Behavioral/Systems/Cognitive

Multivoxel Pattern Analysis Reveals Increased Memory Targeting and Reduced Use of Retrieved Details during Single-Agenda Source Monitoring

Susan G. R. McDuff,[1] Hillary C. Frankel,[3] and Kenneth A. Norman[1,2]

[1]Department of Psychology, and [2]Princeton Neuroscience Institute, Princeton University, Princeton, New Jersey 08540, and [3]Harvard Medical School, Boston, Massachusetts 02115

We used multivoxel pattern analysis (MVPA) of functional MRI (fMRI) data to gain insight into how subjects' retrieval agendas influence source memory judgments (was item X studied using source Y?). In Experiment 1, we used a single-agenda test where subjects judged whether items were studied with the targeted source or not. In Experiment 2, we used a multiagenda test where subjects judged whether items were studied using the targeted source, studied using a different source, or nonstudied. To evaluate the differences between single- and multiagenda source monitoring, we trained a classifier to detect source-specific fMRI activity at study, and then we applied the classifier to data from the test phase. We focused on trials where the targeted source and the actual source differed, so we could use MVPA to track neural activity associated with both the targeted source and the actual source. Our results indicate that single-agenda monitoring was associated with increased focus on the targeted source (as evidenced by increased targeted-source activity, relative to baseline) and reduced use of information relating to the actual, nontarget source. In the multiagenda experiment, high levels of actual-source activity were associated with increased correct rejections, suggesting that subjects were using recollection of actual-source information to avoid source memory errors. In the single-agenda experiment, there were comparable levels of actual-source activity (suggesting that recollection was taking place), but the relationship between actual-source activity and behavior was absent (suggesting that subjects were failing to make proper use of this information).

Key words: **memory; long-term memory; fMRI; source memory; retrieval strategies; pattern classification**

Introduction

Research on source memory, the ability to recall the conditions under which a memory was acquired, has increasingly come to focus on how agendas at the time of retrieval can influence which mnemonic features are retrieved and used to make source judgments (Johnson et al., 1993; Mitchell et al., 2008). One factor that has been shown to influence source memory judgments is the number of sources mentioned in the test instructions: Subjects are more likely to falsely attribute items to a source when that source is the only one mentioned at test (single-agenda source monitoring) compared to when multiple sources are mentioned at test (multiagenda source monitoring) (Lindsay and Johnson, 1989; Zaragoza and Koshmider, 1989; Dodson and Johnson, 1993; Henkel et al., 2000).

The goal of our study is to map out differences in how subjects make source judgments on single-agenda versus multiagenda tests. One possible difference relates to how subjects cue memory. Results from single-agenda tests suggest that, on these tests, subjects attempt to constrain retrieval to the targeted source by activating source-specific information from the study phase (Herron and Rugg, 2003a; see also Rugg, 2004; Jacoby et al., 2005a,b); according to the encoding specificity principle (Tulving and Thomson, 1973), activating targeted-source information should boost recall of the targeted source and reduce retrieval of nontarget source information. Subjects may be more prone to apply this kind of constraint on single-agenda versus multiagenda tests. Another possible difference relates to decision-making: Johnson and Raye (2000) make a distinction between activation (retrieval) of information and how strongly subjects "weight" (use) retrieved information. If subjects focus on the target source during single-agenda tests, they may retrieve information pertaining to nontarget sources but fail to properly use this information when making their source judgments.

To evaluate these hypotheses, we ran functional MRI (fMRI) studies where we manipulated use of single-agenda versus multiagenda instructions. In these studies, we used multivoxel pattern analysis (MVPA) (Norman et al., 2006) to measure activation (at test) of source-specific patterns of brain activity from the study phase. We focused on trials where the targeted source and the actual source differed (i.e., subjects were asked "Was this word studied with source X?" when it was actually studied with source Y). We hypothesized that different approaches to memory

Received July 30, 2008; revised Nov. 22, 2008; accepted Dec. 3, 2008.

This work was supported by National Institute of Mental Health Grant P50 MH062196 to K.A.N. and a National Science Foundation Graduate Research Fellowship to S.G.R.M. We thank Marcia Johnson and an anonymous reviewer for their comments on this manuscript.

Correspondence should be addressed to Kenneth A. Norman, Department of Psychology, Princeton University, Green Hall, Washington Road, Princeton, NJ 08540. E-mail: knorman@princeton.edu.

DOI:10.1523/JNEUROSCI.3587-08.2009

 0270-6474/09/290508-09$15.00/0

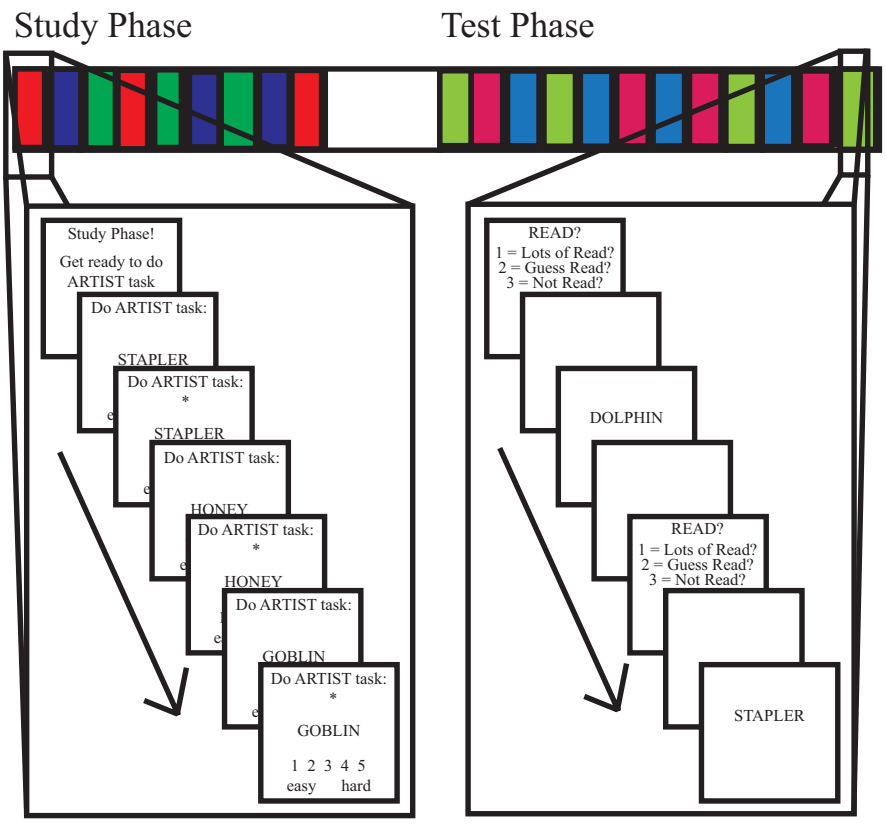


Figure 1. Illustration of the experimental procedure used during a single scanning run, where each run consisted of a study phase and a test phase (separated by a 1-min serial response task). The left side of the figure illustrates the sequence of events during the study phase. The study phase consisted of miniblocks (lasting 3 trials) during which subjects performed a particular encoding task (artist, function, or read). The lower-left part of the figure shows the trial sequence during an artist-task miniblock. On each study trial, subjects were instructed to perform the specified encoding task for 2 s (without overtly responding); after 2 s, an asterisk appeared, at which point subjects had 2 s to enter their 1-to-5 response. The right side of the figure illustrates the sequence of events during the test phase. The test phase consisted of miniblocks (lasting 3 trials) during which subjects were asked to target a particular encoding task. The lower-right part of the figure shows two example test trials from a read-task miniblock. Subjects were presented with a mixture of old items from the study phase and new items. On each test trial, subjects were given a task cue, followed by a test word; upon presentation of the test word, subjects had to indicate whether they had studied that word with the targeted task. The test screens shown here are from Experiment 1 (the test instructions differed slightly for Experiment 2).

cuing and decision-making would be associated with distinct patterns of activation. If subjects are more prone to constrain retrieval (by activating targeted-source information) in the single-agenda case, this should show up as an increase in neural activation of the targeted-source pattern in the single-agenda condition. Also, we can evaluate how well subjects are using recollection of the actual (nontarget) source by looking at the relationship between neural activation of the actual-source pattern and behavior. If subjects are using actual-source information to make their source decisions, high levels of actual-source activation (indicating recollection of the actual, nontarget source) should be associated with increased correct rejections. We predicted that actual-source activation would be more closely tied to behavior in the multiagenda (vs the single-agenda) experiment.

Materials and Methods

Subjects. Eleven people participated in Experiment 1 (four female, ages 20–26). Twelve people participated in Experiment 2 (six female, ages 19–28). Subjects were drawn from the graduate and undergraduate student community at Princeton University and received financial compensation for their participation. The experiments were run sequentially (first Experiment 1, then Experiment 2).

Materials. Experimental stimuli consisted of 216 noun words drawn from the MRC database (Coltheart, 1981; Wilson, 1988). The words that we used in the experiment were all between 4 and 9 letters in length ($M = 5.33$) and had a Kucera and Francis frequency rating of between 1 and 50 ($M = 17.52$). The familiarity rating of the words was between 500 and 620 ($M = 541.84$), the concreteness rating was between 550 and 670 ($M = 592.22$), and the imagery rating was between 490 and 659 ($M = 585.48$).

The words were presented to subjects on the computer via a projection system that reflected the images onto a mirror above subjects' eyes in the bore of the magnet. Subjects studied a total of 162 words. All 162 of these words were also presented on the source memory test, mixed in with 54 new words. The E-Prime software package (Psychology Software Tools) was used to present stimuli and collect responses.

Overview of experiments. The behavioral paradigm that we used was an exclusion test (Jacoby, 1991). Subjects were asked to study nouns; each word was encoded using either the artist encoding task, the function encoding task, or the read encoding task (the tasks are described below). During the test phase, subjects viewed all studied items in addition to new, unstudied items. On each trial, subjects were given a task cue ("Artist?", "Function?", or "Read?"), followed by a blank screen (lasting 3–7 s), followed by the test word. When the test word appeared, subjects had to indicate whether that word was studied using the targeted task; Experiment 1 and Experiment 2 used slightly different test instructions (described below). Test trials in this paradigm be divided up into three types: congruent trials, where the test word was studied using the targeted source; incongruent trials, where the test word was studied using a nontarget source; and new-item trials, where the test word did not appear at study.

Experiment 1 used single-agenda instructions. In this experiment, subjects were instructed to press one button to indicate with high confidence that the test word was studied using the targeted task, a second button to indicate with low confidence that the test word was studied using the targeted task, and a third button to indicate that the test word was not studied using the targeted task (subjects rarely used the "low-confidence yes" response, so we collapsed together "high-confidence yes" and low-confidence yes responses into a single "yes" response category when analyzing the data). Experiment 2 was identical to Experiment 1, except it used multiagenda instructions: For each test word, subjects were instructed to press one button to indicate that the test word was studied using the targeted task, a second button to indicate that the test word was studied using a different task, and a third button to indicate that the test word was new (nonstudied). Note that, here, subjects had to discriminate between three classes of items (targeted task, different task, and new), whereas in Experiment 1 subjects only had to discriminate between two classes of items (targeted and not targeted, where "not targeted" encompassed items studied using a different task and new items).

Detailed procedure. Both Experiment 1 and Experiment 2 consisted of six runs, where each run consisted of a study phase and then a test phase that probed subjects' memory for the immediately preceding study phase. The sequence of events during a run is illustrated in Figure 1 and described below (the test instructions shown in the figure are from Experiment 1).

During the study phase, words were presented, one at a time. For each

word, subjects encoded that word using the artist task, the function task, or the read task. The artist and function tasks were adapted from Dzulkifli and Wilding (2005) (see also Johnson et al., 1997a), and the read task was adapted from Davachi et al. (2003). For the artist task, subjects were asked to rate how easy it would be to draw each object, on a scale of 1–5 where 1 is easy and 5 is hard. For the function task, subjects were asked to think of functions for each object, and then to press a key corresponding to the number of functions they were able to generate (1–5). For the read task, subjects silently read words backwards to themselves (the words were displayed forwards, not backwards) and rated how difficult it was to do so (where 1 is easy and 5 is hard). For all tasks, subjects entered their responses on a keypad while lying in the scanner.

Each study phase consisted of 27 trials that were evenly split across the 3 encoding tasks (i.e., subjects studied 9 words using the artist task, 9 words using the function task, and 9 words using the read task). Trials were arranged in miniblocks of 3 trials, where all trials in the miniblock used the same encoding task. For each miniblock, subjects were first given a task cue (lasting 6 s) that notified them which task they would be performing. The task cue was followed by 3 words presented for 4 s each. For each word, the 4-s presentation period was broken into "study" and "response" phases as follows. For the first two seconds, subjects saw only the word and the rating scale that they would use. During the last two seconds, a small star appeared above the word, and subjects had to enter their numerical (1–5) rating using the keypad. After each study phase, subjects spent 1 min completing a basic serial response task, where numbers appeared on the screen and subjects were required to press buttons corresponding to these numbers. The test phase immediately followed the end of the serial response task.

During the test phase, subjects were presented with a mixture of all 27 items they had seen at study and 9 new words. For each word, subjects were first given a task cue that specified one of the three encoding tasks (e.g., Artist?). We will refer to this task as the targeted task. The task cue was presented for 1 s, and then a blank screen was presented for variable-length delay period, lasting 3, 5, or 7 s (for each test phase, 24 test trials had a 3-s delay, 9 trials had a 5-s delay, and 3 trials had a 7-s delay). After the delay, the test word was presented.

As mentioned above, the instructions for responding to the test word were different for Experiment 1 and Experiment 2. In Experiment 1, subjects were required to indicate whether the item was studied with the targeted task or not. In Experiment 2, subjects were required to indicate whether the item was studied using the targeted task, a different task, or was new. In both experiments, subjects had two seconds to enter their response; if they did not respond in time, the trial timed out (and "no response" was recorded in the data file). After each test trial, there was a variable-length delay (24 test trials had a 2-s delay, 9 trials had a 4-s delay, and 3 trials had a 6-s delay). During the test period, subjects switched retrieval orientations every 3 trials. That is, they received 3 trials in a row where they were asked to target one encoding task; then, for the next 3 trials, they were asked to target a different encoding task, and so on. Exactly half of the test trials were incongruent, 25% were congruent trials, and 25% were new trials. Each of the three encoding tasks served equally often as the targeted task.

fMRI data acquisition. The fMRI data were acquired on a Siemens Allegra 3 Tesla scanner at the Center for the Study of Brain, Mind, and Behavior at Princeton University. Anatomical brain images were acquired with an MP-RAGE sequence consisting of the following parameters: 176 sagittally oriented slices, repetition time (TR) = 2500 ms; echo time (TE) = 4.38 ms; voxel size = 1.0 × 1.0 × 1.0 mm; flip angle = 78°; field of view (FOV) = 256 mm. Functional images were acquired with an EPI sequence where 34 sagittal slices covering the whole brain were collected every 2 s (TR length). TE = 30 ms; voxel size = 3.0 × 3.0 × 3.9 mm; flip angle = 75°; FOV = 192 mm. Anatomical images were acquired at the start of the session. The main part of the experiment consisted of six functional runs. As described earlier, each run encompassed a single study phase and test phase; 292 images were collected per run.

fMRI data analysis: Preprocessing of fMRI data. We used the Analysis of Funtional NeuroImages (AFNI) fMRI data analysis software package (Cox, 1996) to preprocess the data. First, all functional images were coregistered with the first functional scan, and signal spikes were removed. A motion correction algorithm was applied to the data to remove any artifacts associated with head motion. Within each run, linear and quadratic trends were removed to remove the effects of scanner drift. No spatial smoothing was applied to the dataset.

