Dynamic changes in neural representations underlie the repetition effect on false memory

Xuhao Shao [a,b,c,d], Chuansheng Chen [e], Elizabeth F. Loftus [e], Gui Xue [a], Bi Zhu [a,b,c,d,*]

[a] State Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University, Beijing 100875, China
[b] Institute of Developmental Psychology, Beijing Normal University, Beijing 100875, China
[c] Beijing Key Laboratory of Brain Imaging and Connectomics, Beijing Normal University, Beijing 100875, China
[d] IDG/McGovern Institute for Brain Research, Beijing Normal University, Beijing 100875, China
[e] Department of Psychological Science, University of California, Irvine, CA 92697, United States

ARTICLE INFO

ABSTRACT

Restudying word lists (e.g., dream, awake, and bed) strengthens true memory of the studied words and reduces false memory for unstudied but semantically related lures (e.g., sleep). Yet, the neural mechanisms involved in this repetition effect on false memory remain unclear. Possible mechanisms involve item-specific and semantic neural representations at encoding, and the memory strength between encoding and retrieval. This study first replicated the behavioral results (Exp. 1) and then investigated various neural mechanisms by using slow event-related functional magnetic resonance imaging (fMRI) and representational similarity analysis (Exp. 2). Behavioral results confirmed that restudy improved true memory and reduced false memory. The fMRI results showed that restudy induced item-specific neural representations at encoding in the left occipital pole, but reduced neural overlap between semantic representations at encoding in the left temporal pole. Individual differences in these two encoding neural mechanisms were correlated with the behavioral measure of false memory, with greater restudy-induced representational changes at encoding (item-specific neural representations and reduced neural overlap between semantic representations) being associated with lower false memory. Moreover, restudy enhanced the memory strength between encoding and retrieval in the visuoparietal cortex but reduced it in the frontal cortex. These findings suggest that dynamic changes in neural representations underlie the repetition effect on false memory, supporting a dual-coding neural framework.

1. Introduction

Restudying the same materials strengthens true memory. This classic repetition effect on memory was first studied systematically by Ebbinghaus (1880). Since then, researchers have gained much understanding of this effect and its cognitive and neural mechanisms (Clark & Gronlund, 1996; Feustel et al., 1983; Miyashita, 2019). More recently, researchers have extended this line of research into the repetition effect on false memory. A commonly used false memory paradigm is the Deese-Roediger-McDermott (DRM) paradigm, in which participants are asked to learn semantically related words such as "dream", "awake", and "bed", and then are tested. Participants may correctly retrieve studied words such as "dream" (i.e., true memory), but also incorrectly recall or recognize unstudied but semantically related lures such as "sleep" (i.e., false memory) (Roediger & McDermott, 1995). Restudying DRM word lists generally has been found to reduce false memory (Benjamin, 2001;

Jou et al., 2004; Pitarque et al., 2018; Seamon et al., 2002), but the relevant cognitive and neural mechanisms are not well-understood.

Two theories are particularly relevant to the repetition effect of true and false memories, including the fuzzy trace theory (Holliday et al., 2008; Reyna & Brainerd, 1995) and the global matching model (Arndt & Hirshman, 1998; Clark & Gronlund, 1996). The fuzzy trace theory proposes two types of memory traces. Verbatim trace contains memory for the physical appearance of the specific item (i.e., item-specific representations), whereas gist trace contains memory for the semantic relatedness between items (i.e., the overlap between semantic representations). According to the fuzzy trace theory, true memory is driven by both verbatim and gist traces, whereas false memory is driven by gist traces. Restudy improves verbatim traces at encoding, which in turn leads to a greater discriminability between verbatim and gist traces at retrieval (Benjamin, 2001; Pitarque et al., 2018; Seamon et al., 2002).

At the neural level, the dual-coding framework maps the two types of memory traces to their neural processes (Bi, 2021; Wang et al., 2020). Verbatim traces have corresponding item-specific neural representa-

---

tions, measured as encoding-related neural pattern similarity (EnPS) across repetitions for the same word versus different words, whereas gist traces reflect the neural overlap between semantic representations, measured as encoding-related neural global semantic similarity (EnGSS) of semantically related words. Restudy has been found to enhance item-specific neural representations in the visual cortex during encoding. Specifically, the visual cortex shows greater EnPS across study repetitions of the same items than that for different items, and greater EnPS for recalled items than for forgotten items (Ward et al., 2013; Xue et al., 2010). Recent fMRI studies have also found that EnGSS in the left temporal pole was positively associated with behavioral indices of false memory (Chadwick et al., 2016; Zhu et al., 2019). However, it is unknown how restudy would affect EnGSS. In sum, restudy is believed to affect two encoding-related processes (as indexed by EnPS and EnGSS), which would work together to reduce false memory (Arndt & Reder, 2003; Hunt & Einstein, 1981).

