Neuron, Volume *63*

Supplemental Data

Recollection, Familiarity, and Cortical Reinstatement:

A Multivoxel Pattern Analysis

Jeffrey D. Johnson, Susan G.R. McDuff, Michael D. Rugg, and Kenneth A. Norman

Table of Contents

1. Accuracy analyses based on Artist and Function data

As described in the main text, subjects made very few Remember responses to items from the Read task, raising the concern that (due to lack of data) estimates of neural reinstatement these trials would be very noisy, and thus might obscure other effects present in our data. As such, we conducted the analyses of classifier output magnitude both on the data from all three tasks and on the data restricted to the Artist and Function tasks (see Figure 4). The classifier *accuracy* analyses presented in the main paper (Figure 3) used test data from all three tasks. Here, for completeness, we present the results of classifier accuracy analyses focusing only on Artist and Function test trials. We expected the general pattern of findings to be qualitatively similar to those reported in the main text—an expectation that was met.

Classifier accuracy for test items from a particular task was operationalized as the probability that the value at that output node was greater than that for the other two task nodes. In the present analyses, this accuracy measure was computed separately for the Artist and Function test trials, and then averaged across the tasks. The measure was computed by collapsing over all test responses and also separately for each response category (Remember, Sure Old, and Other). Figure S1 shows the resulting accuracy values for a series of TRs, beginning with item onset (TR 1). Classifier accuracy was tested against chance (.33) with pair-wise t-tests at each TR (correcting for multiple comparisons with an overall $p < .05$; Holm, 1979). Accuracy was significantly different from chance at TRs 2 through 7 for all responses collapsed (min. t_{15} = 3.77, $p < .0025$) and for Remember responses (min. $t_{15} = 3.26$, $p < .01$). These accuracy measures for Remember and collapsed responses appeared to be higher and more stable than those including Read items, likely due to the low rate of Remember responses contributing to the Read condition. For Sure Old responses, accuracy differed from chance at TR 5 ($t_{15} = 3.42$, $p < .005$). Although these Sure Old differences were slightly noisier than those reported in the main text, they remain consistent with our findings of reinstatement for such responses. Finally, there were no significant differences associated with the accuracy measure for Other responses.

2. Results of GLM-based analyses

For comparison with the results of the searchlight-based classifications, we conducted a parallel GLM-based analysis. As is conventional for this type of analysis, the fMRI data were first smoothed with an 8 mm FWHM Gaussian kernel. The study phase was modeled by three separate onset regressors of interest (one for each study task). The test phase was modeled by 12 regressors of interest: one for each of the four item types (Artist, Function, Read, and new) crossed with the three response categories (Remember, Sure Old, and Other). An additional regressor for each experimental phase was used to account for incorrect/absent/multiple responses. The six nuisance regressors generated from spatial realignment were also included in the GLM, as was a constant for each of the study and test blocks.

The GLM analysis employed a two-stage mixed effects model. In the first stage, the neural activity due to the study and test words was modeled by delta (impulse) functions at each item onset. The ensuing BOLD response was modeled by convolving the impulse functions with a canonical hemodynamic response function (HRF) and its temporal and dispersion derivatives (Friston et al., 1998). The convolved time courses were downsampled at the midpoint of each TR to form the covariates of the GLM. The parameters for each covariate and the hyper-parameters governing the error covariance were estimated using a restricted maximum likelihood method. Nonsphericity of the error covariance was accommodated by an AR(1) model, in which the temporal autocorrelation was estimated by pooling over supra-threshold voxels (Friston et al., 2002). In the second stage of analysis, the aforementioned parameter estimates were subjected to linear contrasts (consisting of one-sample t-tests) where subjects were treated as a random effect.

As we have previously argued (Johnson & Rugg, 2007), two types of contrasts are necessary to investigate reinstatement with a GLM-based analysis: a task-related (or content-related) contrast from the study phase and an analogous contrast from the test phase. Here, we contrasted the Artist and Function conditions for the study phase data and for each of the three response categories in the test phase. The Read condition was excluded from analysis due to low trial numbers for each response. Reinstatement effects were determined with an inclusive masking procedure (Johnson & Rugg, 2007), in which the intersection of the directional study phase comparison (e.g., Artist > Function) and the corresponding directional test phase comparison (e.g., Artist > Function, for Remember responses) was determined. This procedure was then repeated for the reverse contrast and for the other test response categories. We did not restrict the outcomes of these contrasts to voxels that showed differences according to response (e.g., Remember > Sure Old), because the main focus of the GLM analysis was to determine whether reinstatement was at all present for each response category.

