CONTENT REINSTATEMENT AND SOURCE CONFIDENCE DURING EPISODIC MEMORY RETRIEVAL

A Thesis presented to

the Faculty of the Graduate School

at the University of Missouri-Columbia

In Partial Fulfillment

of the Requirements for the Degree

Master of Arts

by

EMILY K. LEIKER

Dr. Jeffrey D. Johnson, Thesis Supervisor

DECEMBER 2014

The undersigned, appointed by the Dean of the Graduate School, have examined the thesis entitled

CONTENT REINSTATEMENT AND SOURCE CONFIDENCE DURING EPISODIC MEMORY RETRIEVAL

presented by Emily K. Leiker,

a candidate for the degree of Master of Arts,

and hereby certify that, in their opinion, it is worthy of acceptance.

Assistant Professor Jeffrey D. Johnson

Associate Professor David Beversdorf

Associate Professor Shawn E. Christ

Middlebush Professor Jeff Rouder

ACKNOWLEDGEMENTS

I would first like to thank my advisor, Dr. Jeff Johnson, for his guidance during the completion of my thesis, and for his enduring support throughout my graduate education. I am truly grateful for his advice and encouragement throughout this project, and for his incredible patience and understanding along the way. I would also like to thank the members of my thesis committee, Dr. David Beversdorf, Dr. Shawn Christ, and Dr. Jeff Rouder, for their helpful comments and insightful questions, and for engaging me in a thought-provoking discourse that was both challenging and instructive. In addition, I would like to express my sincere gratitude to the Life Sciences Fellowship Program, for the financial support that has granted me the freedom to focus solely on research and coursework over the last several years. Finally, thank you to all of my friends, family and colleagues who provided words of encouragement, advice, and a listening ear over the course of this project – your kindness and support helped make this possible.

TABLE OF CONTENTS

LIST OF FIGURES

Figure Page

LIST OF TABLES

Table Page

LIST OF ABBREVIATIONS

ANOVA:	Analysis of variance
BOLD:	Blood-oxygen-level dependent
fMRI:	Functional magnetic resonance imaging
EPI:	Echo-planar imaging
ERS:	Event-related similarity
FOV:	Field of view
GLM:	General linear model
HRF:	Hemodynamic response function
K:	Know
MVPA:	Multi-voxel pattern analysis
PFC:	Prefrontal cortex
R:	Remember
R/K:	Remember/know
ReML:	Restricted maximum-likelihood
ROI:	Region of interest
RT:	Response time
TE:	Echo time
TR:	Repetition time

CONTENT REINSTATEMENT AND SOURCE CONFIDENCE

DURING EPISODIC MEMORY RETRIEVAL

Emily K. Leiker

Jeffrey D. Johnson, Thesis Supervisor

ABSTRACT

The retrieval of qualitative information from episodic memory ("recollection") is thought to be supported by hippocampally-mediated reinstatement of the neurocognitive processes and representations activated during encoding. Several functional magnetic resonance imaging (*fMRI*) studies have provided evidence for this hypothesis by demonstrating stronger reinstatement when participants report recollecting specific details compared to when recollection fails, and when more episodic information is available for recollection. However, the precise nature of the relationship between recollection and reinstatement remains largely unexplored, particularly in regard to the extent to which participants might monitor the reinstated information to make their memory decision. The current study addressed this issue by examining the relationship between a direct behavioral measure of recollection quality – confidence ratings about source memory judgments – and the magnitude of neural reinstatement during retrieval. Participants viewed a series of words in the context of three encoding tasks, then completed a memory test with a two-step response procedure, in which they first identified the encoding task (source) previously completed for a given word, then rated their confidence in that source judgment. fMRI data were acquired during encoding and retrieval phases, and subjected to pattern classification analyses to obtain an index of

reinstatement. The reinstatement effects were examined according to the behavioral measure and neural correlates of source confidence. The findings are considered in regard to how regions such as left posterior parietal cortex might monitor the reactivated episodic information to guide decisions about retrieval quality.

INTRODUCTION

Episodic retrieval refers to the process by which information about previous experiences is retrieved from memory. Numerous computational models of memory suggest that episodic retrieval relies on a mechanism known as *cortical reinstatement*, in which the pattern of neural activity elicited during initial event exposure is reactivated during retrieval of information about that event (Alvarez & Squire, 1994; McClelland, McNaughton, & O'Reilly, 1995; Hasselmo & Wyble, 1997; Rolls, 2000; Shastri, 2002). According to this account, an event elicits a distributed pattern of cortical activity at the time of encoding, which is indexed and stored as a unique, sparse representation by the hippocampus (Marr, 1971; Teyler & DiScenna, 1986; Norman & O'Reilly, 2003). Activation of the hippocampal representation in response to a subsequently-encountered retrieval cue then leads to the reinstatement (or reactivation) of the encoding-related pattern of cortical activity for the initial event, which in turn allows additional information about the event to be retrieved. Although the involvement of reinstatement during episodic retrieval is well supported by several recent neuroimaging studies (for reviews, see Rugg, Johnson, Park, & Uncapher, 2008; Danker & Anderson, 2010; Rissman & Wagner, 2012), the role of reinstatement in retrieval-based decisions remains somewhat unclear. In the current study, we addressed this issue by investigating how reinstatement co-varies both with subjective (behavioral) accounts of retrieving specific episodic content and the neural correlates of such retrieval.

Neuroimaging studies of reinstatement during episodic memory typically employ a behavioral design consisting of two phases: 1) an encoding phase in which stimuli are

presented in the context of different conditions (e.g., stimulus modality or the task required), and 2) a retrieval phase in which the stimuli are re-presented for the purpose of eliciting reinstatement about the respective encoding condition. The different conditions employed during the encoding phase are designed to encourage participants to engage distinct cognitive operations and representations, which should result in different patterns of cortical activation. A key feature of the retrieval phase is the presentation of simple retrieval cues that are devoid of any information regarding the encoding condition in which the stimulus previously appeared. To assess whether brain activity elicited during encoding is reactivated during retrieval, fMRI data is acquired during both phases. The data are then analyzed to identify the degree of overlap (or similarity) in condition-related activity across the two phases, which is taken as evidence that encoding-related representations and processes were reinstated during retrieval.

Several neuroimaging studies have employed standard, univariate analyses of fMRI data to identify encoding-retrieval overlap that is indicative of reinstatement (e.g., Wheeler, Petersen, & Buckner, 2000; Kahn, Davachi, & Wagner, 2004; Johnson & Rugg, 2007; for review, see Danker & Anderson, 2010). In one study by Johnson and Rugg (2007), participants were presented with words in the context of two distinct encoding tasks. Words in one condition were superimposed on a landscape image and required participants to imagine the object to which the word referred somewhere in the landscape. Words in the other condition appeared on a solid gray background and required the generation of a sentence incorporating the word. On a later memory test, participants were presented with the simple word cues and made judgments about each word according to a standard "remember/know" (*R/K*) procedure (Tulving, 1985). In this

procedure, participants distinguish between items for which they remember (*R*) specific details (e.g., something about the superimposed landscape picture, their generated sentence, or any other detail from encoding); items they know (*K*) were previously encountered, on the basis of a strong sense of familiarity, but which are not accompanied by any specific details; and items they judge as new (not encountered during encoding). Analysis of the fMRI data from the encoding phase identified multiple brain regions where activity was greater for one condition than the other (and vice versa). Importantly, the regions showing encoding-related differences exhibited analogous effects during the retrieval phase. This overlap provided evidence of the involvement of reinstatement during episodic retrieval. Moreover, the regions associated with reinstatement also exhibited greater activity for R than K judgments. This latter finding suggests that reinstatement plays a role in the retrieval decision, contributing more to recollection- as opposed to familiarity-based memory, consistent with the proposal of computational models of reinstatement (Hasselmo & Wyble, 1997; Norman & O'Reilly, 2003).

Whereas the findings of Johnson and Rugg (2007) supported a relationship between reinstatement effects and recollection, it was unclear whether the effects were restricted to recollection judgments, or if they were instead graded across different retrieval judgments, as would be expected if reinstatement guided the retrieval decision. In a follow-up fMRI study, Johnson, McDuff, Rugg, and Norman (2009) addressed this distinction. During the encoding phase of the study, participants viewed a series of words and completed one of three different tasks for each word. One task probed participants to imagine an artist drawing the object denoted by the word, another task required that participants generate possible functions that the object could serve, and the third task

involved covertly pronouncing the word backwards (hereafter, the *Artist*, *Function*, and *Read* tasks, respectively). On a later memory test, old and new words were presented in the context of a modified R/K procedure (Woodruff, Johnson, Uncapher, & Rugg, 2005; Yonelinas, Otten, Shaw, & Rugg, 2005). The criteria for the R response in this procedure was the same as that of the standard R/K procedure. When participants did not remember details, however, they were to indicate their confidence that the word was old or new using a four-point scale (ranging from "highly-confident old" to "highly-confident new"). Importantly, the modified procedure provided sufficient numbers of trials to test for reinstatement when item recognition was presumably strong and absent of recollection (i.e., for "highly-confident old" judgments).

Another important feature of the study by Johnson et al. (2009) is its use of multivariate – rather than univariate – analyses of fMRI data, which have become increasingly common for investigating encoding-related reinstatement during retrieval (e.g., Polyn, Natu, Cohen, & Norman, 2005; McDuff, Frankel, & Norman, 2009; Kuhl, Rissman, Chun, & Wagner, 2011; for review, see Rissman & Wagner, 2012). This increase in popularity is primarily due to the sensitivity of multivariate analyses to detect distributed representations of episodic content that may be either weak at the local (regional) level or variable across participants (for further discussion, see Mur, Bandettini, & Kriegeskorte, 2009; Jimura & Poldrack, 2012). Johnson et al. (2009) employed a multivariate technique called *multi-voxel pattern analysis* (MVPA; for reviews, see Haynes & Rees, 2006; Norman, Polyn, Detre, & Haxby, 2006; Tong & Pratte, 2012) which involved training a pattern classifier, with fMRI data from the encoding phase, to detect differences in neural activity according to the three tasks. The

trained classifier was then presented with fMRI data from the retrieval phase and evaluated on its ability to determine the prior encoding task associated with each individual item. The greater the similarity between the pattern of neural activity elicited for an item during encoding and the pattern elicited at retrieval, the more accurate the classifier should be at identifying the correct task. Classifier accuracy therefore provided an index of the degree to which encoding-related neural activity was reinstated. With this approach, Johnson et al. (2009) observed a graded pattern of reinstatement across the different memory judgments. Classifier performance was highest for R responses, at an intermediate level (though still above chance) for "highly-confident old" responses, and lowest for the remaining responses (including low-confidence judgments and misses). This graded pattern challenged the idea that R and K responses (the latter corresponding most closely with "highly-confident old" responses) distinctively reflect the subjective retrieval experiences of conscious recollection versus an acontextual feeling of familiarity (Yonelinas, 2002). The findings instead suggest that the subjective quality of these retrieval experiences could depend, in part, on whether the level of cortical reinstatement surpasses a decision threshold set by the participant. Reinstatement might therefore provide a direct measure of the episodic information that becomes consciously available to guide the retrieval decision.

