Eye-Movement Reinstatement and Neural Reactivation: Testing the Hebbian Theory of Mental Imagery

Michael B. Bone[1,2,*], Marie St-Laurent[1], Christa Dang[1], Douglas A. McQuiggan[1], Jennifer D. Ryan[1,2], Bradley R. Buchsbaum[1,2]

Summary

Half a century ago, Donald Hebb posited that mental imagery is a constructive process that emulates perception. Specifically, Hebb claimed that visual imagery results from the reactivation of neural activity associated with viewing images. He also argued that neural reactivation and imagery benefit from the re-enactment of eye movement patterns that first occurred at viewing (i.e., fixation reinstatement). To investigate these claims, we applied multivariate pattern analyses to functional MRI (fMRI) and eye tracking data collected while healthy participants repeatedly viewed and visualized complex images. We observed that the specificity of neural reactivation correlated positively with vivid imagery and with detailed memory for the stimulus images. Moreover, neural reactivation correlated positively with fixation reinstatement – even when analyses were constrained to fMRI signal from the occipital lobe. These findings support the conception of mental imagery as a simulation of perception, and provide evidence of the supportive role of eye-movement in neural reactivation.

[1] Rotman Research Institute at Baycrest; Toronto, Ontario, M6A 2E1; Canada
[2] Department of Psychology; University of Toronto; Toronto, Ontario, M5S 1A1; Canada
[*] Lead Contact. Correspondence: michael.bone@mail.utoronto.ca

Introduction

The idea that mental imagery involves the reactivation of neural representations engaged at perception has now been firmly established (e.g. Ishai et al., 2002; Slotnick, Thompson, and Kosslyn, 2005; Polyn et al., 2005; Buchsbaum et al. 2012; Naselaris et al., 2015; Johnson and Johnson, 2014; Cabeza, Ritchey and Wing, 2015). To date, much of the work on the neural basis of visual imagery has examined the phenomenon as if mental images were visual snapshots appearing in their totality to a passive inner observer (but see Chen et al., 2017). However, mental imagery is an active, constructive process (Bartlett, 1932; Hassabis and Maguire, 2009) that is subject to the very kinds of capacity limitations that constrain perception and working memory (Hesslow, 2011), leading some to propose that people engage with mental images in much the same way as they explore the sensory world—using eye-movements to shift the focus of attention to different parts of a mental image (Hebb, 1968; Noton and Stark, 1971; Brandt and Stark, 1997; Richardson and Spivey, 2000; Laeng and Teodorescu, 2002; Johansson, Holsanova and Holmqvist, 2006; Johansson et al., 2012; Wynn et al., 2016). To date, however, there is scant neuroscientific evidence showing that eye-movement patterns are related to the neural representations that support mental imagery for complex visual scenes.

In a seminal paper, Donald O. Hebb (1968) proposed a theory of mental imagery comprising three core claims: 1) imagery results from the reactivation of neural activity associated with the sequential perception of "part-images" (i.e. the constituent, spatially organized, elements of a mental image); 2) analogous to the role of saccades and fixations during

perception, eye movements during imagery temporally organize the neural reinstatement of such "part-images", thereby facilitating imagery by reducing interference between different image parts; and, 3) the vividness and detail of mental imagery is dependent on the order (first-, second-, third-order, etc.) of neuronal cell assemblies undergoing reactivation, such that reactivation extending into lower order regions would elicit greater subjective vividness than reactivation limited to higher-order visual areas.

Hebb's first claim that imagery requires the reinstatement of perceptual neural activity has received considerable empirical support over the last decade. The advent of multi-voxel pattern analysis (MVPA; Haxby, 2012) has especially facilitated the assessment of neural reactivation, which is when stimulus-specific activity patterns elicited at perception are reactivated during retrieval (Rissman and Wagner, 2012; Danker and Anderson, 2010). Researchers have consistently reported substantial similarities between neural regions activated by visual imagery and visual perception (e.g., Ganis, Thompson, and Kosslyn, 2004; Slotnick, Thompson, and Kosslyn, 2005; Polyn et al., 2005), and there is now significant evidence that measures of neural reinstatement reflect the content (Buchsbaum, et al. 2012; Naselaris et al. 2015; Johnson and Johnson, 2014; Cabeza, Ritchey and Wing, 2015; St-Laurent et al., 2014) and vividness (Cui et al. 2007; St-Laurent, Abdi, and Buchsbaum, 2015; Johnson et al. 2015; Dijkstra, Bosch and van Gerven, 2017) of mental imagery.

Hebb's third claim that vivid mental imagery is the result of neural reinstatement within early visual areas (e.g. V1) has also received some neuroscientific support, although evidence is more limited. The hierarchical organization of the visual cortex is well understood (Felleman and Van Essen, 1991; Logothetis and Sheinberg, 1996; Vogels and Orban, 1994; DiCarlo, Zoccolan

and Rust, 2012), but the precise manner in which mental imagery is embedded in this representational structure is still a matter of debate (Kosslyn, Thompson, and Ganis, 2006; Pylyshyn, 2002). Recently, Naselaris and colleagues (2015) showed that visualizing an image leads to the activation of low-level visual features specific to that image within early visual areas V1 and V2 (see also Slotnick, Thompson, and Kosslyn, 2005; Thirion et al. 2006). Some tentative evidence that reactivation within early visual areas correlates with the *vividness* of mental imagery has also emerged (e.g., Lee, Kravitz, and Baker, 2012, although the results were grouped together for the striate and extrastriate cortices, leaving the relation between reinstatement within V1/V2 and vividness unresolved).

In contrast to Hebb's other two claims, support for his claim that eye movements facilitate neural reactivation during imagery remains largely at the behavioral level. Research indicates that stimulus-specific spatiotemporal fixation patterns elicited during perception are reinstated during retrieval (Noton and Stark, 1971; Brandt and Stark, 1997; Laeng and Teodorescu, 2002; Altmann, 2004; Laeng at al., 2014)—even in complete darkness (Johansson, Holsanova and Holmqvist, 2006). Furthermore, this phenomenon of fixation reinstatement appears to facilitate mental imagery (Laeng and Teodorescu, 2002; Johansson et al., 2012; Johansson and Johansson, 2014; Laeng at al., 2014)—although some countervailing evidence exists (Richardson and Spivey, 2000; Spivey and Geng, 2001; Martarelli and Mast, 2013). If eye-movements facilitate mental imagery by coordinating shifts of attention to the elements of a remembered visual scene, then it follows that *eye-movement reinstatement* should be associated with *neural reactivation* of distributed memory representations. To date, however, there is little neuroscientific evidence supporting this foundational claim regarding the link between eye

movement and imagery.

The goal of the present study was therefore to examine how neural reactivation evoked during mental imagery was related to concurrently measured eye-movement patterns. To capture neural reactivation and eye-movement reinstatement, we collected functional MRI (*f*MRI) and eye tracking data simultaneously while 17 healthy participants viewed and visualized a set of complex colored photographs. In the encoding (perception) condition, participants were repeatedly shown a set of 14 images identified by a unique title and were instructed to remember them in detail. Participants then visualized these images in the retrieval (mental imagery) condition. While this aspect of the experiment is not the focus of the current report, our paradigm was also designed to examine how recency of stimulus presentation influenced neural reactivation patterns. Thus, each retrieval trial began with a sequence of three images (from the set of 14) shown in rapid succession, followed by a cue (title) that identified an image from the set. Participants visualized the image that matched the title, and then rated the vividness of their mental image (Figure 1, In-Scan Task). The recency of the image to be visualized was manipulated in four conditions: long term memory (LTM), wherein the visualized image was not among the three-image sequence; and working memory 1, 2 and 3 (WM1, WM2, WM3), wherein the visualized image was presented in the first, second or third position in the three-image sequence, respectively. A post-scan task completed immediately after scanning (Figure 1, Post-Scan Task) served as a behavioral measure of memory acuity. As in the in-scan retrieval condition, participants were shown a sequence of three images (from the in-scan stimulus set) in rapid succession, immediately followed by an image from the set that was either intact or modified (Figure 2). Participants were required to determine whether a subtle change had been

made to the image.

We applied MVPA to the *f*MRI signal to quantify the specificity of neural reactivation during mental imagery. We also developed a multivariate spatial similarity analysis which we applied to the eye tracking data to quantify image-specific patterns of fixation reinstatement. Based on Hebb's claim that fixation reinstatement should contribute to neural reactivation, we hypothesized that the two metrics should correlate positively, and that the correlation should be strongest at corresponding retrieval time points (i.e. when comparing fixation reinstatement at retrieval-time *x* with neural reinstatement at retrieval-time *x*). Moreover, we hypothesized that individuals capable of conjuring up detailed memories for stimulus items may rely more heavily on eye-movement. If so, we expected post-scan behavioral memory performance to be consistent with in-scan metrics of fixation reinstatement as well as neural memory representation. We also examined Hebb's claim that reactivation within early visual areas should contribute positively to the vividness of mental imagery. For this, we correlated perceived vividness with neural reactivation in pre-defined visual cortical regions that included the occipital pole and calcarine sulcus. Finally, we assessed whether neural reactivation and fixation reinstatement were affected by how recently the image was presented before visualization; that is, we examined whether these measures were affected by the presence and serial position of the target image within the three-image sequence preceding the visualization period (i.e. the recency condition).

Our results revealed widespread neural reactivation throughout the time period allocated for visualization. Of interest, imagery vividness ratings correlated positively with reactivation in regions that included the occipital lobe, the ventral and dorsal visual cortex, as well as the calcarine sulcus. Of central importance to our study, neural reactivation was found to correlate

positively with fixation reinstatement, and the correlation was strongest when comparing corresponding time points from retrieval trials. To our knowledge, these results provide the first neuroscientific evidence for Hebb's claim regarding the role of eye movement in mental imagery, as well as support for modern theories of fixation reinstatement, which posit a critical role for eye-movements in the facilitation of memory retrieval (Ferreira, Apel, and Henderson, 2008; Richardson et al., 2009).