One issue with comparing the GLM and classification results is that the statistical thresholds used for one type of analysis might not be appropriate for the other type. As described in the main text, the searchlight analyses employed voxel-wise height thresholds of $p < .01$ along with SPM's cluster-wise correction procedure (Worsley et al., 1996). This correction gave rise to clusters that survived $p < .05$ and were at least approximately 30 voxels in size. To determine the significant effects from the GLM analyses, we first used a threshold-setting procedure comparable to that used for our searchlight analyses. In this procedure, the study phase and test phase contrasts were separately thresholded on a voxel-wise basis at $p < .036$, giving rise to a combined threshold of $p < .01$ (Lazar et al., 2002). The results of the study phase contrasts alone are shown in Figure S2 (thresholded more conservatively at $p < .01$ for display only). As can be seen in the figure, there are significant clusters of voxels for both directions of contrast. These effects, much like the importance maps shown in Figure 2 of the main paper, ensure that there are a number of cortical regions capable of being reactivated during the test phase. (Note that the maps in Figure S2 and Figure 2 were constructed in very different ways and, as a consequence, should not be directly compared to one another.)

Next, regions demonstrating reinstatement were identified by inclusively masking the outcomes of the above study contrasts (at $p < .036$) with the associated task-related effects at test (also at $p < .036$). The results of this analysis are shown in Figure S3. Using this procedure to determine the Artist > Function overlap for the study phase and Remember responses gave rise to a single large cluster in medial parietal cortex (encompassing posterior cingulate, precuneus, and retrosplenial cortex), which survived a cluster-level threshold of $p < .05$. Interestingly, this cluster was in the vicinity of the reinstatement effects we observed in the searchlight analyses to be selective for Remember responses. There were no significant effects at this threshold ($p < .05$, cluster-level) for the reverse contrast (Function > Artist) or for the contrasts involving the remaining two response categories. Thus, it appears that the searchlight analyses were more sensitive than the GLM analyses at detecting reinstatement.

In a second attempt at detecting differences according to study task, we contrasted the Artist and Function data from the test phase, using a voxel-wise threshold of $p < .01$ and a cluster-wise threshold of $p < .05$ (Worsley et al., 1996). These contrasts can be considered to be more liberal than both our GLM and searchlight analyses for two reasons. First, we were able to focus all of the statistical power on the test-phase contrasts, rather than requiring that there was study-test overlap. Second, we tested each direction of the Artist-Function contrast at $p < .01$, which results in a bidirectional threshold of $p < .02$. From this procedure, we again observed significant Artist > Function effects for the Remember response category, which included regions that were additional to the overlapping effects shown above (see Figure S4). However, there were again no significant effects for Sure Old or Other responses.

Finally, we were interested in the involvement of the hippocampus in the present study. Under the assumption that the hippocampus plays a central role in reinstatement (Norman & O’Reilly, 2003), we expected the region to exhibit enhanced activity for responses that were associated with larger reinstatement effects. To further explore these and nearby effects, we broadened our search to also include the surrounding medial temporal lobe (MTL) cortex. The analysis was restricted to this region with a mask that was drawn on the across-subjects mean anatomical image (using MRIcron software; http://www.mricron.com). Within the MTL, a voxel-wise threshold of $p < .001$ and a cluster-wise threshold of $p < .05$ were used (Worsley et al., 1996). First, we contrasted the Remember and Sure Old responses (collapsing over Artist and Function), revealing a single region in right hippocampus as shown in Figure S5. This result is consistent with our reported findings of greater reinstatement for Remember responses. However, contrasting Sure Old and Other responses revealed no MTL effects (as did a contrast between Sure Old responses and correctly-rejected New items). Although this null result provides no support that reinstatement for Sure Old responses relies on the hippocampus, the result is likely due to the counteracting effects of novelty-related encoding processes (see main text for further discussion).

