NeuroImage 132 (2016) 11–23

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Reverse inference of memory retrieval processes underlying metacognitive monitoring of learning using multivariate pattern analysis

Peter Stiers [a],*, Luciana Falbo [a], Alexandros Goulas [a], Tamara van Gog [b], Anique de Bruin [c]

[a] Department of Neuropsychology and Psychopharmacology, Maastricht University, Maastricht, The Netherlands
[b] Department of Educational Psychology, Erasmus University Rotterdam, The Netherlands
[c] Department of Educational Research & Development, Maastricht University, The Netherlands

ARTICLE INFO

Article history:
Received 15 March 2015
Accepted 7 February 2016
Available online 13 February 2016

Keywords:
Working memory
Judgement of learning
Metacognition
Functional MRI
Educational psychology

ABSTRACT

Monitoring of learning is only accurate at some time after learning. It is thought that immediate monitoring is based on working memory, whereas later monitoring requires re-activation of stored items, yielding accurate judgements. Such interpretations are difficult to test because they require reverse inference, which presupposes specificity of brain activity for the hidden cognitive processes. We investigated whether multivariate pattern classification can provide this specificity. We used a word recall task to create single trial examples of immediate and long term retrieval and trained a learning algorithm to discriminate them. Next, participants performed a similar task involving monitoring instead of recall. The recall-trained classifier recognized the retrieval patterns underlying immediate and long term monitoring and classified delayed monitoring examples as long-term retrieval. This result demonstrates the feasibility of decoding cognitive processes, instead of their content.

Introduction

A key aspiration of psychology is to understand complex human behaviour in terms of its constituent psychological processes. The metacognitive ability to monitor one's own state of learning, for example, is thought to be an essential aspect of academic learning (Koriat and Goldsmith, 1996; Thiede and Dunlosky, 1999). It is assumed to involve, a try-out re-activation of the learned material, informing the student about the acquisition status (Nelson and Dunlosky, 1991; Rhodes and Tauber, 2011; Thiede and Anderson, 2003). This assumption is supported, on the one hand, by a loss in relative accuracy of monitoring when performed immediately after the learning (Nelson and Dunlosky, 1991; Rhodes and Tauber, 2011; Thiede et al., 2003; Thiede et al., 2005). In that case, the judgement is thought to be based on information still active from the encoding, independent of the quality of storage in long term memory. On the other hand, restricting the opportunity for retrieval eliminates the higher relative accuracy of delayed monitoring (Dunlosky and Nelson, 1992).

While these findings lend support to the re-activation theory of the delayed judgement of learning effect (Rhodes and Tauber, 2011) the hypothesized processes are only indirectly open to empirical observation, through behavioural measures. Neuroimaging, however, holds the promise of making the hidden processes more directly identifiable through their neurophysiological markers. Long term memory retrieval, for instance, has been associated with increased activity in the hippocampus and in the lateral and medial parietal cortex (Cabeza et al., 2012; Daselaar et al., 2009; Huijbers et al., 2012; Kirwan and Stark, 2004; Okada et al., 2012; Vannini et al., 2011). Likewise, working memory retrieval has been associated with stronger activity in ventral lateral prefrontal cortex, anterior superior frontal gyrus and lateral temporal cortex (Nee and Jonides, 2011, 2013; Oztekin et al., 2009). Such an activation pattern may be used as a biological marker for the underlying process, allowing to conclude that, for instance, memory retrieval took place every time the characteristic pattern is observed. This type of deduction, known as “reverse inference” (Aguirre, 2003), holds the promise of an alternative route to the dissection of complex behaviour.

A first challenge for reverse inference of long-term memory retrieval during monitoring of learning is the superimposed processes related to metacognitive monitoring and the judging response, which are likely to distort the overall brain activation pattern. Metacognition has, for instance, been shown to elicit specific activation in subregions of medial and orbital prefrontal cortex (Chua et al., 2009; Do Lam et al., 2012; Kao et al., 2005). Hence, an effective reverse inference procedure must be able to distinguish activity specific of the target processes from everything else.