Whereas the fuzzy trace theory focuses on the two encoding-related processes, the global matching model emphasizes the match between representations during encoding and those during retrieval (Arndt & Hirshman, 1998; Clark & Gronlund, 1996). According to this model, DRM false memory is a result of many partial matches between the representations of lures at retrieval and those of studied items at encoding. Recent fMRI studies have shown that episodic memory strength (measured as encoding-retrieval neural global pattern similarity, ER-nGPS) in the visual cortex and the hippocampus supports true memory but not false memory, whereas semantic memory strength (ER-nGPS) in the left inferior frontal gyrus supports both true and false memories (Ye et al., 2016; Zhu et al., 2019). Moreover, ER-nGPS in the visual cortex and the hippocampus may help to distinguish between true and false memories, and thereby reducing false memory. Thus far, no study has directly examined whether restudy reduces false memory by altering the ER-nGPS.

To fill the gaps in the literature, the current study examined systematically the three plausible neural mechanisms involved in the repetition effect on false memory. We performed a behavioral experiment and a functional magnetic resonance imaging (fMRI) experiment using the DRM false memory paradigm with a slow-event design. A slow-event design (i.e., 12 seconds per trial in the current study, including 3 seconds for the presentation of each word) was needed to estimate the blood oxygen level-dependent signal pattern associated with each trial in the neuroimaging experiment (Zhu et al., 2019). In the behavioral experiment (Exp. 1), a between-subject design was used to verify that repetition reduces false memory under this slow-event design. Participants were randomly assigned to either the restudy condition (i.e., participants saw each word list three times) or the control condition (i.e., participants saw each word list only once). After a 24-hour delay, participants took a visual recognition test judging studied targets (e.g., dream), unstudied but semantically related lures (e.g., sleep), and unstudied unrelated foils (e.g., pen). We expected to replicate the repetition effect that participants in the restudy condition would show higher true memory for targets and lower false memory for lures than would those in the control condition. Exp. 2 used an independent sample to examine the neural mechanisms involved in the repetition effect on false memory. As in the restudy condition in Exp. 1, participants in Exp. 2 learned each word list three times during encoding and performed a recognition test during retrieval 24 hours later. However, unlike Exp. 1, participants in Exp. 2 completed both the encoding and retrieval phases in the MRI scanner. The initial learning was used as the baseline condition to be compared with subsequent learning (i.e., the first versus the second and the third study), so no control group was included.

Representational similarity analysis was used to test the following three hypotheses. Hypothesis 1 was that restudy would reduce false memory by strengthening item-specific neural representations in the visual cortex, which were measured by the EnPS across study repetitions for the same word as compared to that for different words within the same word list. Hypothesis 2 was that restudy would reduce false memory by weakening neural overlap between semantic representations

in the left temporal pole, which were measured by changes in EnGSS (from the initial study to the subsequent study). We further explored whether individual differences in restudy-induced changes in neural representations at encoding would be negatively correlated with the behavioral indices of false memory. Hypothesis 3 concerned the role of global matching between encoding and retrieval. Hypothesis 3a was that restudy would enhance episodic memory strength by increasing ER-nGPS in the visual cortex and the hippocampus. Hypothesis 3b was that restudy would reduce semantic memory strength by reducing ER-nGPS in the left inferior frontal gyrus. Although both the left temporal pole and left inferior frontal gyrus are involved in semantic processing for false memory, the former seems to be more involved in the encoding phase (Chadwick et al., 2016; Zhu et al., 2019), whereas the latter has been found to play a role in the global matching between encoding and retrieval (Vatansever et al., 2021; Ye et al., 2016; Zhu et al., 2019). Therefore, Hypothesis 3b focused on the left inferior frontal gyrus.