3. Additional details of the searchlight classification results

For the outcomes of the searchlight map contrasts reported in the main text, Table S1 provides detailed characteristics (e.g., peaks, cluster sizes, and statistics) of the significant effects.

4. Analysis of behavioral performance during the study phase

Response times (RTs) from the study phase were submitted to a one-way ANOVA (task: Artist, Function, and Read), which gave rise to a significant main effect ($F_{1.8,26.6} = 10.41$, $p < .005$; degrees of freedom corrected according to Greenhouse & Geisser, 1959). Follow-up t-tests revealed that RTs to Artist items (M = 898 msec, SEM = 54; taken from response cue onset) were shorter than those to Function and Read items (M = 1111 and 1024 msec, SEM = 56 and 38, respectively; minimum $t_{15} = 2.89$, $p < .025$).

5. Study-phase classification used to facilitate voxel selection

As described in the main text, an additional classifier based on only study phase data was used to determine the optimal number of voxels for the whole-brain classification. It is important to stress that by using only the study phase data for both training and testing of this classifier, the generalizability of our primary classification (as reported in the main text) to the independent set of test phase data is not compromised.

The study-phase classifier was trained and tested using a ‘leave-one-out’ cross-validation procedure (see Hastie et al., 2001), whereby the data from two of the three study blocks were used for training, and the data from the remaining study block were used for testing. This procedure was repeated for each combination of study blocks, resulting in three training-testing iterations. The fMRI data from only those TRs that were assigned to a study task (through the binarization method described earlier) were used as training and testing patterns.

Voxel selection began by setting up a generalized linear model (GLM) incorporating regressors corresponding to the convolved time course for each study task along with the nuisance regressors generated from spatial realignment. A separate GLM was defined for each pair of study phase blocks (due to cross-validation). ANOVA of the parameter estimates for the three task conditions generated an F-value at each voxel. The F-values were sorted in descending order, and voxels exhibiting the largest values were selected to be used as input data. This selection was carried out over a range of between 10 and 10000 selected voxels (out of a total M = 35859, SD = 3353, per subject).

Classifier accuracy was computed based on the data from the study phase block that was excluded from training. The accuracy values from each cross-validation iteration were then averaged into a single accuracy score for each number of selected voxels for each subject. The group-wise mean (±SEM) scores across the range of selected voxels are shown in Figure S6. As can be seen in the figure, classifier accuracy peaked at around 1000 voxels with a value of .76, and decreased as more voxels were included (for similar results, see McDuff et al., 2009). Notably, the accuracy of the study-study classifier was reliably above chance (.33) at each level of voxel selection (min. $t_{15} = 12.16$, $p < .001$). In addition to providing an optimal number of selected voxels, these results also confirmed that the three study tasks could be distinguished on the basis of the fMRI data.

The optimal number of 1000 voxels was carried forward to the whole-brain MVPA that is reported in the main text. We used the same number of voxels for each subject, rather than tailoring the number individually to each subject. Note that the whole-brain classifications reported in the main text involve re-selecting voxels based on a single GLM of all three blocks of study phase data. Thus, those selected voxels could differ (although probably slightly) from the voxels selected by the three separate GLMs used here (one for each cross-validation iteration).

6. Supplemental references

Friston, K.J., Fletcher, P., Josephs, O., Holmes, A., Rugg, M.D., Turner, R. (1998). Event-related fMRI: characterizing differential responses. NeuroImage *7*, 30-40.

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

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

Holm, S. (1979). A simple sequentially rejective multiple test procedure. Scand. J. Stat. *6*, 65-70.

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

Lazar, N.A., Luna, B., Sweeney, J.A., Eddy, W.F. (2002). Combining brains: a survey of methods for statistical pooling of information. NeuroImage *16*, 538-550.

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

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

Worsley, K.J., Marrett, S., Neelin, P., Vandal, A.C., Friston, K.J., Evans, A.C. (1996). A unified statistical approach for determining significant signals in images of cerebral activation. Hum. Brain Mapp. *4*, 58-73.

