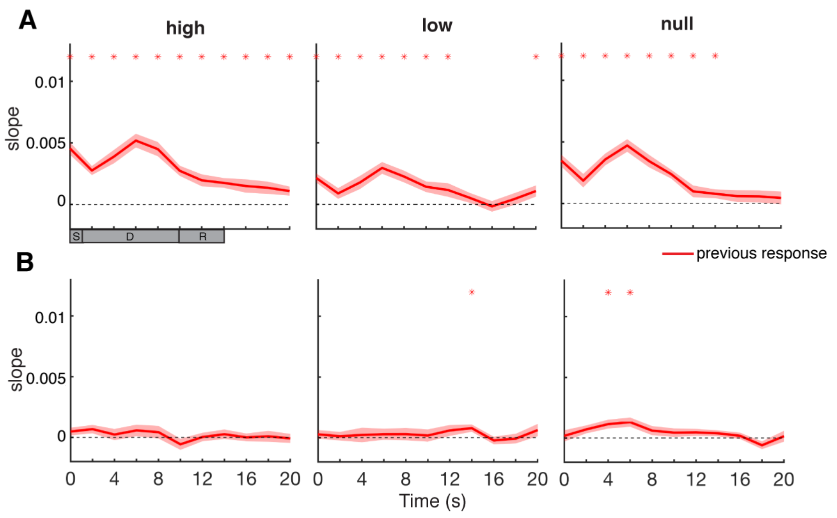

Figure 9. Reconstructions of previous-trial response in early visual cortex and IPS in the *high*, *low*, and *null* conditions.

A. Time course of reconstructions of previous-trial response (indexed by the slope of reconstruction) in early visual cortex; **B**. Time course of reconstructions of previous-trial response in IPS, from left to right: in *high*, *low*, and *null* condition. Red asterisks at the top of each plot denote significance of response reconstructions. Event labels below the x-axis represent the Sample (S), Delay (D), and Response (R) periods, respectively. All *p*-values were corrected with False Discovery Rate (FDR) across conditions and time points. Shaded areas denote ± 1 SEM. *$p < 0.05$.

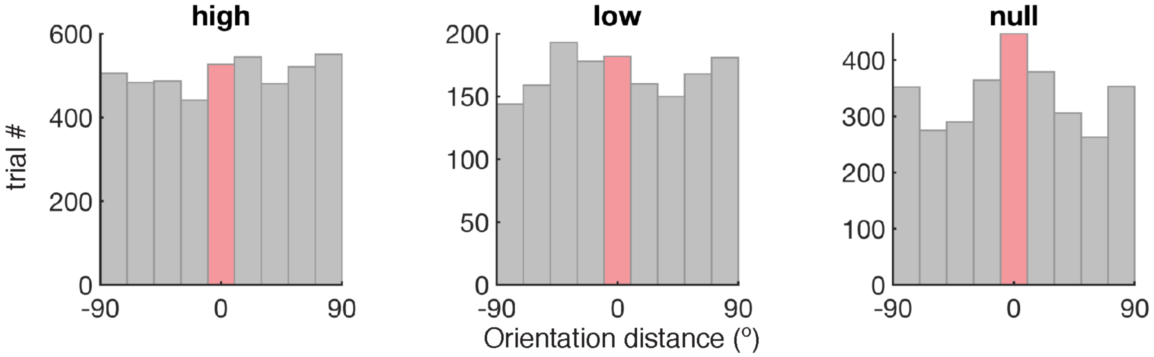

Figure 10. Distribution of the distance between response on the current trial and that on the previous trial

The distribution of the distance between response on the current trial and that on the previous trial, in the *high*, *low*, and *null* conditions, indicated by the gray histograms (bin size = 20º). The red bar at the center indicates the bin with the smallest distance (<= 10º).

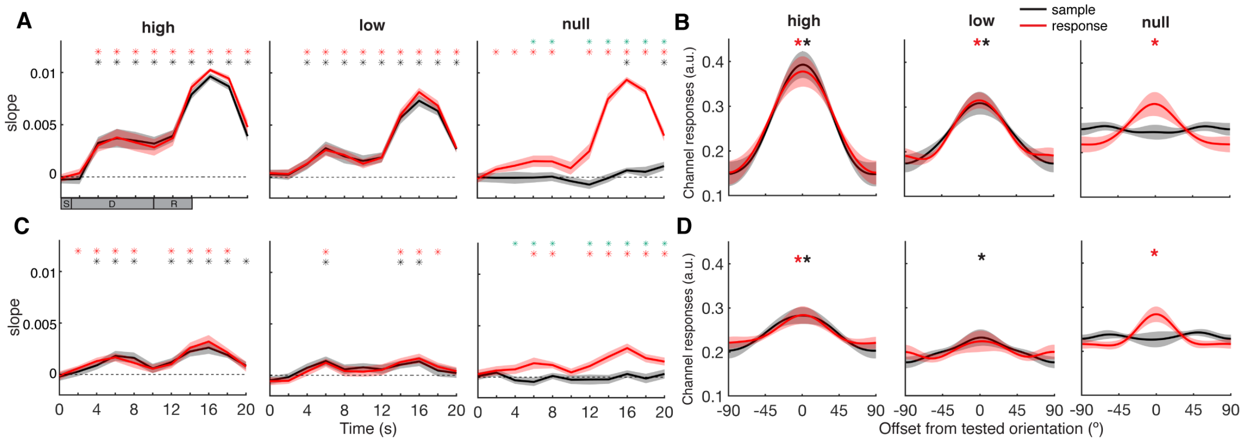

Figure 11. Orientation reconstructions in early visual cortex and IPS in the *high*, *low*, and *null* conditions, after excluding highlighted trials in Figure 10.

A. Time course of orientation reconstructions (indexed by the slope of reconstruction) in early visual cortex, after excluding trials from the bin with the smallest distance in Figure 10, from left to right: in *high*, *low*, and *null* condition. **B**. Orientation reconstruction in the late delay period (8-10 s after trial onset) in early visual cortex. **C**. Time course of orientation reconstructions (indexed by the slope of reconstruction) in IPS, from left to right: in *high*, *low*, and *null* condition. **D**. Orientation reconstruction in the late delay period in IPS. Black and red lines correspond to sample and response, respectively. Black and red asterisks at the top of each plot denote significance of reconstructions, for sample and response, respectively. Green asterisks denote significance of difference between response and sample reconstructions. Event labels below the x-axis represent the Sample (S), Delay (D), and Response (R) periods, respectively. All *p*-values were corrected with False Discovery Rate (FDR) across conditions and time points. Shaded areas denote ± 1 SEM. $*p < 0.05$.