2. Methods

2.1. Participants

In the behavioral study (Exp. 1), 66 participants were randomly assigned into two groups: 32 participants in the restudy group (21 females and 11 males, mean age = 21 ± 2 years) and 34 participants in the control group (24 females and 10 males, mean age = 22 ± 1 years). The fMRI study (Exp. 2) included an independent sample of 32 participants and had only the restudy condition (21 females and 11 males, mean age = 21 ± 2 years). These sample sizes were determined by power analyses based on the effect sizes found in previous studies. The effect sizes (Cohen's d) ranged from 0.80 to 1.34 for the repetition effect on behavioral indices of false memory (Benjamin, 2001; Pitarque et al., 2018). The effect sizes (Cohen's d) were 0.67 for item-specific neural representations at encoding across repetitions in the visual cortex (Xue et al., 2010), 0.71~1.26 for neural overlap between semantic representations at encoding in the left temporal pole (Chadwick et al., 2016), and 0.62 for the difference in the encoding-retrieval neural global pattern similarity between true and false memories with one-shot study in the visual cortex (Ye et al., 2016). For Exp. 1 that compared the control and the restudy groups, power analysis using G*Power 3.1 (independent samples t-test, minimal effect size = 0.80, α = 0.05, and power = 0.80) found the minimal required sample size to be 56 (i.e., 28 for each group). For Exp. 2 that compared across repetitions at encoding, power analysis (paired sample t-test, minimal effect size = 0.67, α = 0.05, and power = 0.80) found the minimal required sample size to be 20. For Exp. 2 that compared neural indices between true and false memories at retrieval, power analysis (paired sample t-test, minimal effect size = 0.62, α = 0.05, and power = 0.80) found the minimal required sample size to be 23. The three groups of participants (restudy group in Exp. 1, control group in Exp. 1, and restudy group in Exp. 2) did not differ in gender composition and mean age (ps > 0.19). All participants were right-handed Chinese college students, who had a normal or corrected-to-normal vision and no history of psychiatric or neurological disorders. Participants provided written informed consent before the experiment. This study was approved by the Institutional Review Board of Beijing Normal University, China.

2.2. Materials

Nine lists of 12 two-character Chinese words were used (Ye et al., 2016; Zhu et al., 2019). Each word list describes one theme. At encoding, each list included eight of the 12 words to be studied (e.g., dream, awake, bed, doze, yawn, snore, drowsy, and blanket). For the recognition test, a total of 108 words were used, including 36 target words (four of the eight studied words from each list, e.g., dream, bed, yawn, and drowsy), 36 semantically related but unstudied words as lures (four from each list, e.g., sleep, rest, tired, and pillow), and 36 words as foils
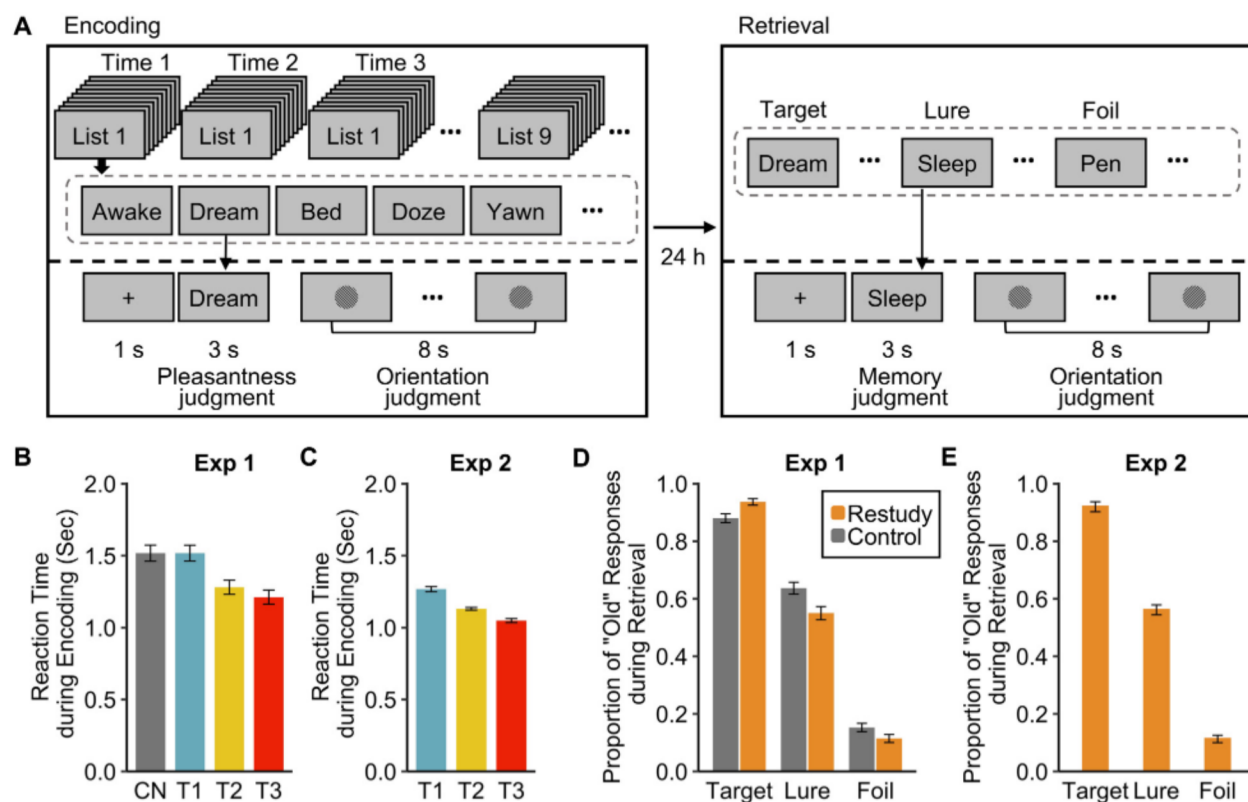
Fig. 1. Experimental procedure and behavioral results. (A) Experimental procedure in the restudy condition. At encoding, 72 words (nine lists of eight semantically related words) were studied. Each word list was presented three times. At retrieval, participants were asked to make a memory judgment on each of 108 words (36 targets, 36 lures, and 36 foils). (B-C) Each repetition significantly reduced the reaction time during encoding (Time 1 [T1] > Time 2 [T2] > Time 3 [T3]) in Exp. 1 (B) and Exp. 2 (C). Exp. 1 was a behavioral experiment comparing two conditions of restudy and control (CN, studied each word list only once). Exp. 2 was an fMRI experiment using an independent sample with only the restudy condition. Sec = seconds. (D-E) The proportion of "old" responses during retrieval in Exp. 1 (D) and Exp. 2 (E). Error bars indicate standard errors of the mean in Exp. 1 (B and D) and within-participant standard errors in Exp. 2 (C and E).