Under the assumption outlined above – that the level of reinstatement correlates with subjective distinctions in retrieval quality – a further prediction is that reinstatement should co-vary with neural activity in regions that are also sensitive to the amount of episodic information retrieved. A number of recent neuroimaging studies have demonstrated that activity in regions such as left inferior parietal cortex and hippocampus

increases with the retrieval of additional details (e.g., Vilberg & Rugg, 2007, 2009a, 2009b; Guerin & Miller, 2011; Rugg et al., 2012). Some of these studies have manipulated the presentation duration of items during encoding (Vilberg & Rugg, 2009a; Guerin & Miller, 2011), which presumably affects the amount of episodic information available for later retrieval, whereas others have relied on direct reports from participants about the number of details retrieved (Vilberg & Rugg, 2007, 2009b). In a study by Vilberg and Rugg (2009a), for example, participants were presented with stimuli consisting of multiple pictures of objects and outdoor scenes for one of two durations at encoding (1 and 6 seconds). On a subsequent recognition memory test, single instances of the objects were presented in the context of a standard R/K procedure. In addition to observing a recollection effect in left inferior parietal cortex, such that activity was elevated for R versus K judgments (also see Wheeler & Buckner, 2004; Woodruff et al., 2005; Yonelinas et al., 2005), Vilberg and Rugg (2009a) identified a posterior region of this parietal cluster that was further sensitive to the encoding-duration manipulation. Specifically, activity was greater for recollected items from the longer (6 s) duration relative to that for the shorter duration. A follow-up memory test administered after the R/K test phase confirmed that participants could verbally report more details for items from the longer duration. Together, these findings have led to the suggestion that posterior parietal cortex tracks the accumulation of recollected information in service of the memory task (also see Vilberg & Rugg, 2007, 2009b; Guerin & Miller, 2011).

Leiker and Johnson (2014) recently extended the findings described above by investigating the relationship between the magnitude of reinstatement and the amount-sensitive retrieval activity in left posterior parietal cortex. In line with previous studies of

the amount of retrieved information, a behavioral manipulation was used in which words were presented for two different durations at encoding (4 or 8 s; Vilberg & Rugg, 2009a; Guerin & Miller, 2011). To encourage participants to encode different types of information, three encoding tasks were employed orthogonal to the duration manipulation. Two of the encoding tasks were identical to the Artist and Function tasks previously described (Johnson et al., 2009; McDuff et al., 2009); the third task, hereafter referred to as the *Cost* task, required participants to think about the relative cost of each item and was used to ameliorate the low levels of memory performance previously associated with the *Read* task (see Johnson et al., 2009). The retrieval test consisted of the modified R/K procedure discussed earlier. As in Johnson et al. (2009), MVPA was conducted on the fMRI data to track the episodic content specific to the three tasks across encoding and retrieval. The analyses confirmed that neural patterns elicited during retrieval resembled those occurring during encoding, providing evidence of reinstatement. Furthermore, the magnitude of reinstatement differed according to the previous presentation duration of items at encoding. As shown in Figure 1, retrieval of items from the longer encoding duration elicited a greater degree of reinstatement than items from the shorter duration. Crucially, this reinstatement effect was mirrored by the expected amount-related difference in activity in left posterior parietal cortex, with greater activity in this region occurring during retrieval of items from the longer duration (as in Vilberg & Rugg, 2009a; Guerin & Miller, 2011; also see Vilberg & Rugg, 2007, 2009b). Together, these findings are consistent with the hypothesis that left posterior parietal cortex is sensitive to the accumulation of reinstated episodic information during retrieval.

The Current Study

Although amount-related differences in reinstatement magnitude were concurrent with changes in posterior parietal activity in our previous study (Leiker & Johnson, 2014), uncertainty regarding the extent to which participants monitored the reinstated information to inform their memory decision was a significant limitation. In particular, participants in the previous study were given only a single recollection-based response option at test, and thus we could not make any claims about whether they were aware of variation in the amount of information recollected on a trial-by-trial basis. In light of this, the current study was designed to investigate the extent to which the magnitude of neural reinstatement might directly contribute to the subjective outcome of memory retrieval attempts.

Our participants first viewed a series of words in the context of three encoding tasks: Artist, Function, and Cost (see Figure 2A; also see Leiker & Johnson, 2014). At test, participants completed a two-step response procedure for old and new words, as shown in Figure 2B. The first step required participants to identify which encoding task (source) was previously completed for a given word, or that the word was new. In the second response step, participants made judgments on a three-point scale to indicate their level of confidence in making the initial source judgment. Presumably, the source confidence judgments required participants to focus on and assess the retrieval of specific, task-related information that accompanied each test item. This procedure was therefore more ideal than the encoding-duration manipulation of the previous study for parametrically tracking the amount (or strength) of encoding-related reinstatement.

As in our previous studies (Johnson & Rugg, 2007; Johnson et al., 2009; Leiker & Johnson, 2014), fMRI data were acquired during both the encoding and retrieval phase. MVPA was used to track informational content about the three encoding tasks across the encoding and retrieval phases, providing a measure of reinstatement associated with each test item. Based on the prior findings of stronger reinstatement when more information is recollected (Leiker & Johnson, 2014), we predicted that the magnitude of reinstatement would increase with increasing source confidence. That is, we expected reinstatement would be greatest for high-confidence responses and lowest for low-confidence responses. As an extension of this hypothesis, our analyses also focused on assessing the degree of encoding-related reinstatement in relation to activity in the brain regions that appear to track the amount (or strength) of retrieved information. Judgments of source memory confidence have been previously shown to give rise to graded effects in left posterior parietal cortex similar to the changes in activity arising from manipulations of amount (e.g., Hayes, Buchler, Stokes, Kragel, & Cabeza, 2011; Yu, Johnson, & Rugg, 2012a). Based on these findings, activity in parietal regions, as well as in hippocampus (Rugg et al., 2012; Yu et al., 2012a, 2012b), was expected to correlate positively with the magnitude of encoding-related reinstatement. Correlational analyses, both at a within-participant (trial-by-trial) level and across participants, were performed to address this relationship. Positive findings from these analyses would provide compelling support for the hypothesis that the levels of activity in posterior parietal cortex and hippocampus signify the tracking of reactivated episodic information in service of retrieval judgments.

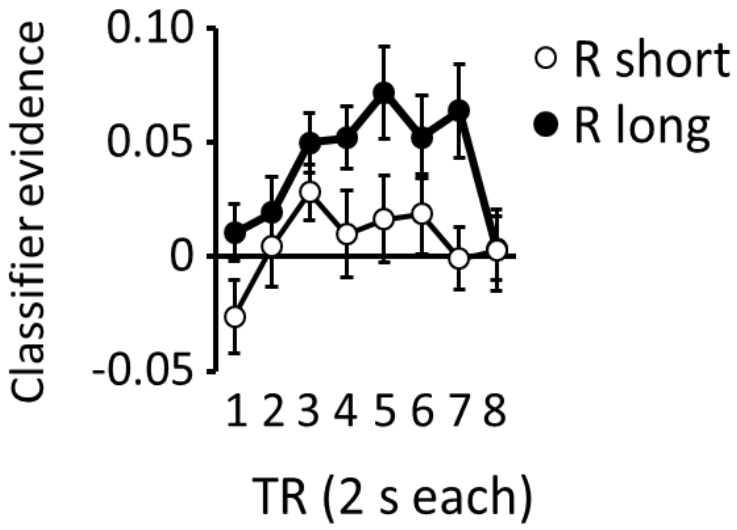

Figure 1. Classifier evidence for reinstatement of encoding-related neural activity during recollection, reported in Leiker and Johnson (2014). Evidence is computed as the difference between classifier output for the correct encoding task and the mean classifier output for the two remaining (incorrect) tasks (chance = 0). Time courses of mean classifier evidence (error bars: ±SEM) are displayed for test items from the long and short encoding durations, beginning with item onset (TR 1). TR = repetition time, 2-s each. R = remember.

METHOD

Participants

Twenty-one volunteers were recruited from the University of Missouri (MU) student population and received either course credit or monetary compensation for their participation. Participants were self-reported to be right-handed, native-English speakers, with normal or corrected-to-normal vision, no history of neurological disease, and no other MRI contraindications. Data from five participants were excluded from all analyses: one participant was excluded for not completing the experiment, another for not following instructions for the response buttons, and three others were excluded due to insufficient numbers of trials. The final sample consisted of 16 participants (9 males, 7 females) with a mean age of 19 years. Informed consent was obtained from all participants in accordance with the guidelines of the MU Health Sciences Institutional Review Board.

Stimuli and Design

The stimuli consisted of a pool of 308 words drawn from the MRC database (Coltheart, 1981; Wilson, 1988; http://websites.psychology.uwa.edu.au/school/MRCDatabase/uwa_mrc.htm). Each word met the following selection criteria: four to nine letters in length (M = 5.5, SD = 1.3), a written frequency between one and 50 per million (M = 16.9, SD = 13.1; Kucera & Francis, 1967), and scores of at least 500 on scales of familiarity (M = 581.9, SD = 34.6), concreteness (M = 539.1, SD = 27.5), and imagability (M = 581.6, SD = 31.2). All stimuli appeared in white uppercase 36-point

Arial font on the black background of a screen, which was positioned at the head of the magnet bore and viewed through a mirror placed in front of participants' eyes.

For each participant, 216 words were randomly selected from the stimulus pool for presentation during the encoding phase. These words were randomly assigned to one of three encoding blocks (72 words each) and to one of three encoding tasks within each block (resulting in 24 items per block/task combination). Of these encoding stimuli, 162 were re-presented as *old* items during the retrieval phase, along with 54 words randomly selected from the pool to serve as *new* items (not studied). The retrieval phase was also divided into three blocks (72 words each), with equivalent numbers of stimuli from each encoding block/task combination presented in each retrieval block. The remaining thirty-eight words from the pool were used during instruction and practice phases.