Multivoxel pattern analysis: Overview. We used MVPA to measure activation (at test) of source-specific patterns of fMRI activity from the study phase. The MVPA approach to analyzing fMRI data involves training a pattern classifier to detect multivoxel patterns of fMRI data corresponding to particular cognitive states (for reviews, see Haynes and Rees, 2006; Norman et al., 2006). By aggregating the information that is present in multiple voxels' responses, MVPA achieves a higher level of sensitivity to the subject's cognitive state than standard mass-univariate approaches. In our study, the increased sensitivity of MVPA makes it possible to track fluctuations in activation of the targeted source and also fluctuations in source-specific recollection across single brain scans (where each scan was acquired over a 2 s period). These fluctuations, in turn, can be related to subjects' behavior.

Our MVPA analysis procedure closely resembled the procedure that we used previously by Polyn et al. (2005). The procedure was composed of three steps: First, we selected voxels to use in the classification analysis. Second, we trained a classifier to discriminate between study-phase brain patterns corresponding to subjects performing the artist, function, and read encoding tasks. Third, to measure activation of study-phase patterns, we applied the trained classifier to single (2-s) brain scans from the test phase. The output of the classifier gives us a graded index of how well the test pattern matches the artist, function, and read brain patterns from the study phase. As is typical for MVPA analyses, the analysis procedure was performed within individual subjects (i.e., voxel selection, classifier training, and classifier testing were all performed on data from the same subject).

Multivoxel pattern analysis: Details. To select voxels for the classification analysis, we ran a mass-univariate General Linear Model analysis in AFNI, and we found the 1000 voxels (across the whole brain) that were most strongly affected by the encoding task manipulation (artist vs function vs read) at study. After completing voxel selection, the functional data from these 1000 voxels were loaded into MATLAB (Mathworks) using the Princeton MVPA Toolkit (http://www.csbmb.princeton.edu/mvpa). All of the subsequent classification steps used the MVPA Toolkit. First, we *z*-scored the functional data separately for each voxel and each run, to ensure that we had a normalized activation value across runs. Next, we trained a simple neural network classifier (looking just at the selected voxels) to discriminate between single brain scans acquired while subjects were performing the artist, function, and read encoding tasks at study. The neural network consisted of two layers: an input layer with 1000 units (corresponding to each of the 1000 selected voxels), and an output layer with 3 units (one per encoding task). Each input unit was connected in a feedforward manner to all 3 output units; these connection weights define a function that maps between voxel activity values and encoding task. Neural network training was implemented using the backpropagation algorithm, which iteratively adjusts connection weights to minimize prediction error when mapping between inputs and outputs (Rumelhart et al., 1996). After this training process, we used the classifier to evaluate a series of test patterns (single brain scans) that had not been presented during classifier training. For most of the analyses reported here, we selected voxels and trained the classifier using all 6 runs of study-phase data, and then we applied the classifier to all 6 runs of test-phase data. We also present the results of analyses where we selected voxels and trained the classifier using data from 5 of 6 study-phase runs, and then we applied the classifier to data from the sixth ("left out") study-phase run. Additional details regarding our MVPA analysis methods are provided in the supplemental materials, available at www.jneurosci.org, including classifier "importance maps" that graphically depict which voxels the classifier used to distinguish between the study phase task conditions.

Logic of MVPA analyses. We focused our MVPA analyses on new-item and incongruent test trials. For these analyses, we binned the task-specific classifier outputs for each trial according to the role each task played on that trial. For new-item trials, the classifier outputs were

Table 1. Summary of predictions for single- and multiagenda source monitoring

Test type	Targeted-task activation (TT– OT)	Constraint (correlation between TT and AT)	Relationship between AT and behavior
Single-agenda (Experiment 1)	High	Yes, negative correlation between TT and AT	None
Multiagenda (Experiment 2)	Low	Negative correlation between TT and AT may be curtailed by restriction in the range of TT	Positive (high levels of AT predict correct rejections)

binned according to whether a particular task was the targeted task (TT) on that trial or one of the other tasks (OT) on that trial. For incongruent-item trials, the classifier outputs were binned according to whether a particular task was the TT on that trial, the actual task (AT) that was performed on that item at study, or the other task (OT). For example, if subjects were asked Artist? but the item was originally studied using the Function task, then TT = artist, AT = function, and OT = read. As described below, the OT can be used as a baseline when measuring activity related to the targeted task and the actual task. Note that the mapping of tasks (artist, function, read) onto conditions (TT, AT, and OT) varies from trial to trial.

Our analysis procedure is founded on two key claims. The first claim is that we can measure memory targeting by looking at the difference between TT and OT activity. We hypothesized that subjects would attempt to constrain retrieval to the targeted task by performing the targeted task on the test word (Jacoby et al., 2005a, 2005b), and that subjects would be more prone to do this in the single-agenda (vs multiagenda) condition. If this is the case, we should see a selective increase in TT activity relative to OT activity, and this increase should be larger in Experiment 1 than in Experiment 2.

The second claim is that, on incongruent trials, recollection of the actual task should lead to an increase in AT activity relative to OT activity. Prior research has established that recollection of memories from a particular source is associated with activation of source-specific patterns of activity from the study phase (Nyberg et al., 2000; Wheeler et al., 2000; Vaidya et al., 2002; Wheeler and Buckner, 2003; Kahn et al., 2004; Smith et al., 2004; Johnson and Rugg, 2004, 2007; Woodruff et al., 2005). Thus, in our exclusion paradigm, we would expect strong recollection of the actual source on incongruent trials to be associated with strong AT activity (relative to OT activity). Likewise, weak recollection of the actual source should be associated with weak AT activity (relative to OT activity). Note that our MVPA approach to dissociating targeted-task versus actual-task activity only works for incongruent trials. On congruent trials, the targeted task and the actual task are the same, so there is no way to tease apart activity relating to targeting versus recollection of the actual source; as such, we did not include congruent trials in our analysis. Also, we acknowledge that, in principle, other factors besides actual-task recollection could affect AT activity. For example, subjects might enact a strategy of performing all three tasks at test, to see which one fits best with the test word. A key point in this regard is that nonselectively performing all three tasks will affect TT, AT, and OT activity equally, so this strategy cannot be used to explain differences between TT and AT activity (on the one hand) and OT activity (on the other).

Importantly, the idea that actual-task activity indexes recollection makes it possible to assess the relationship between targeted-task activity and recollection of the actual task. If targeting of one task reduces recollection of other tasks, then – across incongruent trials – high levels of TT activity (indicating strong constraint) should be associated with low levels of AT activity (indicating low recollection of the actual task), resulting in a negative correlation between TT and AT activity. We expected that this negative correlation would be easier to observe in Experiment 1 than in Experiment 2, insofar as we expected TT activity to be lower overall in Experiment 2; this (anticipated) restriction in the range of TT should reduce the size of the correlation.

The link between actual-task activity and recollection also makes it possible to determine whether recollection of the actual (nontarget) task is differentially used in single- and multiagenda source monitoring. The key idea here is that AT recollection is not, on its own, sufficient to trigger a correct rejection. AT recollection needs to happen early in the trial (otherwise, subjects will not be able to respond before the 2-s deadline) and, more importantly, subjects need to attend to this early trial recollection to benefit from it (Johnson and Raye, 2000). We can measure how strongly subjects are attending to AT recollection by measuring the relationship between early trial AT activity and behavioral accuracy. If subjects are attending to AT recollection, high levels of AT activity early in the trial (indicating that AT recollection occurred, and that it occurred early enough to influence responding) should be associated with increased correct rejections. If subjects are not attending to AT recollection, this relationship between early trial AT activity and correct rejections should be absent. We expected that subjects would devote more scrutiny to AT recollection in the multiagenda test than in the single-agenda test; as such, we predicted that the relationship between early trial AT activity and behavior would be stronger in Experiment 2 than in Experiment 1. Table 1 presents a summary of the predictions described above.

Results

Classification of task-related activity during the study phase

All of our MVPA analyses depend on the idea that we can train a classifier to successfully detect patterns of brain activity associated with performing the artist, function, and read encoding tasks. To assess our ability to discriminate between task-specific encoding states, we trained the classifier on study-phase data from 5 of 6 scanner runs, and measured the classifier's ability to correctly predict (for each individual scan) which encoding task the subject was performing on the remaining study run. The correspondence between the classifier's predictions and the actual encoding task (as indexed by correlation) was significantly above chance for each individual subject. The average percentage correct classification of individual brain volumes (chance = 33.3%) was 79.1% for Experiment 1 (SEM = 3.7%) and 73.9% for Experiment 2 (SEM = 3.0%); classification accuracy was not significantly different across experiments, $t_{(21)} = 1.10$, $p > 0.05$. For further details on study-phase classification, see the supplemental materials, available at www.jneurosci.org.

Classification of task-related activity at test

For all subsequent analyses, the classifier was trained on study-phase data from all 6 scanner runs, and was applied to test-phase data from all 6 scanner runs. For each new-item and incongruent test trial, we measured the activity of each classifier output unit (artist, function, and read) for 7 successive scans (lasting 2 s each), starting with the scan when the test word was presented. As discussed earlier, classifier outputs from new-item and incongruent-item trials were binned according to whether that task was the targeted task, the actual task performed on the item at study, or the other task.

For all of the results presented below, dependent measures (e.g., classifier output for a particular condition) were computed separately for each individual subject. Figures and tables show the mean and SEM (across subjects) of these per-subject measures. We used two-tailed *t* tests (applied to these per-subject measures) to assess whether effects were reliable across subjects.

Analysis of targeted-task activity

To evaluate our prediction that the amount of targeted task activity would be higher in Experiment 1 versus Experiment 2, we

computed the average amount of TT activity in both studies. Figure 2*A* shows the event-related classifier output averages for both experiments. The left side of the figure plots average classifier output for TT and OT on new-item trials, and the right side of the figure plots average classifier output for TT, AT, and OT on incongruent trials. Figure 2*B* plots baseline-corrected targeted-task activity (TT–OT) for new-item trials and incongruent trials, as a function of Experiment (1 vs 2). For all of the subplots in Figure 2, classifier output is shown for 7 successive scans, starting with the scan when the test word was presented.

In Experiment 1, for both new-item and incongruent trials, TT activity was significantly higher than OT activity at multiple time points. This provides strong evidence that subjects were activating the TT representation at test. The supplemental materials, available at www.jneurosci.org, contain further analyses exploring the timing of TT activity. These additional analyses, which control for "spill-over" of TT activity from the preceding trial, demonstrate that TT activity was triggered by the test word, as opposed to the task cue that preceded the test word; these timing results fit with the idea, mentioned earlier, that TT activity reflects subjects performing the targeted task on the test word. In Experiment 2, the TT–OT difference was also significant for some time points, but numerically the TT–OT difference scores were smaller in Experiment 2 than in Experiment 1. When we directly compared TT–OT difference scores across experiments, we found that the difference was significant for incongruent trials at time points 3 and 5, and the difference was significant for new-item trials at time point 4; when we combined new-item and incongruent trials, the cross-experiment difference in TT–OT was significant at time points 3, 4, and 5.

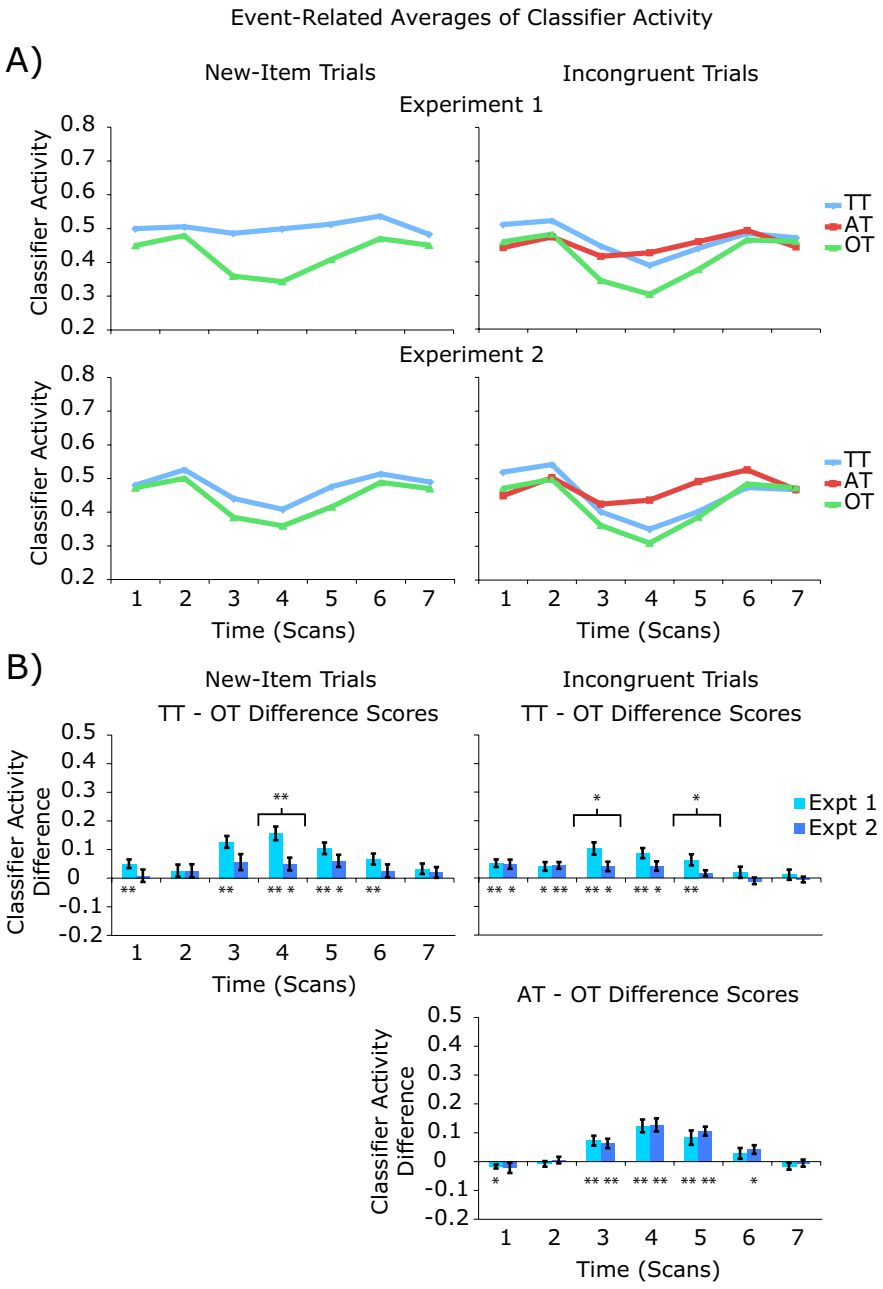


Figure 2. Event-related averages of classifier output in Experiment 1 and Experiment 2. Classifier output is shown for 7 successive scans, starting with the scan when the test word was presented. The left side of the figure shows classifier output for new-item trials, and the right side of the figure shows classifier output for incongruent-item trials. The graphs in ***A*** show raw classifier outputs, and the graphs in ***B*** show difference scores (TT–OT and AT–OT). Error bars in ***B*** indicate the SEM (across subjects) of the difference score. Individual bars marked with asterisks are significantly different from zero; pairs of bars marked with asterisks are significantly different from each other. *$p < 0.05$; **$p < 0.01$.