that were unstudied and semantically unrelated to any of the nine lists (e.g., pen, recycle, factory, and television). Study and test items were presented visually on the computer screen.

2.3. Experimental design

Experiment 1 was a behavioral study that used a between-subjects design to compare the memory performance between two conditions (i.e., restudy vs. control). Experiment 2 was an fMRI study that used an independent sample to explore the neural representations at encoding and retrieval under the restudy condition only. Otherwise, experimental materials and design were the same for the restudy condition in these two experiments.

At encoding, participants in the restudy condition in two experiments were exposed to each word list three times within a run, referred to hereafter as Time 1 (T1), Time 2 (T2), and Time 3 (T3) (Fig. 1A). There were nine runs, with one run for each word list. Participants in the control condition in Exp. 1 were exposed to each word list only once. Before studying a word list at each study time (three times for the restudy condition and once for the control condition), a 2-second visual cue was used to identify the list with a number (e.g., "List 1"). The presentation order of the nine word lists, as well as the presentation order of the eight studied words within each list for each study time, was randomized across participants. A slow event-related design was used (12 seconds for each trial). Each trial started with a 1-second fixation point. Next, a studied word was presented visually for 3 seconds. To help participants to remember these words, they were asked to make a pleasantness judgment on each word by pressing 1 of 4 buttons (1 = "very pleasant", 2 = "mildly pleasant", 3 = "mildly unpleasant", 4 = "very unpleasant")

as quickly and accurately as possible within 3 seconds. Participants used their index and middle fingers on both hands to respond, with half of the participants using one type of correspondence (left middle finger = 1, left index finger = 2, right index finger = 3, and left middle finger = 4) and the other half using the opposite pattern (left middle finger = 4, left index finger = 3, right index finger = 2, and left middle finger = 1). Participants practiced pressing the buttons until they were familiar with them. In the next 8 seconds, participants were asked to perform a self-paced perceptual judgment task, to prevent further processing of this word. Participants were asked to judge the orientation of a Gabor image, tilting 45 degrees to either left or right, by pressing 1 of 2 buttons (button 1 for left tilt, button 4 for right tilt) as quickly and accurately as possible.

At retrieval (24 hours after study), participants took a recognition test by judging 108 words (i.e., 36 targets, 36 lures, and 36 foils). These words were presented in a pseudorandomized order over three runs. The slow event-related design (12 seconds for each trial) was also used for the retrieval phase (i.e., 1-second fixation, 3-second visual presentation of a word for memory judgment, and 8-second self-paced perceptual judgment task). For memory judgment, participants were asked to judge whether they had studied these words one day earlier, by pressing 1 of 4 buttons (1 = "definitely new", 2 = "probably new", 3 = "probably old", 4 = "definitely old") within 3 seconds. The response buttons were the same as for the pleasantness judgment at encoding mentioned above.