Table S1. Results of the searchlight contrasts testing where reinstatement was equivalent for Remember and Sure Old responses and where reinstatement was selective for Remember responses. Direct contrasts of responses were conducted on the output values of the correct classifier node, whereas the remaining contrasts compared accuracy values to chance. All effects survived cluster-level correction of p < .05. Coordinates are in Talairach space. L = left.

Contrast description	Region	# of voxels	Peak coordinates			Peak Z-score
			x	y	z	
Equivalent reinstatement for Remember and Sure Old responses: Remember$_{acc}$ > .33 (p < .01), inclusively masked with Sure Old$_{acc}$ > .33 (p < .01), exclusively masked with Remember vs. Sure Old (p < .1)	L lateral temporal cortex	66	-57	-45	-9	4.70
	L superior frontal gyrus	33	-42	-3	54	4.37
	L inferior frontal gyrus	48	-51	18	21	4.30
Selective reinstatement for Remember responses: Remember$_{acc}$ > .33 (p < .01), inclusively masked with Remember > Sure Old (p < .01), exclusively masked with Sure Old$_{acc}$ vs. .33 (p < .1)	L retrosplenial cortex	40	-6	-57	15	3.43
	L posterior cingulate	36	-12	-66	39	3.38

Figure S1. Mean classifier accuracy (±SEM) collapsed across all response categories and separated by response category, based on only items from the Artist and Function tasks. Time point (TR) 1 corresponds to test item onset. Shaded bars indicate the TRs during which accuracy was significantly above chance (.33; correcting for multiple comparisons).