Analysis of the relationship between targeted-task and actual-task activity

According to the encoding specificity principle, TT activity should lead to reduced recollection of AT information on incongruent trials. As a first-pass measure, we compared the overall level of baseline-corrected AT activity (i.e., AT–OT) in the two experiments; as discussed above, this difference score provides an index of the degree of AT recollection that is taking place on incongruent trials. If TT activity suppresses AT recollection, then AT activity should be lower in Experiment 1 (where TT activity was relatively high) than in Experiment 2 (where TT activity was relatively low). Contrary to this prediction, we found that AT activity rose significantly above baseline in both experiments, and that the level of baseline-corrected AT activity was virtually identical across experiments (Fig. 2*B*, bottom).

To further investigate the relationship between TT and AT activity, we ran a more sensitive within-subjects analysis where we correlated (across incongruent trials) the level of TT activity with the level of AT activity. If TT activity suppresses AT recollection, then we would expect to see a negative correlation within individual subjects (assuming that there is adequate across-trial variability in TT activity). The TT–AT correlation was computed separately for each time point (scan) in the trial, starting with the scan when the test word was presented (e.g., we correlated TT activity at time point 1 with AT activity at time point 1; we correlated TT activity at time point 2 with AT activity at time point 2; and so on).

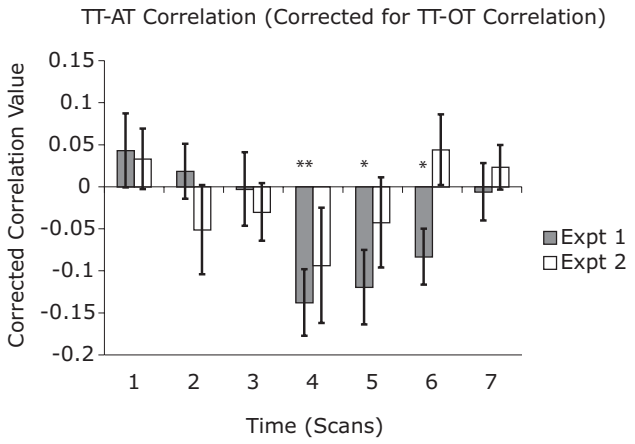


Figure 3. Correlation between TT and AT classifier activity (across incongruent trials) in Experiment 1 and Experiment 2. Correlation values were computed separately for each time point (scan) in the trial, starting with the scan when the test word was presented. TT–AT correlation values were corrected by subtracting out the corresponding TT–OT correlation value; this corrected score indicates whether TT activity was more strongly correlated with AT activity than OT activity. Error bars indicate the SEM (across subjects) of the corrected correlation measure. Individual bars marked with asterisks are significantly different from zero. $*p < 0.05$; $**p < 0.01$.

One complicating factor in this analysis is that classifiers can register a negative correlation between cognitive states even if (at a "process" level) the cognitive states are not related to each other. Intuitively, the more that one pattern is present in the brain, the less any other pattern will be present. With the neural network classifiers that we are using, the process is not completely zero-sum (i.e., it is possible to increase one classifier output without reducing other outputs), but we commonly observe some degree of negative correlation. To deal with this issue, we also computed the correlation between the TT and OT classifier outputs; then we subtracted out the TT–OT correlation from the TT–AT correlation. This measure factors out the "baseline" level of negative correlation (which should apply equally to TT–OT and TT–AT) and makes it possible to test whether TT activity is more negatively correlated with AT activity than with activity of the third (irrelevant) task.

Figure 3 shows the average "corrected correlation value" (across subjects) for each time point, for both Experiment 1 and Experiment 2. After correcting for the TT–OT correlation, there was a significant negative correlation between TT and AT activity at multiple time points (4, 5, and 6) in Experiment 1. In contrast, the correlation between TT and AT activity was not significant for any time point in Experiment 2. These results fit with our prediction that, when subjects attempt to constrain recall by activating the targeted task, TT activity will reduce recall of memories from other sources. The lack of a significant correlation for Experiment 2 can be explained in terms of the lower overall level of (baseline-corrected) TT activity in that experiment, which effectively restricts the range of TT and squelches the correlation. Importantly, the correlational nature of these results prevents us from making strong causal inferences. The observed negative correlation in Experiment 1 could be caused either by TT activity blocking AT recollection, or by AT recollection displacing TT activity (intuitively, strong recollection of the actual task will make it difficult to focus on the targeted task). The two possibilities are not mutually exclusive, and it seems likely that both of these situations occur to some extent.

Table 2. Experiment 1: Mean proportions of trials where subjects responded "yes," responded "no," or failed to respond in time as a function of trial type (congruent, incongruent, and new)

	Yes	No	Failed to respond
Congruent	0.69 (0.06)	0.15 (0.03)	0.16 (0.06)
Incongruent	0.09 (0.03)	0.77 (0.05)	0.14 (0.05)
New	0.06 (0.02)	0.83 (0.04)	0.12 (0.04)

Numbers in parentheses indicate SEM.

Table 3. Experiment 2: Mean proportion of trials where subjects responded "same task," responded "different task," responded "new," or failed to respond in time as a function of trial type (congruent, incongruent, and new)

	Same task	Different task	New	Failed to respond
Congruent	0.50 (0.05)	0.11 (0.04)	0.05 (0.02)	0.35 (0.07)
Incongruent	0.07 (0.02)	0.48 (0.05)	0.08 (0.02)	0.37 (0.08)
New	0.01 (0.01)	0.07 (0.04)	0.57 (0.07)	0.35 (0.07)

Numbers in parentheses indicate SEM.

Analysis of the relationship between actual-task activity and behavior

The finding that AT activity was above-baseline in both experiments allows us to look at how AT recollection affected behavior in Experiment 1 versus Experiment 2; as discussed earlier, we can use the relationship between AT activity and behavior on incongruent trials to assess the weight that subjects are giving to AT recollection when making source memory decisions. Behavioral results for Experiments 1 and 2 are presented in Tables 2 and 3, respectively (see the supplemental materials, available at www.jneurosci.org, for additional behavioral analyses). As in previous comparisons of single-agenda versus multiagenda tests (Lindsay and Johnson, 1989), false alarms were higher in the single-agenda experiment; this trend was significant for new items, $t_{(21)} = 2.28$, $p = 0.03$, but not for incongruent items, $t_{(21)} = 0.56$, $p > 0.05$. To get an overall sense of the relationship between AT activity and behavior, we plotted event-related averages of baseline-corrected AT scores (i.e., AT–OT) for correct rejections and errors, in both Experiment 1 and Experiment 2 (note that both incorrect responses and failures to respond in time were counted as errors). The results of these analyses are shown in Figure 4*A*.

We also used an area-under-the-receiver-operating-characteristic-curve (AUC) measure (Fawcett, 2006) to sensitively index how well AT activity discriminates between correct rejection and error trials. Specifically, the AUC measure indexes the overlap between the observed distributions of AT activity scores associated with correct rejections versus errors; AUC provides extra information (beyond looking at means and SDs) because it factors in the entire shape of the distribution. The AUC analysis was run separately for each time point (scan) in the trial, starting with the scan when the test word was presented. AUC scores range from 0 to 1, where 0.5 indicates chance discrimination. AUC scores >0.5 indicate that AT activity was associated with increased correct rejections, and AUC scores <0.5 indicate that AT activity was associated with increased errors. If subjects are using AT recollection when making their source memory judgments, AUC scores should be >0.5. Note that we used AT (alone) instead of AT–OT as our trial-by-trial measure of recollection when computing AUC. Subtracting out OT is valuable for demonstrating that (on average) the actual task is activated more strongly than the other task. However, when measuring recol-

lection on a trial-by-trial basis, subtracting out OT adds noise relative to using AT alone (since the proportion of OT variance that is not shared with AT is large, relative to the proportion of variance that is shared with AT). The AUC scores for Experiments 1 and 2 are shown in Figure 4*B*.

The results of these analyses show clear differences between Experiment 1 and Experiment 2. In Experiment 1, there was an overall trend for AT activity to be negatively associated with behavioral accuracy: For all but one time point, AT–OT was numerically higher for errors than for correct rejections; this trend was significant at the end of the trial, at time point 6. In contrast, in Experiment 2, early trial AT activity was positively associated with behavioral accuracy: AT–OT was significantly higher for correct rejections than errors at time point 2, and the AUC measure was significantly >0.5 at this time point. When we directly compared the AUC scores from the two experiments, we found that AUC scores were significantly higher for Experiment 2 versus 1 (indicating a more positive relationship between AT activity and correct rejections) at time points 2, 6, and 7.

To summarize, our key prediction regarding the relationship between AT activity and behavior was confirmed: Early trial AT activity was associated with correct rejections in Experiment 2 but not Experiment 1. The other main finding from this analysis, the relationship between late-trial AT and errors in Experiment 1, was unexpected, and merits further discussion. One possible interpretation of this result is that subjects in Experiment 1 were treating recollection of any information (even AT information) as evidence for the targeted task; for an example of how subjects can misattribute retrieved information from one source as evidence for another source, see Henkel et al. (2000). However, the timing of the effect in Experiment 1 argues against this interpretation: If AT recollection were actually causing errors, then we would expect to see this effect early in the trial, but the association between AT activity and errors was only present late in the trial (~10–12 s after stimulus onset). The timing of this effect suggests that increased AT activity on error trials in Experiment 1 may reflect postdecisional processing (i.e., subjects recalling and thinking about the actual task after they made an error) as opposed to predecisional processing; informally, subjects often reported that they would respond yes to an item on an incongruent trial and then, immediately afterward, they would realize that the item was from the wrong source (Van Zandt and Maldonado-Molina, 2004). Errors in both experiments may be attributable to factors that are "invisible" to the classifier (e.g., item familiarity) as opposed to task-specific activity (for additional discussion of this point, see the Limitations of our analysis procedure section below).

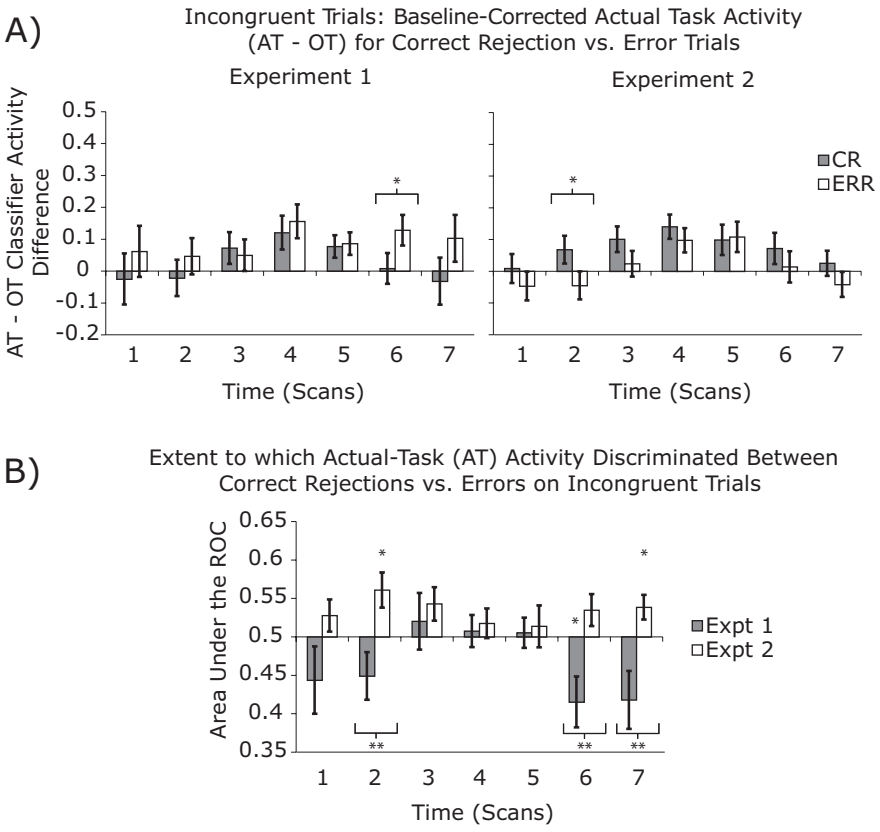


Figure 4. Analyses of the relationship between actual-task activity and behavior on incongruent trials in Experiments 1 and 2. ***A***, shows event-related averages of baseline-corrected AT activity (AT–OT) for Experiments 1 and 2, split by whether subjects made a correct rejection (CR) or an error (ERR). In ***B***, we used AUC measure (Fawcett, 2006) to index how well AT activity discriminated between correct rejection and error trials. AUC scores range from 0 to 1, where 0.5 indicates chance discrimination. AUC scores >0.5 indicate that AT activity was associated with increased correct rejections, and AUC scores <0.5 indicate that AT activity was associated with increased errors. Error bars in ***A*** indicate the SEM (across subjects) of the CR versus ERR difference for that time point (note that CR and ERR bars were not individually tested vs zero). Error bars in ***B*** indicate the SEM (across subjects) of the AUC score. In ***A***, pairs of bars marked with asterisks are significantly different from each other. In ***B***, individual bars marked with asterisks are significantly different from chance (0.5), and pairs of bars marked with asterisks are significantly different from each other. $*p < 0.05$; $**p < 0.01$.

Discussion

The goal of this study was to use neural data to gain psychological insight into how subjects make source memory judgments when they are asked to consider one source (single-agenda) versus when they are asked to consider multiple sources (multiagenda). Our first prediction was that subjects would be more likely to perform the targeted encoding task at test given single-agenda versus multiagenda instructions. Our MVPA results support this prediction: Targeted-task activation was significantly higher (relative to baseline) in Experiment 1 than in Experiment 2. We also hypothesized that activation of the targeted task would be associated with reduced recollection of the actual task on incongruent trials. Support for this claim was mixed: The level of actual-task activation (relative to baseline) was similar across experiments despite the difference in targeted-task activation. However, a more sensitive within-subjects analysis revealed that TT and AT activity were negatively correlated within individual subjects (Fig. 3). Our final prediction was that subjects would make better use of AT recollection on a multiagenda task, compared with a single-agenda task. Our results support this prediction: In Experiment 2, high levels of AT activation were associated with increased correct rejections on incongruent trials, but this relationship was not present in Experiment 1 (despite similar overall

levels of AT activation). To our knowledge, this is the first demonstration that subjects can retrieve diagnostic source information but nonetheless fail to use this information when making their source judgments.