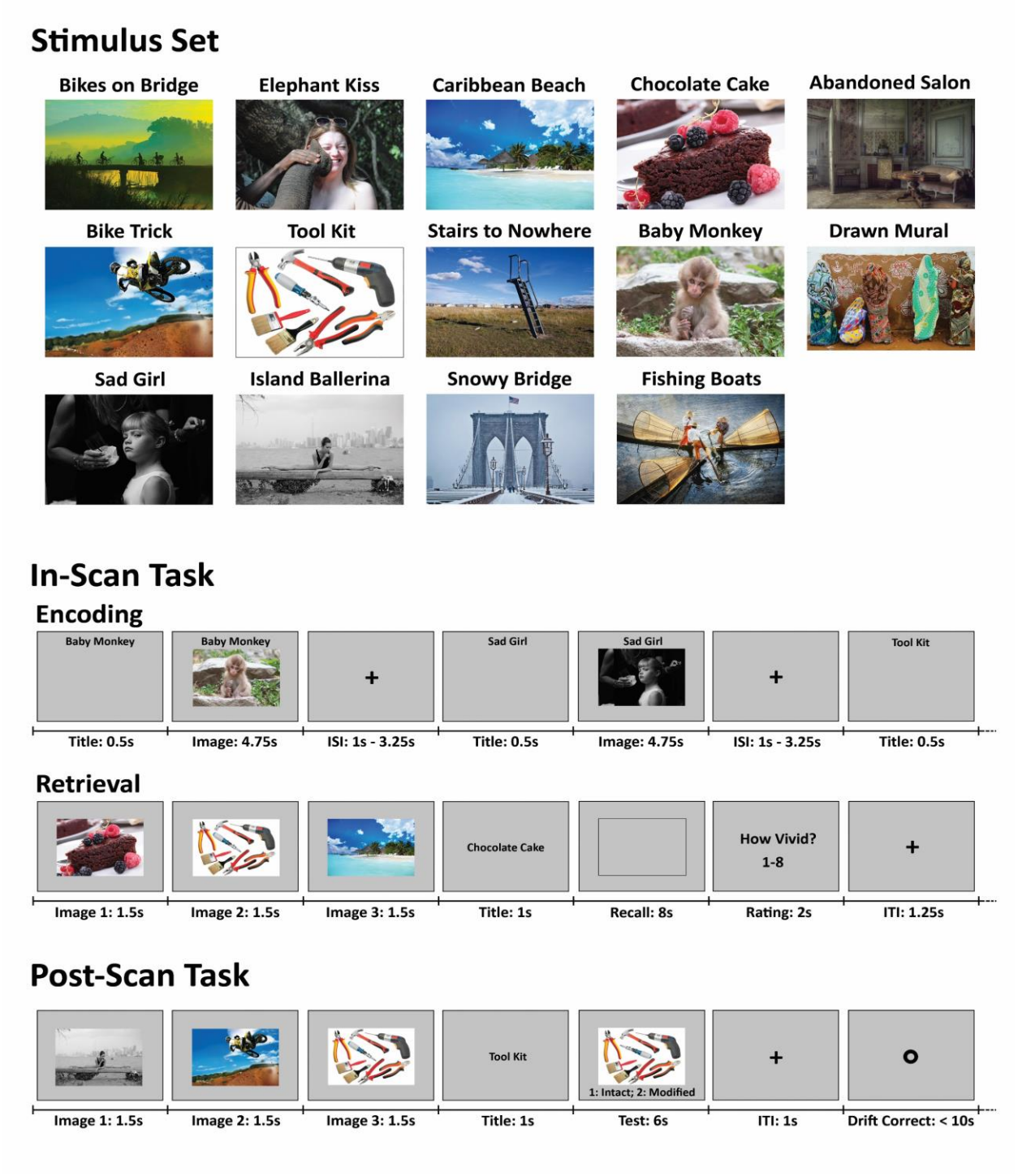

Figure 1. Image Stimuli and Task Procedures. See Methods for an in-depth description of the tasks.

Figure 2. Post-Scan Image Modification. One example of the small modifications that participants were asked to detect during the post-scan behavioral task. Shown images were either modified (right) or identical to the original image (left) held in memory. In this case, the twig the monkey is holding has been lengthened.

Results

Fixation Reinstatement during Image Recall

We first asked whether patterns of eye fixations made during image encoding are reinstated when the image is visualized at recall. We tested this by computing pairwise similarity measures of fixation patterns captured at encoding and recall, which is a form of "representational similarity analysis" (Kriegeskorte, Mur and Bandettini, 2008) applied to eye-movements. Importantly, participants tend to "contract" their patterns of fixations towards the center of the screen during visualization relative to encoding (Brandt and Stark, 1997; Gbadamosi and Zangemeister, 2001; Johansson, Holsanova and Holmqvist, 2011). To account for this tendency, we tested three separate reinstatement measures that we refer to as: "original", "expanded", and "angle". All three measures quantify the similarity of the spatial distribution of fixations between

encoding and retrieval trials by calculating the difference between two-dimensional fixation density maps (Pomplun, Ritter and Velichkovsky, 1996; Wooding, 2002). The measures differ in the spatial transformation applied to the fixations before similarity is calculated: "original" has no transformation, "expanded" increases the distance of retrieval fixations from the center of the image by a constant ratio optimized for each participant, and "angle" radially projects fixations onto a centered circle—effectively removing information relating to the fixations' distance from the center (see Methods and Supplementary Figure S1 for more detail). For all three measures, a value greater than zero indicates some reinstatement, and a value of one indicates perfect reinstatement.

We calculated fixation reinstatement across trials by pooling fixations from all retrieval trials with the same subject-image pair. A random effects bootstrap analysis (see Methods for a description of all bootstrap methods) with subject and item (image) as random effects found all three measures of fixation reinstatement to be significant (Figure 3A). The "angle" measure was found to produce the highest similarity score (~60% greater than the "original" measure), but the difference between measures, calculated with a paired-samples variant of the random effects bootstrap procedure, was not significant after correcting for multiple comparisons using false discovery rate (FDR) (Supplementary Table S2). Nevertheless, the fact that the angle measure demonstrated the numerically greatest sensitivity to fixation reinstatement despite its lower representational capacity (i.e. its one-dimensional representation of fixation position along a centered circle's circumference) provides evidence that it successfully controls for the central contraction of fixations during visualization. Consequently, all subsequent fixation reinstatement calculations reported here were performed using the "angle" measure.

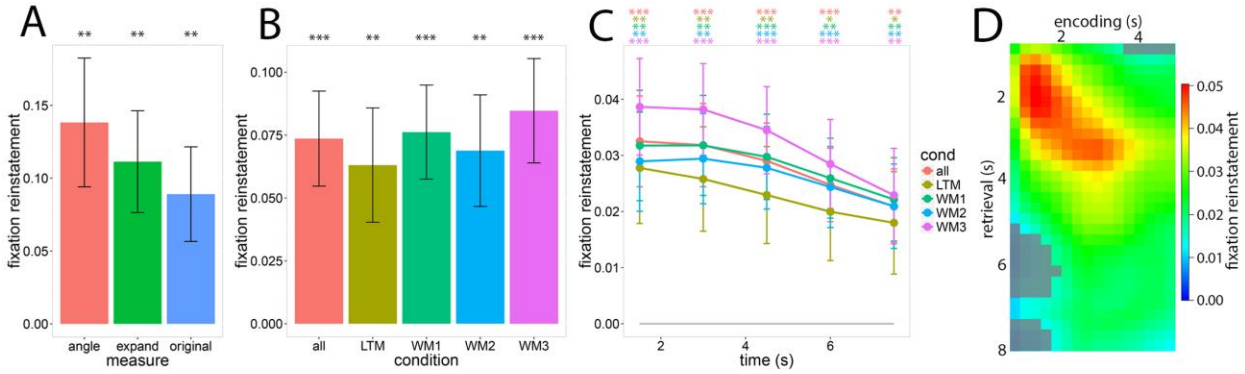

Figure 3. Fixation Reinstatement. Data are represented as mean ± 1 SEM; FDR corrected one-tailed p-value: · < .1, * < .05, ** < .01, *** < .001. **A)** Fixation reinstatement for each measure (expand = expanded). All other fixation reinstatement results are derived from the "angle" measure. **B)** Fixation reinstatement for each recency condition. **C)** Fixation reinstatement for each recency condition divided into retrieval period temporal windows. **D)** Heatmap of fixation reinstatement divided into temporal windows, wherein the columns are encoding windows and the rows are retrieval windows (width=1s, stride=0.25s). Greyed-out regions were not significantly greater than zero ($p > 0.05$; not corrected for multiple comparisons). See also Supplementary Table S1.

To assess the relationship between neural reactivation and fixation reinstatement, we calculated trial-unique fixation reinstatement scores (i.e. a separate fixation reinstatement value was calculated for each retrieval trial rather than for each subject-image combination) based on the "angle" measure. These scores were significantly greater than zero in all recency conditions after FDR correction with a one-tailed alpha of .05 (Figure 3B)**,** providing convincing evidence that participants spatially recreated image-specific fixation patterns during imagery.

We assessed the temporal dynamics of fixation reinstatement by dividing the 8-second visualization period of retrieval trials into 5 evenly divided time-windows (see Methods). Fixation reinstatement was found to be significantly greater than zero at all time points in all recency conditions, after correcting for multiple comparisons using FDR with a one-tailed alpha set to .05 (Figure 3C). A two-way analysis of variance (ANOVA) with repeated measures over

time and recency condition, corrected for departures from sphericity using the Greenhouse-Geisser approach, indicated that the main effect of retrieval time was significant [$F(1.12, 17.85) = 8.44$, $p = .008$]. However, the main effect of recency condition only reached trend level [$F(2.54, 40.57) = 2.78$, $p = .06$], and the interaction effect between recency condition and time was not significant [$F(2.49, 39.65) = 1.15$, $p = .33$].

In the above analyses, all encoding fixations were collapsed over time so that we could examine spatial, but not temporal, eye-movement reinstatement. To examine whether image-specific patterns of eye movement evolved similarly over time during encoding and retrieval, we split encoding and retrieval trials into 16 and 29 temporal windows, respectively, and we computed the similarity between all combinations of encoding and retrieval temporal windows [See Methods; windows had a 1s duration (width) and adjacent windows were offset by .25s (stride)]. As can be seen in Figure 3D, almost all pairings of encoding and retrieval temporal windows show significant similarity ($p < 0.05$, one-tailed; not corrected for multiple comparisons). There also appears to be a ~0.5 second delay in the onset of fixation reinstatement during mental imagery, with reinstatement peaking around 0.75-2.75 seconds for encoding trials and around 1.25-3.25 seconds for retrieval trials. This delay may reflect the time needed to access a mental image from the cue. Importantly, greater reinstatement values seemed to cluster around the approximately 45-degree diagonal trend extending from the top-left corner, suggesting similar progressions of image-specific fixation patterns at encoding and retrieval. To test the significance of these qualitative observations, we examined whether the diagonal bins in the temporal fixation reinstatement analysis showed greater reinstatement than the off-diagonal bins. We tested this spatiotemporal effect with a linear mixed-effects (LME) model with fixation

reinstatement as the dependent variable (DV), encoding and retrieval times as independent variables (IV; 1-16 scalar valued: only the first 16 retrieval time windows were used to match encoding), the absolute difference between encoding and retrieval times as an IV, and participant as a random effect. Statistical assessments were performed using bootstrap analyses. We found that the absolute difference between encoding and retrieval times—a measure of the distance from the diagonal—correlated negatively with fixation reinstatement ($r = -.224$, $p < .0001$), thereby providing empirical evidence for the spatiotemporal reinstatement of fixations during mental imagery (Brandt and Stark, 1997; Laeng and Teodorescu, 2002; Johansson, Holsanova and Holmqvist, 2006).