This brings us to a second and more fundamental challenge. Several authors have criticized reverse inference on more principle grounds. This critique focuses on the notion of process-specific activity (Aguirre, 2003; Christoff and Owen, 2006; D'Esposito et al., 1998; Poldrack, 2006, 2011; Fox and Friston, 2012). They argue that a

* Corresponding author at: Maastricht University, Faculty of Psychology and Neuroscience, Department of Neuropsychology and Psychopharmacology, P.O. Box 616, 6200 MD Maastricht, The Netherlands.
E-mail address: peter.stiers@maastrichtuniversity.nl (P. Stiers).

http://dx.doi.org/10.1016/j.neuroimage.2016.02.008
1053-8119/© 2016 Elsevier Inc. All rights reserved.

consistent activity increase in particular brain structures during execution of a particular cognitive process does not tell us whether the structures are indicative of or selectively engaged by the processes under study. As an illustration of this, the well-established attribution of long term memory storage to the hippocampus was criticized recently by demonstrations of activation in this structure also during working memory retrieval (Nee et al., 2008; Nee and Jonides, 2011; Postle, 2006; Ranganath and Blumenfeld, 2005). Likewise, the traditional attribution of working memory processes to the ventral–lateral prefrontal cortex (e.g., Fuster, 1989; Fuster and Alexander, 1971; Goldman-Rakic, 1987; Miller et al., 1996; Ptito et al., 1995), has been challenged by studies showing that activity in this part of prefrontal cortex reflects attention control mechanisms that are not specific to working memory (D'Esposito et al., 1998; D'Esposito and Postle, 1999; Nee et al., 2008; Passingham et al., 2000; Postle, 2006). It should be clear from this that lacking knowledge of the specificity of neural structures or brain activation patterns for particular cognitive processes imposes limits to reverse inference. Practically, the selectivity and specificity of activation patterns for a particular process can be investigated using available large-scale brain activation databases (e.g., BrainMap.org, NeuroSynth.org). Such databases allow to estimate the probability of activation given the execution of tasks thought to activate the process and the execution of tasks that should not activate the process (e.g., Chang et al., 2013; Poldrack, 2006).

While databases of published data provide a practical by-pass for our limited knowledge, they also point towards a third problem of reverse inference, recently addressed by Hutzler (2014). Ideally, the functional signature for a particular cognitive process should allow inferring the involvement of the process in any context. In reality, however, its validity is restricted to the contexts used to establish the characteristic signature, and validity beyond these contexts needs to be established empirically. Hutzler (2014) showed that this limitation can be turned into an advantage. By explicitly taking the context of the task under study into account, justified inferences can be made about processes taking place within this specific context. This was investigated for the (left) fusiform face area, which is known to activate during both face recognition and reading tasks. However, when the studied task involves visually presented words (e.g., a reading task without pictures of faces) it is safe to infer that activity in this area marks processing of word images. Consequently, a quantitative data-base driven meta-analysis of experiments using visual-verbal tasks can be sufficient to yield the voxels that are uniquely associated with the process in this type of tasks. The advantage here is that specific reverse inference questions about processes underlying particular task paradigms can get a quantitative answer, without the requirement to answer the most general question of the unique functional signature of the processes under all possible contexts.

In the present paper we follow this line of restriction to address the problem of the hypothesized long-term retrieval process underlying the typical delayed judgement of learning task paradigm. However, we do not rely on meta-analysis of already existing data to delineate voxel activation patterns with significant predictive power of reverse inference. Instead, we make use of the typical task paradigm to collect new data and use multivariate pattern classification as our method to find the indicative activation pattern. It was Poldrack (2011) who suggested that multivariate pattern classification could provide a formal means to implement reverse inference, because these methods quantitatively estimate the degree to which a pattern of brain activation is predictive of the engagement of a specific cognitive process. These methods use brain activation maps derived under two (or more) prototypical conditions as examples to train a statistical machine learning algorithm to find the optimal pattern to distinguish the example classes. The trained classifier is subsequently used to make predictions about the activation patterns in a new set of similar examples (O'Toole et al., 2007; Pereira et al., 2009). These techniques have been used to predict which stimulus classes participants were viewing, or imagining (Haxby et al., 2001; Kamitani and Tong, 2005; Lewis-Peacock and Postle, 2008; Lewis-Peacock et al., 2012). The techniques were also successful in predicting more cognitive aspects of behaviour, such as the intention to perform one or the other task (Haynes et al., 2007), the stimulus–response mapping rules in a task (Woolgar et al., 2011), and which of a set of tasks was being performed (Poldrack et al., 2009; Stiers et al., 2010).