The key methodological innovation underlying these findings was our use of pattern classifiers to track the appearance of task-specific activity during retrieval. As discussed earlier, MVPA increases sensitivity to the comings and goings of cognitive states by aggregating the information that is present in multiple voxels. This increase in sensitivity allowed us to derive meaningful measures of TT, AT and OT activity for each incongruent trial. Importantly, these MVPA analyses should be viewed as complementing (not replacing) standard voxel-based General Linear Model analyses. While MVPA is useful for addressing questions about what information is present in the subject's head at a particular point in time, whole-brain MVPA analyses are less useful for mapping out which brain regions are involved in particular cognitive processes (for a discussion of pitfalls associated with using whole-brain MVPA for brain mapping, see Norman et al., 2006). In the supplemental materials, available at www.jneurosci.org, we present voxel-based analyses exploring which brain regions discriminate between different tasks at study and which brain regions discriminate between correct versus incorrect responses to incongruent items at test.

Relationship to previous neural studies of agenda-dependent memory

Our results add to the growing body of neural evidence supporting agenda-dependent memory, the idea that subjects' goals at the time of retrieval can impact what information comes to mind and how subjects use this information (Mitchell et al., 2008; for examples of relevant neuroimaging studies, see Johnson et al., 1997b; Ranganath et al., 2000; Rugg and Wilding, 2000; Dobbins and Wagner, 2005; Dobbins and Han, 2006; for reviews of relevant studies, see Rugg, 2004; Simons, 2009). To our knowledge, there has only been one previous neuroimaging study that directly compared single-agenda to multiagenda source monitoring: Raye et al. (2000) (Experiment 1C) compared a single-agenda test ("Was the item studied as a picture?") to a multiagenda test ("Was the item studied as a picture or studied auditorily?") and found differences in frontal activity on the two types of tests. This finding indicates that processing is different in the two conditions but it does not indicate whether the difference relates to memory cuing or to the evaluation of retrieved information (or both).

While direct comparisons of single-agenda and multiagenda tests are scarce, there have been numerous imaging studies that speak to our hypotheses (set forth in the Introduction) about how subjects approach single-agenda tests. For example, a recent study by Woodruff et al. (2006) used a single-agenda source memory test and, like our study, found that subjects activate information relating to the targeted source. In Woodruff et al. (2006), subjects studied picture and word stimuli mixed together. At test, subjects were asked to target items from a particular source (e.g., pictures). For each test item, subjects were asked to say "yes" to items that were studied using the targeted source, and to say "no" otherwise. Woodruff et al. (2006) focused their fMRI analysis on new-item trials. They found that brain activity on new-item trials differed as a function of whether subjects were targeting picture versus word memories. Furthermore, brain activity patterns associated with targeting picture versus word memories were similar to brain activity patterns associated with studying pictures versus words, respectively (for a similar result, see Hornberger et al., 2006).

Additional relevant evidence comes from single-agenda studies that have compared ERPs on congruent and incongruent trials. Specifically, these studies have looked at the effect of retrieval orientation on the parietal old/new ERP effect, an ERP correlate of recollection (for discussion of this ERP effect, see Rugg et al., 2000; Rugg and Curran, 2007). Many of these studies have found that the parietal old/new effect is larger for congruent than for incongruent trials, suggesting that subjects have some ability to prevent recollection of information that mismatches the targeted source (Herron and Rugg, 2003a,b; Dzulkifli and Wilding, 2005; Herron and Wilding, 2005; Dzulkifli et al., 2006; for similar results from a slightly different paradigm, see Dywan et al., 1998, 2001, 2002).

The fMRI and ERP studies reviewed in this section provide some support for the idea that (on single-agenda tests) subjects attempt to constrain retrieval to the targeted source: The fMRI studies found activation of the targeted source, and the ERP studies found reduced recollection of nontarget memories. However, these studies did not address the relationship between activation of the targeted source and recollection of nontarget memories. In our study, we were able to address this relationship by simultaneously measuring targeted-task activation and actual-task activation, and then correlating these measures across trials. Also, the fMRI and ERP studies reviewed above did not measure the relationship between recollection of nontarget memories and behavior. In our study, we demonstrated a significant link between actual-task activity and behavioral accuracy in Experiment 2 (but not in Experiment 1), and we used this link to argue that subjects make better use of nontarget recollection in Experiment 2 versus Experiment 1.

Limitations of our analysis procedure

Importantly, the classifier was trained to detect patterns of activity that discriminate between the three tasks. This training procedure gives the classifier the ability to detect recollection of task-specific details, but it does not give the classifier the ability to detect recollection of nondiagnostic details (i.e., details shared by all three tasks) or feelings of familiarity. For evidence that subjects are influenced by nondiagnostic forms of memory on exclusion tests, see Dobbins and McCarthy (2008). Another limitation of our analysis procedure is that it focuses on activation (at test) of patterns from the study phase. As such, the analysis procedure will not detect processes that are engaged only at test (not at study).

Future directions

Our long-term goal is to exploit the sensitivity of MVPA to examine how memory cuing and decision-making processes vary (across subject populations, and as a function of situational factors). For example, recent results from Jacoby et al. (2005b) and Velanova et al. (2007) suggest that (on single-agenda tests with tasks as sources) elderly adults are less likely than young adults to perform the targeted task at test. Also, ERP studies have identified several manipulations that affect how strongly subjects orient to the targeted task, e.g., reducing the memorability of the targeted source (Herron and Rugg, 2003a; Dzulkifli et al., 2006; but see Herron and Wilding, 2005) and varying the targeted source unpredictably from trial to trial at test (vs using a blocked design) (Johnson and Rugg, 2006). We plan to explore these (and other) factors using variants of the design used here.

Conclusions

Our MVPA approach provides a new kind of evidence regarding how information is processed during memory retrieval. Using this technique, we compared retrieval processing during single-agenda (in Experiment 1) and multiagenda source monitoring (in Experiment 2). We observed that single-agenda source monitoring is associated with increased memory targeting and reduced use of retrieved

diagnostic details. Going forward, the ability to separately track targeted-task and actual-task activity should help us to develop more nuanced theories of how subjects cue memory, how cues interact with stored memory traces, how subjects make memory decisions, and how these processes go awry in subjects with memory disorders.

References

Coltheart M (1981) The MRC psycholinguistic database. Q J Exp Psychol A 33A:497–505.

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

Davachi L, Mitchell JP, Wagner AD (2003) Multiple routes to memory: distinct medial temporal lobe processes build item and source memories. Proc Natl Acad Sci U S A 100:2157–2162.

Dobbins IG, Han S (2006) Cue- versus probe-dependent prefrontal cortex activity during contextual remembering. J Cogn Neurosci 18:1439–1452.

Dobbins IG, McCarthy D (2008) Cue framing effects in source remembering: A memory misattribution model. Mem Cognit 36:104–118.

Dobbins IG, Wagner AD (2005) Domain-general and domain-sensitive prefrontal mechanisms for recollecting events and detecting novelty. Cereb Cortex 15:1768–1778.

Dodson CS, Johnson MK (1993) Rate of false source attributions depends on how questions are asked. Am J Psychol 106:541–557.

Dywan J, Segalowitz SJ, Webster L (1998) Source monitoring: ERP evidence for greater reactivity to nontarget information in older adults. Brain Cogn 36:390–430.

Dywan J, Segalowitz SJ, Webster L, Hendry K, Harding J (2001) Event-related potential evidence for age-related differences in attentional allocation during a source monitoring task. Dev Neuropsychol 19:99–120.

Dywan J, Segalowitz S, Arsenault A (2002) Electrophysiological response during source memory decisions in older and younger adults. Brain Cogn 49:322–340.

Dzulkifli MA, Wilding EL (2005) Electrophysiological indices of strategic episodic retrieval processing. Neuropsychologia 43:1152–1162.

Dzulkifli MA, Herron JE, Wilding EL (2006) Memory retrieval processing: neural indices of processes supporting episodic retrieval. Neuropsychologia 44:1120–1130.

Fawcett T (2006) An introduction to ROC analysis. Pattern Recognit Lett 27:861–874.

Haynes JD, Rees G (2006) Decoding mental states from brain activity in humans. Nat Rev Neurosci 7:523–534.

Henkel LA, Franklin N, Johnson MK (2000) Cross-modal source monitoring confusions between perceived and imagined events. J Exp Psychol Learn Mem Cogn 26:321–335.

Herron JE, Rugg MD (2003a) Strategic influences on recollection in the exclusion task: electrophysiological evidence. Psychon Bull Rev 10:703–710.

Herron JE, Rugg MD (2003b) Retrieval orientation and the control of recollection. J Cogn Neurosci 15:843–854.

Herron JE, Wilding EL (2005) An electrophysiological investigation of factors facilitating strategic recollection. J Cogn Neurosci 17:777–787.

Hornberger M, Rugg MD, Henson RN (2006) fMRI correlates of retrieval orientation. Neuropsychologia 44:1425–1436.

Jacoby LL (1991) A process dissociation framework: separating automatic from intentional uses of memory. J Mem Lang 30:513–541.

Jacoby LL, Shimizu Y, Daniels KA, Rhodes MG (2005a) Modes of cognitive control in recognition and source memory: depth of retrieval. Psychon Bull Rev 12:852–857.

Jacoby LL, Shimizu Y, Velanova K, Rhodes MG (2005b) Age differences in depth of retrieval: memory for foils. J Mem Lang 52:493–504.

Johnson JD, Rugg MD (2006) Electrophysiological correlates of retrieval processing: effects of consistent versus inconsistent retrieval demands. J Cogn Neurosci 18:1531–1544.

Johnson JD, Rugg MD (2007) Recollection and the reinstatement of encoding-related cortical activity. Cereb Cortex 17:2507–2515.

Johnson MK, Raye CL (2000) Cognitive and brain mechanisms of false memories and beliefs. In: Memory, brain, and belief (Schacter DL, Scarry E, eds), pp 35–86. Cambridge, MA: Harvard UP.

Johnson MK, Hashtroudi S, Lindsay DS (1993) Source monitoring. Psychol Bull 114:3–28.

Johnson MK, Kounios J, Nolde SF (1997a) Electrophysiological brain activity and memory source monitoring. Neuroreport 8:1317–1320.

Johnson MK, Nolde SF, Mather M, Kounios J, Schacter DL, Curran T (1997b) The similarity of brain activity associated with true and false recognition memory depends on test format. Psychol Sci 8:250–257.

Kahn I, Davachi L, Wagner AD (2004) Functional-neuroanatomic correlates of recollection: implications for models of recognition memory. J Neurosci 24:4172–4180.

Lindsay DS, Johnson MK (1989) The eyewitness suggestibility effect and memory for source. Mem Cognit 17:349–358.

Mitchell KJ, Raye CL, McGuire JT, Frankel H, Greene EJ, Johnson MK (2008) Neuroimaging evidence for agenda-dependent monitoring of different features during short-term source memory tests. J Exp Psychol Lean Mem Cogn 34:780–790.

Norman KA, Polyn SM, Detre GJ, Haxby JV (2006) Beyond mind-reading: multi-voxel pattern analysis of fMRI data. Trends Cogn Sci 10:424–430.

Nyberg L, Habib R, McIntosh AR, Tulving E (2000) Reactivation of encoding-related brain activity during memory retrieval. Proc Natl Acad Sci U S A 97:11120–11124.

Polyn SM, Natu VS, Cohen JD, Norman KA (2005) Category-specific cortical activity precedes retrieval during memory search. Science 310:1963–1966.

Ranganath C, Johnson MK, D'Esposito M (2000) Left anterior prefrontal activation increases with demands to recall specific perceptual information. J Neurosci 20:RC108.

Raye CL, Johnson MK, Mitchell KJ, Nolde SF, D'Esposito M (2000) fMRI investigations of left and right PFC contributions to episodic remembering. Psychobiology 28:197–206.

Rugg MD (2004) Retrieval processing in human memory: electrophysiological and fMRI evidence. In: The cognitive neurosciences, 3rd Ed (Gazzaniga M, ed), Cambridge, MA: MIT.

Rugg MD, Curran T (2007) Event-related potentials and recognition memory. Trends Cogn Sci 11:251–257.

Rugg MD, Wilding EL (2000) Retrieval processing and episodic memory. Trends Cogn Sci 4:108–115.

Rugg MD, Allan K, Birch CS (2000) Electrophysiological evidence for the modulation of retrieval orientation by depth of study processing. J Cogn Neurosci 12:664–678.

Rumelhart D, Durbin R, Golden R, and Chauvin Y (1996) Backpropagation: the basic theory. In: Backpropagation: theory, architectures, and applications (Chauvin Y, Rumelhart D, eds), pp 1–34. Mahwah, NJ: Erlbaum.

Simons JS (2009) Constraints on cognitive theory from neuroimaging studies of source memory. In: Neuroimaging of human memory: linking cognitive process to neural systems (Roesler F, Ranganath C, Roder B, Kluwe RH, eds), pp 403–424. New York: Oxford UP, in press.

Smith AP, Henson RN, Dolan RJ, Rugg MD (2004) fMRI correlates of the episodic retrieval of emotional contexts. Neuroimage 22:868–878.

Tulving E, Thomson DM (1973) Encoding specificity and retrieval processes in episodic memory. Psychol Rev 80:352–373.

Vaidya CJ, Zhao M, Desmond JE, Gabrieli JD (2002) Evidence for cortical encoding specificity in episodic memory: memory-induced re-activation of picture processing areas. Neuropsychologia 40:2136–2143.

Van Zandt T, Maldonado-Molina MM (2004) Response reversals in recognition memory. J Exp Psychol Learn Mem Cogn 30:1147–1166.

Velanova K, Lustig C, Jacoby LL, Buckner RL (2007) Evidence for frontally mediated controlled processing differences in older adults. Cereb Cortex 17:1033–1046.

Wheeler ME, Buckner RL (2003) Functional dissociation among components of remembering: control, perceived oldness, and content. J Neurosci 23:3869–3880.

Wheeler ME, Buckner RL (2004) Functional-anatomic correlates of remembering and knowing. Neuroimage 21:1337–1349.

Wheeler ME, Petersen SE, Buckner RL (2000) Memory's echo: vivid remembering reactivates sensory-specific cortex. Proc Natl Acad Sci U S A 97:11125–11129.

Wilson M (1988) The MRC psycholinguistic database: machine readable dictionary, Version 20. Behav Res Methods Instrum Comput 20:6–11.