Neural Reactivation During Image Recall

For the assessment of neural reactivation, we trained a multivariate pattern classifier to discriminate each of the 14 images using data from the three runs during which participants performed the encoding task (Figure 1, In-Scan Encoding). The trained classifier was then applied to the data from the retrieval task (Figure 1, In-Scan Retrieval) to yield a timepoint-by-timepoint estimate of classifier evidence over the course of the visualization window ("cross-decoding"; Kriegeskorte et al. 2011; Lewis-Peacock et al. 2012). Due to the sluggishness of the BOLD response, stimulus-evoked activity associated with the presentation of the three-image sequence is temporally overlapping with activity related to visualizing the cued item (see Figure 1, In-Scan Retrieval), which could artificially inflate classification accuracy. To isolate activity patterns elicited by mental visualization, we therefore calculated an "adjusted classifier confidence". We calculated a baseline classifier performance using trials for which an image shown in a certain position within the three-image sequence (e.g., "Baby Monkey", second

position) was *not* cued for visualization (i.e., a different image was cued). We then subtracted this baseline performance from classifier performance for trials in which that same image shown in the same position *was* the visualized image (See Methods and Supplementary Figure S2).

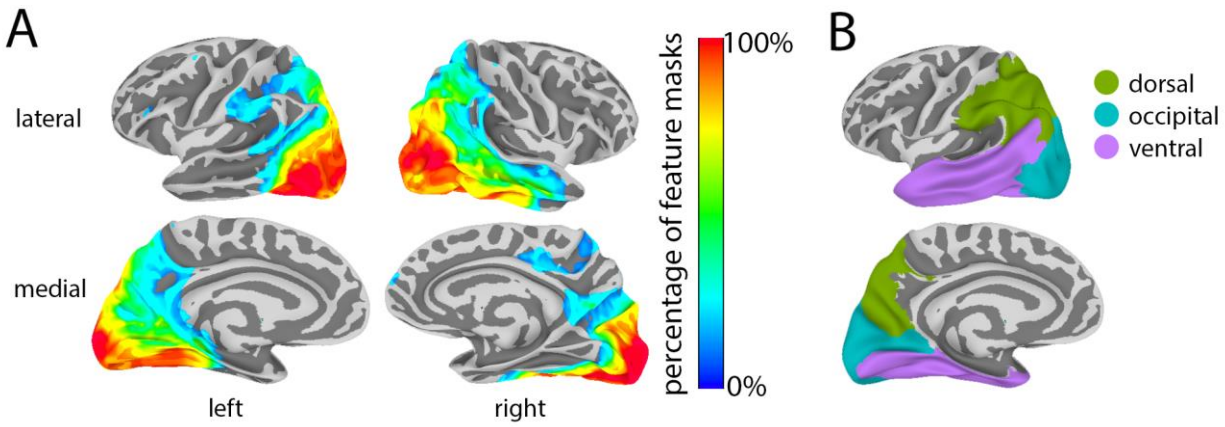

Figure 4. Surface maps of feature and two-stream ROI masks. A) Percentage of subject-specific feature masks that contain each voxel. Thresholded at 10%. **B)** Two-stream ROI masks. See Supplementary Table S5 for a list of the FreeSurfer ROIs that compose each region.

Neural reactivation was calculated across the whole-brain using feature selection (Figure 4A). To separately assess early visual cortex and regions associated with spatial overt attention/eye movements (i.e. intraparietal cortex) while minimizing multiple comparisons, the feature-selected regions were divided into dorsal, occipital, and ventral cortical regions of interest (ROIs) based upon Two-Streams hypothesis (Goodale and Milner, 1992) (Figure 4B; see Methods section for details). Using random effects bootstrap analyses with 10000 random samples over both subjects and items (where each item is one of the 14 stimulus images; see images in Figure 1), average reactivation across the visualization period was found to be significant in all four "recency" (LTM, WM1, WM2, WM3) conditions (Figure 5A, 5B) and within the dorsal, ventral and occipital ROIs (Figure 5C) after correcting for multiple

comparisons using FDR with alpha set to .05 (one-tailed). This result is consistent with previous findings of neural reactivation during mental imagery (Buchsbaum et al. 2012; St-Laurent et al. 2015; Johnson and Johnson, 2014; Cabeza, Ritchey and Wing, 2015) in the context of a novel paradigm. However, reinstatement within the early visual cortex, represented by the calcarine sulcus and occipital pole, was not significant when averaged across the eight second visualization period (Supplementary Table S1).

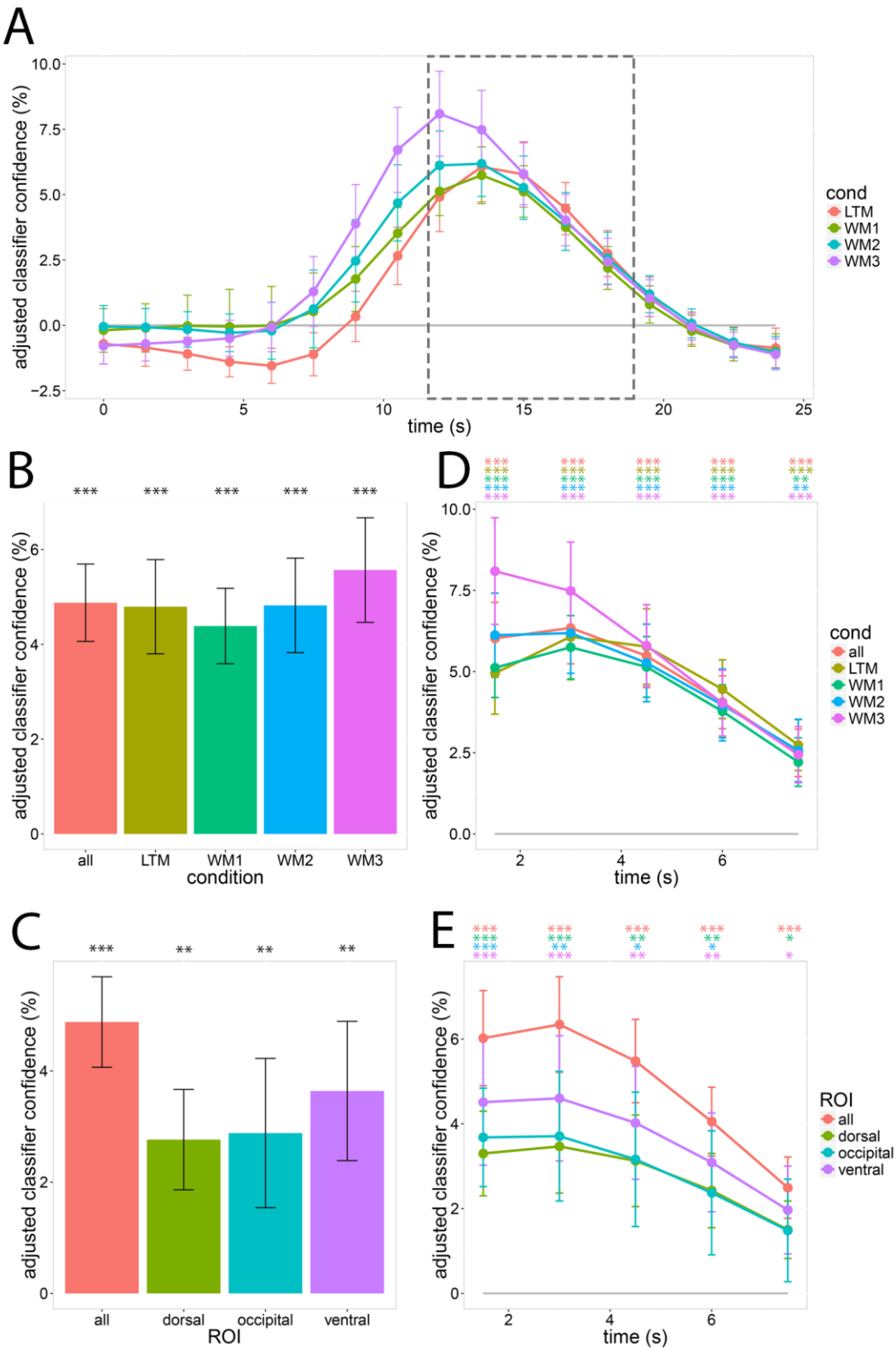

Figure 5. Neural Reinstatement. Data are represented as mean ± 1 SEM; FDR corrected one-tailed p-value: · < .1, * < .05, ** < .01, *** < .001. **A)** Adjusted classifier confidence—a measure of neural reinstatement—for each recency condition over time. The visualization period, offset by 6 seconds to account for hemodynamic response delay, is indicated by the grey dashed box. **B)** Visualization period neural reinstatement for each recency condition. The "all" category, which was included in multiple graphs as a point of reference, refers to the full-brain measure which included all four levels of recency. **C)** Visualization period neural reinstatement for each ROI. As in B, the "all" category refers to the full-brain measure which included all four recency conditions. **D)** Neural reinstatement for each recency condition divided into visualization period temporal windows. **E)** Neural reinstatement for each ROI divided into visualization period temporal windows. See also Supplementary Table S2.

Figures 5D and 5E illustrate the neural reactivation results across the eight second visualization period. Neural reactivation at all time points, recency conditions and ROIs was found to be significantly greater than zero after correcting for multiple comparisons, except for the fifth occipital time point (6-7.5 sec). For the ROI (dorsal, occipital, ventral) data, we performed a three-way (ROI by recency by time) repeated-measures ANOVA. The main effects of retrieval time ($F(1.36, 21.74) = 35.25$, $p < .0001$), ROI ($F(1.87, 29.88) = 3.60$, $p = .0425$), and the recency by time interaction ($F(3.08, 49.22) = 4.01$, $p = .0119$) were significant, while all other effects were not (all *p*s > .15). The significant time effect supports the qualitative observation of a decrease in neural reactivation over time (Figure 5D and 5E), while the recency by time interaction appears to be the result of greater early neural reactivation in the working memory conditions—particularly WM3. Together, these analyses confirm that visual imagery results from the reactivation of perceptual neural activity, and suggest that reactivation is a dynamic process affected by the recency of the visualized image's presentation.

Relation between Neural Reactivation and Fixation Reinstatement

Having observed that neural reactivation and fixation reinstatement were both present during our

imagery task, we then examined the relationship between the two phenomena. To calculate the correlation between neural reactivation and fixation reinstatement, it was necessary to model several fixed and random factors—including participant, recency condition (LTM, WM1, etc.), recalled image, and recall number (the number of times the current trial's target image had been previously recalled)—so we used an LME model. In an analysis of the data from all retrieval trials, we modeled neural reactivation (trial-specific adjusted classifier performance) as a dependent variable (DV), fixation reinstatement (trial-specific fixation map similarity measure) and recall number as scalar independent variables (IV), recency condition as a categorical IV, and participant and image as crossed random effects (random-intercept only, due to model complexity limitations). Statistical assessments were performed using bootstrap analyses.