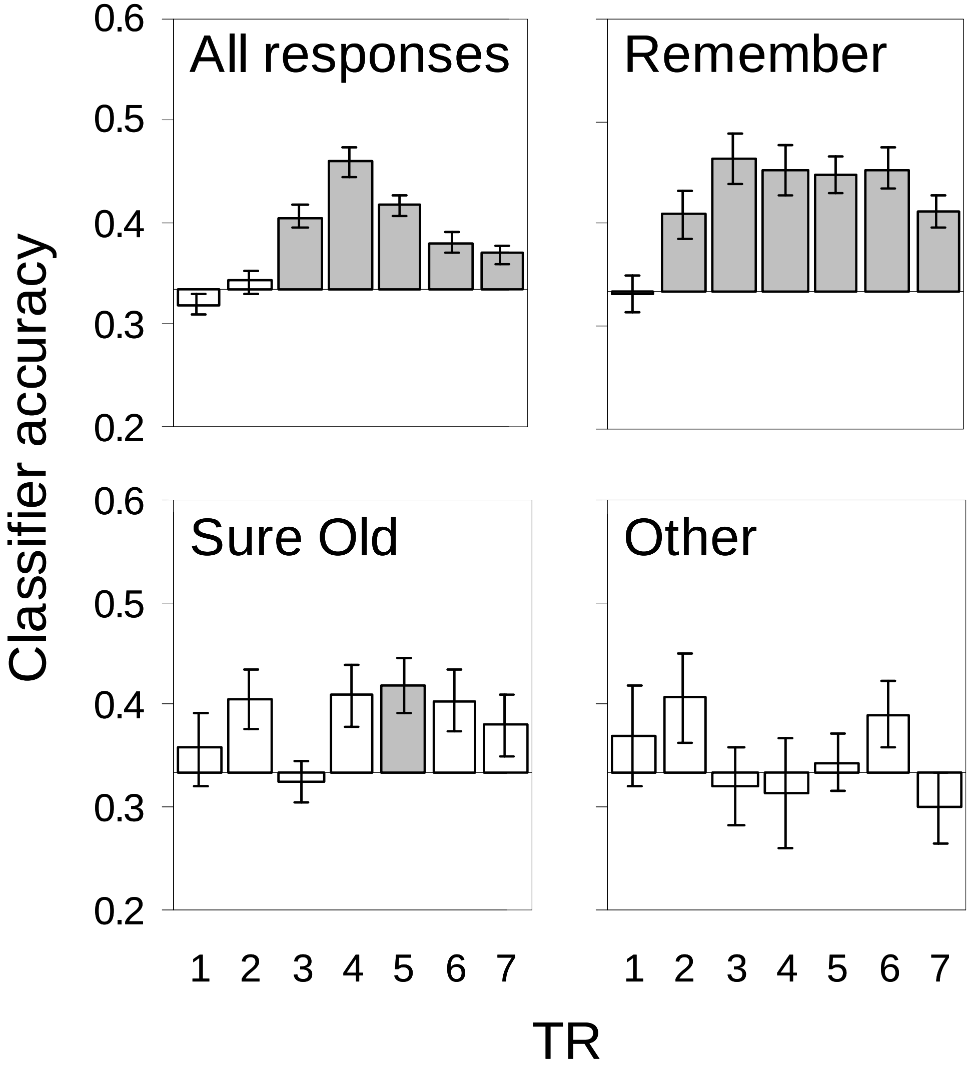


Figure S2. Results of GLM analyses showing task-related differences during the study phase. The effects are thresholded at p < .036 (with no voxel extent restriction) in order to subsequently mask with the analogous retrieval-related effects. See Supplemental Material text for details of the masking procedure. All effects are overlaid on the mean normalized anatomical data.

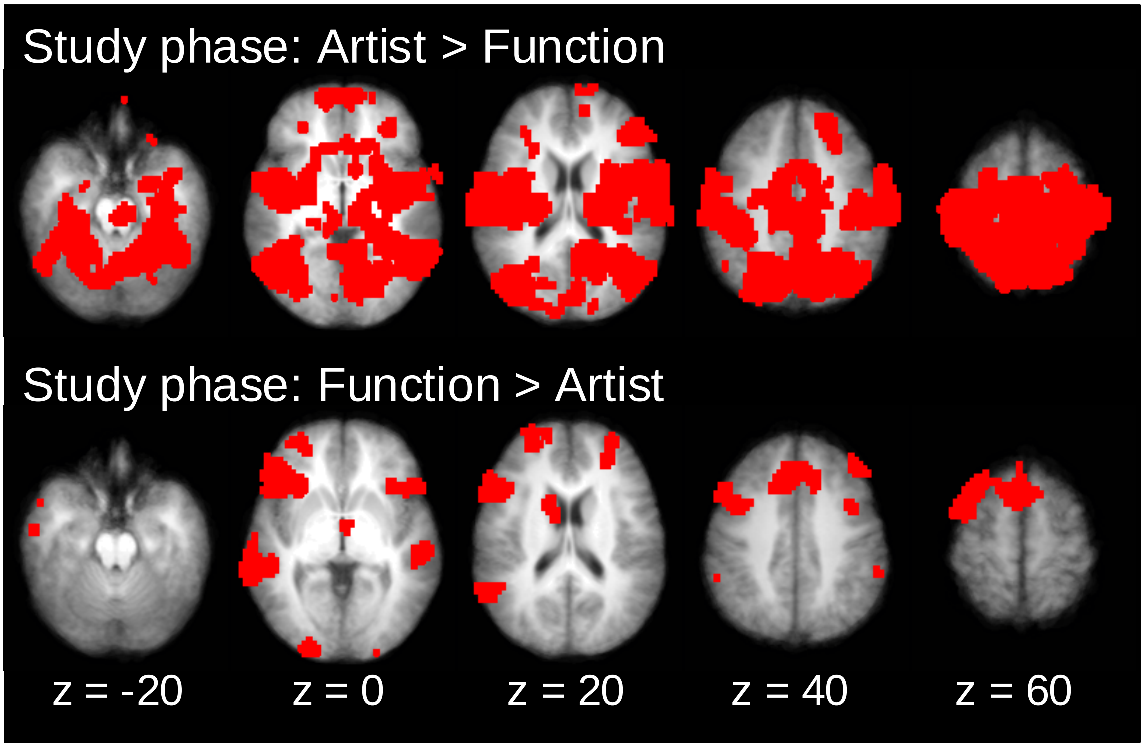

Study phase: Artist > Function
Study phase: Function > Artist
z = -20
z = 0
z = 20
z = 40
z = 60

Figure S3. Results of GLM analyses showing reinstatement for Remember responses. The effect depicted here (692 voxels, peak Talairach coordinates: 9 -30 54, peak Z-score: 3.68) involved greater activity for items from the Artist compared to Function tasks. The effect survived a cluster-wise corrected threshold of p < .05 and is overlaid on the mean normalized anatomical data. See Supplemental Material text for details of masking procedure and individual thresholds.