In this study, the slow-event design was used because it provides better estimates of single-trial blood-oxygen-level-dependent (BOLD) response by reducing the temporal overlap between adjacent trials (Choupan et al., 2020; Mumford et al., 2012). This procedure has been well-validated (Gordon et al., 2014; Liu, 2012; Nahab et al., 2009;

Thompson & Siegle, 2009; Visser et al., 2016; Visser et al., 2013; Xue et al., 2010; Xue et al., 2013; Zheng et al., 2021) and has been used specifically in previous neural pattern similarity research on memory (Visser et al., 2013; Xue et al., 2010; Zheng et al., 2021). Between trials, we used a perceptual judgment task, instead of a rest condition, because such a specific task is a better baseline than the rest condition, the latter of which may involve ambiguous mental activities such as day dreaming, self-reflecting, and problem solving (Stark & Squire, 2001). This moderately difficult perceptual judgment task should halt further processing of each word and hence to reduce the autocorrelation of BOLD signals between words.

2.4. Behavioral analysis

For the recognition performance, the proportions of "old" responses (the number of items scored as 3 [probably old] or 4 [definitely old] divided by 36 [the total number of items for each type]) for targets, lures, and foils were used as indices of true memory, false memory, and false alarms to foils, respectively. Independent samples t tests were used to compare the mean proportions of "old" responses between the restudy and control groups in Exp. 1, and between the restudy group in Exp. 2 and the control group in Exp. 1. The data were also analyzed for high-confidence (i.e., definitely) and low-confidence (i.e., probably) responses, separately. To assess reaction time during recognition, we calculated separate reaction times for hits to targets (Target Hits), false alarms to lures (Lure FA), correct rejections to lures (Lure CR), and correct rejections to foils (Foil CR). Reaction times were not calculated for misses to targets and false alarms to foils, because they were rare. Independent samples t tests were used to compare the reaction times at initial learning (i.e., Time 1) between the restudy and control groups in Exp. 1. Paired sample t-tests were used to compare the reaction times between the three study times at encoding in each restudy group of the two experiments. Mixed ANOVA was used to compare the reaction times at retrieval between the restudy and control groups in Exp. 1 (i.e., 2 groups [restudy vs. control] by 4 types of responses [Target Hits, Lure FA, Lure CR, and Foil CR]). Finally, repeated-measures ANOVA was used to compare the reaction times at retrieval among the four types of responses in Exp. 2.

2.5. MRI acquisition and preprocessing

Brain images were acquired on a Siemens 3T Magnetom Trio scanner at Beijing Normal University Brain Imaging Center. Functional images were acquired using a T2*-weighted single-shot gradient EPI sequence (TR/TE/θ = 2000 ms/25 ms/90°, FOV = 192 × 192 mm, matrix = 64 × 64, slice thickness = 3 mm). Forty-one contiguous axial slices parallel to the AC-PC line were obtained. Structural images were acquired using a T1-weighted, 3D, gradient-echo pulse-sequence (T1/TR/TE/θ= 800 ms/2,530 ms/3.09 ms/10°, FOV = 256 × 256 mm, matrix = 256 × 256, 208 slices, slice thickness = 1mm).

Imaging data were preprocessed using fMRIPrep v1.4.0 (Esteban et al., 2019). Structural images were corrected for intensity non-uniformity and skull-stripped. Spatial normalization to the MNI152NLin2009cAsym template was performed through nonlinear registration. Functional data were coregistered to the corresponding structural data by boundary-based registration with 9 degrees of freedom. Then, they were slice-time corrected and motion corrected. The motion-correcting transformations, functional-to-structural transformation, and structural-to-template warp were concatenated and applied in a single step with antsApplyTransforms using Lanczos interpolation. The BOLD time-series were resampled to surfaces on the MNI152NLin2009cAsym template. The frame-wise displacement was calculated in Nipype for each functional run. Spatial smoothing was performed using a 5-mm full-width-half-maximum Gaussian kernel. Then, a nonlinear high-pass filter with a 100-second cutoff was used to filter the data temporally.

2.6. Single-item response estimation

Using the General Linear Model (GLM) (FSL v5.0.9), we computed the t-statistic map for each of the 72 studied words during encoding for each study time and 108 words during retrieval, separately. The unsmoothed data were used. To obtain reliable estimates of single-trial responses, the least-squares separate method was used (Mumford et al., 2012). Each single-trial model using the GLM included five regressors: the trial of interest, all other trials, six motion parameters, frame-wise displacement, and reaction time. The presentation of each word was modeled as an impulse at the onset of the stimulus, with a duration of 3 seconds. It was convolved with a double-gamma hemodynamic response function. The perceptual judgment and the fixation were not included in the model, but were treated as the implicit baseline. The estimated t value for each trial of each participant (t value = beta value / square root of variance), which was more reliable compared with the beta value, was used to compute the neural pattern similarity in the following analysis (Walther et al., 2016).
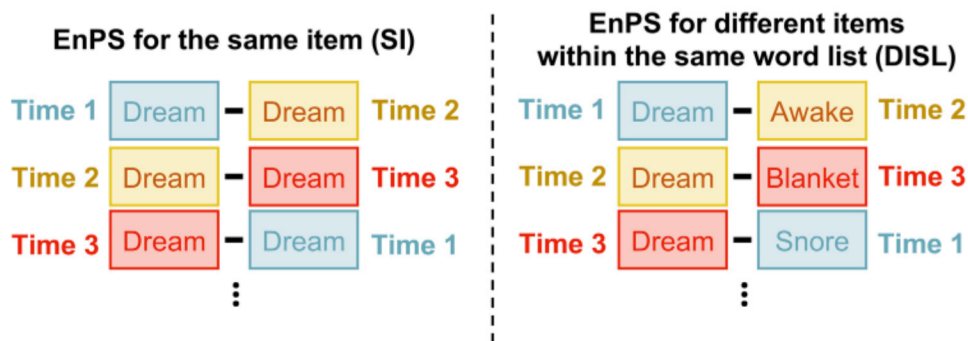
2.7. Representational similarity analysis