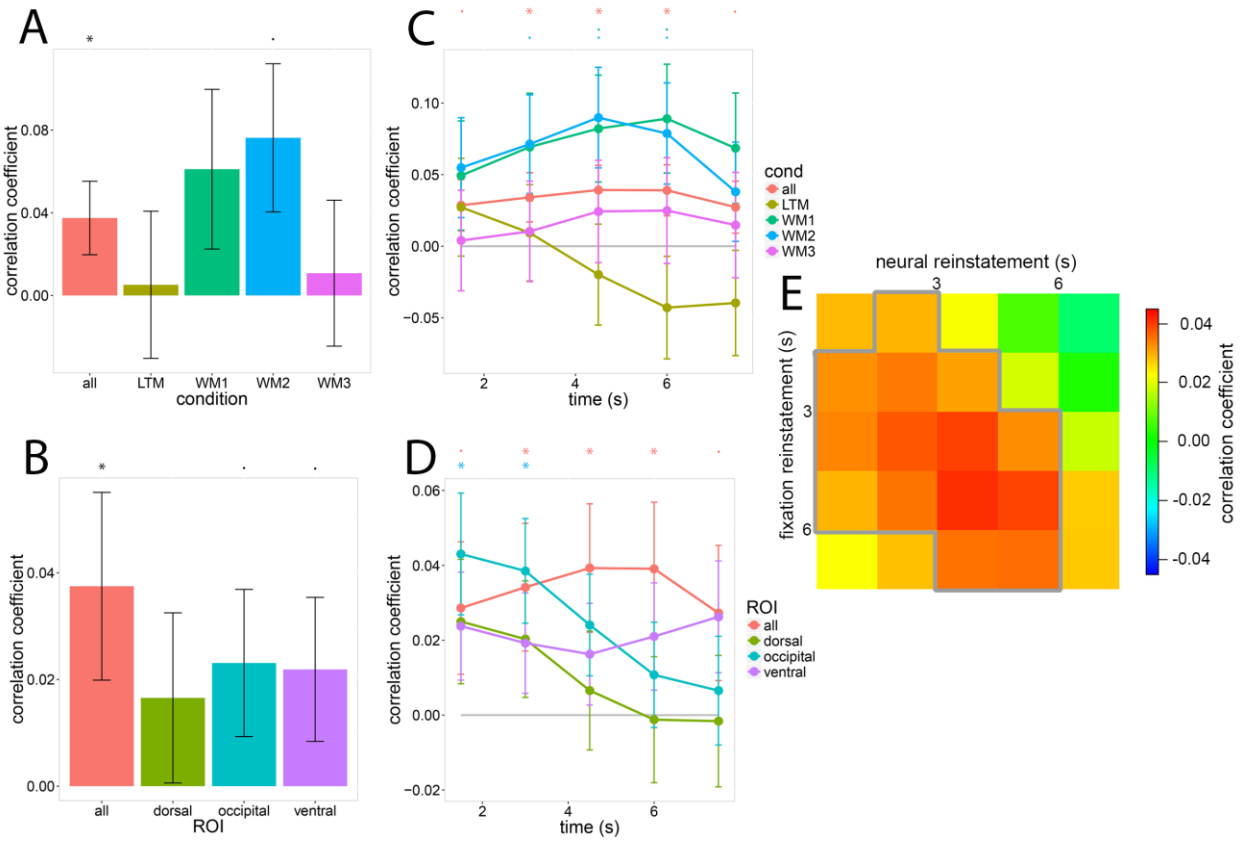

Figure 6. Correlation Between Fixation Reinstatement and Neural Reactivation. Data are represented as correlation coefficient ± 1 SE; FDR corrected one-tailed p-value: · < .1, * < .05, ** < .01, *** < .001. **A)** The correlation between fixation reinstatement and neural reactivation for each recency condition. The "all" category, which was included in multiple graphs as a point of reference, refers to the full-brain measure that included all recency conditions. **B)** The correlation between fixation reinstatement and neural reactivation for each ROI. **C)** The correlation between fixation reinstatement and neural reactivation for each recency condition divided into retrieval-period temporal windows. **D)** The correlation between fixation reinstatement and neural reactivation for each ROI divided into retrieval-period temporal windows. **E)** The correlation between fixation reinstatement and neural reactivation divided into retrieval-period temporal windows, wherein the columns are neural reactivation windows and the rows are fixation reinstatement windows. The grey border surrounds windows that were significantly greater than zero ($p > 0.05$; not corrected for multiple comparisons). See also Supplementary Table S3.

Figures 6A and 6B illustrate the correlation between fixation reinstatement and neural reactivation. After correcting for multiple comparisons (FDR with one-tailed alpha set to .05), fixation reinstatement correlated significantly with reactivation within the feature-selected full-brain when trials from all recency conditions were included (the "all" measure). Correlations specific to recency conditions or limited to signal from specific ROIs were not significant. The results for the "all" measure is consistent with Hebb's claim that eye movements facilitate the neural reinstatement of part-images during mental imagery—offering the first direct evidence of a link between fixation reinstatement and neural reactivation.

Figures 6C and 6D show the correlation between fixation reinstatement and neural reactivation across the eight-second visualization period. For the feature-selected full-brain correlation including all recency conditions, labeled "all", the correlation peaked approximately in the middle of the visualization period with the second, third and fourth time windows significantly greater than zero. Correlations specific to ROIs and recency conditions were not significantly greater than zero in all time points after correcting for multiple comparisons, except for the first (0-1.5 sec) and second (1.5-3 sec) occipital time points (occipital 0-1.5 sec: r = .043,

p = .03; occipital 1.5-3 sec: r = .039, p = .03). No significant effects were uncovered by a three-way (ROI by recency condition by time) repeated-measures ANOVA performed on the ROI-specific (dorsal, occipital, ventral) correlation data (all *p*s > .11).

Figure 6E shows the relationship between fixation reinstatement and full-brain neural reactivation over time. If eye movements during imagery temporally organize the neural reinstatement of part-images, we hypothesized that the correlation between fixation reinstatement and neural reactivation would be strongest when both measures overlapped in time, i.e. neural reactivation at time *x* should correlate most strongly with fixation reinstatement at time *x*. Qualitatively, the diagonal trend from top-left to bottom-right in figure 6E supports this hypothesis. To test this observation, we first calculated separate correlations between fixation reinstatement and neural reactivation for each time-point combination and each participant. Each correlation was calculated using the LME approach described above, with the exception that participant was not included as a random effect. We then performed an LME analysis with correlations between fixation reinstatement and neural reactivation as the DV, fixation reinstatement time and neural reactivation time (1-5 scalar valued) as IVs, the absolute difference between fixation reinstatement and neural reactivation times as an IV, and participant as a random effect. Statistical assessments were performed using bootstrap analyses. We found that the absolute difference between fixation reinstatement and neural reactivation times correlated negatively with the correlation between fixation reinstatement and neural reactivation (r = -.096, p = .038). In other words, fixation reinstatement and neural reactivation measures were more consistent with each other when taken from time bins that were closer in time, indicating a temporal relationship between the two measures.

Post-Scan Memory Task Performance and Vividness Ratings

The final analyses investigated post-scan memory task performance, vividness ratings and their relation to neural reactivation and fixation reinstatement. Our goal was two-fold: 1) assess whether trials that received high vividness ratings (a subjective measure of imagery) also ranked highly on fixation reinstatement and neural reactivation measures, and 2) determine whether individuals with more detailed memories (those who performed better on the post-scan behavioral memory test) had more specific memory representations (as revealed by in-scan neural reactivation) and relied more heavily on eye-movement recapitulation during imagery.

The post-scan memory task was designed to be difficult, but participants performed above chance, with each individual providing more correct than incorrect answers (% correct: mean = 64.8, p(less than or equal to chance at 50%) < .0001; statistics calculated with bootstrap analyses). To determine whether individuals with good post-scan memory performance (% correct) also obtained high fixation reinstatement and neural reactivation scores, we first computed average fixation reinstatement and neural reactivation scores for each participant, and then we correlated these values with the participants' memory performance. We covaried out head motion using the maximum displacement (mm) for each subject within the scanner using standard multiple regression. Bootstrap analyses were used to calculate the statistics. Post-scan memory performance correlated strongly with neural reactivation ($r = .624$, $p = .0003$, one-tailed), and correlated moderately with fixation reinstatement, although this result did not reach significance (correlation coefficient = .401, $p = .096$, one-tailed).

Given previous findings of a positive correlation between vividness ratings and neural

reinstatement (Lee, Kravitz, and Baker, 2012; St-Laurent, Abdi, and Buchsbaum, 2015), we set out to replicate these results, and also to assess whether fixation reinstatement correlated with vividness in the same manner. The within-subject correlations were calculated with a LME model on data from all retrieval trials, wherein either neural reactivation or fixation reinstatement was the dependent variable (DV). Vividness rating and recall number were entered as scalar independent variables (IV), recency condition was a categorical IV, and participant and image were crossed random effects (random-intercept only, due to model complexity limitations). Statistical assessments were performed using bootstrap analyses. Consistent with previous findings, vividness ratings (1-8 scale wherein 1 is very-low and 8 is very-high; mean = 5.57, SD = 1.42) correlated positively with full-brain measures of neural reinstatement (Figure 7A), indicating that image-specific patterns of neural reactivation—an index of memory representation—were more specific during trials perceived as more vivid by the participants. Vividness also correlated with reactivation within the ventral, dorsal and occipital ROIs (Figure 7A and 7C). A two-way (ROI by time) repeated-measures ANOVA revealed that the effects of ROI, time and their interaction were not significant (ROI: $F(1.81, 28.90) = .86$, $p = .42$; time: $F(2.23, 17.87) = 2.46$, $p = .13$; ROI-time interaction: $F(2.04, 32.71) = 2.39$, $p = .11$). Against our hypothesis, no significant positive correlation was observed between vividness and fixation reinstatement ($r = .018$, $p = .21$, one-tailed).