To support reverse inference of cognitive processes, however, the training examples need to reflect as closely as possible the theoretically relevant difference: cognitive processes, rather than cognitive contents (stimulus classes, response classes, task rules, etc.). Under these circumstances, the training set constitutes an ostensive definition of the brain functioning patterns that distinguish the two cognitive processes. The multivariate pattern analysis translates this defining set into a high-dimensional statistical pattern, which can be applied to brain activation examples generated during tasks where the underlying processes are unknown. Thus, in the multivariate pattern classification approach, instead of relying on large-scale data bases (Poldrack, 2006, 2011) or task-specific meta-analyses (Hutzler, 2014), new data are collected that are specific to the process and paradigm of interest and the critical alternative processes, and the predictive patterns are generated from these data. The selectivity and specificity of the pattern for inferring the process of interest can then be computed from the classification accuracies in the reference task.

The first aim of this study was to investigate the feasibility of multivariate pattern classification for reverse inferring. We adopted a single participant approach (Formisano et al., 2008), because people may use different strategies to learn and evaluate their state of learning, and consequently manifest idiosyncratic activation patterns. Randomization statistics were applied to establish above chance classification performance. To make sure that results are not subject dependent we repeated the analysis independently in five different individual data sets. We tested the feasibility of reverse inference in two steps. In the first experiment the decoding of immediate and long term retrieval was validated using an overt cued recall task for word pairs. The aim was to show that the classification procedure had sufficient selectivity and specificity to correctly infer the known retrieval processes underlying overt recall. In the second experiment reverse inference was critical put to test by having participants perform judgements of learning of word pairs, instead of overt recalling them. Reverse inference would be established if judgements of items stored in long term memory prior to the task are recognized as long term memory retrievals, while identical judgements made immediately after encoding of the content are not.

The second aim of our study was to investigate the long term memory retrieval interpretation of the delayed judgement of learning effect. Conditional to the confirmation that multivariate pattern analysis allows reverse inference of long-term memory retrieval underlying monitoring of long term learning, the re-activation hypothesis would predict that the classifier, trained on immediate and long term overt recall activation patterns, would also recognize in the delayed monitoring trials the long term memory retrieval pattern.

Materials and methods

Participants

Six healthy right-handed volunteers (2 males, mean age 26.94 (3.69) years) took part in the study after giving their written informed consent. The study was approved by the local Ethical Committee. Participants were recruited from the university community and screened for psychological and medical problems, right-handedness and absence of contra-indications for exposure to magnetic field. Due to technical failure data recorded from one participant during the monitoring task were lost. Hence, for this task data from only 5 subjects were available for analysis.

Task description

Participants performed two tasks requiring learning and either recalling or monitoring word pairs (Fig. 1). Each task consisted of an alternating sequence of encoding trials and either recall (memory task) or judgement of learning (monitoring task) trials. Stimuli were concrete Dutch nouns, 1–5 syllables long, that were pseudo-randomly paired, avoiding categorical association between words of a pair. Words were presented in black, in the middle of a grey screen. Presentation of word pairs (encoding) or cue words (recall/monitoring) was scanner triggered and lasted 6 s, followed by 2–10 s of fixation screen to optimize deconvolution of trial type specific hemodynamic responses. The recall and the monitoring task, while administered in different runs, differed only in the response that participants had to make, which was either to verbalize the second word of a pair, or to verbalize the likelihood (6-point scale) that it would be recalled later. The monitoring scale ranged from "absolutely sure I will not recall this item" (score 1) to "absolutely sure I will recall this item" (score 6).