Behavioral Procedure

Prior to entering the scanner, participants received instructions and completed a short practice version of the encoding phase. Once in the scanner, participants completed the three encoding blocks, followed by an anatomical scan, instructions and practice for the retrieval phase, and the three retrieval blocks. Instructions and practice for the retrieval phase were delayed until immediately prior to that phase to prevent any influence on encoding strategy.

For the encoding phase, participants were informed that they would be presented with a series of words and would need to think about each word in the context of one of three tasks. Participants also received instructions to make responses about each task on a four-point scale by pressing buttons with their right index through little fingers. A

schematic of this procedure is shown in Figure 2A. The three tasks – referred to as the Artist, Function, and Cost tasks – were selected for their ability to elicit elaborate and distinct processing (also see Johnson et al., 2009; McDuff et al., 2009; Leiker & Johnson, 2014). The Artist task required participants to imagine how an artist would draw the object denoted by the word and to rate the difficulty of drawing that object from 1 (”easy”) to 4 (“hard”). The Function task required participants to think of as many different functions as they could for the object and to respond with the number of functions generated from “1” to “4”. The Cost task required participants to think about the relative cost of the object and to rate the cost from 1 (“low”) to 4 (“high”). To facilitate the identification of distinct patterns of brain activity elicited by the three tasks, encoding stimuli were grouped into mini-blocks in which a particular task was completed for four consecutive words. Mini-blocks began with a 3-s instructional display indicating the task to be completed and the response options for the upcoming words. The instructional display remained on screen throughout the mini-block. Each encoding word was then centrally displayed for 3 s, with an asterisk appearing above the word for the final second of display to indicate that participants should make their response. The next word in the mini-block immediately followed. Upon completion of each mini-block, a central fixation cross appeared for 2, 4, or 6 s (12, 4, and 2 instances, respectively, in each encoding block) until the next instructional display. The mini-blocks were pseudo-randomly ordered to prevent consecutive completion of the same task.

For the retrieval phase, participants completed a memory test for a series of intermixed words that either appeared in the prior encoding phase (*old*) or did not (*new*). The memory test employed a two-step response procedure, as shown in Figure 2B. The

first step required participants to make a four-alternative source-memory choice indicating the encoding task previously performed for a given word or that the word was new ("A", "F", "C", or "N", respectively described to participants as "Artist", "Function", "Cost", and "New"). The ordering of these responses was counterbalanced across participants, such that half of the participants used the index through pinky finger of their right hand to indicate "A", "F", "C", and "N", while the other half used the pinky through index finger of their left hand to indicate "N", "C", "F", and "A" . For the first step, each test item was centrally displayed for 3 s, with the four response options simultaneously appearing at the bottom of the screen, and participants were instructed to make their response during this time. If participants indicated an encoding task for a given test item (regardless of whether it was the correct task), the second step of the response procedure proceeded. In this step, the test item disappeared and a new set of response options appeared along the bottom of the screen for 3 s. During this time, participants made one of three button-press responses to indicate their relative confidence ("1", "2", or "3", respectively described to participants as "low", "moderate", and "high") for the preceding source-memory decision. For this response, participants were instructed to use the hand opposite the one they used to make the source memory decision. The ordering of these responses was also counterbalanced, with half of the participants using the ring through index finger of their left hand to respond "1", "2", or "3", and the other half of participants using the index through ring finger of their right hand to respond "3", "2", or "1". Judgments indicating that a test item was new for the first step, however, were followed by the presentation of a centrally displayed fixation cross instead of the aforementioned confidence-rating options for the same amount of time (3 s). The central

fixation marker occurring between trials was presented for a randomly-chosen interval of 3, 5, or 7 s (48, 16, and 8 instances, respectively, in each test block). Responses occurring beyond any of the designated time windows were not analyzed.

MRI Data Acquisition and Preprocessing

Whole-brain MRI data were obtained at the MU Brain Imaging Center on a 3-Tesla Siemens Magnetom TIM Trio scanner equipped with an 8-channel head coil (Siemens Medical Solutions, Erlangen, Germany). fMRI data were acquired using an echo-planar imaging (EPI) pulse sequence sensitive to blood-oxygen-level dependent (BOLD) contrast (T2*-weighted, 2-s TR, 30-ms TE, 90° flip angle). Each fMRI volume consisted of 32 axial slices (3-mm thick, 1-mm gap, ascending interleaved acquisition) with an in-plane resolution of 3 × 3 mm (192-mm FOV, 64 × 64 matrix). The fMRI data were acquired in six separate runs, corresponding to the three encoding-phase blocks (220 volumes each) and the three retrieval-phase blocks (332 volumes each). T1-weighted anatomical data were acquired sagittally with an MP-RAGE pulse sequence (176 slices, 256-mm FOV, 1-mm isotropic voxels).

The fMRI data were pre-processed with the SPM8 toolbox (Wellcome Trust Centre for Neuroimaging, London, UK; http://www.fil.ion.ucl.ac.uk/spm) in MATLAB (R2012a; MathWorks, Natick, MA) prior to analysis. The data were spatially realigned to the first volume of the first run and then to the mean volume across runs. Differences in slice-acquisition time were corrected by temporally shifting (via sinc interpolation) the time series of each voxel to the TR midpoint. The functional and anatomical data were then co-registered. A unified segmentation procedure (Ashburner & Friston, 2005) was

used to segment the anatomical data into gray and white matter and to deform the resulting images to a set of standard tissue probability maps (International Consortium for Brain Mapping; http://www.loni.ucla.edu/ICBM/). The fMRI data were then normalized with the deformation parameters determined by the segmentation procedure and resampled into 3-mm isotropic voxels. For the univariate analyses, the normalized data were smoothed by an 8-mm full-width half-maximum Gaussian kernel, and the time series in each voxel was high-pass filtered at 1/128 Hz and scaled to a grand mean of 100 (over all voxels). For the multivariate analyses, the normalized data were kept unsmoothed, and the time series for each voxel were linearly and quadratically detrended and z-scored within scanning runs.

Analysis of the fMRI Data

The fMRI analyses employed a combination of univariate and multivariate approaches. Group-based univariate analyses were used to identify regions sensitive to the confidence of source-memory judgments. Multivariate analyses, conducted on an individual-participant basis, involved training a pattern classifier to detect differences in neural activity across the three tasks at encoding and testing the classifier on data from the retrieval phase. Classifier performance therefore depended on the extent to which there were similarities in the neural activity patterns across encoding and retrieval, providing an index of reinstatement. The reinstatement measure was then assessed according to activity in the regions sensitive to source confidence (from the univariate analyses), allowing us to evaluate our hypotheses about the relationship between those regions and the reinstatement measure.

Univariate Analyses. Univariate analyses were performed only on data from the test phase, using a 2-stage mixed effects model in SPM8. For the first stage, items during the test phase were divided into four conditions of interest: old words eliciting correct source judgments followed by high confidence responses, old words eliciting correct source judgments followed by moderate and low confidence responses, old words eliciting incorrect source or new (miss) responses, and new words that were correctly rejected. The neural activity elicited by each test word was modeled as a delta function (i.e., impulse event) at stimulus onset. The resulting condition-wise functions were convolved with a canonical hemodynamic response function (HRF) to model the ensuing BOLD response, then downsampled at the midpoint of each scan to form covariates (defined below) in a General Linear Model (GLM). The parameters for each covariate and the hyper-parameters governing the error covariance were estimated using a restricted maximum-likelihood (ReML) method. Non-sphericity of the error covariance was accommodated by an AR(1) model, in which the temporal autocorrelation was estimated by pooling over suprathreshold voxels (Friston et al., 2002). In addition, covariates were used to account for trials with multiple or omitted responses, motion-related effects (determined during image realignment), and across-session (constant) effects.

The second stage of analysis involved contrasting of the aforementioned parameter estimates, treating participants as a random effect. The primary contrast of interest was the effect of greater activity for high-confidence compared to moderate- and low-confidence source judgments (cf. Yu et al., 2012a, 2012b). Other contrasts (e.g., high > moderate > low confidence) were also examined but did not reveal any additional

regions that were of interest. All contrasts consisted of one-sample t-tests, and were thresholded for 10 or more contiguous voxels surviving $p < 0.001$, unless otherwise noted.

Multivariate Analyses. Multivariate analyses were primarily conducted with the Princeton Multi-Voxel Pattern Analysis (MVPA) toolbox (The Princeton Neuroscience Institute, Princeton, NJ; https://code.google.com/p/princeton-mvpa-toolbox/), with additional functionality implemented in SPM8 and custom MATLAB code. These analyses followed a pattern-classification procedure that was carried out for each participant individually, and the results were then averaged across participants. The classifiers employed here were based on regularized (L2) logistic regression, in which weights for multiple input features of a model were simultaneously estimated and then summed to generate the model output. A pattern of input features was comprised of the activity (intensity) levels of individual voxels obtained from a single time point (TR) of fMRI data. Each pattern was labeled according to the experimental condition (i.e. the encoding task) with which it corresponded, in order to train the classifier to discriminate between conditions. With respect to the onset of each encoding word (corresponding to what we hereafter refer to as the first TR), the fMRI patterns from 4-8 s post-stimulus onset (the third and fourth TRs) were labeled according to the task completed. This shift in TRs roughly corresponds to the delayed peak of the canonical HRF (also see Johnson et al., 2009; Rissman, Greely, & Wagner, 2010; Kuhl, Rissman, & Wagner, 2012) and resulted in 8 input patterns for each miniblock. A feature-selection procedure was also used to select the 5000 voxels exhibiting the largest F-values from an ANOVA

contrasting the three tasks (conducted separately for each cross-validation iteration), thereby reducing the influence of less informative voxels. A regularization (L2) value of 100 was also used for each classification.

The trained classifier was evaluated by presenting it with fMRI data from the retrieval phase and assessing its ability to predict the prior encoding task condition (Artist, Function, or Cost) associated with each test item. Classifier performance was assessed with two measures: accuracy and evidence. Whereas accuracy corresponds to a simplified index of performance over multiple trials, classifier evidence provides a more-graded measure of the magnitude of performance. These two measures are described in further detail in the corresponding sections of the Results. In addition to assessing these measures at the fourth and fifth TRs following each test item (i.e. where classifier performance should peak), we also constructed peri-stimulus time courses of performance beginning with item onset (TR 1) and lasting for 8 TRs (similar to what is shown in Figure 1 for our previous study).

Finally, to identify those voxels that were influential to the classification, *importance maps* were created by multiplying the trained weight for a given voxel by its average activity during the encoding phase. Voxels with positive values for both the activity and weight were assigned positive importance values, voxels with negative values for activity and weight were assigned negative importance values, and those with opposite-signed activity and weight were assigned importance values of zero (Johnson et al., 2009; McDuff et al., 2009; cf. Polyn et al., 2005). Because these importance maps are purely descriptive, they are relegated to the Appendix.