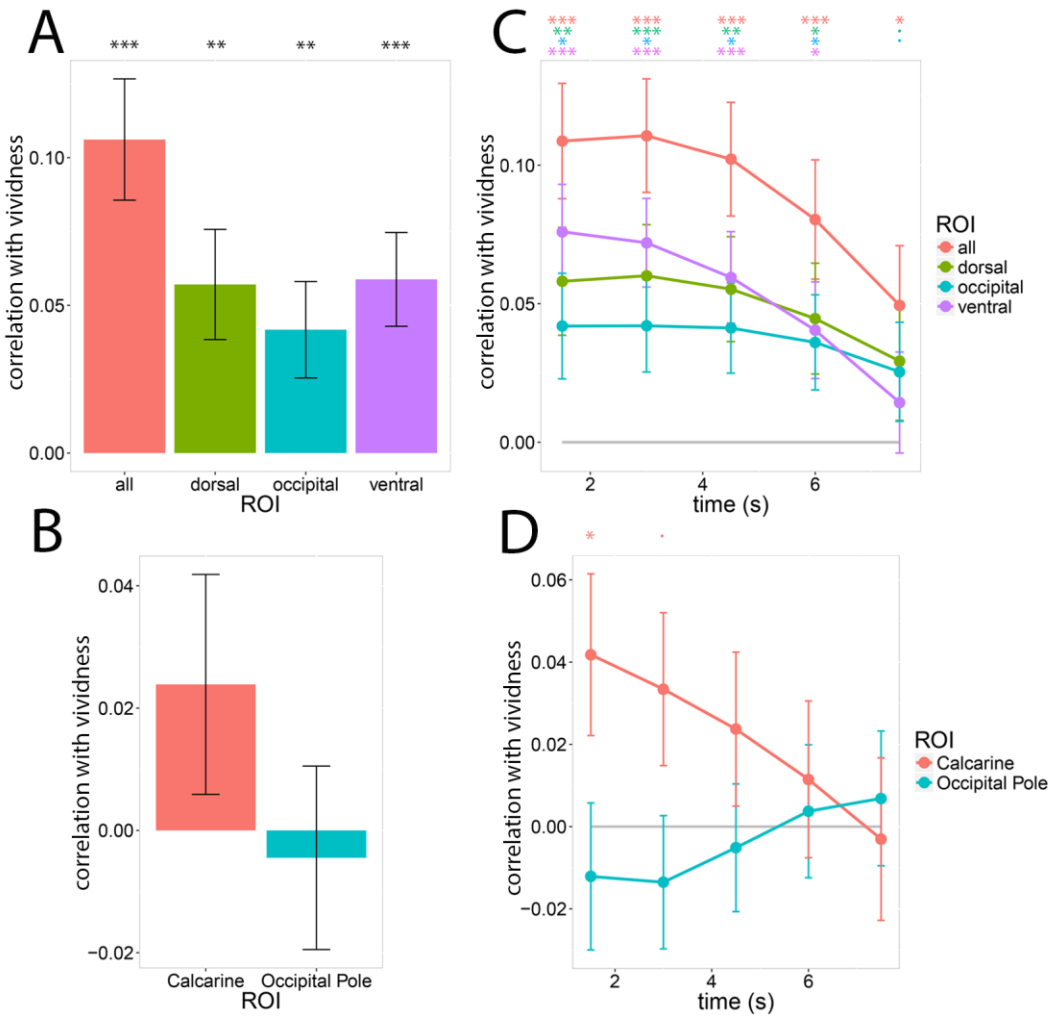

Figure 7. Correlation Between Vividness Rating and Neural Reactivation. Data are represented as correlation coefficient ± 1 SE; FDR corrected one-tailed p-value: · < .1, * < .05, ** < .01, *** < .001. **A)** The correlation between vividness rating and neural reactivation for each ROI. The "all" category refers to the full-brain measure which included all recency conditions. **B)** The correlation between vividness rating and neural reactivation for the calcarine sulcus and occipital pole. **C)** The correlation between vividness rating and neural reactivation for each ROI divided into retrieval-period temporal windows. **D)** The correlation between vividness rating and neural reactivation for the calcarine sulcus and occipital pole divided into retrieval-period temporal windows. Multiple comparison correction was applied sequentially, starting at the first time-point. See also Supplementary Table S4.

We also tested Hebb's claim that neural reactivation in early visual areas elicited more vivid visual mental imagery. Looking specifically at the signal from early visual ROIs, namely

the occipital pole and calcarine sulcus, we found no significant correlations between reactivation and vividness after FDR correction (Figure 7B and 7D). A two-way (ROI by time) repeated-measures ANOVA revealed that the effects of ROI, time and their interaction were non-significant (ROI: $F(1, 16) = 1.90$, $p = .18$; time: $F(1.46, 23.37) = 1.25$, $p = .29$; ROI by time interaction: $F(1.33, 21.33) = 2.00$, $p = .16$). Because neural reinstatement decreased approximately linearly over retrieval time (Figure 5), an ANOVA—which does not assume any relation between time points—may be underpowered. To address this issue, we ran an LME model that assumed a linear relation between time points. In this model, the correlation between vividness and neural reinstatement, calculated for each subject-ROI-time combination, was the DV; ROI [calcarine = 1, occipital pole = -1], time and their interaction were IVs; and participant was a random effect. The main effect of ROI and the ROI-time interaction were significant, indicating that the correlation between neural reinstatement and vividness was significantly stronger within the calcarine sulcus than the occipital pole—particularly near the start of the visualization period (ROI: coefficient = .183, $p = .0058$; time: coefficient = -.109, $p = .09$; ROI-time interaction: coefficient = -.146, $p = .024$; calculated via bootstrap analyses). Based upon this finding, we re-analyzed the correlation between vividness and neural reactivation over time by assessing each time window sequentially, starting from the beginning of the visualization period and including all previous time windows in a multiple comparison analysis using FDR. Using this method, we found the first (0-1.5 sec) visualization time window for the calcarine sulcus to be significant (0-1.5 sec: $r = .041$, $p = .03$, one-tailed), whereas all other windows for both ROIs were not significant.

To determine whether these results were limited by our ability to detect neural

reactivation in these early visual regions, we assessed neural reactivation over time within the occipital pole and calcarine sulcus. We performed random effects (subjects and items) bootstrap analyses for each retrieval time point—controlling for multiple comparisons using sequential FDR. Only the first visualization time window (0-1.5 sec) was found to be significant for the calcarine sulcus (0-1.5 sec: adjusted classifier confidence (%) = 1.51, p = .03, one-tailed), mirroring the correlation results.

These results document the spatiotemporal relationship between neural reactivation and the perceived vividness of mental images. While we observed significant correlations between vividness and reactivation across the visual cortex, we found limited evidence in support of Hebb's claim that reactivation in early visual cortices leads to vivid mental imagery. That being said, our capacity to detect reactivation in early visual cortices was hindered by methodological limitations which we address in the discussion.

Discussion

Neural Reactivation and Eye-Movement Reinstatement

The primary goal of the current study was to test whether eye movements contribute to the creation of mental images by examining the relationship between fixation reinstatement and neural reactivation. Our results provide significant evidence in favor of Hebb's (1968) claim that visual imagery is sequentially constructed in a manner analogous to vision, and that eye movements help coordinate the construction of mental images. We observed a significant positive correlation between a novel measure of fixation reinstatement that accounts for the

contraction of fixations during imagery, and neural reactivation. This correlation increased when fixation reinstatement and neural reactivation metrics were calculated for time points that were closer in time, demonstrating that the two phenomena peaked in synchrony, and establishing a link between eye movement and the neural mechanism of mental imagery.

Previous research has only assessed the link between fixation reinstatement and mental imagery using behavioral measures of imagery rather than neural reactivation. For example, Laeng and Teodorescu (2002), and Johansson et al. (2012), found that the degree of fixation reinstatement predicted behavioral performance on an imagery task. Thus, our findings provide the first direct neuroimaging evidence for Hebb's claim and the currently dominant fixation reinstatement theories (Brandt and Stark, 1997; Spivey and Geng, 2001; Ferreira, Apel, and Henderson, 2008; Richardson et al., 2009).

Our analyses also addressed the relationship between fixation reinstatement, neural reactivation and behavioral memory performance. Based on findings linking fixation reinstatement (Laeng and Teodorescu, 2002; Johansson et al., 2012) and neural reactivation (Johnson and Johnson, 2014; Cabeza, Ritchey and Wing, 2015; St-Laurent et al., 2014) to memory performance, we predicted that both in-scan fixation reinstatement and neural reactivation would correlate with performance on the post-scan memory task. We also predicted similar patterns of correlations with in-scan ratings of imagery vividness. Our results were partially congruent with these predictions. We observed that neural reactivation correlated strongly with both objective and subjective behavioral measures of memory performance, but that fixation reinstatement was a poor predictor of either form of behavior.

Research into the relationship between fixation reinstatement and memory acuity has been mixed (Hale and Simpson, 1971; Richardson and Spivey, 2000; Spivey and Geng, 2001; Laeng and Teodorescu, 2002; Johansson et al. 2012; Laeng at al., 2014). For example, when fixations were constrained to a region that either did or did not correspond to the previous location of objects to be recalled, Johansson and Johansson (2014) found that memory performance was superior in the "corresponding" condition, whereas Martarelli and Mast (2013) did not. These inconsistent results may be due to differences in the features to be recalled: spatial features (orientation and relative position) in Johansson and Johansson (2014), and primarily non-spatial features (e.g. color) in Martarelli and Mast (2013). Consistent with this interpretation, de Vito et al. (2014) demonstrated that incongruent eye movements preferentially disrupt spatial recollection. Our objective measure of memory performance, the post-scan change detection task, included both spatial (e.g. size, position) and non-spatial (e.g. color, object identity) image modifications. Therefore, we may have observed a larger correlation between in-scan fixation reinstatement and post-scan change detection if the task only had spatial modifications. Similarly, subjective vividness ratings reflected an overall impression of the crispness of the mental image, rather than its spatial features. Another alternative to consider is that our "angle" measure of fixation reinstatement may not have fully control for the contraction of fixations during retrieval relative to encoding—a phenomenon that has been shown to correlate with individual differences in spatial imagery (Johansson, Holsanova and Holmqvist, 2011). The "angle" measure controls for a participant's average contraction of fixations towards the center of the screen, but this correction could only partially control for situations when participants contracted their fixations towards multiple points. Future work should explore using a variant of the "angle" method that can model multiple contraction points per image.

In summary, we provided the first evidence that fixation reinstatement is linked to neural reactivation, thereby supporting one of the pillars of Hebb's theory of imagery, as well as current fixation reinstatement theories which also posit that reciprocal facilitation occurs between fixation reinstatement and internal memory representations (Ferreira, Apel, and Henderson, 2008; Richardson et al., 2009). Nonetheless, we must consider alternative interpretations of our results. It is possible that the observed correlation between fixation reinstatement and neural reactivation was predominantly driven by neural correlates of eye position, rather than mental imagery. Positive correlations based on signal limited to the occipital lobe mitigate against this possibility, as this ROI excludes areas strongly associated with eye movement control, such as the frontal eye fields and posterior intraparietal sulcus (Blanke and Seeck, 2003; Williams and Smith, 2010). Discrepant results across recency conditions also raise the possibility that the correlation between fixation reinstatement and neural reactivation holds only in situations wherein working memory is employed, as opposed to long-term memory. Although no significant differences were found between conditions, a qualitative inspection of the results (Figures 6A and 6C) indicates that fixation reinstatement may only correlate with full-brain neural reactivation in the working memory recency conditions with significant visual interference, i.e. WM1 and WM2, but not WM3 or LTM. Future studies should further investigate how patterns of eye movement and neural reactivation differ between working and long-term memory, and whether they are influenced by interference.

Finally, while our correlational findings reveal a relationship between eye movement and imagery, they cannot conclusively determine the causality and directionality of this relationship. To directly address this unresolved issue, future research could take advantage of the high

temporal resolution of techniques such as magnetoencephalography to link distinct patterns of neural activity to specific portions of seen and imagined complex images. If reciprocal facilitation occurs between fixation reinstatement and internal memory representations (Ferreira, Apel, and Henderson, 2008; Richardson et al., 2009) then neural reactivation should predict, and be predicted by, eye movements towards the location associated with the neural activity pattern.

Imagery Vividness and Reinstatement within Early Visual Cortical Areas

We also tested Hebb's claim that highly vivid mental imagery requires cortical reactivation within early visual areas, i.e. V1 and V2. As such, we hypothesized that reactivation within the occipital pole and the calcarine sulcus would correlate positively with vividness. Consistent with previous findings (Lee, Kravitz, and Baker, 2012; St-Laurent, Abdi, and Buchsbaum, 2015; Johnson et al. 2015), we observed correlations between vividness and reactivation within dorsal, ventral and occipital ROIs that were sustained throughout the retrieval trial. Looking specifically at early visual areas, a significant correlation was observed between vividness ratings and reinstatement within the calcarine sulcus (the brain region wherein V1 is concentrated; DeYoe et al., 1996), but not the occipital pole, in the first 1.5 seconds of visualization.