Woodruff CC, Johnson JD, Uncapher MR, Rugg MD (2005) Content-specificity of the neural correlates of recollection. Neuropsychologia 43:1022–1032.

Woodruff CC, Uncapher MR, Rugg MD (2006) Neural correlates of differential retrieval orientation: Sustained and item-related components. Neuropsychologia 44:3000–3010.

Zaragoza MS, Koshmider JW 3rd (1989) Misled subjects may know more than their performance implies. J Exp Psychol Learn Mem Cogn 15:246–255.

SUPPLEMENTAL MATERIALS for "Multi-voxel pattern analysis reveals increased memory targeting and reduced use of retrieved details during single-agenda source monitoring" by S. G. R. McDuff, H. C. Frankel, and K. A. Norman

TABLE OF CONTENTS

1. Multi-voxel pattern analysis methods

As described in the main text, our multi-voxel pattern analysis (*MVPA*) analysis procedure was carried out on each subject separately, using fMRI data collected from the whole brain. The analysis procedure involved three steps: *Voxel selection* (to narrow down the number of voxels that are fed into the classifier), *classifier training*, and

generalization testing. In this section, we describe general aspects of our procedure relating to voxel selection and classification. In the next two sections (*Section 2* and *Section 3*, below), we discuss specific details of how we used MVPA to analyze study-phase and test-phase data.

Voxel selection

The goal of our voxel selection procedure was to isolate, for each subject, the voxels whose activity differed most strongly across the three encoding tasks at study (Polyn et al., 2005). For discussion of the benefits of doing voxel selection prior to classification, see Norman et al. (2006) and Mitchell et al. (2004). To enact voxel selection, we ran a multiple regression analysis on data from the study phase using AFNI's 3dDeconvolve program (for general discussion of AFNI, see Cox, 1996; for specific AFNI software routines, see http://afni.nimh.nih.gov/afni). In this analysis, we predicted each voxel's time course using regressors corresponding to each of the three encoding tasks (artist, function, and read) plus six nuisance motion-correction parameter vectors that were generated by the AFNI motion correction algorithm, 3dvolreg. We created each encoding task regressor by first creating a "boxcar" regressor corresponding to the study-phase time points when a word was onscreen and the subject was performing the specified encoding task (e.g., artist) on the word. Then, to account for temporal dispersion in the hemodynamic response, we used AFNI's *waver* function to convolve the boxcar regressors with a gamma-variate hemodynamic response function. For each voxel, we specified a set of three linear contrasts to test for differences in the beta-weights

associated with the artist, function, and read tasks (the contrast weights were 2 -1 -1, -1 2 -1, and -1 -1 2). Finally, we computed the F-statistic for the combination of these three contrasts. This F-statistic indicates how strongly the voxel's activity differed depending on whether subjects were performing artist vs. function vs. read encoding.

Our voxel selection procedure involved taking, for each subject, the N voxels (out of approximately 50,000 total) that had the largest F statistic for that subject. These N voxels were then used in the pattern classification analyses for that subject. To select a value for N, we ran our study-phase classification analysis procedure (see *Section 2* below) using different numbers of voxels, and we chose the number of voxels that yielded the highest level of study-phase classification accuracy (i.e., accuracy in classifying which task subjects were performing at study, when the classifier was trained on other study trials). For simplicity, we decided to use the same number of voxels for all subjects, instead of individually tuning the number of voxels to maximize study-phase accuracy for each subject. Supplemental Figure 1 shows how study-phase classification accuracy (averaged across subjects from both Experiment 1 and Experiment 2) varied as a function of the number of voxels included in the analysis. The peak of this curve was 1,000 voxels, so we set the number of voxels equal to 1,000 for all of our classification analyses.

==========

Insert Supplemental Figure 1 About Here

==========

It is important to emphasize that our voxel selection procedure only used data from the study phase of the experiment; it did not factor in test phase data in any way. As such, the results of our train-on-study-phase, generalize-to-test-phase analyses (described in the main text and in *Section 3* below) reflect the classifier's ability to generalize to entirely new data that were not used for either voxel selection or classifier training.

Neural network classifier

Our classification analyses were implemented in MATLAB, using the Princeton MVPA Toolbox (http://www.csbmb.princeton.edu/mvpa) and the MATLAB Neural Networks Toolbox (Mathworks, Natick MA). The classification procedure was run separately for each subject. Before running the classifier, we z-scored the functional data separately for each voxel and each run (Polyn et al., 2005). We used a two-layer neural network classifier (i.e., input and output layers only; no hidden layer). There were 1,000 units in the input layer, corresponding to the 1,000 voxels that passed through our GLM-based voxel selection procedure for that subject. The output layer contained three output units, one for each of the encoding tasks (artist, function, and read). The network was purely feed-forward with full connections between the input and output layers, and a sigmoid transfer function was used on the output layer (Polyn et al., 2005). As in Polyn et al. (2005), the classifier was trained using the conjugate gradient descent version of the backpropagation algorithm, and a cross-entropy function was used to calculate prediction error during training (for further discussion of backpropagation see, e.g., Bishop, 1995; Duda et al., 2001; LeCun et al., 1998; Rumelhart et al., 1996).

Analysis procedure

For all of our classification analyses, we divided up the fMRI data into a *training set* and a *generalization testing set* (see *Section 2* and *Section 3* below for description of the different ways in which we specified the training and testing sets). First, we did voxel selection on the training set. Next, we initialized the classifier weights to random values. After that, we trained the classifier on data from the study phase. Specifically, the classifier was given training patterns corresponding to individual study-phase brain scans (where one scan was collected every two seconds). For each scan, we clamped the pattern of voxel activity values from that scan onto the input layer. To specify the target ("correct") output value for each study-phase pattern, we took the hemodynamically convolved task regressors described in the *Voxel selection* section above, rescaled the regressors into the zero-to-one range, and then binarized them such that values above 0.5 were set to 1 and values below 0.5 were set to 0. Time points where all three tasks had a target output value of zero (indicating that no task was strongly active at that time point) were not included in the classification analysis. For all of the remaining time points, there was a single "correct answer" for each input pattern (i.e., one of the three output units had a target value of 1, and the other two output units had a target values of zero). Weights were updated after each training trial using backpropagation.

After training, we used the network to classify individual brain scans from the generalization testing set. For each test pattern, we recorded the activity of each of the three output units. These activity values indicate how well the test pattern matches the

artist, function, and read brain states that were present in the training data set. To average out variability associated with random initial weight settings, we repeated this classification procedure 100 times, and we averaged together the classifier output values (for each test pattern) across these 100 classifier iterations.

2. Study-phase classification analyses

As discussed in the main text, all of our MVPA analyses depend on the idea that we can classify which task subjects are performing at study. To address this question, we used a *leave-one-out* generalization procedure where we took study-phase data from 5 out of the 6 functional runs and did voxel selection and classifier training on these five runs; we then applied the trained classifier to study-phase data from the sixth ("left out") run. We iterated this procedure six times, leaving out a different run each time (see Hastie et al., 2001 for additional discussion of the leave-one-out procedure).

Percent correct analyses

We assessed the performance of the classifier by comparing the classifier's response to the binary encoding task regressors described in the *Analysis procedure* section above. For each individual scan, we labeled that scan as being "correct" if the classifier output unit associated with the correct task was more active than the classifier output units associated with the other two tasks. For each subject, we computed an overall percent correct value.

The significance of these percent correct values was assessed for each individual subject using a non-parametric statistical procedure; for a detailed description of this procedure, see the Supporting Online Material associated with Polyn et al (2005). We used a wavelet-based signal decomposition procedure to generate surrogate classifier output time-courses that had the same spectral characteristics as the original time-courses (Bullmore et al., 2001). For each of the actual classifier output time-courses (one per task), we computed ten thousand surrogate classifier output time-courses. These surrogate time-courses were used to create distributions of percent correct scores. By comparing the actual percent correct score to these distributions, we were able to generate a *p* value: the proportion of surrogate percent correct scores exceeding the percent correct score that was obtained in the experiment. These p-values indicate the probability that the observed percent correct score would have been obtained by chance (i.e., assuming that there was no real correspondence between classifier output and the task being performed).

==========

Insert Supplemental Table 1 About Here

==========

The scores in Supplemental Table 1 indicate that classification was significantly above chance (where chance = 1/3 = 33%) for every subject. Note that these accuracy scores reflect a "best case" scenario, insofar as we tuned our voxel selection parameters to maximize average study-phase classification accuracy (see *Section 1* above). However, we should note that classifier accuracy was reasonably robust to changes in the

number of selected voxels. Supplemental Figure 1 plots how classification accuracy (averaged across subjects from both experiments) varied as a function of the number of included voxels. The figure shows that including anywhere from 500 to 5000 voxels yielded near-peak average classification accuracy. The only time when classifier accuracy dipped far below its peak was in the 10-voxel condition (and classifier accuracy was still well above chance in this condition).

3. Measuring activity at test

To measure activation of task-specific study-phase patterns at test, we did voxel selection and trained the classifier based on study-phase data from all six runs. We then applied the trained classifier to test-phase data from all six runs. To analyze these results, we created *event-related averages* showing classifier output for new-item trials and incongruent trials. For each new-item and incongruent test trial, we measured the activity of each classifier output unit (artist, function, and read) for 7 successive scans (lasting 2 seconds each), starting with the scan when the test word was presented. For new-item trials, the classifier outputs were binned according to whether a particular task was the targeted task (TT) on that trial, or one of the other tasks (OT) on that trial. For incongruent-item trials, the classifier outputs were binned according to whether a particular task was the targeted task (TT) on that trial, the actual task (AT) that was performed on that item at study, or the other task (OT). Our primary analyses of these event-related averages are presented in the *Results* section of the main text. Here, we present extra analyses that address the timing of targeted-task activity.

Analyzing the timing of targeted-task activity

In the main text, we hypothesized that subjects in Experiment 1 would attempt to constrain retrieval at test by *performing the targeted task on the test word* (for additional discussion of this idea, see Jacoby et al., 2005). This claim has strong implications for the timing of activation of the targeted task (TT). If we assume that TT activity commences when the test word is presented, and we assume that TT activity triggers a standard hemodynamic response, this implies that activation of the TT representation (operationalized in terms of the TT – OT difference) should start out at zero on the scan when the test word is presented, and that it should peak approximately three scans later. Contrary to this view, in Experiment 1, the TT – OT difference was significant for the scan when the test word was presented (time point 1 on the event-related averages) for both new-item and incongruent trials (this trend was also present, albeit to a lesser extent, in Experiment 2; see Figure 2 in the main paper). There are two possible explanations for this finding: One possibility is that subjects started to activate their TT representation in response to the task cue (e.g., “Artist?”), instead of waiting for the presentation of the test word. Another possibility relates to the fact that test trials were arranged into 3-trial miniblocks, where each trial in a miniblock used the same TT task. As such, TT activity at the outset of a trial could reflect “spill-over” of TT activity from the preceding trial.

==========

Insert Supplemental Figure 2 About Here

==========

To eliminate the possibility of spill-over effects, we re-analyzed the data from new-item and incongruent trials in Experiment 1, limiting the analysis to trials/conditions where neither the targeted task nor the other task served as the targeted task on the previous trial. This ensures that the TT – OT difference is uncontaminated by lingering TT activity. Limiting the analysis in this fashion involves discarding the second and third trials from each miniblock (since, as mentioned above, the targeted task on these trials was the same as the targeted task on the previous trial). It also involves discarding some fraction of the first-in-miniblock data, since the "other" task on these trials sometimes matched the targeted task on the previous trial. To compensate for this loss of data, we pooled together data from new-item and incongruent trials for this analysis (instead of analyzing new-item and incongruent trials separately). The results of this analysis are shown in Supplemental Figure 2. The left-hand side of the figure shows an event-related average of classifier output, time-locked to the presentation of the test word (at time point 1). The figure shows that, when the possibility of spill-over is eliminated, the TT – OT difference was very close to zero for time point 1 (the scan where the test word was presented), and the TT – OT difference peaked at time points 3, 4, and 5. The timing of this response fits with the idea, expressed above, that subjects activated the TT representation in response to the test word (not beforehand). The right-hand side of Supplemental Figure 2 shows an event-related average of classifier output, time-locked to the presentation of the *task cue* (at time point 1). As discussed in the *Materials and Methods* section of the main text, the task cue preceded the test word by either 2, 3, or 4

time points. The graph here collapses across these conditions, so the test word appeared at either time point 3, 4, or 5 on the graph. The figure illustrates that there is no event-related TT activity response triggered by the task probe – TT activity does not start to appear until much later (after the test word has appeared).

4. Group General Linear Model analysis: Task-specific study-phase activity

Our finding of well-above-chance scan-by-scan classification of encoding task states (see *Section 2* above) indicates that the three encoding tasks were associated with distinct patterns of neural activity. We ran two analyses aimed at characterizing which brain regions showed discriminative activity. In the first analysis (described in this section), we used a standard mass-univariate GLM analysis to identify brain regions whose activity reliably discriminated between the encoding tasks during the study phase. In the second analysis (described in *Section 5*) we created classifier "importance maps" showing which voxels contributed most strongly to the classifier's ability to detect the three task states. The GLM analysis and the importance-map analysis included subjects from both Experiment 1 and Experiment 2.

The specific goal of the GLM multiple regression analysis described here was to identify clusters of voxels that (across subjects) were *reliably more active* or *reliably less active* during a particular encoding task at study, relative to the other two tasks. The preprocessing steps and GLM regressors used in this analysis were identical to the parameters that we used for the GLM analyses described in the *Voxel selection* section. As with the voxel selection analyses, we ran the multiple regression analysis using

AFNI's 3dDeconvolve program. The only differences between this analysis and the GLM used for voxel selection are as follows: 1) In this analysis, we report results at the group level, whereas voxel selection was done based on effects at the individual-subjects level; and 2) in this analysis, we report task-specific effects (i.e., for each task, which voxels showed differential activity for that task vs. the other tasks), whereas voxel selection was done based on the *overall* (main) effect of encoding task on voxel activity.

The multiple regression analysis produced a set of beta weights indicating the extent to which a particular voxel's activity correlated with a condition of interest for individual subjects. Prior to running the group analysis, we warped these beta weights from each individual subject's brain space into Talairach space (using AFNI's program, @auto_tlrc) and smoothed them using a 4mm Gaussian blur (with AFNI's program 3dmerge). To identify regions that were differentially activated (or deactivated) by a particular task at study, we conducted a 2-way ANOVA (using AFNI's 3dANOVA2) on the beta weights from the study phase (encoding task: artist, function, and read). We treated task (artist, function, read) as a fixed effect, and subject as a random effect. As in our voxel selection procedure, we ran contrasts on the artist, function, and read beta weights to identify voxels that were differentially active during the different encoding tasks. For example, the contrast weights used to specify voxels that were differentially active for the artist task (vs. the function and read tasks) were 2, -1, -1 for artist, function, and read.