The same searchlight method, thresholds, and definitions of regions of interest (ROIs) were used for all representational similarity analyses in the current study. The searchlight analysis was conducted in the standard space. For each voxel, signals were extracted from a cubic ROI containing 125 surrounding voxels (i.e., 5 × 5 × 5 voxels cubic). They were then used to calculate Pearson's correlations between words, which were transformed into Fisher's Z scores. These neural pattern similarity scores were then averaged separately for each type of trials and contrasted at the subject level. They were fed into a random-effects model for group analysis. Since no first-level variance was available, an ordinary least squares model was used. Group images were thresholded (Z > 2.3 and a cluster probability of $p < 0.05$) and corrected for whole-brain multiple comparisons using Gaussian random field theory. Based on the whole-brain searchlight results, multiple ROIs were then defined by including the voxels in each cluster showing suprathreshold values for each contrast. Moreover, because they were located within large clusters of brain regions, some ROIs (i.e., the left temporal pole, left ventrolateral occipital cortex, left dorsolateral occipital cortex, left superior parietal lobe, and left hippocampus) were further defined by masking with their anatomical structures based on the Harvard-Oxford probabilistic atlas (threshold at 25% probability). In each of these ROIs, the mean values of similarity indices were extracted and analyzed at the subject level. More details of this searchlight method have been reported in previous studies (Ye et al., 2016; Zhu et al., 2019).
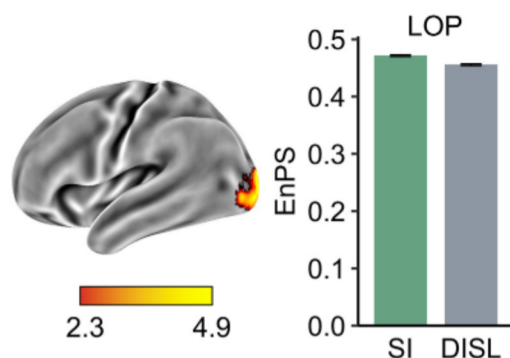
2.8. Encoding-related neural pattern similarity (EnPS)

To assess restudy-induced item-specific neural representations, we calculated and compared the EnPS across repetitions for the same item (SI) and different items within the same word list (DISL) (Fig. 2A). The EnPS for SI and for DISL were calculated by averaging the Fisher's Z scores for SI (i.e., 3 pairs [Time 1 & Time 2, Time 2 & Time 3, and Time 3 & Time 1] × 72 words = 216 pairs of correlations in total) and for DISL (i.e., 3 pairs [Time 1 & Time 2, Time 2 & Time 3, and Time 3 & Time 1] × 72 words = 216 pairs of correlations in total), respectively. The EnPS-DISL potentially involved 72 words × 7 different words × 3 time combinations = 1512 correlations. To match the number of correlations for EnPS-SI, we randomly chose one of the remaining seven studied words within the same word list to produce 3 correlations (e.g., Time 1 Dream with Time 2 Awake, Time 2 Dream with Time 3 Blanket, and Time 3 Dream with Time 1 Snore), which yielded 72 × 3 = 216 correlations. This procedure eliminated the potential problem of an unbalanced numbers of correlations (Xue et al., 2010). Nevertheless, we conducted an additional analysis using all 1512 correlations and presented the results in the Supplementary Materials.
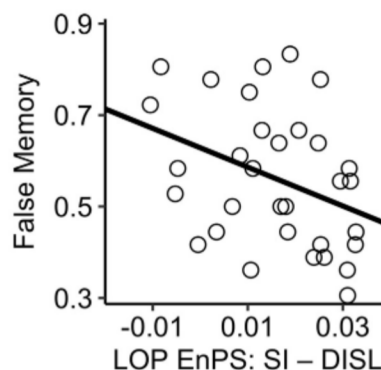
Fig. 2. Restudy induced item-specific neural representations across repetitions at encoding. (A) These representations were indexed by the encoding neural pattern similarity across repetitions (EnPS) for the same items (SI) as compared with that for different items within the same word list (DISL). (B) EnPS in the left occipital pole (LOP). Error bars indicate within-participant standard errors. (C) The restudy-induced EnPS differences between SI and DISL in the left occipital pole were negatively correlated with the behavioral index of false memory.