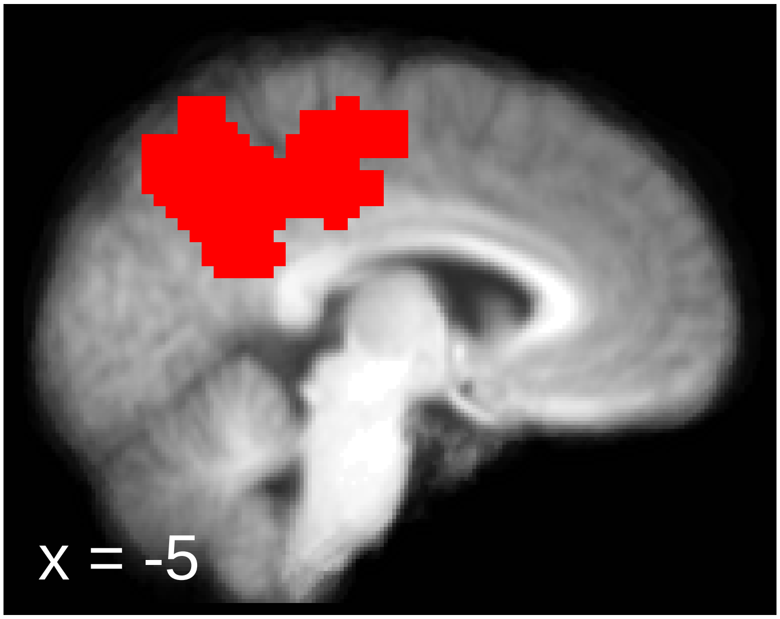


Figure S4. Results of GLM analyses showing a task-related effect (Artist > Function) for Remember responses. The effects were evident in left lateral parietal (507 voxels, peak coordinates: -45 -60 54, peak Z-score: 5.00), medial parietal (219 voxels, peak coordinates: -3 -66 48, peak Z-score: 3.80), medial frontal (189 voxels, peak coordinates: -9 54 45, peak Z-score: 3.10), and right lateral parietal (184 voxels, peak coordinates: 39 -45 45, peak Z-score: 2.99). All effects survived a cluster-wise corrected threshold of p < .05 and are overlaid on the mean normalized anatomical data. Coordinates are in Talairach space. L = left.

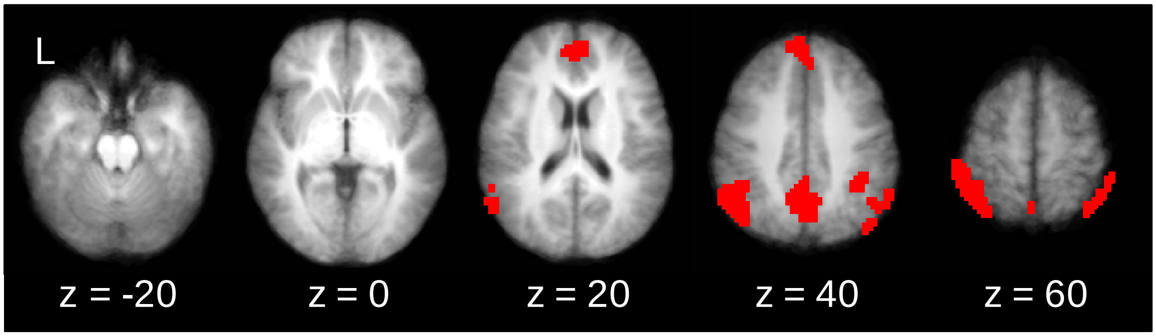


Figure S5. Results of GLM analyses showing greater activity for Remember compared to Sure Old responses (collapsed over the Artist and Function tasks). The effect depicted here (24 voxels, peak Talairach coordinates: 33 -36 -6, peak Z-score: 4.26) was localized to the right hippocampus and survived a cluster-wise corrected threshold of p < .05.