The simplest explanation for the null result within the occipital pole is that vivid mental images can be conjured up without its contribution to neural reactivation. However, other factors need to be considered. First, St-Laurent, Abdi, and Buchsbaum (2015) found that activity levels within the occipital pole correlated strongly and positively with the perceived vividness of videos mentally replayed from memory, which suggests that this area contributes to the perceived

vividness of mental imagery. Second, we only observed evidence of reactivation during the first 1.5 seconds of visualization within the calcarine sulcus, and not within the occipital pole, mirroring our correlation results. This finding suggests that the observed correlation between reactivation within early visual regions and vividness was limited by our ability to detect reactivation within these regions.

Research by Naselaris et al. (2015) provides strong evidence of reactivation of neural patterns associated with low-level visual features within the early visual cortex during scene imagery. One significant methodological difference between this study and our own is that the authors asked their participants to fixate centrally throughout their task, thereby eliminating the natural eye-movements that occur during mental imagery—which were the explicit focus of our study. This significant constraint on the participants' fixations would have eliminated the variance caused by the image's neural representation shifting across the retinotopically-organized early visual cortex due to eye movements, but at the cost of being able to study the functional role of eye-movements during imagery (Johansson et al., 2012). Note that the occipital pole and posterior calcarine sulcus are predominantly responsible for central vision, which has high spatial resolution, whereas the anterior calcarine sulcus is predominantly responsible for peripheral vision, which has relativity low spatial resolution (DeYoe et al., 1996). Consequently, visual representations within the calcarine sulcus should be less sensitive to eye movements than visual representations within the occipital pole, which is consistent with our results. We therefore suspect that free eye-movements may have caused our null reactivation finding within the occipital pole. By extension, our methods could not adequately quantify the correlation between vividness and reactivation within the occipital pole and calcarine sulcus, and limited our capacity

to test Hebb's claim that early visual cortical reactivation leads to vivid imagery. To preserve ecological validity, future research concerning neural reactivation during mental imagery should avoid artificial constraints on fixations, and instead develop and utilize measures of neural reactivation that explicitly model the effect of eye movements on neural activity within the visual cortex.

Conclusions

In conclusion, the results from this study support the three major claims of the Hebbian theory of mental imagery: 1) imagery involves the reinstatement of perceptual neural activity; 2) reinstatement of fixations during imagery facilitates neural reinstatement; 3) the vividness of mental imagery is associated with reactivation within early visual areas (calcarine sulcus). The findings reported here provide a promising avenue to establish how fixations contribute to the neural processes underlying mental imagery. Future work should clarify the fine-scale temporal relationship between eye-movement reinstatement and memory reactivation in a way that can unravel the causal connection between these interacting neural processes.

Author Contributions

Conceptualization, M.B.B., B.R.B., M.S.; Methodology, M.B.B., B.R.B., M.S.; Software, M.B.B., B.R.B.; Formal Analysis, M.B.B., B.R.B., M.S.; Investigation, M.S., C.D., D.A.M.; Resources, C.D.; Data Curation, M.B.B., B.R.B.; Writing – Original Draft, M.B.B., M.S.; Writing – Review and Editing, M.B.B., B.R.B., M.S., J.D.R., D.A.M.; Visualization, M.B.B., B.R.B., M.S.; Supervision, B.R.B., J.D.R.

Acknowledgements

We thank Morris Moscovitch and Jordana Wynn for their insightful discussions. This work was supported by a NSERC Discovery Award. There are no conflicts of interest.

References

Ahdesmäki, M., Zuber, V., Gibb, S., and Strimmer, K. (2014). Shrinkage Discriminant Analysis and CAT Score Variable Selection. URL: http://strimmerlab.org/software/sda.

Altmann, G. T. (2004). Language-mediated eye movements in the absence of a visual world: The 'blank screen paradigm'. *Cognition*, 93(2), B79-B87.

Bartlett, F. C. (1932). Remembering: An experimental and social study. *Cambridge: Cambridge University*.

Bates, D., Maechler, M., Bolker, B., Walker S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1-48.

Blanke, O., and Seeck, M. (2003). Direction of saccadic and smooth eye movements induced by electrical stimulation of the human frontal eye field: effect of orbital position. *Experimental brain research*, 150(2), 174-183.

Blazhenkova, O., and Kozhevnikov, M. (2009). The new object-spatial-verbal cognitive style model: Theory and measurement. *Applied cognitive psychology*, 23(5), 638-663.

Brandt, S. A., and Stark, L. W. (1997). Spontaneous eye movements during visual imagery

reflect the content of the visual scene. *Journal of cognitive neuroscience*, 9(1), 27-38.

Buchsbaum, B. R., Lemire-Rodger, S., Fang, C., and Abdi, H. (2012). The neural basis of vivid memory is patterned on perception. *Journal of Cognitive Neuroscience*, *24*(9), 1867-1883.

Cabeza, R., Ritchey, M., and Wing, E. A. (2015). Reinstatement of Individual Past Events Revealed by the Similarity of Distributed Activation Patterns during Encoding and Retrieval. *Journal of cognitive Neuroscience*, 27(4), 679-691.

Chen, J., Leong, Y. C., Honey, C. J., Yong, C. H., Norman, K. A., and Hasson, U. (2017). Shared memories reveal shared structure in neural activity across individuals. *Nature Neuroscience*, 20(1), 115-125.

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

Cui, X., Jeter, C. B., Yang, D., Montague, P. R., and Eagleman, D. M. (2007). Vividness of mental imagery: individual variability can be measured objectively. *Vision research*, *47*(4), 474-478.

Brandt, S. A., and Stark, L. W. (1997). Spontaneous eye movements during visual imagery reflect the content of the visual scene. *Journal of cognitive neuroscience*, 9(1), 27-38.

DeYoe, E. A., Carman, G. J., Bandettini, P., Glickman, S., Wieser, J., Cox, R., Miller, D., and Neitz, J. (1996). Mapping striate and extrastriate visual areas in human cerebral cortex. *Proceedings of the National Academy of Sciences*, *93*(6), 2382-2386.

de Vito, S., Buonocore, A., Bonnefon, J. F., and Della Sala, S. (2014). Eye movements disrupt spatial but not visual mental imagery. *Cognitive processing*, 15(4), 543-549.

DiCarlo, J. J., Zoccolan, D., and Rust, N. C. (2012). How does the brain solve visual object recognition?. *Neuron*, 73(3), 415-434.

Dijkstra, N., Bosch, S., and van Gerven, M. A. (2017). Vividness of Visual Imagery Depends on the Neural Overlap with Perception in Visual Areas. *Journal of Neuroscience*, 3022-16.

Felleman, D.J., and Van Essen, D.C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, 1(1), 1–47.

Ferreira, F., Apel, J., and Henderson, J. M. (2008). Taking a new look at looking at nothing. *Trends in cognitive sciences*, 12(11), 405-410.

Gbadamosi, J., and Zangemeister, W. H. (2001). Visual imagery in hemianopic patients. *Journal of Cognitive Neuroscience*, 13(7), 855-866.

Ganis, G., Thompson, W. L., and Kosslyn, S. M. (2004). Brain areas underlying visual mental imagery and visual perception: an fMRI study. *Cognitive Brain Research*, *20*(2), 226-241.

Goodale, M. A., and Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends in neurosciences*, 15(1), 20-25.

Hale, S. M., and Simpson, H. M. (1971). Effects of eye movements on rate of discovery and vividness of visual images. *Perception and Psychophysics*, 9(2B), 242–246.

Hassabis, D., and Maguire, E. A. (2009). The construction system of the brain. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 364(1521), 1263-1271.

Haxby, J. V. (2012). Multivariate pattern analysis of fMRI: the early beginnings. *Neuroimage*, 62(2), 852-855.

Hebb, D. O. (1968). Concerning imagery. *Psychological review*, 75(6), 466-477.

Hesslow, G. (2011). The current status of the simulation theory of cognition. *Brain Research*, 1428(27), 71-79.

Hoover, M. A., and Richardson, D. C. (2008). When facts go down the rabbit hole: Contrasting features and objecthood as indexes to memory. *Cognition*, 108(2), 533-542.

Ishai, A., Haxby, J. V., and Ungerleider, L. G. (2002). Visual imagery of famous faces: effects of memory and attention revealed by fMRI. *Neuroimage*, 17(4), 1729-1741.

Johansson, R., Holsanova, J., Dewhurst, R., and Holmqvist, K. (2012). Eye Movements During Scene Recollection Have a Functional Role, but They Are Not Reinstatements of Those Produced During Encoding. *Journal of Experimental Psychology: Human perception and performance*, 38(5), 1289.

Johansson, R., Holsanova, J., and Holmqvist, K. (2006). Pictures and spoken descriptions elicit similar eye movements during mental imagery, both in light and in complete darkness. *Cognitive Science*, 30(6), 1053-1079.

Johansson, R., Holsanova, J., and Holmqvist, K. (2011). The Dispersion of Eye Movements During Visual Imagery is Related to Individual Differences in Spatial Imagery Ability. *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society*, 1200-1205.

Johansson, R., and Johansson, M. (2014). Look here, eye movements play a functional role in memory retrieval. *Psychological Science, 25*(1), 236-242.

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

Johnson, M. K., Kuhl, B. A., Mitchell, K. J., Ankudowich, E., and Durbin, K. A. (2015). Age-related differences in the neural basis of the subjective vividness of memories: Evidence from multivoxel pattern classification. *Cognitive, Affective, and Behavioral Neuroscience*, *15*(3), 644-661.

Kosslyn, S. M., Thompson,W. L., and Ganis, G. (2006). The case for mental imagery (New York, NY: Oxford University Press).

Kriegeskorte, N., Mur, M., and Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2, 4.

Kriegeskorte, N. (2011). Pattern-information analysis: from stimulus decoding to computational-model testing. *Neuroimage*, 56(2), 411-421.

Laeng, B., and Teodorescu, D. S. (2002). Eye scanpaths during visual imagery reenact those of perception of the same visual scene. *Cognitive Science*, 26, 207-231.

Laeng, B., Bloem, I. M., D'Ascenzo, S., and Tommasi, L. (2014). Scrutinizing visual images: The role of gaze in mental imagery and memory. *Cognition*, *131*(2), 263-283.

Lee, S. H., Kravitz, D. J., and Baker, C. I. (2012). Disentangling visual imagery and perception of real-world objects. *Neuroimage*, *59*(4), 4064-4073.

Lewis-Peacock, J. A., Drysdale, A. T., Oberauer, K., and Postle, B. R. (2012). Neural evidence for a distinction between short-term memory and the focus of attention. *Journal of Cognitive Neuroscience*, 24(1), 61-79.

Logothetis, N.K., and Sheinberg, D.L. (1996). Visual object recognition. *Annual Review of Neuroscience*, 19, 577–621.

Martarelli, C. S., and Mast, F. W. (2013). Eye movements during long-term pictorial recall. *Psychological research*, *77*(3), 303-309.

Naselaris, T., Olman, C. A., Stansbury, D. E., Ugurbil, K., and Gallant, J. L. (2015). A voxel-wise encoding model for early visual areas decodes mental images of remembered scenes. *Neuroimage*, 105, 215-228.