The ANOVA yielded contrast maps that showed the extent to which each individual voxel's activity was differentially modulated by each encoding task at study. We used these contrast maps to identify clusters of voxels whose activity (as a group) was differentially modulated by the artist, function, and read tasks, respectively. To correct

for the problem of multiple comparisons, we used AFNI's program AlphaSim to identify the number of contiguous voxels (the size of the cluster) necessary to reduce the probability of falsely detecting a cluster of that size to 0.001. The clusters were chosen to include voxels whose contrast result was either positive or negative and exceeded $t > 5.68$ ($p = 0.00001$). The volume of the cluster had to exceed 243 microliters with a minimum connectivity radius of 6 mm. The clusters that were found using the above restrictions are reported in Supplemental Table 2. Coordinates are reported in the Talairach-Tournoux (T-T) Atlas coordinate space. The focus point of each cluster is reported (LPI format). The "direction of activation change" column indicates whether a particular cluster showed *increased* (+) or *decreased* (-) activity for the specified task, relative to the other two tasks.

==========

Insert Supplemental Table 2 About Here

==========

Given that the primary purpose of this paper is not to identify discriminative brain regions, we will not dwell on the results of this GLM analysis. Nonetheless, it is worth noting that our group GLM results appear to be consistent with the results of other fMRI studies that have used similar tasks. For example, numerous studies have observed left inferior frontal gyrus (LIFG) activity in tasks that require controlled retrieval from semantic memory and/or selection among multiple meanings (for a recent review, see Badre and Wagner, 2007). This fits well with our finding that LIFG was more active for

the function task (which places strong demands on controlled semantic retrieval and selection) than for the other two tasks (which place weaker demands on controlled semantic retrieval and selection).

5. Classifier importance maps

A more direct way of gaining insight into how the classifier is discriminating between encoding tasks is to look at the classifier's weights (after it has been trained on study-phase data). Specifically, we wanted to use classifier weight information to establish which voxels were *most important in activating each task's output unit, when that task was present*. For example, for scans associated with the artist task, which voxels played the largest role in (correctly) activating the artist unit? In our neural network classifier, the net contribution of a voxel to activating a task unit is a function of the voxel's activation, multiplied by the weight between that voxel and the task unit. Logically speaking, there are two ways for a voxel to make a net positive contribution to activating a particular task unit:

1) The voxel could have a positive z-scored average activation value (indicating that it was more active than usual) for scans associated with that task, and it could have a positive weight to that task unit. Voxels meeting this criterion were assigned a *positive* importance value $imp_{ij} = w_{ij} * avg_{ij}$, where w_{ij} is the weight between input unit i (corresponding to voxel i) and output unit j (corresponding to task j), and avg_{ij} is the average activation of input unit i while subjects were performing task j.

2) The voxel could have a negative z-scored average activation value (indicating that it was less active than usual) for scans associated with that task, and it could have a negative weight to that task unit. In this case, the "double negative" combination of negative activation and negative weight results in a net positive contribution. Voxels meeting this criterion were assigned a *negative* importance value $imp_{ij} = -w_{ij} * avg_{ij}$.

Voxels where the sign of w_{ij} differed from the sign of avg_{ij} (indicating a net negative contribution of that voxel to detecting that task state) were assigned an importance value of zero. Importance maps were computed using the above equations for each individual subject. Crucially, note that (with these equations) both positive and negative importance values indicate a net positive contribution of that voxel to activating the task unit (when that task is present). The sign of the importance value indicates whether the voxel contributes via a characteristic deactivation that is picked up by the classifier (via a negative weight), or a characteristic activation that is picked up by the classifier (via a positive weight). Computing importance values in this way makes it easier to compare importance maps to the GLM results discussed in *Section 4* (which indicate, for each task/cluster combination, whether the cluster is more or less active for that task, compared to other tasks). Note that this procedure for computing importance values differs from the procedure used by Polyn et al. (2005), which measured whether each voxel *i* made a net positive or negative contribution to the activation of output unit *j*, but did not indicate whether voxels making net positive contributions did so because they were *more* or *less* active than usual during condition *j*.

After creating importance maps for each subject, these individual-subject importance maps were transformed to match a common template in Talairach space using AFNI's automatic Talairaching functionality (@auto_tlrc). We then used AFNI's program 3dmerge to apply a Gaussian blur with a full width at half maximum of 4mm to the Talairached importance maps for each individual subject. This blurring makes it easier to identify commonalities in importance maps across subjects. To create group average importance maps, AFNI's program 3dmerge was used to compute the mean importance value across all subjects for each voxel/task combination. The structural image in Talairach space from subject 8 was used as an underlay for the montage, which was created with the AFNI software package. The resulting importance maps are shown in Supplemental Figure 3.

==========

Insert Supplemental Figure 3 About Here

==========

The importance maps reveal distinct patterns of activations and deactivations that (across subjects) were associated with each encoding task. The regions labeled as important in Supplemental Figure 3 appear to be a superset of the regions identified by the GLM analysis (see Supplemental Table 2). The fact that regions found by the GLM appear in the classifier importance maps is not surprising, insofar as we used a GLM analysis to select which voxels were fed into the classifier for each subject. The main difference between the classifier importance maps and the GLM clusters is that the

classification analysis is more liberal in choosing which voxels to include. In addition to incorporating significant "peak clusters", the classifier maps also incorporate voxels that were in the "top 1000 voxels" for some subjects but (for whatever reason) did not reach conventional levels of significance in the group analysis. For discussion of how classifiers benefit by accumulating weak information from voxels that do not meet conventional significance levels, see Norman et al. (2006), Haynes and Rees (2006), and Kamitani and Tong (2005).

6. Group General Linear Model analysis: Regions that predict correct rejections of incongruent items

In the main paper, we used MVPA techniques to show that the amount of AT activity measured in Experiment 2 (during multi-agenda source monitoring) was related to the probability of correctly rejecting an incongruent item, but this relationship was not present in Experiment 1 (during single-agenda source monitoring). In this section, we use a GLM analysis to explore whether there were any *specific brain regions* whose activity predicted correct rejections of incongruent items, and (if so) whether the relationship between activity and behavior in these regions interacted with our use of single-agenda vs. multi-agenda instructions.

Towards this end, we ran GLM multiple regression analyses on Experiments 1 and 2 separately to identify regions of the brain that discriminate between correct and incorrect responses on incongruent trials. The preprocessing steps used in this analysis were identical to the parameters that we used for the GLM analyses described in *Section 4* of

this supplemental report (*Group General Linear Model analysis: Task-specific study-phase activity*). As before, we ran the multiple regression analysis using AFNI's 3dDeconvolve program. We used all of the regressors that we used in the study-phase GLM analysis (i.e., regressors corresponding to each of the three encoding tasks, plus six nuisance motion-correction parameter vectors), and we also included six regressors specifying the different response types made during the test phase: correct and incorrect responses to incongruent items, correct and incorrect responses to congruent items, and correct and incorrect responses to new items.

The multiple regression analysis produced a set of beta weights indicating the extent to which a particular voxel's activity correlated with a condition of interest for individual subjects. To identify regions that were differentially activated (or deactivated) by a particular response type at test, we conducted a 2-way ANOVA (using AFNI's 3dANOVA2) on the beta weights from the test phase (incongruent response type: correct and incorrect). We treated response (correct, incorrect) as a fixed effect, and subject as a random effect. We ran contrasts on the correct and incorrect beta weights to identify voxels that were differentially active when the subject made correct vs. incorrect judgments on incongruent trials.

The ANOVA yielded contrast maps that showed the extent to which each individual voxel's activity was differentially modulated by incongruent response type. We used these contrast maps to identify clusters of voxels whose activity (as a group) was differentially modulated by the correct and incorrect responding. The clusters were chosen to include voxels whose contrast result was either positive or negative and exceeded $t > 5.90$ ($p = 0.0001$). The three clusters that were found are reported in

Supplemental Table 3. Coordinates are reported in the Talairach-Tournoux (T-T) Atlas coordinate space. The focus point of each cluster is reported (LPI format). The "direction of activation change" column indicates whether a particular cluster showed *increased* (+) or *decreased* (-) activity for the specified task, relative to the other two tasks.

==========

Insert Supplemental Table 3 About Here

==========

To follow up on these results, we took the peak voxels from each of the three clusters identified in the above analysis, and we ran an independent samples t-test comparing the individual subject beta weights from Experiment 1 vs. Experiment 2. The t-test revealed that the reported region in the hippocampus discriminated more strongly between correct and incorrect trials in Experiment 2 vs. Experiment 1, $t(21) = 3.20$, $p < 0.01$. This result would not survive multiple-comparisons corrections, but it is suggestive of there being a difference in how hippocampal activity relates to behavior in Experiment 1 vs. Experiment 2.

In summary: The results of this GLM analysis echo the results of our MVPA analysis. In both cases, we found brain activity that discriminates between correct and incorrect responding to incongruent items in Experiment 2 (multi-agenda source monitoring), but not Experiment 1 (single-agenda source monitoring). Importantly, the GLM result is ambiguous when considered on its own: The fact that hippocampal activity was related to behavior in Experiment 2 does not tell us what subjects were remembering on those trials, or even if they are remembering anything (insofar as the hippocampus is activated

by encoding as well as retrieval; see, e.g., Stark and Squire, 2001). The results of the MVPA analysis help to fill in these details: Specifically, the MVPA analysis tells us that retrieval of information about the actual task performed on the item at study is driving correct rejections in Experiment 2 (but not in Experiment 1).

7. Accuracy differences as a function of targeted task depth-of-processing

==========

Insert Supplemental Tables 4 and 5 About Here

==========

In this section, we explore how depth-of-processing of the targeted task affected subjects' memory performance. Prior work by Marsh and Hicks (1998) and others (e.g., Dobbins and McCarthy, 2008) has demonstrated that – on exclusion tasks – subjects respond more accurately when they are asked to target deeply encoded items vs. shallowly encoded items. In the Marsh and Hicks (1998) study, subjects either targeted items they generated from anagrams (deep encoding) or items they read (shallow encoding). Subjects in that study showed a higher hit rate and a lower false alarm rate (on both incongruent trials and new-item trials) when asked to target the "generate" condition than the "read" condition.

In our study, we can address the effects of targeting deeply vs. shallowly encoded items by comparing the target-artist condition (deep) to the target-read condition (shallow), or by comparing the target-function condition (deep) to the target-read

condition (shallow). Supplemental Table 4 presents the behavioral results from Experiment 1, broken down according to which task was targeted at test (artist, function, or read). The overall pattern of results replicates the pattern observed by Marsh and Hicks (1998): For both comparisons (target-artist vs. target-read, and target-function vs. target-read), hits were numerically greater for the deep condition than the shallow condition, and false alarms (on both incongruent trials and new-item trials) were numerically lower for the deep condition than the shallow condition. Three of the aforementioned differences were significant at $p < .05$: artist hits > read hits; function incongruent false alarms < read incongruent false alarms; artist new-item false alarms < read new-item false alarms; the other differences had p values > .05. For discussion of how to interpret this pattern of results, see Dobbins and McCarthy (2008).

Supplemental Table 5 shows the behavioral results from Experiment 2, broken down according to which task was targeted at test. As in Experiment 1, artist hits (i.e., correct "same task" responses) were significantly greater than read hits in Experiment 2, $t(11) = 5.17$, $p < .01$. There was a trend for function hits to be greater than read hits but it was not significant, $t(11) = 1.65$, $p > .05$. In Experiment 2, there were no significant effects of depth of processing on false alarms to incongruent items (i.e., incorrect "same task" responses); neither artist nor function false alarms to incongruent items differed significantly from read false alarms. Likewise, there were no significant effects of depth of processing on false alarms to new items (i.e., incorrect "same task" responses); neither artist nor function false alarms to new items differed significantly from read false alarms.

8. Accuracy differences as a function of actual task depth-of-processing

==========

Insert Supplemental Tables 6 and 7 About Here

==========

==========

Insert Supplemental Figure 4 About Here

==========

The previous section explored how depth-of-processing of the targeted task affected performance. This section explores how depth-of-processing of the *actual* task (on incongruent trials) affected performance. We expected that the effect of deep vs. shallow processing would depend on *how well subjects were making use of retrieved information* about the actual (non-target) task. If subjects make proper use of actual-task recollection (i.e., they treat it as evidence that item was *not* studied with the targeted task), accuracy should be better for deep vs. shallow items; deep items are more likely to trigger actual-task recollection (see Yonelinas, 2002 for a review), and thus will be more likely to trigger a correct rejection.

Supplemental Table 6 shows the behavioral results for incongruent trials in Experiment 1, split by actual task. Numerically, the data show a reversed depth-of-processing effect: Accuracy was better for the (shallow) read task than the (deep) artist and function tasks. The function vs. read difference was significant, $t(10) = -2.90$, $p = 0.02$ but the artist vs. read difference was not, $t(10) = -.81$, $p > .05$. This trend towards a reversed depth of processing effect suggests that subjects in Experiment 1 were *not* using

actual-task recollection to reject incongruent items (otherwise, the opposite pattern of results would have been observed).

Supplemental Table 7 shows the behavioral results for incongruent trials in Experiment 2, split by actual task. In this study, there was a normal depth-of-processing effect: Accuracy was better for the (deep) artist and function tasks than the (shallow) read task. The artist vs. read difference was significant, $t(11) = 3.46$, $p = 0.005$ but the function vs. read difference was not, $t(11) = 1.83$, $p = 0.09$. These results fit with the idea that subjects in Experiment 2 were utilizing actual-task recollection to (correctly) reject incongruent items.

These depth-of-processing effects, coupled with the fact that AT classifier activity was much higher for deep vs. shallow tasks in Experiment 2 (average AT classifier activity was .57 for artist, .59 for function, and .35 for read at time point 2), suggest a possible confound: Insofar as deep tasks were associated with *higher accuracy* than shallow tasks, and deep tasks were also associated with *higher AT activity*, then the observed relationship between AT activity and accuracy in Experiment 2 (at time point 2) may reflect task differences in AT activity (i.e., deep tasks = high accuracy and high AT; shallow tasks = low accuracy and low AT) as opposed to a within-task relationship between AT activity and accuracy. If this were the case, it would not contradict our hypothesis (insofar as it still shows a relationship between AT and accuracy), but it would compromise our claim that we can detect behaviorally meaningful differences in AT activity across trials (within conditions).

To address the concern that depth-of-processing differences were driving our effect, we re-ran the Experiment 2 area-under-the-ROC (AUC) analysis, looking just at the trials

where the actual task was a deep encoding task (artist or function). Also, to avoid the possibility that differences between the two deep tasks might introduce a confound, we ran the analysis separately for "actual task = artist" and "actual task = function" trials and then averaged these results together (weighting them equally) for each subject.[1] The results of this analysis are shown in Supplemental Figure 4, along with the results of our original AUC analysis from the main paper (Figure 4B) where we did not restrict by task. Importantly, the AUC scores were still just as large in the new analysis as in the original analysis (and AUC for time point 2 was still significantly above .5). This finding indicates that, while there were AT differences across tasks, our ability to predict behavior using AT in Experiment 2 was not a simple artifact of task differences: We were still able to predict behavior when we focused just on AT variance that was present within the two deep-task conditions.[2]

9. RT analyses

==========

Insert Supplemental Table 8 About Here

==========

[1] We were not able to run this analysis for Experiment 1 because the analysis requires at least one artist error and one function error per subject (otherwise AUC is undefined), and this was not the case for all of the subjects in Experiment 1.