To test hypothesis 1 that restudy induces item-specific neural representations, we first examined the EnPS on the contrast (SI minus DISL). Based on the whole-brain searchlight analysis, we defined two visuo-cortical ROIs including the left occipital pole and right ventrolateral occipital cortex by including the voxels showing suprathreshold values for this contrast. The searchlight analysis did not reveal any other brain regions with significant results. In each of these two ROIs, the mean values of EnPS for SI and DISL were extracted. To explore their individual differences, the EnPS differences between SI and DISL in each ROI were used to correlate with the behavioral performance of false memory.

2.9. Encoding-related neural global semantic similarity (EnGSS)

To assess neural pattern overlap between semantically related studied words at each study time, we calculated the EnGSS for each study time by averaging the Fisher's Z scores of the pairwise correlations of the encoding neural activation patterns of all eight studied words in each list and then averaging them across all nine lists at each of three study times, separately (Fig. 3A). To test hypothesis 2 that restudy weakens this neural pattern overlap, we first examined the EnGSS on the contrast (Time 1 minus Time 3) using the whole-brain searchlight method. Due to the key role of the left temporal pole in the neural overlap between semantic representations (Chadwick et al., 2016; Zhu et al., 2019), we defined the ROI of the left temporal pole (LTP) by including the voxels showing suprathreshold values for this contrast, masked by its anatomical structure. In the LTP, the mean values of EnGSS at each study time were extracted and compared using the repeated measures ANOVA and post hoc paired-sample t tests (e.g., Time 1 vs. Time 2 vs. Time 3). To explore their individual differences, the restudy-induced reduction in EnGSS in

the LTP (e.g., LTP EnGSS: Time 1 – Time 2 or Time 1 – Time 3) were used to correlate with the behavioral performance of false memory. In addition, to explore whether restudy would enhance or reduce EnGSS in other brain regions, we used the whole-brain searchlight method to examine the EnGSS on each of the contrasts between different study times (e.g., Time 1 minus Time 3 or Time 3 minus Time 1).

2.10. Encoding-retrieval neural global pattern similarity (ER-nGPS)

To assess memory strength between retrieval and encoding for each study time, we calculated the ER-nGPS for each study time by averaging Fisher's Z scores of neural activation pattern similarity (Pearson r) between each test item during retrieval with all studied items during encoding for each study time (Fig. 4). We modeled four types of trials: Target Hits, Lure FA, Lure CR, and Foil CR. To test hypothesis 3 that restudy changes memory strength in the brain, we first examined the ER-nGPS differences (Target Hits minus Foil CR) on the contrasts (Time 3 vs. Time 1), using the whole-brain searchlight method. For the visuoparietal cortex, we defined three ROIs (i.e., the left ventrolateral occipital cortex, left dorsolateral occipital cortex, and left superior parietal lobe) by including the voxels showing suprathreshold values for the contrast of Time 3 – Time 1, masked with their anatomical structure. Moreover, due to the key role of the hippocampus in memory, we also defined the ROI of the left hippocampus by including the voxels showing suprathreshold values for this contrast, masked with the hippocampal anatomical structure. The right hippocampus was not defined as ROI since it had no voxel showing suprathreshold values for this contrast. We defined the ROI of left inferior frontal gyrus by including the voxels showing suprathreshold values for the contrast of Time 1 – Time

A Encoding Neural Global Semantic Similarity (EnGSS)

B EnGSS in the left temporal pole (LTP)
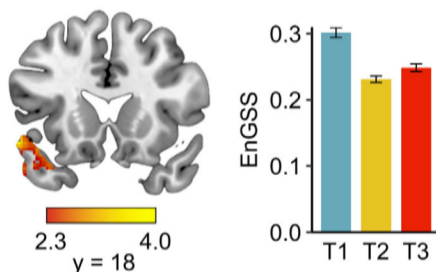
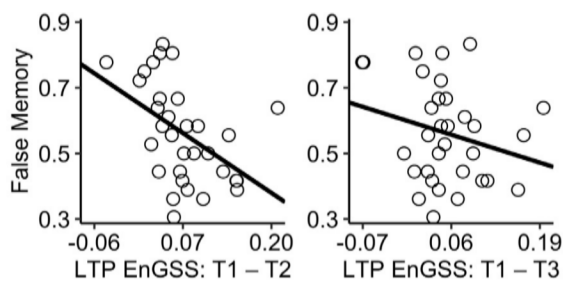## C Brain-behavioral correlations for false memory