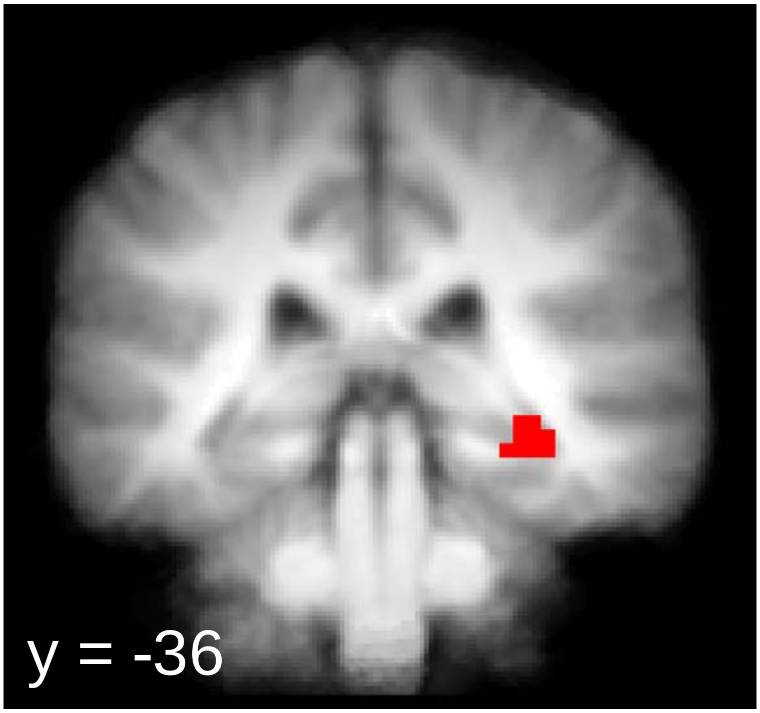
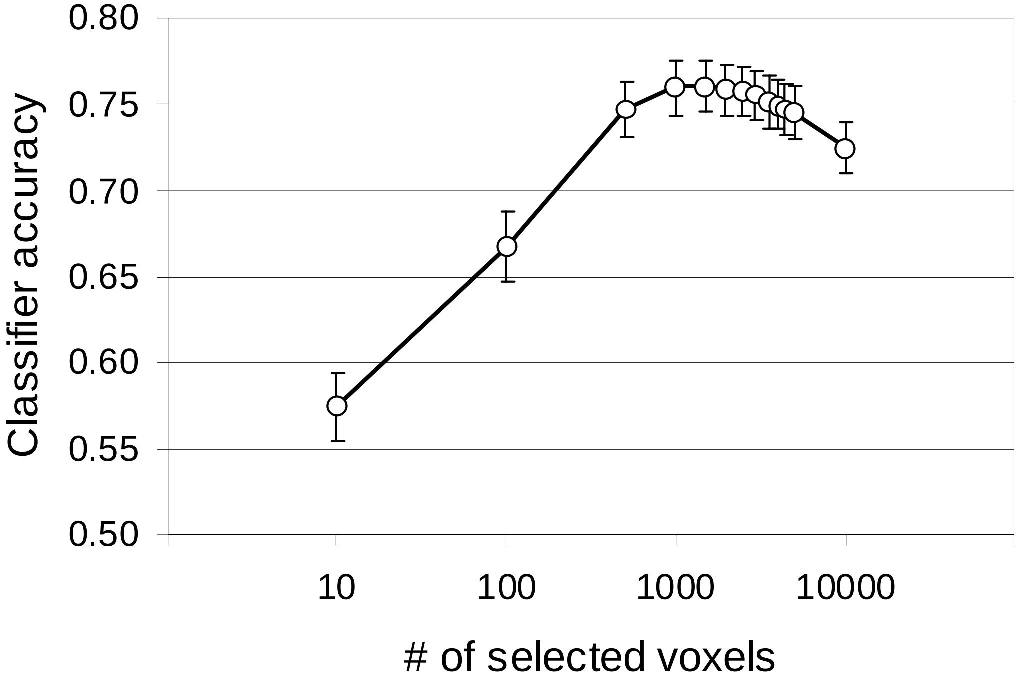


Figure S6. Mean classifier accuracy (±SEM) as a function of how many voxels were used for the classifier based on only study phase data. Accuracy peaked at around 1000 voxels, but was significantly above chance (.33) at all levels.


Classifier accuracy
# of selected voxels