Noton, D., and Stark, L. (1971). Scanpaths in saccadic eye movements while viewing and recognizing patterns. *Vision research*, 11(9), 929-IN8.

Polyn, S. M., Natu, V. S., Cohen, J. D., and Norman, K. A. (2005). Category-specific cortical activity precedes retrieval during memory search. *Science*, 310(5756), 1963-1966.

Pomplun, M., Ritter, H., and Velichkovsky, B. (1996). Disambiguating complex visual

information: towards communication of personal views of a scene. *Perception*, 25(8), 931.

Pylyshyn, Z. W. (2002). Mental imagery: In search of a theory. *Behavioral and brain sciences*, 25(02), 157-182.

Richardson, D. C., Altmann, G. T., Spivey, M. J., and Hoover, M. A. (2009). Much ado about eye movements to nothing: a response to Ferreira et al.: Taking a new look at looking at nothing. *Trends in Cognitive Sciences*, 13(6), 235-236.

Richardson, D. C., and Spivey, M. J. (2000). Representation, space and Hollywood Squares: Looking at things that aren't there anymore. *Cognition*, 76(3), 269-295.

Rissman, J., and Wagner, A. D. (2012). Distributed representations in memory: Insights from functional brain imaging. *Annual Review of Psychology*, 63, 101–128.

Sligte, I. G., Scholte, H. S., and Lamme, V. A. (2009). V4 activity predicts the strength of visual short-term memory representations. *The Journal of Neuroscience*, 29(23), 7432-7438.

Slotnick, S. D., Thompson, W. L., and Kosslyn, S. M. (2005). Visual mental imagery induces retinotopically organized activation of early visual areas. *Cerebral cortex*, 15(10), 1570-1583.

Spivey, M. J., and Geng, J. J. (2001). Oculomotor mechanisms activated by imagery and memory: Eye movements to absent objects. *Psychological research*, 65(4), 235-241.

St-Laurent, M., Abdi, H., and Buchsbaum, B. R. (2015). Distributed Patterns of Reactivation

Predict Vividness of Recollection. *Journal of Cognitive Neuroscience*, 27(10), 2000-2018.

St-Laurent, M., Abdi, H., Bondad, A., and Buchsbaum, B. R. (2014). Memory reactivation in healthy aging: evidence of stimulus-specific dedifferentiation. *The Journal of Neuroscience*, 34(12), 4175.

Thirion, B., Duchesnay, E., Hubbard, E., Dubois, J., Poline, J. B., Lebihan, D., and Dehaene, S. (2006). Inverse retinotopy: inferring the visual content of images from brain activation patterns. *Neuroimage*, 33(4), 1104-1116.

Vogels, R., and Orban, G. A. (1994). Activity of inferior temporal neurons during orientation discrimination with successively presented gratings. *Journal of Neurophysiology*, 71(4), 1428–1451.

Williams, A. L., and Smith, A. T. (2010). Representation of eye position in the human parietal cortex. *Journal of Neurophysiology*, 104(4), 2169-2177.

Wooding, D. S. (2002). Eye movements of large populations: II. Deriving regions of interest, coverage, and similarity using fixation maps. *Behavior Research Methods, Instruments, and Computers*, 34(4), 518-528.

Wynn, J. S., Bone, M. B., Dragan, M. C., Hoffman, K. L., Buchsbaum, B. R., and Ryan, J. D. (2016). Selective scanpath repetition during memory-guided visual search. *Visual Cognition*, 24(1), 15-37.

Methods

Participants

Twenty-three right-handed young adults (6 males and 17 females, 20-30 years old [mean: 24.1], 14-21 years of education [mean: 16.9]) with normal or corrected-to-normal vision and no history of neurological or psychiatric disease were recruited through the Baycrest subject pool, tested and paid for their participation per a protocol approved by the Rotman Research Institute's Ethics Board. Subjects were either native or fluent English speakers and had no contraindications for MRI. Data from six of these participants were excluded from the final analyses for the following reasons: excessive head motion (2), poor eye tracking signal (1), misunderstood instructions (1), fell asleep (2). Thus, seventeen participants were included in the study (5 males and 12 females, 20-28 years old [mean: 23.8], 15-21 years of education [mean: 17.1]).

Stimuli

Nineteen complex coloured photographs were gathered from online sources and resized to 757 by 522 pixels in Adobe Photoshop. Five images were used for practice, and the remaining 14 were used during the in-scan and post-scan tasks (Figure 1). Each image was paired with a short descriptive title in 30-point Courier New font during in-scan encoding; this title served as a retrieval cue during the in-scan and post-scan memory tasks. Four different "modified" versions of each image were also created using Adobe Photoshop for a post-scan memory test: a minor

local element of the image was either added, removed or transformed in a way that was realistic and congruent with the image (Figure 2).

Procedure

In-Scan

Before undergoing MRI, participants were trained on a practice version of the task incorporating five practice images. Inside the scanner, participants completed three encoding runs and six retrieval runs of functional MRI. To keep participants engaged with the task, we interspaced the encoding and the retrieval runs (each encoding run was followed by two retrieval runs). A high-resolution structural scan was acquired between the 6th (retrieval) and 7th (encoding) functional runs, which provided a mid-task break. Eye-tracking data was acquired during all functional runs.

Encoding runs were 7m 18s long. Each run started with 10s of warm up during which instructions were displayed on-screen. Each trial began with a title shown in the top portion of the screen (0.5s; font = Courier New, font size = 30), followed by the appearance of the matching image in the center of the screen (4.75s; the title remained visible above the image). Images occupied 757 by 522 pixels of a 1024 by 768 pixel screen. Between trials, a cross-hair appeared in the center of the screen (font size = 50) for either 1s, 1.75s, 2.5s or 3.25s. Participants were instructed to pay attention to each image and to encode as many details as possible so that they could visualize the images as precisely as possible during the retrieval task. During the second and third encoding runs, participants were encouraged to pick up details they

had missed and to integrate them into their memory representation. Each image was shown four times per run, for a total of 12 encoding trials per image throughout the experiment. Within each run, the entire set of images was shown in a randomized order before the set could be shown again (e.g. each image needed to be shown twice before an image could be presented for the third time).

Retrieval runs were 8m 17s long, starting with 13 seconds of warm up during which instructions appeared on-screen. Each trial began with three 757 by 522 images shown in succession in the center of the screen for 1.5s each. Then, an image title appeared in the center of the screen for 1s (font = Courier New, font size = 30). For most trials, this title matched one of the three images in the sequence. The first, second and third image from the sequence were each cued during working memory conditions 1, 2 and 3, respectively (WM1, WM2, and WM3). WM1, WM2 and WM3 trials each corresponded to 1/4 of the total number of trials. In the remaining 1/4 of trials, the title corresponded to an image from the stimulus set that was not included in the sequence (the long-term memory condition, LTM). After 1s, the title was replaced by an empty rectangular box shown in the center of the screen (8s), and whose edges corresponded to the edges of the stimulus images (757 by 522 pixels). Participants were instructed to visualize the image that corresponded to the title as accurately and in as much detail as they could within the confines of the box. Once the box disappeared, participants were prompted to rate the vividness of their mental image on a 1-8 scale (2s) using two four-button fiber optic response boxes (one in each hand; 1 = left little finger; 8 = right little finger). Between each trial, a cross-hair (font size = 50) appeared in the center of the screen for 1.25s. Participants were instructed to attribute ratings of 4 or 5 for trials whose vividness felt “average for them”.

There were 28 trials per run (seven trials in each condition: WM1, WM2, WM3 and LTM), and 42 trials per condition for the entire scan.

Post-Scan

A post-scan test was conducted shortly after scanning to obtain behavioral measures of memory specificity as a function of task condition for the same 14 images encoded and retrieved inside the scanner. For each original image, four modified versions were created (Figure 2) which were used as difficult recognition probes to test each individual's memory acuity for the 14 images. Participants were instructed on the new task and completed a practice that included the five practice images shown during pre-scan training. The task involved four consecutive retrieval blocks separated by short breaks and, if needed, eye-tracking recalibration. For each trial, three images (757 by 522 pixels) from the set were presented consecutively in the center of a 1024 by 768 pixel screen for 1.5s each. Then, in a manner analogous to the in-scan retrieval task, an image title appeared in the center of the screen (1s; font = Courier New, font size = 30) that either matched the first (WM1), second (WM2) or third (WM3) image from the sequence, or that corresponded to an image from the set that was not included in the sequence (LTM; 1/4 of trials were assigned to each condition). The title was followed immediately by a version of the corresponding image that was either intact or modified. Participants were given 6s to determine whether the image was intact or modified using a keyboard button press (right hand; 1 = intact, 2 = modified). After 6s, the image was replaced by a 1s fixation cross (font size = 50) during which participants' response could still be recorded. The images shown in the 3-image sequence were always intact. Each of the four modified versions of an image appeared only once in the experiment (for a single trial), each in a different condition. During the inter-trial interval,

participants were required to fixate on the inner portion of a small circle in the center of the screen. The experimenter pressed a button to correct for drifts in calibration and to trigger the onset of the next trial. Participants were informed they could move their gaze freely during the rest of the trial.

For each original image, four modified versions were created (Figure 2) that were arbitrarily labeled modified images 1 to 4. Across participants, we counterbalanced the conditions in which an image was tested within each block, the condition to which an image's modified version was attributed, and the block in which a modified image's version appeared.

Setup and Data Acquisition

Participants were scanned with a 3.0-T Siemens MAGNETOM Trio MRI scanner using a 12-channel head coil system. A high-resolution gradient-echo multi-slice T1-weighted scan coplanar with the echo-planar imaging scans (EPIs) was first acquired for localization. Functional images were acquired using a two-shot gradient-echo T2*-weighted EPI sequence sensitive to BOLD contrast (22.5 x 22.5 cm field of view with a 96 x 96 matrix size, resulting in an in-plane resolution of 2.35 x 2.35 mm for each of 26 3.5-mm axial slices with a 0.5-mm interslice gap; repetition time = 1.5 sec; echo time = 27ms; flip angle = 62 degrees). A high-resolution whole-brain magnetization prepared rapid gradient echo (MP-RAGE) 3-D T1 weighted scan (160 slices of 1mm thickness, 19.2 x 25.6 cm field of view) was also acquired for anatomical localization.

Both the in-scan and the post-scan task were programmed with Experiment Builder version 1.10.1025 (SR Research Ltd., Mississauga, Ontario, Canada). In the scanner, stimuli and button press responses were presented and recorded using EyeLink 1000 (SR Research Ltd.,

Mississauga, Ontario, Canada). Visual stimuli were projected onto a screen behind the scanner made visible to the participant through a mirror mounted on the head coil. In-scan monocular eye movements were recorded with an EyeLink 1000 infrared video-graphic camera equipped with a telephoto lens (sampling rate 1000Hz) set up inside the scanner bore behind the participant's head. The camera picked up the pupil and corneal reflection from the right eye viewed from the flat surface mirror attached inside the radio frequency coil. Nine-point eye movement calibration was performed immediately before the first functional run. If needed, manual drift correction was performed mid-scan immediately prior to the onset of the next trial, and calibration was re-done in-between subsequent runs.