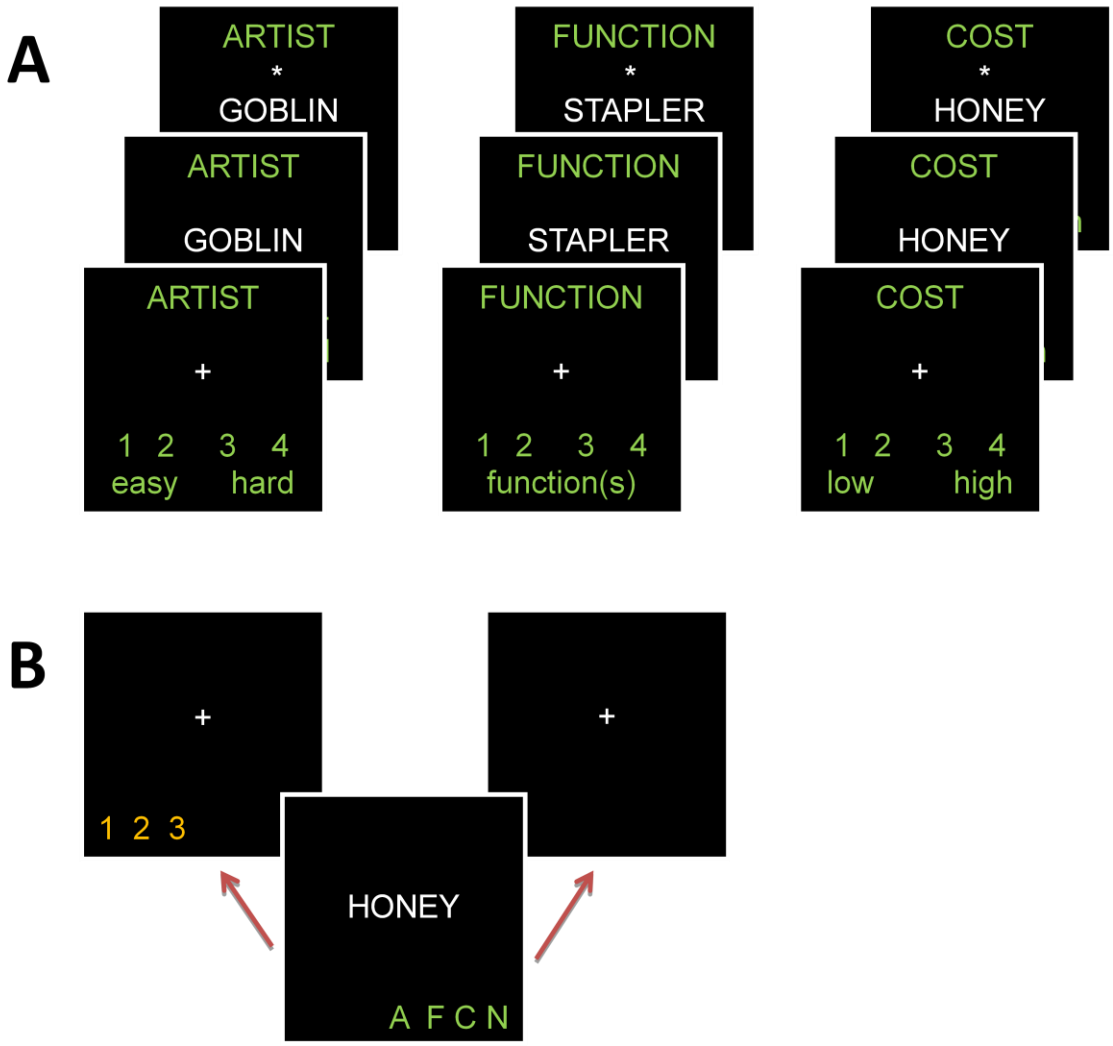

Figure 2. Schematics of the task sequence and example stimuli for the encoding and retrieval phase. (*A*) During the encoding phase, participants viewed a series of words in the context of three encoding tasks (Artist, Function, Cost). Depending on the task for which a given word was presented, participants were instructed either to imagine how an artist would draw that item and rate the difficulty of drawing the item (Artist), think of as many functions as possible for the item and indicate the number of functions they generated (Function), or think about the relative cost of the item and rate that cost (Cost). The sequence of trials during each of the encoding tasks began with the presentation of

the respective task instructions and response options, along with a central fixation cross for 3 s. Next, words were presented in place of the fixation cross for 3 s each. A star appeared above each word for the final 1 s of display to indicate that a response should be made. (*B*) During the retrieval phase, participants viewed a series of randomly intermixed old and new words, and responded in a two-step procedure: first, participants indicated the source (encoding task) for a given word, or that a word was new; next, if participants designated the source for an item, they then indicated their confidence in that source decision (participants did not make confidence decisions following a new response). Each trial began with the presentation of an old or new word, along with the source (A, F, C) and new (N) response options for 3 s. Following a source response, confidence ratings (1, 2, and 3, described respectively to participants as low, moderate, and high confidence) were then displayed for 3 s (shown left). Participants were required to make their responses within the 3-s timeframe for each step (source and confidence decisions).Following a new response, a fixation cross appeared for 3 s and no additional response was required (shown right).

RESULTS

Behavioral Results

The behavioral data from the retrieval phase were first analyzed according to item recognition, which entailed collapsing across the first-step test responses to disregard whether source memory was correct or incorrect. The proportions of correct item-recognition responses and the associated response times (RTs) are provided in Table 1. As shown, participants were highly and comparably accurate at recognizing items from each of the three prior encoding-task conditions and rejecting new items. A one-way ANOVA of the mean proportions of correct responses indicated no significant differences across the four item conditions ($p = .17$). For the mean RTs associated with these correct responses, an ANOVA revealed a significant effect of item condition ($F(3,45) = 23.22$, $p < .001$). Follow-up pair-wise tests revealed that RTs were shorter for new items compared to each of the other conditions (all $t(15) > 2.20$, $p < .05$) and also shorter for items from the artist condition relative to those from the function and cost conditions ($t(15) = 5.82$, $p < .001$; $t(15) = 5.03$, $p < .001$, respectively).

The correct item-recognition responses to old items were next segregated according to whether they indicated the correct versus incorrect source judgment. The proportion and RT data for these responses are also provided in Table 1. As shown in the table, participants responded with the correct source judgment on the majority of trials. Due to the infrequency of incorrect source responses and new responses to old items (misses), the remaining behavioral analyses were restricted to correct source judgments. A one-way ANOVA of the mean proportions gave rise to a significant effect of prior task

($F(2,30) = 6.9$, $p < .005$). Source recognition was higher for the artist items relative to items from the other task conditions (vs. function: $t(15) = 2.77$, $p < .05$; vs. cost: $t(15) = 4.06$, $p < .005$). An ANOVA of the correct-source RTs also revealed a significant effect of prior task ($F(2,30) = 15.41$, $p < .001$), indicating that RTs for items from the artist condition were shorter than those for the other conditions (vs. function: $t(15) = 4.17$, $p < .001$; vs. cost: $t(15) = 5.07$, $p < .001$).

We next assessed the confidence ratings associated with correct source judgments, the mean proportions of which are shown in Table 2. These data were analyzed with a two-way ANOVA that included factors of prior task (artist, function, cost) and confidence (high, moderate, low). The analysis revealed a significant main effect of confidence ($F(2,30) = 67.11$, $p < .001$) and a significant interaction ($F(4,60) = 8.03$, $p < .001$). To interpret the interaction, we examined the effect of prior task separately within each confidence level. An ANOVA of the proportions of high-confidence responses revealed a significant task effect ($F(2,30) = 11.58$, $p < .001$), indicating a larger proportion of high-confidence responses for items from the artist task compared to those from the function and cost tasks ($t(15) = 4.91$, $p < .001$; $t(15) = 2.50$, $p < .025$, respectively). There was also a significant effect of task for moderate-confidence judgments ($F(2,30) = 8.33$, $p < .005$), which indicated a larger proportion of moderate-confidence responses for items from the function task as opposed to the artist task ($t(15) = 4.53$, $p < .001$). The ANOVA of low-confidence judgments did not reveal a significant effect ($p = .86$).

Finally, we did not analyze the RTs associated with the source-confidence judgments, as these data were deemed uninformative given that participants could have

anticipated the timing of the cue to make their confidence judgment (due to a lack of temporal jitter separating the first and second response steps). We instead segregated the mean RTs for the correct-source judgments (first-response step) according to the confidence response that followed. Due to the low frequencies of moderate- and low-confidence responses, we collapsed the RT data over these two response types. The RTs for the collapsed moderate- and low-confidence responses (M = 2074 ms, SD = 205) were significantly longer than those associated with high-confidence judgments (M = 1740 ms, SD = 149; $t(15) = 8.78$, $p < .001$).

fMRI Results

Reinstatement and Source Confidence. The fMRI analysis first employed MVPA to examine the reinstatement of patterns of neural activity from encoding at the time of retrieval. Using fMRI data from the encoding phase, we trained a whole-brain pattern classifier for each participant to distinguish between the three encoding tasks (Artist, Function, and Cost). Following training, the classifier was independently evaluated with the data from the retrieval phase. Classifier performance at correctly identifying the prior encoding task for a given test item should reflect the degree to which neural activity associated with that task is reactivated (reinstated) during retrieval.

To provide an initial assessment of reinstatement, we obtained a simple measure of how accurate the classifier was at identifying the prior encoding task associated with test items. Time courses of classifier accuracy were constructed starting with the onset of each test item (the first TR) and extending for seven additional TRs (14 s), allowing us to capture the expected delay in the hemodynamic response associated with each item.

Accuracy should be near chance (33%, given that there were three tasks) at item onset, peak around 8 seconds later, and finally return to chance for the next trial (due to items being randomly drawn from the three tasks). Figure 3A shows the accuracy results for items that were designated with correct source judgments (collapsed across the subsequent confidence judgments). As shown, the time course of classifier accuracy followed the delayed progression that we anticipated, peaking at the fourth TR. To avoid inflating the family-wise (Type 1) error rate that would likely result from testing the data against chance at each TR, we limited our analysis to data averaged over the fourth and fifth TRs, based on previous studies demonstrating peak effects in that time period (e.g., Johnson et al., 2009; Kuhl et al., 2012). Classifier accuracy averaged over those TRs reached 42%, which was significantly greater than chance ($t(15)= 6.06$, $p < .001$, one-tailed).