In both tasks the time lag between the encoding of a word pair and the subsequent recall or judgement of learning response was manipulated in order to create three conditions. In immediate trials the word pair to recall or monitor was from the immediately preceding encoding trial. In delayed trials there were four encodings and four recalls/judgements (or from 80 to 102 s) between the presentation of the cue word and the actual encoding of the word pair. Lastly, in the long term condition the word pairs were memorized minimally 24 h before the session, tested before the scanning session to assure correct recall, and retrieved or judged in the scanner.

A task comprised 3 runs, each with 39 encoding trials and 13 recall/monitoring trials per time lag type (immediate, delayed, long term). Since long-term trials were learned outside the scanner, "dummy" encoding trials had to be included. These word pairs were not cued for recall or monitoring. For the monitoring task, after each run correct memory formation of the word pairs judged was tested with a cued recall paradigm, in the scanner, but without scanning. This test also includes dummy encoding word pairs.

Prior training

Participants were trained on speaking with minimal head movement in a dummy scanner while listening to recorded EPI sound. Under conditions of loud noise participants tend to increase the intensity of their voice by speaking louder and for longer, causing major and prolonged movements (Tourville et al., 2008), which affect the quality of MRI measurements during vocalization. In order to limit artefacts due to overt speech production, participants were trained to speak with minimal facial muscular activity. Post-hoc analysis of actual head movement measures during the acquisition of the data showed no significant difference in volume-to-volume displacement during the first two volumes of non-speech (encoding) and (recall/monitoring) speech trials.

Participants were also trained in using visual strategies to encode word pairs. Interactive imagery has shown to be a valid technique for word pair tasks (Dunlosky and Hertzog, 2001). They practised in making up an image or scene involving the two words during the 6 s of presentation and to rest afterwards

Overt verbal responses in the scanner

We used a cued recall task, which is commonly applied in judgement of learning studies, instead of a recognition memory task that is more common in fMRI studies, because the former avoids the confounding of fully endogenous retrieval and familiarity and therefore provides a more straightforward operationalization of the cognitive retrieval process assumed to underlie metacognitive monitoring of learning. To avoid differences in response modality in the two tasks, participants were requested to vocalize the probability of recalling the word in the monitoring task as well. The verbal responses during scanning were recorded for off-line analysis with a fibre optic microphone (FOM 1190 OPTIMIC™) using the GoldWave v5.05 software. The resulting audio files (4410 Hz sampling, 38 dB, 16 bit stereo) were processed in mono format with the Scanner Noise Cancellation Tool (Cusack et al., 2005). The cleaned audiofiles were subsequently used for evaluating the responses and establishing response reaction times.