[2] When we limited the analysis to "actual task = read" trials, we did not find a positive relationship between AT activation and correct rejections at time point 2. To the contrary, AT activation was nonsignificantly higher for errors than for correct rejections. This null effect may be attributable to a floor effect on AT recollection (and classifier activation) for the read task. Practically speaking, this finding suggests that researchers interested in showing a relationship between AT activation and behavior should focus on deep (as opposed to shallow) tasks.

Supplemental Table 8 shows reaction times as a function of response type and trial type (congruent, incongruent, and new) for both Experiment 1 and Experiment 2. Note that trials where subjects did not respond within the allotted 2 seconds were omitted from the RT analysis, since we did not record a response on these trials. These RT results provide a source of converging evidence regarding subjects' use (or lack of use) of actual-task recollection. Several studies have found that scrutinizing retrieved information is a time-consuming process (see, e.g., Gronlund and Ratcliff, 1989; Hintzman and Curran, 1994; Rotello and Heit, 2000); subjects need to wait for recollected information to come to mind, and they need to compare recollected information with their representation of the target source. If subjects rely on this (slow) process to correctly reject familiar lures, this should slow down correct rejection RTs; in particular, it should differentially slow down correct rejection responses compared to hits (which can sometimes be triggered by fast familiarity, in addition to slower, more deliberative processes). RT data from Malmberg (2008) support this idea: Malmberg used an associative recognition paradigm where subjects were relying heavily on recollected information to reject familiar lures; in this study, he found that correct rejections of related lures were significantly slower than hits.

The above results imply that we can use the RT difference between incongruent correct rejections (CRs) and hits to assess whether subjects are scrutinizing retrieved details: The more that subjects scrutinize retrieved details, the larger this RT difference should be. In particular, the claim (from the main paper) that subjects are making use of recollected details in Experiment 2 but not Experiment 1 suggests that the (incongruent)

CR – hit RT difference should be larger in Experiment 2 than in Experiment 1. To test this hypothesis, we computed CR – hit RT difference scores for each subject in both experiments, and then we ran a t-test to see whether these difference scores were different across experiments. Numerically, incongruent correct rejections were 57 msec slower than hits in Experiment 2, but correct rejections were 26 msec *faster* than hits in Experiment 1. As predicted, the CR – hit RT difference was significantly larger in Experiment 2 than in Experiment 1, $t(23) = 2.08$, $p < .05$. This finding supports our hypothesis that subjects were scrutinizing retrieved details more carefully in Experiment 2 vs. Experiment 1.

REFERENCES

Badre D, and Wagner AD (2007) Left ventrolateral prefrontal cortex and the cognitive control of memory. Neuropsychologia 45:2883-2901.

Bishop C (1995) Neural Networks for Pattern Recognition. New York: Oxford University Press.

Bullmore E, Long C, Suckling J, Fadili J, Calvert G, Zelaya F, Carpenter TA, and Brammer M (2001) Colored noise and computational inference in neurophysiological (fMRI) time series analysis: resampling methods in time and wavelet domains. Hum Brain Mapp 12:61-78.

Cox R (1996) AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. Computers and Biomedical Research 29.

Dobbins I, and McCarthy D (2008) Cue framing effects in source remembering: A memory misattribution model. Memory and Cognition 36:104-118.

Duda R, Hart P, and Stork D (2001) Pattern Classification, 2nd Edition. New York: Wiley.

Gronlund S, and Ratcliff R (1989) Time course of item and associative information: implications for global memory models. J Exp Psychol Learn Mem Cogn 15:846-858.

Hastie T, Tibshirani R, and Friedman J (2001) The Elements of Statistical Learning. New York: Springer.

Haynes J, and Rees G (2006) Decoding mental states from brain activity in humans. Nat Rev Neurosci 7:523-534.

Hintzman DL, and Curran T (1994) Retrieval dynamics of recognition and frequency judgments - evidence for separate processes of familiarity and recall. J Mem Lang 33:1-18.

Jacoby LL, Shimizu Y, Daniels KA, and Rhodes MG (2005) Modes of cognitive control in recognition and source memory: Depth of retrieval. Psychon Bull Rev 12, 852-857.

Kamitani Y, and Tong F (2005) Decoding the visual and subjective contents of the human brain. Nat Neurosci 8:679-685.

LeCun Y, Bottou L, Orr G, and Mueller K (1998) Efficient BackProp. Lecture Notes in Computer Science 1524:9.

Malmberg K (2008) Towards an understanding of individual differences in episodic memory: Modeling the dynamics of recognition memory. In: The Psychology of Learning and Motivation: Skill and Strategy in Memory Use (Benjamin, A and Ross, B eds), pp 313-349. San Diego: Academic Press.,

Marsh R, and Hicks J (1998) Test formats change source-monitoring decision processes. Journal of Experimental Psychology: Learning, Memory, and Cognition 24:1137-1151.

Mitchell T, Hutchinson R, Niculescu S, Pereira F, Wang X, Just M, and Newman S (2004) Learning to decode cognitive states from brain images. Machine Learning 57:145-175.

Norman K, Polyn S, Detre G, and Haxby J (2006) Beyond mind-reading: multi-voxel pattern analysis of fMRI data. Trends Cogn Sci 10:424-430.

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

Rotello C, and Heit E (2000) Associative recognition: a case of recall-to-reject processing. Mem Cognit 28:907-922.

Rumelhart D, Durbin R, Golden R, and Chauvin Y (1996) Backpropagation: the basic theory. In: Backpropagation: theory, architectures, and applications (Chauvin Y and Rumelhart D, eds), pp 1-34. Mahwah, NJ: Erlbaum.

Stark, CEL and Squire, LR (2001) When zero is not zero: The problem of ambiguous baseline conditions in fMRI. Proc Natl Acad Sci 98:12760-12766.

TABLES

Supplemental Table 1

Experiment 1: Study phase percent correct classification results, 1000 voxels

Subject	Percent Correct	P Value
1	84.96	< 0.0001
2	78.75	< 0.0001
3	56.43	< 0.0001
4	57.07	< 0.0001
5	79.69	< 0.0001
6	88.74	< 0.0001
7	92.99	< 0.0001
8	83.79	< 0.0001
9	84.70	< 0.0001
10	88.73	< 0.0001
11	74.62	< 0.0001

Experiment 2: Study phase percent correct classification results, 1000 voxels

Subject	Percent Correct	P Value
1	80.01	< 0.0001
2	77.17	< 0.0001
3	85.37	< 0.0001

4	64.17	< 0.0001
5	75.52	< 0.0001
6	78.03	< 0.0001
7	78.30	< 0.0001
8	50.50	< 0.0001
9	72.02	< 0.0001
10	81.00	< 0.0001
11	84.59	< 0.0001
12	60.31	< 0.0001

Supplemental Table 2

Voxel clusters differentially activated by the encoding tasks, as identified by the study-phase group GLM analysis (using data from Experiment 1 and Experiment 2)

Study Task	L	P	I	Volume	# Voxels	Direction of Activation Change	Area
Artist	31	-88	-4	2160	80	-	R. Inferior Occipital Gyrus
	-27	-91	-10	1836	68	-	L. Inferior Occipital Gyrus
	32	-62	-26	810	30	-	R. Cerebellum
	-49	-5	47	729	27	-	L. Precentral Gyrus
	-3	1	60	513	19	-	L. Medial Frontal Gyrus
Function	22	-68	40	3321	123	-	R. Precuneus
	-10	49	40	2511	93	+	L. Superior Frontal Gyrus (BA 8)
	-47	28	-1	2430	90	+	L. Inferior Frontal Gyrus
	-5	9	62	1971	73	+	L. Superior Frontal Gyrus (BA 6)
	-45	-43	43	1728	64	-	L. Inferior Parietal Lobule
	49	-59	-15	1323	49	-	R. Fusiform Gyrus
	-27	-69	36	1242	46	-	L. Precuneus
	-56	-16	-8	1080	40	+	L. Middle Temporal Gyrus (BA 21)
	-48	-67	-9	1053	39	-	L. Middle Occipital Gyrus (BA 19)
	-8	-77	-6	648	24	+	L. Lingual Gyrus
	-61	-40	0	648	24	+	L. Middle Temporal Gyrus (BA 21/22)

	-32	-86	0	594	22	-	L. Middle Occipital Gyrus (BA 18)
	-4	49	-1	540	20	+	L. Anterior Cingulate
	44	-0	29	432	16	-	R. Precentral Gyrus
	-44	2	54	432	16	+	L. Middle Frontal Gyrus
	-52	-61	18	324	12	+	L. Superior Temporal Gyrus
	-53	1	35	324	12	-	L. Precentral Gyrus
	33	-84	5	297	11	-	R. Middle Occipital Gyrus
	-7	17	42	270	10	+	L. Cingulate Gyrus
Read	35	-76	-8	8019	297	+	R. Middle Occipital Gyrus
	-36	-82	-6	7020	260	+	L. Inferior Occipital Gyrus
	24	-66	41	6129	227	+	R. Precuneus
	-34	-58	40	4617	171	+	L. Inferior Parietal Lobe
	-50	0	31	1134	42	+	L. Precentral Gyrus
	-11	52	42	945	35	-	L. Superior Frontal Gyrus
	-33	-59	-24	918	34	+	L. Culmen
	-4	49	-2	864	32	-	L. Anterior Cingulate (BA 10)
	46	0	29	567	21	+	R. Precentral Gyrus
	-9	-75	-8	324	12	-	L. Lingual Gyrus
	-3	36	-6	270	10	-	L. Anterior Cingulate (BA 32)
	-46	28	-5	243	9	-	L. Inferior Frontal Gyrus
	40	42	25	243	9	+	R. Middle Frontal Gyrus

Supplemental Table 3

Voxel clusters differentially activated by correct vs. incorrect responding to incongruent items, as identified by the test-phase group GLM analysis (described in Section 6).

Exp	L	P	I	Volume	# Voxels	Direction of Activation Change	Area
1							(no clusters found)
2	28	-5	-1	162	6	+	R. Lentiform Nucleus
	-31	-14	2	108	4	+	L. Lentiform Nucleus
	-16	-26	-18	81	3	+	L. Hippocampus

Supplemental Table 4

Experiment 1: Mean proportions of trials where subjects responded "yes," responded "no," or failed to respond in time as a function of trial type (congruent, incongruent, and new) and which task was targeted at test

	Yes	No	Failed to Respond
Congruent, Artist	0.83 (0.05)	0.07 (0.04)	0.11 (0.04)
Congruent, Function	0.69 (0.07)	0.17 (0.05)	0.14 (0.07)
Congruent, Read	0.55 (0.08)	0.20 (0.04)	0.24 (0.07)
Incongruent, Artist	0.09 (0.03)	0.77 (0.05)	0.14 (0.05)
Incongruent, Function	0.07 (0.02)	0.81 (0.05)	0.12 (0.05)
Incongruent, Read	0.12 (0.03)	0.71 (0.06)	0.17 (0.05)
New, Artist	0.03 (0.02)	0.88 (0.04)	0.10 (0.04)
New, Function	0.04 (0.02)	0.87 (0.04)	0.09 (0.04)
New, Read	0.10 (0.03)	0.74 (0.06)	0.17 (0.04)

Note: Numbers in parentheses indicate the standard error of the mean.

Supplemental Table 5

Experiment 2: Mean proportion of trials where subjects responded "same task", responded "different task", responded "new", or failed to respond in time as a function of trial type (congruent, incongruent, and new) and which task was targeted at test

	Same Task	Different Task	New	Failed to Respond
Congruent, Artist	0.68 (0.04)	0.06 (0.02)	0.02 (0.01)	0.25 (0.05)
Congruent, Function	0.47 (0.07)	0.13 (0.05)	0.01 (0.01)	0.39 (0.08)
Congruent, Read	0.34 (0.06)	0.13 (0.05)	0.11 (0.06)	0.43 (0.08)
Incongruent, Artist	0.08 (0.02)	0.43 (0.06)	0.09 (0.02)	0.40 (0.08)
Incongruent, Function	0.07 (0.03)	0.52 (0.06)	0.09 (0.03)	0.33 (0.07)
Incongruent, Read	0.08 (0.03)	0.50 (0.06)	0.06 (0.02)	0.37 (0.07)
New, Artist	0.01 (0.01)	0.07 (0.04)	0.60 (0.07)	0.31 (0.07)
New, Function	0.01 (0.01)	0.10 (0.06)	0.56 (0.08)	0.32 (0.08)
New, Read	0.00 (0.00)	0.05 (0.03)	0.55 (0.08)	0.40 (0.08)

Note: Numbers in parentheses indicate the standard error of the mean.

Supplemental Table 6

Experiment 1: Mean proportions of incongruent trials where subjects responded "yes", responded "no", or failed to respond in time, as a function of the actual task performed on the test word at study.

Actual Task	Yes	No	Timeout
Artist	0.09 (0.03)	0.78 (0.06)	0.13 (0.04)
Function	0.14 (0.03)	0.70 (0.06)	0.16 (0.05)
Read	0.05 (0.02)	0.81 (0.06)	0.14 (0.06)

Note: Numbers in parentheses indicate the standard error of the mean.

Supplemental Table 7

Experiment 2: Mean proportions of incongruent trials where subjects responded "same task", responded "different task", responded "new", or failed to respond in time, as a function of the actual task performed on the test word at study.

Actual Task	Same Task	Different Task	New	Timeout
Artist	0.07 (0.02)	0.59 (0.05)	0.03 (0.01)	0.30 (0.06)
Function	0.09 (0.03)	0.48 (0.07)	0.06 (0.02)	0.37 (0.08)
Read	0.06 (0.02)	0.37 (0.07)	0.14 (0.05)	0.43 (0.09)

Note: Numbers in parentheses indicate the standard error of the mean.

Supplemental Table 8

Experiment 1: Mean reaction time in milliseconds, as a function of response (yes, no) and trial type (congruent, incongruent, and new)

	Yes	No
Congruent	1367 (42)	1415 (74)
Incongruent	1564 (49)	1341 (52)
New	1593 (78)	1190 (53)

Note: Numbers in parentheses indicate the standard error of the mean. All 11 subjects contributed to each cell, except for the new/yes cell (where 9 subjects contributed).

Experiment 2: Mean reaction time in milliseconds, as a function of response (same task, different task, new) and trial type (congruent, incongruent, and new)

	Same Task	Different Task	New
Congruent	1496 (40)	1549 (38)	1516 (62)
Incongruent	1579 (55)	1553 (37)	1476 (39)
New	1116 (395)	1651 (69)	1400 (52)

Note: Numbers in parentheses indicate the standard error of the mean. All 12 subjects contributed to each cell, except for the following cells: congruent/new (7 subjects contributed), incongruent/new (10 subjects contributed), new/same (3 subjects contributed), and new/different (7 subjects contributed).