We next turned our analysis to investigating whether encoding-related reinstatement differed according to the confidence with which participants reported retrieving source information. To assess the magnitude of reinstatement, we computed the difference between classifier output for the correct encoding task for a test item and the mean classifier output for the two remaining (incorrect) encoding tasks. This measure of *classifier evidence* therefore reflects the strength that the classifier preferred the correct encoding task. As preference for the correct task increases, classifier evidence will rise above zero; preference for the incorrect encoding tasks will correspond to a negative evidence value, while a value of zero indicates chance performance.

To investigate how reinstatement magnitude might vary with source confidence, we obtained separate measurements of classifier evidence for the different confidence

responses. As described earlier for the behavioral results, low numbers of items received either moderate- or low-confidence judgments. Thus, we collapsed the results for these items into a single category, which we hereafter refer to as "low" confidence. This procedure ensured that each category of confidence (low and high) was associated with at least 12 trials per participant. (The results of a subsidiary analysis in which each confidence level was examined individually are provided in the Appendix.) Time courses of classifier evidence for items judged with high and low confidence, obtained for eight consecutive TRs starting with item onset, are displayed in Figure 3B. As shown, evidence for items designated with high confidence appeared to exceed that for low confidence. As was done with classifier accuracy, we collapsed the evidence measure over the peak period of the fourth and fifth TRs for statistical analysis. As anticipated, classifier evidence for high-confidence significantly exceeded chance ($t(15) = 6.42$, $p < .001$, one-tailed). In contrast, evidence associated with low confidence was not greater than chance ($p = .09$, one-tailed). Importantly, classifier evidence was significantly greater for items designated with high confidence than those designated with low confidence ($t(15) = 2.62$, $p < .01$, one-tailed).

Reinstatement and Neural Correlates of Source Confidence. Having demonstrated variation in the magnitude of reinstatement according to source memory confidence, we next sought to relate the reinstatement effects to regions where the level of activity is sensitive to source confidence. Regions of this type were identified with a univariate (GLM-based) analysis of greater activity for items designated with high compared to low source confidence, collapsing over the previous encoding task

conditions. The outcome of this analysis (thresholded at $p < .001$ for 10 contiguous voxels) is shown in Figure 4A and detailed in Table 3. As shown, several regions – including left posterior parietal cortex, medial prefrontal cortex (PFC), and posterior cingulate – exhibited greater activity for items associated with high confidence (also see Yu et al., 2012a, 2012b).

One notable region missing from the foregoing results is the hippocampus, in which activity has also been previously shown to correlate with graded changes in episodic retrieval (Rugg et al., 2012; Yu et al., 2012b). To further test for effects in this region, we repeated the above analyses using an anatomically-defined mask of bilateral hippocampus (Tzourio-et al., 2002). Doing so allowed us to reduce the family-wise error rate and thus use a more liberal threshold of $p < .005$ for 5 contiguous voxels (also see Vilberg & Rugg, 2007; Rugg et al., 2012). The results of this analysis, shown in Figure 4B, comprised clusters in left (20 voxels; peak coordinates: -27, -16, -17; peak $z = 3.10$) and right (12 voxels; peak coordinates: 42, -16, -17; peak $z = 3.15$) anterior hippocampus where activity was greater for high- than low-confidence responses.

The outcome of the foregoing analysis was next used to construct regions of interest (ROIs) that could be used in correlational analyses with the effects of reinstatement magnitude. In addition to treating the voxels identified by the univariate analysis as a whole (i.e. one ROI including all of the voxels reported in Table 3), we also extracted the time-course data from five separate ROIs – corresponding to left posterior parietal cortex, medial PFC, posterior cingulate, and bilateral hippocampus – that have been consistently identified in previous studies as sensitive to recollection (Yu et al., 2012a, 2012b; Rugg et al., 2012). The extracted time courses for these regions are

displayed in Figure 5. These time-course data should exhibit effects that are analogous to the GLM-based parameter estimates used to identify these regions (i.e. high > low confidence). Indeed, averaging over the fourth and fifth TRs gave rise to a significant effect for each of these ROIs (all $t > 2.44$, $p < .05$), with the exception of left hippocampus ($p = .1$). We note further that the correlational analyses relating reinstatement magnitude to each of these measures – the peak time course and the parameter estimates – gave rise to similar results. Only the results based on the time courses are reported here, as those data were subjected to the same pre-processing methods as the data used for the classification (reinstatement) analysis.

The correlational analyses first took an across-participant approach, in which the mean activity from each ROI described above and the mean reinstatement magnitude (as defined by the measure of classifier evidence) was extracted for each participant. This analysis was first performed on data collapsed over all trials in which a correct source response was made (without regard to the confidence response). Correlating reinstatement with each ROI resulted in a significant positive relationship for the left posterior parietal region (Spearman's $r = .66$, $p < .01$). These data are shown in Figure 6. None of the other ROIs exhibited a significant correlation (range of r-values: .04 to .24, all $p > .1$). For the left posterior parietal region, secondary correlational analyses performed separately for high- and low-confidence responses yielded a significant result for both high- and low-confidence responses ($r = .68$, $p < .01$; $r = .51$, $p < .05$, respectively). These data are also displayed in Figure 6. Further analyses, in which we accounted for the overall level of activity in the voxels exhibiting reinstatement effects (i.e. included in the classification analyses), yielded the same pattern of effects, with a

significant positive relationship evident between reinstatement and left posterior parietal activity ($r = .64$, $p < .01$).

Finally, we extended the correlational analyses by examining the correspondence between classifier evidence and ROI activity on a trial-by-trial basis. For these analyses, the individual-trial values of reinstatement magnitude and the level of activity in a given ROI were correlated separately for each participant. The r-values for each participant were then transformed into z-values (Fisher, 1915), and one-sample t-tests were used to compare these values to zero (for a similar approach, see Staresina, Henson, Kriegeskorte, & Alink, 2012). In contrast to the results of the across-participant analyses reported above, a significant negative correlation was identified for the left posterior parietal ROI ($M = -.07$, $SD = .11$; $t(15) = 2.35$, $p < .05$). None of the results for the other ROIs reached significance (range of M z-values: -.04 to -.01, all $p > .2$). As before, partialling-out the overall level of activity in reinstatement voxels yielded the same pattern of results, in which the correlation was significant only for the left posterior parietal region ($r = -.06$, $p < .05$; for all other ROIs, range of r-values: -.02 to .01, all $p > .2$).

Table 1. Mean (SD) correct proportions and corresponding response times (RTs) according to item and source recognition during the retrieval phase.

	Item condition			
	Artist	Function	Cost	New
Item recognition				
Proportion	0.90 (0.09)	0.85 (0.09)	0.85 (0.07)	0.83 (0.15)
RT (ms)	1774 (183)	1974 (169)	1930 (164)	1646 (221)
Source recognition				
Proportion	0.74 (0.16)	0.63 (0.16)	0.61 (0.15)	
RT (ms)	1704 (212)	1944 (184)	1891 (127)	

Table 2. Mean (SD) proportions of each confidence judgment for correct source responses.

	Item condition		
	Artist	Function	Cost
Confidence level			
High	0.57 (0.20)	0.40 (0.21)	0.43 (0.17)
Moderate	0.11 (0.07)	0.17 (0.11)	0.11 (0.09)
Low	0.04 (0.04)	0.04 (0.04)	0.04 (0.03)

Table 3. Regions exhibiting greater activity for high-confidence compared to low-confidence (moderate and low collapsed) source judgments.

Region	Brodmann area	k	Peak *z*	Peak coordinates (x, y, z)
Ventral medial prefrontal cortex	32	489	4.90	0, 23, -11
R white matter (frontal)		32	4.61	27, 5, 22
L angular gyrus	39	65	4.55	-60, -67, 25
L anterior middle temporal gyrus	21	123	4.40	-51, 8, -32
L anterior medial prefrontal cortex	10	62	4.10	-12, 59,7
Posterior cingulate	31	93	4.09	-6, -40, 43
L superior frontal gyrus	9	54	4.08	-6, 62, 37
L superior temporal gyrus	22	52	4.06	-63, -49, 13
R cuneus	19	128	4.00	15, -88, 37
R anterior middle temporal gyrus	21	52	3.98	60, 5, -17
L inferior parietal lobule	40	40	3.87	-63, -40, 37
R angular gyrus	39	11	3.83	57, -70, 25
L cuneus	7	10	3.76	-15, -88, 46
R inferior parietal lobule	40	18	3.70	69, -25, 40
R insula	21	11	3.70	42, -7, -11
R white matter (temporal)		12	3.58	48, -25, -17
L retrosplenial cortex	23	18	3.56	-9, -64, 16
R superior temporal gyrus	21	17	3.55	69, -22, -5
R amygdala	34	12	3.49	12, 2, -14
L angular gyrus	39	10	3.47	39, -67, 25
Occipital pole	18	15	3.40	3, -97, 14
R lateral ventricle		14	3.39	36, -49, 1
L angular gyrus	39	11	3.37	-45, -73, 37
L cuneus	19	18	3.32	-3, -91, 34

Note. L = left, R = right, k = number of voxels. Coordinates are in MNI space.

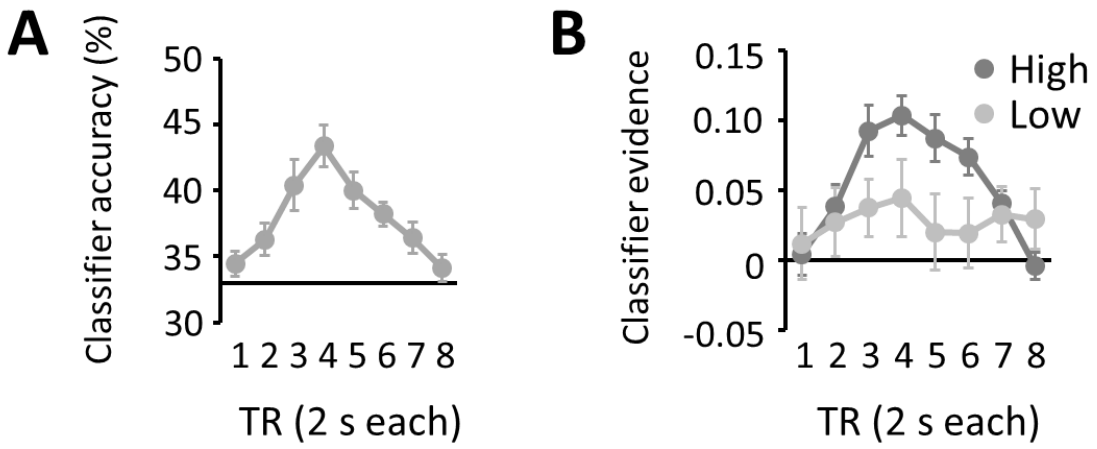

Figure 3. Classifier accuracy and evidence indicating the reinstatement of encoding-related neural activity during source memory retrieval. The first time point (TR 1) corresponds to item onset. Error bars reflect ±SEM. (A) Mean classifier accuracy collapsed over all test items designated with correct source responses. The horizontal axis is placed at chance level of accuracy (33%). (B) Mean classifier evidence for items designated with correct source responses separated according to high and low confidence. Evidence is computed as the difference between classifier output for the correct encoding task and the mean classifier output for the two remaining (incorrect) tasks (chance = 0).