Fig. 3. Restudy reduced the neural overlap between semantic representations at encoding as measured by the encoding neural global semantic similarity (EnGSS) in the left temporal pole. (**A**) EnGSS reflects the encoding neural overlap between studied items in the same word list at each study time. (**B**) EnGSS in the left temporal pole (LTP). Error bars indicate within-participant standard errors. (**C**) The restudy-induced EnGSS differences from Time 1 to Time 2 in the left temporal pole were negatively correlated with the behavioral index of false memory, but the correlation was not significant for the EnGSS reduction from Time 1 to Time 3.

3, which is the only cluster revealed by this contrast. In each ROI, the mean values of ER-nGPS for Target Hits, Lure FA, Lure CR, and Foil CR for each of the three study times were then extracted and analyzed. Next, we examined whether restudy would enhance ER-nGPS differences between true and false memories in each ROI, using the 2 (Target Hits vs. Lure FA) by 2 (study time [Time 1 vs. Time 3 or Time 1 vs. Time 2]) repeated measures ANOVA and post hoc paired-sample t tests.

2.11. Controlling for univariate activations at encoding and retrieval

To ascertain that the results from neural pattern similarity analysis could not simply be accounted for by univariate activations, we added analyses controlling for univariate activations at different study times during encoding and retrieval. We then examined the effects of EnGSS and ER-nGPS after controlling for the univariate activation at encoding and retrieval in each ROI, using mixed-effect regression analyses in R packages lme4 and lmerTest (Kuznetsova et al., 2017). Specifically, the univariate activation was averaged across all voxels in each ROI for each trial and each participant. For each ROI, we modeled the EnGSS or ER-nGPS as the dependent variable, included fixed effects terms to account for the study time (e.g., Time 1 vs. Time 2 or Time 1 vs. Time 3), response type (e.g., Target Hits vs. Foil CR or Target Hits vs. Lure FA), and their interaction terms, and included random effects terms to account for univariate activations at encoding and retrieval and participants' identity.

3. Results

3.1. Behavioral results

Results from Exp. 1 showed that repeated study increased true memory and decreased false memory. Specifically, compared to the control group, the restudy group had higher true memory (i.e., the proportion

of "old" responses for targets) (mean = 94% vs. 88%, $t(64) = 2.91$, $p = 0.005$), lower false memory (i.e., the proportion of "old" responses for lures) (mean = 55% vs. 64%, $t(64) = -2.84$, $p = 0.006$). The two groups did not differ significantly in false alarms to foils (i.e., the proportion of "old" responses for foils) (mean = 11% vs. 15%, $t(64) = -1.89$, $p = 0.06$) (Fig. 1D-E). The behavioral results of Exp. 2 (i.e., the fMRI study with only the restudy condition) showed 92% true memory, 56% false memory, and 11% false alarms to foils. Finally, although the environmental difference between the two experiments (inside vs. outside the noisy fMRI scanner) would prevent us from drawing firm conclusions, it is worth mentioning that compared with the control group in Exp. 1, the restudy group in Exp. 2 had lower false memory ($t(64) = -2.26$, $p = 0.03$). However, these two groups did not differ significantly in true memory ($t(64) = 1.87$, $p = 0.07$) and false alarms to foils ($t(64) = -1.79$, $p = 0.08$).

When the data were analyzed separately for high- and low-confidence responses, we found that the main behavioral results reported above for the repetition effect were driven by high-confidence responses (Table 1). Mixed-design ANOVA (group [restudy vs. control] by confidence level [high vs. low]) was used to examine whether the confidence level of responses moderated the repetition effect in Exp. 1. There were significant interactions between group and confidence level for the proportion of target hits, target misses, and foil correct rejections ($F(1, 64) = 4.50$, 7.69, and 4.36, $ps < 0.04$), but not for lure false alarms, lure correct rejections, and foil false alarms ($ps > 0.08$). For high-confidence responses, compared to the control group in Exp. 1, the restudy group in Exp. 1 had higher target hits ($t(64) = 2.54$, $p = 0.01$), lower lure false alarms ($t(64) = -2.26$, $p = 0.03$), higher lure correct rejections ($t(64) = 2.76$, $p = 0.008$), and higher foil correct rejections ($t(64) = 2.14$, $p = 0.04$), but the two groups did not differ in target misses and foil false alarms ($ps > 0.15$). For low-confidence responses, the two groups did not differ in any indices ($ps > 0.06$), except for target