FIGURE CAPTIONS

Supplemental Figure 1: Plot showing how study-phase classification accuracy (averaged across subjects from Experiment 1 and Experiment 2) varied as a function of the number of voxels included in the classification analysis. Error bars indicate the SEM (across subjects) of the classification accuracy scores. See *Section 1* for details of our voxel selection procedure, and see *Section 2* for details of how we computed study-phase classification accuracy. Classification accuracy peaked at 1,000 voxels. Note that, for the range of values explored here, classification was well above chance (33%) regardless of how many voxels were selected.

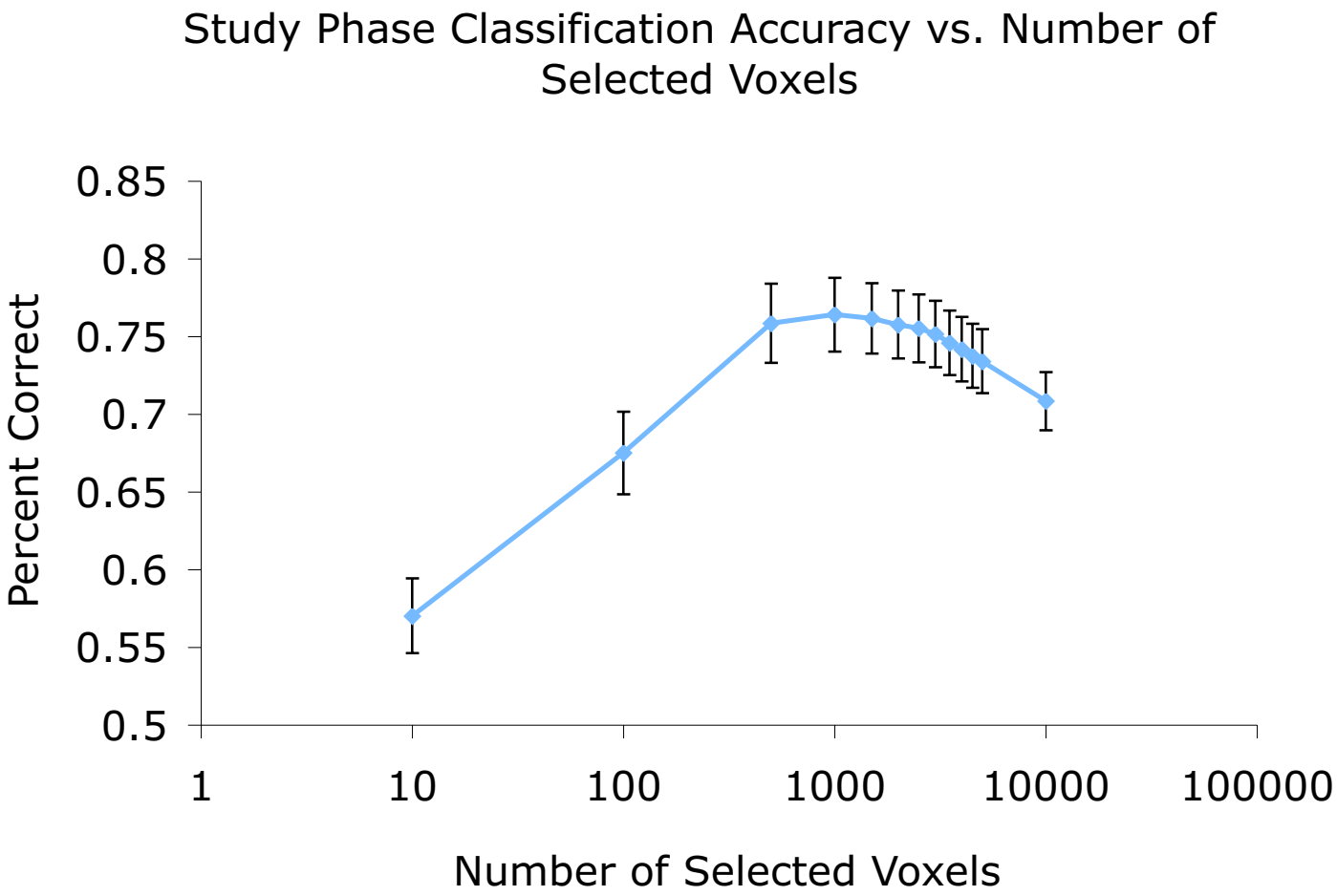
Supplemental Figure 2: Event-related averages of targeted-task (TT) and other-task (OT) classifier output from Experiment 1, after we factor out "spill-over" of TT activity from preceding trials. The plots include data from new-item trials and incongruent trials. To ensure that our measure of TT activity (on this trial) was not contaminated by TT activity from the preceding trial, these plots only include trials/conditions where the targeted task and the other task (on this trial) differed from the targeted task on the preceding trial. The left-hand side of the figure shows classifier output for 7 successive scans, starting with the scan when the test word was presented. The right-hand side of the figure shows classifier output for 7 successive scans, starting with the scan when the *task cue* was presented (note that, on a given trial, the test word was presented either 2, 3, or 4 time points after the task cue). On both the left and the right, the upper graph shows raw

classifier outputs, and the lower graph shows TT – OT difference scores. Error bars on the lower graphs indicate the SEM (across subjects) of the difference score. Scores marked with asterisks are significantly different from zero; one asterisk indicates $p < .05$; two asterisks indicate $p < .01$.

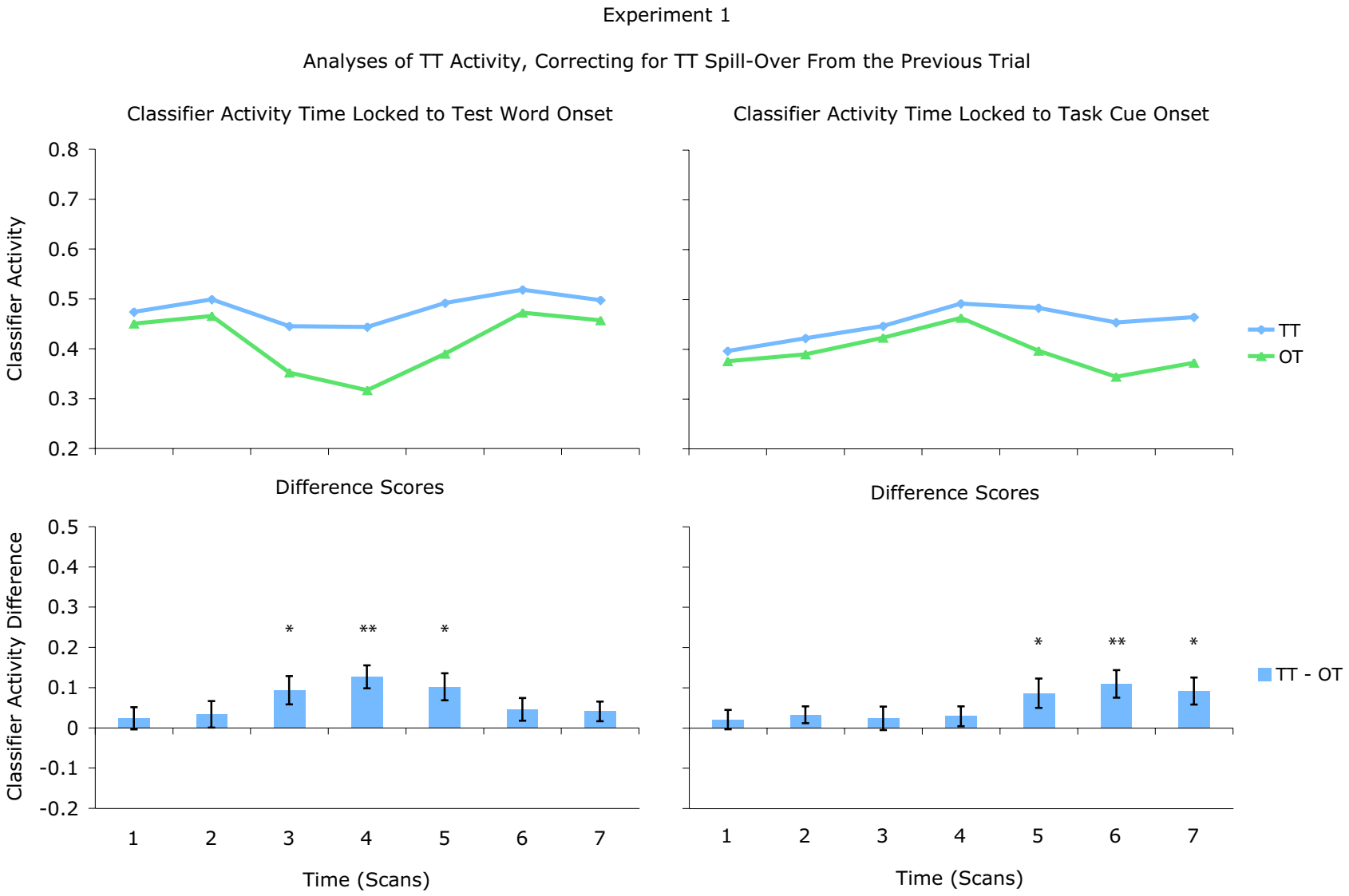
Supplemental Figure 3: Classifier importance maps, showing which voxels were used by the classifier to discriminate between the three tasks. Specifically, the three figures plot the average *importance* of each voxel for each task, where importance is a function of the weight connecting the voxel to the task unit, and the average activity of the voxel when the task is being performed (see *Section 5* for details). Voxels can contribute to detecting a task state by having a positive (above-average) activation value for the task and a positive weight, or by having a negative (below-average) activation value for the task and a negative weight. Voxels in the former category (positive activation, positive weight) were given a positive importance value equal to activation * weight. Voxels in the latter category (negative activation, negative weight) were given a negative importance value equal to – activation * weight. The importance maps shown here were created by computing individual subject importance maps (for subjects from both Experiment 1 and Experiment 2), putting them in Talairach space, applying a 4mm Gaussian blur, and then averaging the maps together. Red patches indicate positive importance values and blue patches indicate negative importance values. The slices in each diagram (going from left-to-right and top-to-bottom) correspond to Z = -12, -2, 8, 18, 28, 38, 48, 58, and 68. The underlay shows anatomical data from a single subject (subject 8, Experiment 1).

Supplemental Figure 4: Analysis of how actual-task activity relates to behavior on incongruent trials in Experiment 2. Area under the curve (AUC) scores > .5 indicate that AT activity at that time point is associated with *increased correct rejections*; for further explanation of the AUC measure see the main paper (Figure 4). "All" = version of the analysis where we lumped all trials together (regardless of the identity of the actual task) – these results were shown in Figure 4 in the main paper. "AF only" = version of the analysis where we computed AUC within the two deep tasks (artist and function) separately, and then averaged these results together. Scores marked with asterisks are significantly different from chance (.5) at $p < .05$.

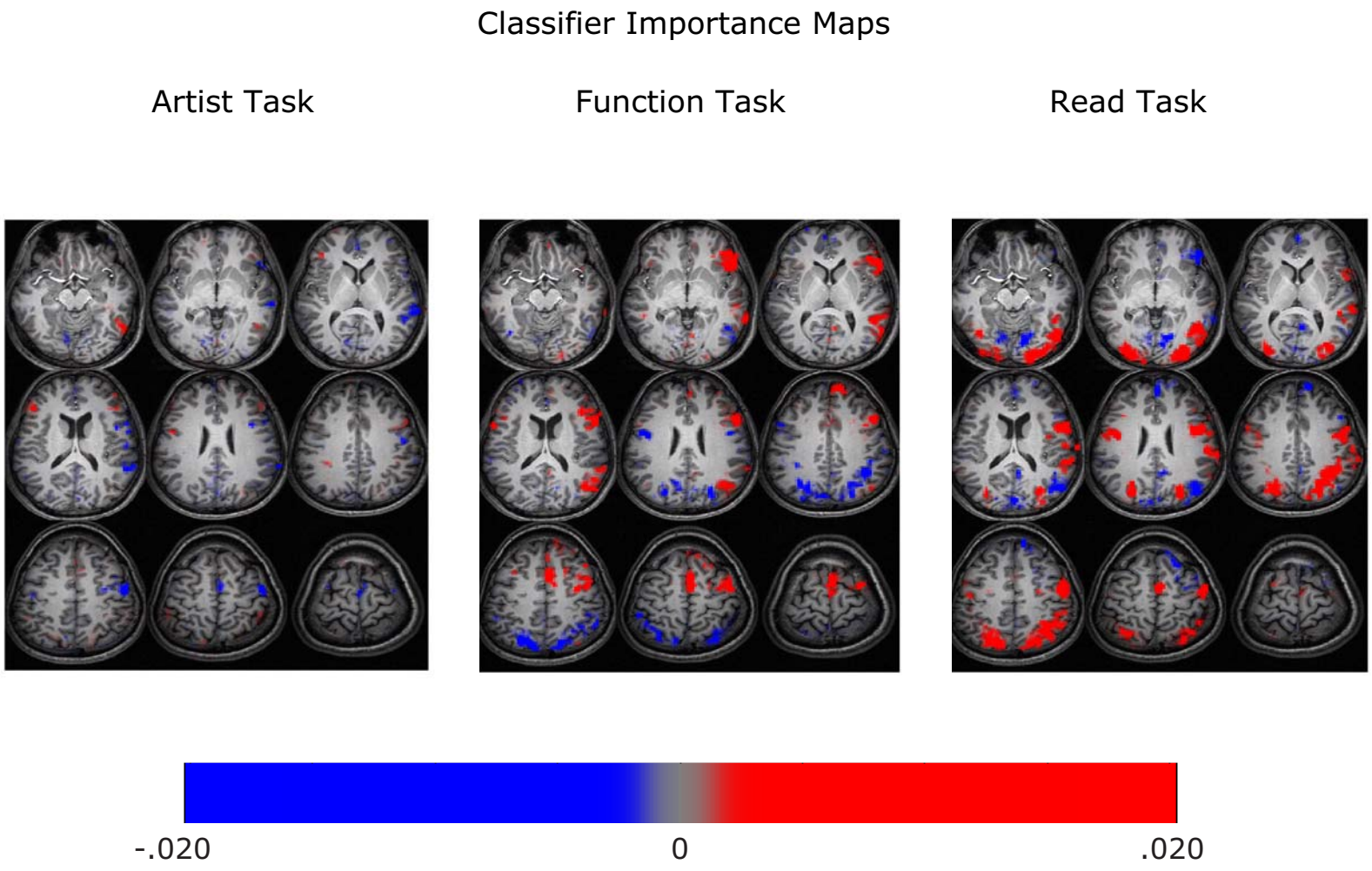
Supplemental Figure 1



Supplemental Figure 2



Supplemental Figure 3



Supplemental Figure 4