Post-scan stimuli were presented on a 19-in. Dell M991 monitor (resolution 1024×768) from a 24-inch distance. Monocular eye movements (the most accurate eye was selected during calibration) were recorded with a head-mounted Eyelink II eye tracker (sample rate 500 Hz) set to detect the pupil only. Eye movement calibration was performed at the beginning of the experiment, and drift correction (>5°), if needed, was performed immediately prior to the onset of each trial.

In-scan and post-scan eye tracking and behavioral data (vividness ratings, accuracy, and response time) were analyzed with Dataviewer version 1.11.1 (SR Research Ltd.). Saccades were determined using the built-in EyeLink saccade-detector heuristic. Acceleration (9500°/s/s) and velocity (30°/sec) thresholds were set to detect saccades greater than 0.5° of visual angle. Blinks were defined as periods in which the saccade-detector signal was missing for three or more samples in a sequence. Fixations were defined as the samples remaining after the categorization of saccades and blinks.

For the post-scan memory task, regions of interest (ROIs) were defined manually a priori for each image. A rectangular shape was drawn over each area of the image where a modification was introduced during the change-detection task, totaling four ROIs per image. Variations in the shape and orientation of these rectangles was dictated by the nature of the change, but strict counterbalancing insured that each variation was assigned to different conditions in a non-biased manner across participants.

*f*MRI and Neural Reactivation Measures

All statistical analyses were first conducted on realigned functional images in native EPI space. Functional images were converted into NIFTI-1 format, motion-corrected and realigned to the average image of the first run with AFNI's (Cox 1996) *3dvolreg* program, and smoothed with a 4-mm FWHM Gaussian kernel. The maximum displacement for each EPI image relative to the reference image was recorded.

For each subject, shrinkage discriminant analysis (SDA, Ahdesmäki, Zuber, Gibb and Strimmer, 2014; http://strimmerlab.org/software/sda) was used to train a pattern classifier to discriminate between the set of 14 images using fMRI data from the encoding runs. The full-brain, "all" ROI, pattern classifier was trained in two steps. First, a multivariate searchlight analysis using an 8mm radius was used to detect regions with above chance classification accuracy. The searchlight classification accuracy maps were then thresholded at $Z > 1.65$ (binomial distribution with chance accuracy = 1/14) to create separate feature masks for each subject (Figure 4A). A second SDA classifier was then trained on the encoding runs using all voxels falling inside the subject's feature mask, producing a final full-brain classifier that could

be used to evaluate image-specific reactivation during the memory task.

For the ROI reinstatement analyses, the subject-specific feature masks (Figure 4A) were divided into "dorsal", "occipital" and "ventral" regions (Figure 4B), based upon Two-Streams hypothesis (Goodale and Milner, 1992)—where "occipital" ROIs are not predominantly associated with one of the streams (see Supplementary Table S5 for a list of the FreeSurfer ROIs that compose each region). Three SDA classifiers per subject, one for each ROI, were then trained on the encoding runs using all voxels falling inside the subject's feature mask and the ROI's mask. In a similar manner, occipital pole and calcarine sulcus ROI analyses were performed with two SDA classifiers per subject, one for each ROI, but they were trained using all voxels within the corresponding FreeSurfer bilateral ROIs.

The SDA pattern classifiers trained on the set of encoding trials were then applied to data from the same brain regions acquired during the mental imagery task. First, the time-series data for each individual memory trial was divided into 16 1.5 second intervals (spanning 0-24 s), where the first interval (0-1.5 s) is aligned to start of the trial, which is defined as the onset of the first of the three images stimuli (see Figure 1). Next, the SDA classifiers were applied to each time-point over all trials of the memory experiment, producing a time-course of classifier confidence for each trial of the experiment. To control for the cortical activation caused by the recency condition (i.e. the initial three images during retrieval trials), we produced an "adjusted classifier confidence" (see Figure S2 for an explanatory diagram), which was calculated by subtracting the average classifier confidence for the image at the same serial position as the visualized image (e.g. position 2 if the current trial has the "WM2" condition) on the trials where the image is in the same position (e.g. position 2) but is *not* retrieved (i.e. the condition is

different than the current trial). Thus, a value greater than 0 indicates neural reinstatement. The adjusted classifier confidence for each time-point was then smoothed by convolving the data with a Gaussian filter (SD = 2 seconds), and a single adjusted classifier confidence was calculated for each trial by averaging across the five time-points corresponding to the visualization period (5.5-13 seconds offset by 6 seconds, i.e. 11.5-19s, to account for hemodynamic delay; the last 0.5 seconds were cut to avoid overlapping the vividness judgement).

Fixation Reinstatement Measures

Fixation reinstatement, i.e. the similarity between fixations during encoding and retrieval, was assessed by determining the difference between fixation density maps (Pomplun, Ritter and Velichkovsky, 1996; Wooding, 2002). For the creation of the fixation density maps, a 3D gaussian distribution was centered on each fixation. The Gaussian's "height" was proportional to the respective fixation's duration and its width was such that one standard deviation was about 1 degree of visual angle, approximating the width of the fovea. For each pixel, the Gaussians' values at that pixel were then summed, and the resulting map was subsequently normalized so that the sum over all pixel values was 1. To speed up the algorithm, the maps were calculated at 1/8th the dimensional length of the original screen.

To determine the similarity between two fixation density maps (encoding and retrieval), we first calculated the absolute difference between the two maps' corresponding pixels, thereby producing a "difference map" wherein a localized high value indicates that the two maps had greatly differing densities in that region. This "difference map" was then summed over all pixels

producing a single value between 0 (the maps were equivalent) and 2 (the maps were completely different).

Multiple studies have shown that the dispersion of fixations is lower when visualizing an image relative to perceiving it, and that this effect varied significantly between individuals (Brandt and Stark, 1997; Gbadamosi and Zangemeister, 2001; Johansson et al., 2006). This observed individual difference was subsequently found by Johansson, Holsanova and Holmqvist (2011) to be linked to differences in spatial imagery based on the OSIVQ measure (Blazhenkova and Kozhevnikov, 2009). Specifically, the researchers found that participants with higher spatial imagery scores had more spatially constrained fixations. Consequently, those with superior spatial imagery could counterintuitively show lower fixation similarity when utilizing the outlined similarity measure. To control for this contraction of fixations during mental imagery, we explored applying two fixation coordinate transformations before fixation similarity was calculated: the "expanded" transformation, and the "angle" transformation (see Figure S1 for an explanatory diagram).

The “expanded” transformation accounts for the possibility of a simple—and consistent—subject-specific contraction of fixations towards the center of the frame during retrieval, relative to encoding. To calculate the transformation, the subject-specific retrieval fixation-coordinate mean (i.e. the center of mass of the subject's cross-image retrieval fixations) is subtracted from all retrieval fixations. The resulting retrieval coordinates are then multiplied by the ratio (encoding/retrieval) of the average distance from the center for both encoding and retrieval, thereby equalizing the mean distance and countering the subject-specific contraction during retrieval. A significant drawback of this method is the assumption of a consistent within-

subject contraction ratio, given that it is likely that any potential contraction varies between and within trials. The "angle" transformation sidesteps this weakness by discarding the distance of fixations from the center entirely. The transformation is calculated by first subtracting the mean, as in the expanded transformation, from both the encoding and retrieval fixation coordinates. The fixation coordinates are then normalized to a length of one, and subsequently multiplied by the cross-subject average fixation distance-from-the-center—calculated from encoding and retrieval data. This transformation radially projects fixations onto a central circle, such that all fixations along a given radial line (extending from the center) are considered equivalent. This equivalence assumption is both a strength and a weakness of the transformation, making the similarity measure robust to varying fixation contraction ratios while leaving it completely insensitive to the fixations' distance from the center.

For the current study, fixation reinstatement was assessed by measuring how similar the spatial distribution of retrieval/visualization fixations is to the spatial distribution of encoding fixations during perception of the visualized image relative to the spatial distribution of encoding fixations during perception of all other images. To accomplish this, we first generated encoding and retrieval maps for each subject-image combination, and retrieval maps for each subject-image-trial combination, by pooling all fixations within the image or frame boundary from the relevant trial(s). We then calculated the similarity between each encoding and retrieval map with a corresponding subject-image pair, thereby acquiring a single similarity value for each combination (the subject-image retrieval maps were used for the comparisons between fixation reinstatement measures and the correlation between "angle" fixation reinstatement and post-scan memory task performance, whereas the subject-image-trial maps for the "angle" measure were

used for all other calculations). Let's call this value S[s,i,t], where lower case s, i and t represent the subject, image and retrieval trial (let t := 0, for subject-image retrieval maps), respectively; e.g. S[1,2,5] is the similarity value for subject 1, image 2, and retrieval trial 5. To control for the possibility of general (i.e. not image specific) viewing tendencies driving the similarity results, we calculated the average similarity between subject-image(-trial) retrieval maps and all encoding maps with the same subject but different images. Let's call this control value Sc[s,i,t]. The ratio between the corresponding similarity and control values was then calculated with the control as the denominator (e.g. S[1,2,5]/Sc[1,2,5]), indicating how similar the retrieval map is to the encoding map from the corresponding image relative to the encoding maps acquired from different images. Finally, this ratio is subtracted from 1 (e.g. 1 - S[1,2,5]/Sc[1,2,5]), so that 0 or less represents “no similarity”, and 1 represents “perfect similarity”.

We also performed two separate temporally windowed fixation reinstatement analyses, which were calculated in the same manner as the non-windowed trial-specific version above, with the following exceptions. For the first measure, which was generated to inspect the temporal properties of fixation reinstatement, 16 encoding and 29 retrieval maps were calculated for each subject-image combination, incorporating all fixations that fell within a given window's time span (window duration/width = 1s, temporal distance between windows/stride = 0.25s). Fixations that straddled the border of a window had their durations limited to the duration spent within the window. Subsequently, the similarity values between all combinations of encoding and retrieval windows – within subject-image – were computed. For the second measure, which was generated to match-up with the 5 temporally-windowed neural reinstatement values, we generated 5 temporally-windowed fixation reinstatement values for each trial. These values were

produced for each trial by calculating the similarity between the 29 retrieval fixation maps and the corresponding non-windowed subject-image encoding fixation map. The resultant 29 fixation similarity values for each trial were reduced to 5 values by convolving the 29 values with 5 Gaussian distributions (means = 0.8, 2.4, 4.0, 5.6, 7.2 sec; SD = 2 sec).

Bootstrap Statistics

All bootstrap statistics were calculated with 10000 samples. For the calculation of correlation statistics using a linear mixed effects (LME) model, bootstrap analyses were calculated with the BootMer function (Bates et al., 2015). For the calculation of mean statistics using a LME model, an array was created with each dimension representing a random effect—in this case, participants (17 rows) and images (14 columns). Each element of the array is the mean value for the element's combination of random effects (e.g. row 3, column 5 contains the mean value for participant 3, image 5). To generate a bootstrap distribution of the mean, 10000 new matrices were generated by randomly sampling the rows and columns of the original matrix with replacement, and then the 10000 means of the matrices' elements were calculated. For the paired-samples variant of the preceding procedure, each element of the array was a difference of means (i.e. the difference between the means generated by two different fixation similarity algorithms).