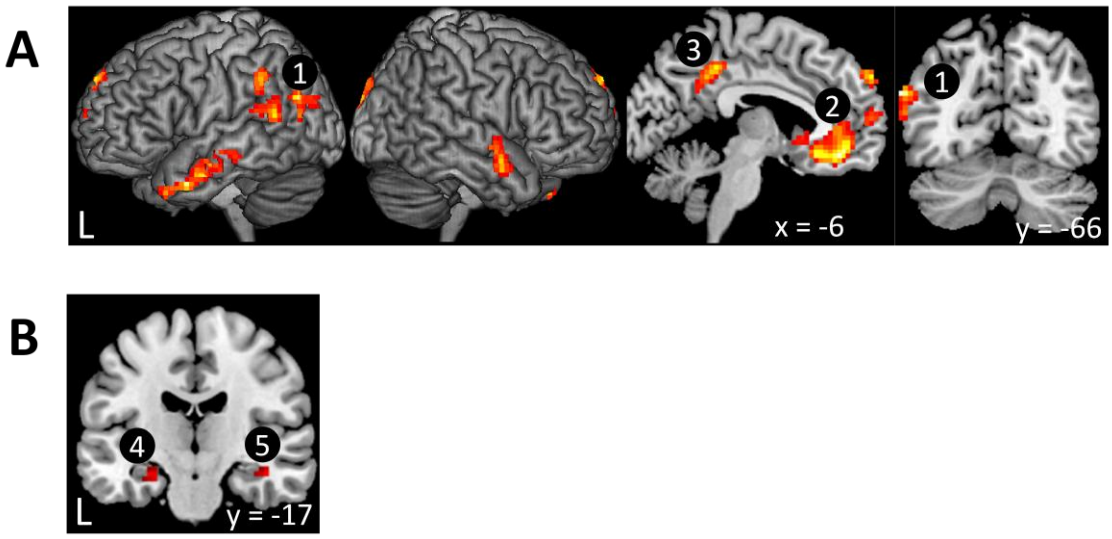

Figure 4. Regions exhibiting source confidence effects (High > Low) for test items designated with correct source responses. The effects are overlaid on renderings and slices of a standard anatomical template. L = left hemisphere. Coordinates are in MNI space. (A) Clusters in (1) left parietal, (2) medial PFC, and (3) posterior cingulate ROIs. (B) Clusters in (4) left and (5) right hippocampus, identified at a lower statistical threshold.

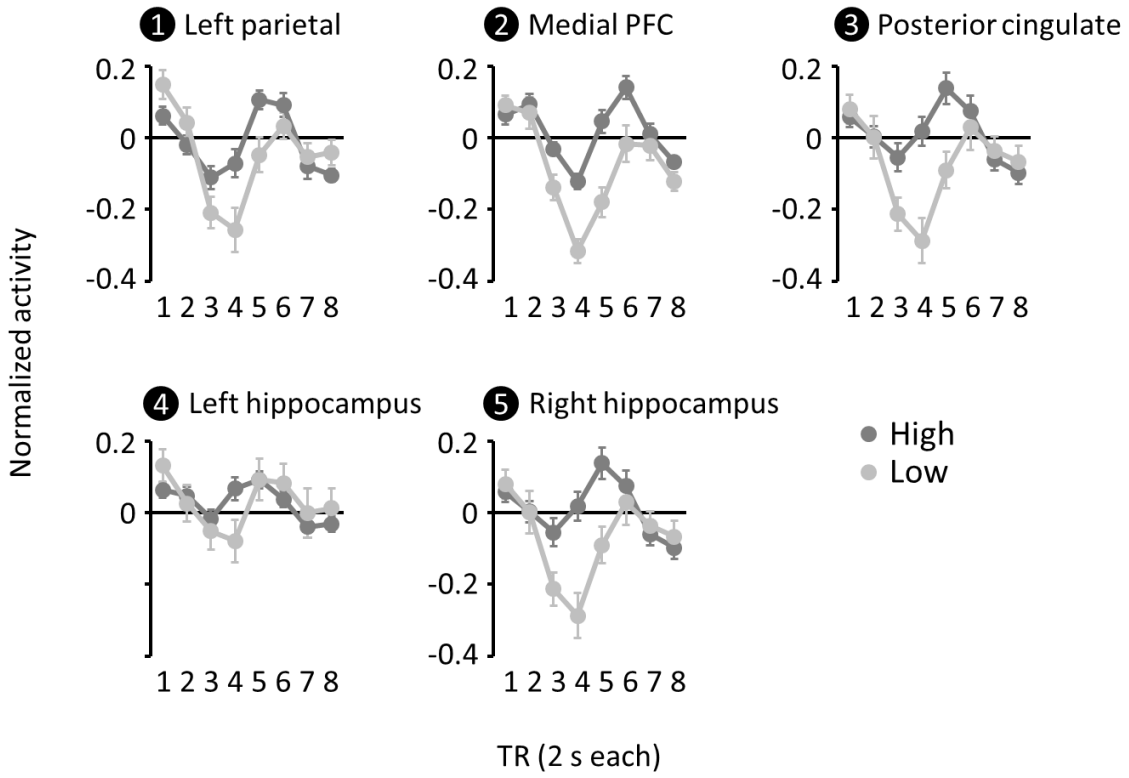

Figure 5. Time courses of activity in select ROIs for test items designated with correct source responses, separated according to high- and low-confidence. Mean activity is displayed for (1) left parietal, (2) medial PFC, (3) posterior cingulate, (4) left hippocampus, and (5) right hippocampus (error bars: ±SEM). Time courses begin with item onset at the first time point (TR 1).

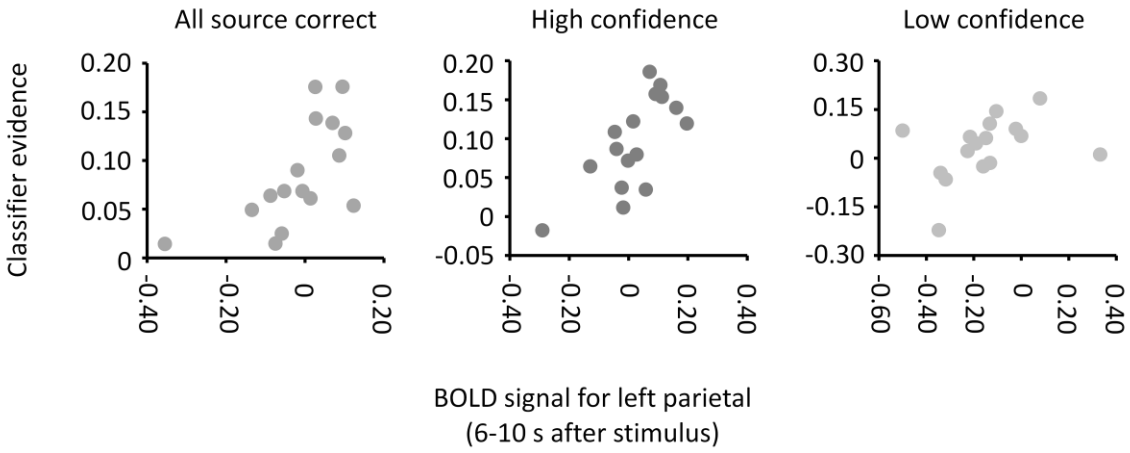

Figure 6. Group-level correlations of classifier evidence with activity in left parietal region. Mean classifier evidence is plotted against BOLD signal in the left parietal ROI (averaged over TRs 4 and 5) for source-correct responses, collapsed over all confidence response categories (*left panel*), and separated according to high- (*center panel*) and low-confidence (*right panel*).

DISCUSSION

The purpose of this study was to investigate how the reinstatement of encoding-related neural activity during memory retrieval relates to participants' assessments of the subjective qualities of retrieved information. We employed pattern-classification analyses of fMRI data to assess the magnitude with which patterns of brain activity associated with encoding were reactivated (reinstated) at the time of retrieval. These reinstatement effects were then related to both behavioral and neural measures of how confident participants were about source memory retrieval. We demonstrate here that reinstatement increases with increasing source confidence, and identify activity in a set of regions that is sensitive to source confidence. Interestingly, the level of activity in left posterior parietal cortex – a region consistently shown to be sensitive to source memory retrieval – exhibited different relationships with the magnitude of reinstatement depending on whether the correlational analyses were carried out at the group level or at the individual trial level. In the following discussion, we attempt to reconcile these seemingly disparate correlational findings, with the goal of incorporating the reinstatement of episodic information into existing theories of how participants use that information in service of making subjective retrieval judgments.

By employing whole-brain MVPA, we successfully classified with above-chance accuracy the prior encoding condition of old items that received correct source designations. Consistent with the findings of previous studies, we interpret these results as evidence of retrieval-related reinstatement of information from encoding (Polyn et al., 2005; Johnson et al., 2009; McDuff et al., 2009; Kuhl et al., 2011). Using classifier

evidence to measure the magnitude of reinstatement according to confidence responses for correct source identifications, we found that the level of reinstatement for high confidence responses significantly exceeded that of moderate- and low-confidence responses. Consistent with Leiker and Johnson (2014), these findings lend further support to the notion that even strong memories involving the recollection of qualitative (e.g., source) information may be based on a graded process or signal, as opposed to relying on an all-or-none process (also see Wixted, 2007; Johnson et al., 2009; Mickes, Wais, & Wixted, 2009; Wixted & Mickes, 2010). The current findings are the first, to our knowledge, to move beyond a simple association between variable levels of reinstatement and different recollection-based memory judgments, to indicate that varying levels of reinstatement may actually inform participants' memory decisions.

Related to the aforementioned findings of changes in reinstatement with respect to source confidence, one result that stands out is the near-chance level of reinstatement for low-confidence (and to some extent, moderate-confidence) source judgments. One possible explanation for this modest result is that, by focusing participants on different levels of source memory, our retrieval task may have operated in a manner analogous to the list-strength effect of behavioral studies of memory (Ratcliff, Clark & Shiffrin, 1990; Shiffrin, Ratcliff, & Clark, 1990; Murnane & Shiffrin, 1991; also see Norman, 2002; Diana & Reder, 2005; Norman, Tepe, Nyhus, & Curran, 2008). By this account, reinstatement may have been especially strong on a subset of trials, leading participants to make high-confidence judgments about source memory. When reinstatement was not as strong, however, participants may have been more inclined to settle on making a lower-confidence response than attempt to reinstate additional information.

An alternative possibility is that the low-confidence response category includes trials for which participants merely guessed the correct source, instead of recollecting source-identifying information. In this case, the inclusion of trials for which reinstatement presumably did not occur (i.e. guesses) would cause the level of reinstatement for the low-confidence source category to appear artificially lower. In follow-up studies, the inclusion of a guessing option for the source judgment (the first response step) might help to segregate such trials from those in which reinstatement is the basis for responding.

In addition to the positive relationship between reinstatement magnitude and the behavioral measure of source memory confidence, we also examined other neural correlates of source memory. Consistent with the findings of previous studies, we identified a set of regions where activity was sensitive to participants' confidence about correct-source judgments, such that activity was greater for high- relative to low-confidence responses. This set of regions included left posterior parietal cortex, medial PFC, and posterior cingulate, which have often been identified as related to source memory and other graded measures of recollection (Vilberg & Rugg, 2007, 2009a, 2009b; Guerin & Miller, 2011; Yu et al., 2012a, 2012b). Additionally, we identified a similar pattern of graded activity in bilateral hippocampus, but at a lower statistical threshold, providing further support for previous findings of hippocampal sensitivity to confidence about recollection (Rugg et al., 2012; Yu et al., 2012b). Notably, the low-confidence response category used to identify source-confidence effects in the current study differed from that of prior studies. Whereas the latter comprised low-confidence source correct trials and trials where the incorrect source was designated, in the current

study we restricted this response category to only those trials where source memory was correct (including moderate- and low-confidence responses). Although this difference in approach was primarily dictated by the number of trials available in each condition, it ultimately resulted in allowing us to strengthen the conclusions drawn in previous studies. That is, we demonstrate here that these regions are interested in source confidence when accuracy is held constant, rather than being confounded with source accuracy (as was the case by including incorrect source judgments in previous studies). Together, the findings of these studies fit with the interpretation that these regions are sensitive to a memory signal that reflects variation in the qualitative information that is retrieved.

Having explored the relationship between source memory confidence and reinstatement, we next turn to the results of the correlational analyses between the magnitude of reinstatement and the aforementioned neural correlates of source memory. These analyses took two forms. The first was a group-level analysis, for which the mean activity from a given ROI and the mean reinstatement magnitude were correlated across participants. The second analysis followed a within-participant approach in which the mean activity from a given ROI was correlated with the mean reinstatement magnitude on a trial-by-trial basis, with the resulting correlation values then averaged across participants. Whereas the group-level analysis was employed to test the overall relationship between regions sensitive to source memory and reinstatement magnitude, the within-participant analysis has the potential to provide insight into how this relationship changes across individual trials. Of the regions described above as consistently sensitive to source memory, only activity in left posterior parietal cortex

exhibited a significant correlation with reinstatement magnitude at the group level. In other words, as the level of parietal activity increased for a given participant, the level of reinstatement increased as well. Such a result fits well with recent investigations identifying the parietal cortex as the only region, out of several recollection-sensitive regions, that is predominantly sensitive to finer gradations in the qualitative aspects of retrieved information. For this reason, the parietal cortex has been suggested to play a role in the maintenance or representation of information for retrieval (Vilberg & Rugg, 2007, 2009a, 2009b; Yu et al., 2012a, Leiker & Johnson, 2014). The results of the group-level analysis outlined above are consistent with such an appraisal.

Despite the congruency of the group correlations with the interpretation outlined above, the results of within-participant (trial-based) correlations paint a different picture. As reinstatement increased on a trial-by-trial basis, the level of parietal activity for that same trial actually decreased. This negative correlation between reinstatement and parietal activity during an individual trial appears to violate the notion that parietal cortex is tracking the accumulation of retrieved information (reinstatement) in service of the memory decision. Instead, such results may indicate a possible tradeoff between parietal activity and reinstatement during a single trial. Such a tradeoff would not necessarily contradict the positive group correlation discussed previously. That is, it is possible that both reinstatement and posterior parietal activity play roles in informing participants' retrieval decisions at the group level, but that only one of them is utilized in the making of an individual memory decision. In other words, although reinstatement and parietal activity might accomplish the same goal, participants might rely on only one or the other to make a single retrieval judgment. Nevertheless, the group-level correlation suggests

that participants who are more likely to rely on one of these processes are also more likely to rely on the other process, giving rise to enhanced memory performance.

One other result from the trial-based correlational analyses deserves mention. Specifically, we failed to observe a correlation between the level of hippocampal activity and reinstatement magnitude. This null result is noteworthy given that multiple studies have demonstrated positive correlations between these two variables (Ritchey, Wing, LaBar, & Cabeza, 2012; Staresina et al., 2012; Gordon, Rissman, Kiani, & Wagner, 2013). For example, Staresina et al. (2012) employed an index of reinstatement referred to as event-related similarity (ERS) in which the voxel-wise pattern of activity for a single item at encoding is correlated with the activity pattern for that item at retrieval. The authors observed reinstatement in parahippocampal cortex, and the level of reinstatement additionally correlated positively with the level of activity in hippocampus. Ritchey et al. (2012) found similar results, in which ERS patterns in occipital cortex and PFC cortices correlated positively with hippocampal activity. Based on the findings of these two studies, one possibility is that hippocampus is interested in event-specific reinstatement, rather than a broader form of task-specific reinstatement (as we investigated). A study by Gordon et al. (2013), however, seems to contradict that possibility: the positive correlation was identified even when classifying groups of items (as opposed to individual items). Upon closer examination, a characteristic common to these three studies is the focus of the encoding tasks on visual aspects of the items—either by presenting richly-detailed visual stimuli (e.g., pictures of rooms and landscapes in Staresina et al., 2012; emotionally-charged pictures in Ritchey et al., 2012) or by instructing participants to visualize such stimuli (famous faces and scenes in Gordon et

al., 2013). It is therefore unsurprising that the resulting reinstatement effects were either restricted to or largely evident in brain regions that are typically involved in visual perception, such as occipital and temporal cortices (PFC shown by Ritchey et al. being an exception). In contrast, only the Artist task of the current study was likely to have focused participants on visual processing, whereas the Function and Cost tasks likely involved more abstract cognitive representations. Thus, our failure to identify a correlation between hippocampus and reinstatement may relate to the degree to which reinstated activity involved visual processing.

To summarize, the current study demonstrates that variation in the neural signal reflecting reinstatement of episodic information during retrieval is related to source-memory confidence. These findings support the notion that neural reinstatement provides a basis for judging the subjective quality (source confidence) of retrieval. We replicate previous findings in which regions sensitive to source memory confidence have been identified, including left posterior parietal cortex, medial PFC, posterior cingulate, and right hippocampus. Furthermore, the current study makes the novel contribution of providing evidence of a relationship between reinstatement magnitude and the level of activity in posterior parietal cortex. Whereas this relationship changed in direction going from the group-based to trial-based levels, it provides a preliminary view of how reinstatement and recollection-based processes contribute to episodic memory retrieval.

REFERENCES

Alvarez, P., & Squire, L. R. (1994). Memory consolidation and the medial temporal lobe: a simple network model. *Proceedings of the National Academy of Sciences of the United States of America*, *91*(15), 7041–5.

Ashburner, J., & Friston, K. J. (2005). Unified segmentation. *NeuroImage*, *26*(3), 839–51.

Coltheart, M. (1981). The MRC psycholinguistic database. Quarterly Journal of Experimental Psychology, 33A, 497-505.

Danker, J. F., & Anderson, J. R. (2010). The ghosts of brain states past: remembering reactivates the brain regions engaged during encoding. *Psychological Bulletin*, *136*(1), 87–102.

Diana, R. A., & Reder, L. M. (2005). The list strength effect: a contextual competition account. *Memory & Cognition*, *33*(7), 1289-1302.

Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, *10*, 507-521.

Friston, K. J., Glaser, D. E., Henson, R. N. A, Kiebel, S., Phillips, C., & Ashburner, J. (2002). Classical and Bayesian inference in neuroimaging: applications. *NeuroImage*, *16*(2), 484–512.

Gordon, A. M., Rissman, J., Kiani, R., & Wagner, A. D. (2013). Cortical reinstatement mediates the relationship between content-specific encoding activity and subsequent recollection decisions. *Cerebral Cortex*.

Guerin, S. A., & Miller, M. B. (2011). Parietal cortex tracks the amount of information retrieved even when it is not the basis of a memory decision. *NeuroImage*, *55*(2), 801–807.

Hasselmo, M. E., & Wyble, B. P. (1997). Free recall and recognition in a network model of the hippocampus: simulating effects of scopolamine on human memory function. *Behavioural Brain Research*, *89*(1-2), 1–34.

Hayes, S. M., Buchler, N., Stokes, J., Kragel, J., & Cabeza, R. (2011). Neural correlates of confidence during item recognition and source memory retrieval: evidence for both dual-process and strength memory theories. *Journal of Cognitive Neuroscience*, *23*(12), 3959-3971.

Haynes, John-Dylan; Rees, G. (2006). Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience*, *7*, 523–534.

Jimura, K., & Poldrack, R. A. (2012). Analyses of regional-average activation and multivoxel pattern information tell complementary stories. *Neuropsychologia*, *50*(4), 544–552.

Johnson, J. D., Mcduff, S. G. R., Rugg, M. D., & Norman, K. A. (2009). Recollection, familiarity, and cortical reinstatement: a multivoxel pattern analysis. *Neuron*, *63*(5), 697–708.

Johnson, J. D., & Rugg, M. D. (2007). Recollection and the reinstatement of encoding-related cortical activity. *Cerebral Cortex*, *17*, 2507–2515.

Kahn, I., Davachi, L., & Wagner, A. D. (2004). Functional-neuroanatomic correlates of recollection: implications for models of recognition memory. *The Journal of Neuroscience , 24*(17), 4172–4180.

Kucera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

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

Kuhl, B. A., Rissman, J., & Wagner, A. D. (2012). Multi-voxel patterns of visual category representation during episodic encoding are predictive of subsequent memory. *Neuropsychologia*, *50*(4), 458–469.

Leiker, E. K., & Johnson, J. D. (2014). Neural reinstatement and the amount of information recollected. *Brain Research*, *1582*, 125-138.

Marr, D. (1971). Simple memory: a theory for archicortex. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *262*(841), 23–81.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, *102*(3), 419–457.

McDuff, S. G. R., Frankel, H. C., & Norman, K. A. (2009). Multivoxel pattern analysis reveals increased memory targeting and reduced use of retrieved details during single-agenda source monitoring. *The Journal of Neuroscience*, *29*(2), 508–516.

Mickes, L., Wais, P. E., & Wixted, J. T. (2009). Recollection is a continuous process: implications for dual-process theories of recognition memory. *Psychological Science*, *20*, 509-515.

Mur, M., Bandettini, P. A., & Kriegeskorte, N. (2009). Revealing representational content with pattern-information fMRI--an introductory guide. *Social Cognitive and Affective Neuroscience*, *4*(1), 101–109.

Murnane, K., & Shiffrin, R. M. (1991). Word repetitions in sentence recognition. *Memory & Cognition*, *19*(2), 119-130.

Norman, K. A. (2002). Differential effects of list strength on recollection and familiarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(6), 1083-1094.

Norman, K. A., & O'Reilly, R. C. (2003). Modeling hippocampal and neocortical contributions to recognition memory: a complementary-learning-systems approach. *Psychological Review*, *110*(4), 611–646.

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, *10*(9), 424–430.

Norman, K. A., Tepe, K., Nyhus, E., & Curran, T. (2008). Event-related potential correlates of interference effects on recognition memory. *Psychonomic Bulletin & Review*, *15*(1), 36-43.

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

Ratcliff, R., Clark, S. E., & Shiffrin, R. M. (1990). List-strength effect: I. data and discussion. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(2), 163-178.

Rissman, J., Greely, H. T., & Wagner, A. D. (2010). Detecting individual memories through the neural decoding of memory states and past experience. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(21), 9849–9854.

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

Ritchey, M., Wing, E. A., LaBar, K. S., & Cabeza, R. (2012). Neural similarity between encoding and retrieval is related to memory via hippocampal interactions. *Cerebral Cortex*, *12*, 2818-2828.

Rolls, E. T. (2000). Hippocampo-cortical and cortico-cortical backprojections. *Hippocampus*, *10*(4), 380–388.

Rugg, M. D., Johnson, J. D., Park, H., & Uncapher, M. (2008). Encoding-retrieval overlap in human episodic memory: a functional neuroimaging perspective. *Progress in Brain Research*, *169*(7), 339-352.

Rugg, M. D., Vilberg, K. L., Mattson, J. T., Yu, S. S., Johnson, J. D., & Suzuki, M. (2012). Item memory, context memory and the hippocampus: fMRI evidence. *Neuropsychologia*, *50*(13), 3070–3079.

Shastri, L. (2002). Episodic memory and cortico-hippocampal interactions. *Trends in Cognitive Sciences*, *6*(4), 162–168.

Shiffrin, R. M., Ratcliff, R., Clark, S. E. (1990). List-strength effect: II. theoretical mechanisms. *Journal of Experimenatl Psychology: Learning, Memory, and Cognition*, *16*(2), 179-193.

Staresina, B. P., Henson, R. N. A., Kriegeskorte, N., & Alink, A. (2012). Episodic reinstatement in the medial temporal lobe. *Journal of Neuroscience*, *32*, 18150-18156.

Teyler, T. J., & DiScenna, P. (1986). The hippocampal memory indexing theory. *Behavioral Neuroscience*, *100*(2), 147–154.

Tong, F., & Pratte, M. S. (2012). Decoding patterns of human brain activity. *Annual Review of Psychology*, *63*, 483–509.

Tulving, E. (1985). Memory and consciousness. *Canadian Psychology*, *26*(1), 1–12.

Vilberg, K. L., & Rugg, M. D. (2007). Dissociation of the neural correlates of recognition memory according to familiarity, recollection, and amount of recollected information. *Neuropsychologia*, *45*(10), 2216–2225.

Vilberg, K. L., & Rugg, M. D. (2009a). Functional significance of retrieval-related activity in lateral parietal cortex: evidence from fMRI and ERPs. *Human Brain Mapping*, *30*(5), 1490–1501.

Vilberg, K. L., & Rugg, M. D. (2009b). Lateral parietal cortex is modulated by amount of recollected verbal information. *Neuroreport*, *20*(14), 1295–1299.

Wheeler, M. E., Petersen, S. E., & Buckner, R. L. (2000). Memory's echo: vivid remembering reactivates sensory-specific cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(20), 11125-11129.

Wheeler, M. E., & Buckner, R. L. (2004). Functional-anatomic correlates of remembering and knowing. *NeuroImage*, *21*(4), 1337–1349.

Wilson, M. D. (1988). The MRC Psycholinguistic Database: Machine Readable Dictionary, Version 2. *Behavioural Research Methods, Instruments and Computers*, *20*(1), 6-11.

Wixted, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review*, *114*, 152-176.

Wixted, J. T., & Mickes, L. (2010). A continuous dual-process model of remember/know judgments. *Psychological Review*, *117*, 1025-1054.

Woodruff, C. C., Johnson, J. D., Uncapher, M. R., & Rugg, M. D. (2005). Content-specificity of the neural correlates of recollection. *Neuropsychologia*, *43*, 1022-1032.

Yonelinas, A. P. (2002). The nature of recollection and familiarity: a review of 30 years of research. *Journal of Memory and Language*, *46*, 441-517.

Yonelinas, A. P., Otten, L. J., Shaw, K. N., & Rugg, M. D. (2005). Separating the brain regions involved in recollection and familiarity in recognition memory. *The Journal of Neuroscience*, *25*, 3002-3008.

Yu, S. S., Johnson, J. D., & Rugg, M. D. (2012a). Dissociation of recollection-related neural activity in ventral lateral parietal cortex. *Cognitive Neuroscience*, *3*(3-4), 142-149.

Yu, S. S., Johnson, J. D., & Rugg, M. D. (2012b). Hippocampal activity during recognition memory co-varies with the accuracy and confidence of source memory judgments. *Hippocampus*, *22*(6), 1429–1437.

APPENDIX

Supplemental Material

Figure A1. Importance maps indicating the voxels that were influential for the classification analysis that assessed encoding-retrieval reinstatement. "Importance" values were computed by multiplying the trained weight for a given voxel by that voxel's average activity during the encoding phase. Voxels with positive activity and weight were assigned positive importance values, whereas those with negative activity and weight were assigned negative importance values. Voxels with opposite-signed activity and weight were rare and are not included in the importance maps shown here. Importance values were computed separately for each participant, with the results then averaged across participants. The resulting maps for each encoding task are displayed here, overlaid on a series of axial slices (z-coordinates are in MNI space; also see orthogonal view of slices). For display purposes, the maps are arbitrarily thresholded at ±0.0002 (red: positive, blue: negative) for at least 20 contiguous voxels. L = left.

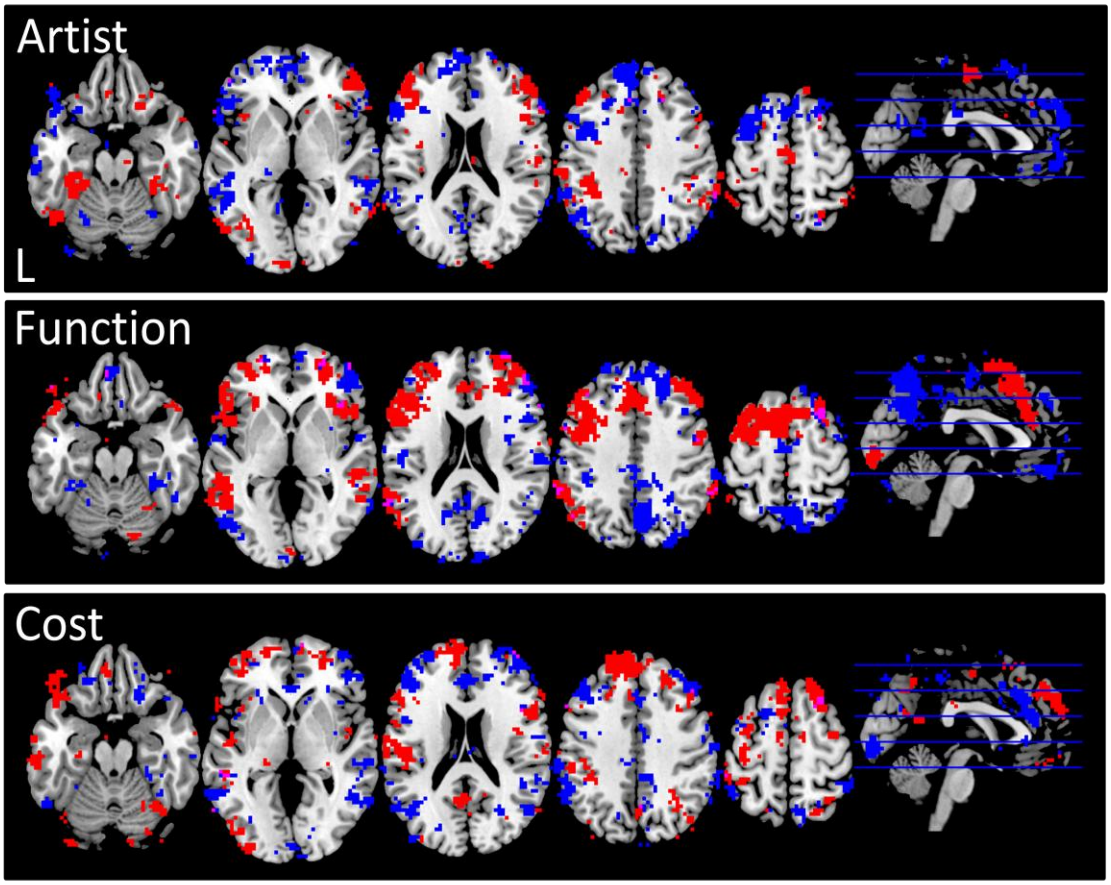
Artist
L
Function
Cost
z = -20
z = 0
z = 20
z = 40
z = 60

Figure A2. Mean classifier evidence for correct source judgments, separated according to each confidence response (high, moderate, low) and collapsed over the peak period of the fourth and fifth TRs. As reported in the text, classifier evidence for high-confidence correct source judgments was significantly greater than chance ($t(15) = 6.42$, $p < .001$). In contrast, classifier evidence for correct source judgments followed by moderate-confidence responses did not significantly differ from chance ($p = .14$). There were not enough subjects with sufficient trials to test low-confidence judgments.