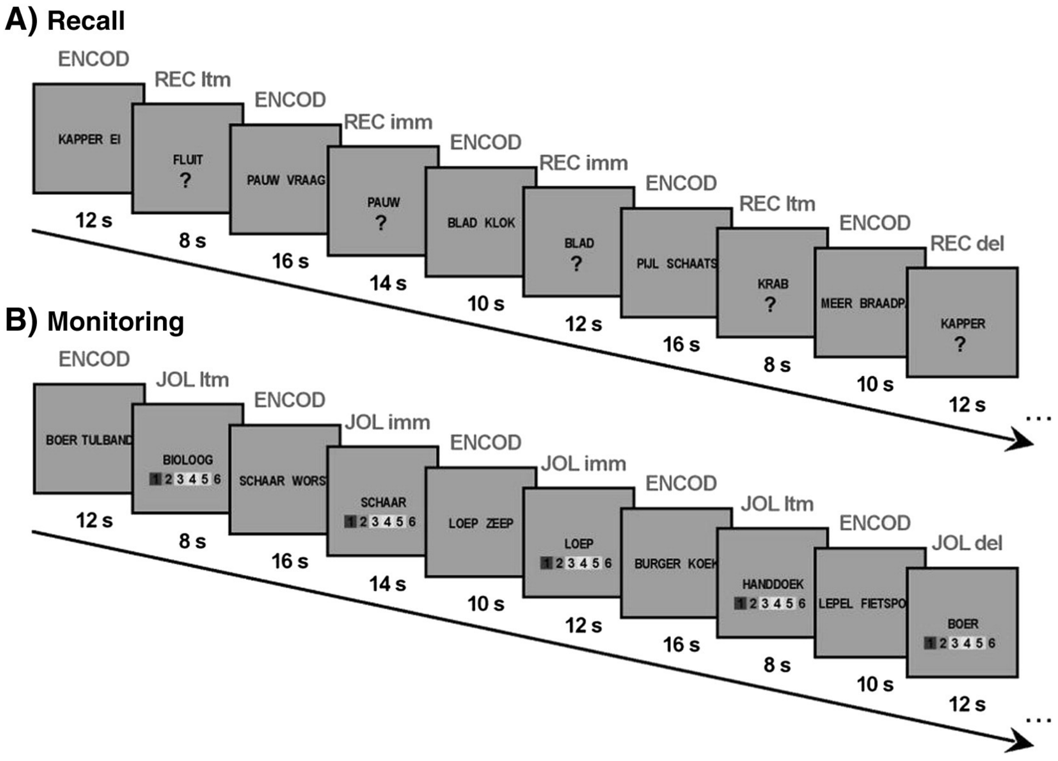

Fig. 1. Illustration of the overt recall (A) and monitoring of learning (B) tasks. See text for further details.

Image acquisition

Data were collected on a Siemens MAGNETOM Allegra 3 T MRI head-only scanner. Head motion was constrained by the use of foam padding. A total number of 32 axial slices covering the whole brain including the cerebellum were acquired by using a T2*-weighted gradient echo planner pulse sequence (TR = 2000 ms, TE = 30 ms, FA = 90°, FOV = 224, slice thickness = 4 mm, matrix size = 64 × 64, flip angle = 90°). Voxel size was 3.5 × 3.5 × 4 mm. A gradient echo image (TR = 704 ms, TE 5.11 and 7.57 m; flip angle = 60°) with the same grid and slice orientation as the functional images was acquired to generate a field map for correcting susceptibility-related distortions in the functional images. A T1-weighted anatomical scan was also acquired (TR = 2250 ms, TE = 2.6 ms, flip angle = 9°, FOV = 256 mm, slice thickness = 1 mm, matrix size = 256 × 256, number of slices = 192). Voxel size was 1 × 1 × 1 mm.

Preprocessing

The data were pre-processed using the SPM5 software (Wellcome Department of Cognitive Neurology, University College, London, UK). Functional volumes were re-aligned, spatially corrected using the field map, slice time corrected, and co-registered with the anatomical scan. The individual T1-weighted anatomical scan was segmented into tissue density maps (grey matter, white matter and CSF) and non-linearly normalized to the MNI template in an integrated procedure. The resulting normalization parameters were applied to the functional data of that participant. In the process, the functional volumes were re-sampled to 2 mm isotropic voxels. Lastly the functional volumes were smoothed with a 6 mm FWHM Gaussian kernel.

Univariate analysis of fMRI task effects

Mass univariate general linear model analyses were performed on the functional data from each participant and each task separately, to produce task condition specific percent signal change maps, and on the pooled data from the different participants to obtain group results. In all these analyses the conditions of interest were the encoding phase and the three different time lag trial types during the recall/monitoring. Condition events were modelled with the theoretical hemodynamic response function and its time and dispersion derivative. Recall trials in which a participant did not answer correctly were modelled in the same way as a fifth condition. To reduce the impact of head motion the affine realignment parameters were added as regressors of no interest. In addition, trials whose induced BOLD signal overlapped with abrupt head movement were modelled as events of no interest (Keulers et al., 2011). Abrupt head motion was defined as an absolute difference between successive volumes in the z-direction of more than 1/10th of the slice thickness (i.e., 0.4 mm), or rotation around the x dimension exceeding 0.26° (i.e., 0.4 mm z-displacement of frontopolar voxels for a rotation around a point midway on the anterior–posterior dimension (88 mm from the most anterior point in Talairach and Tournoux (1988)). Only 0.74% of volume pairs met this criterion for the z-translation and none for the x-rotation. Taking into account the hemodynamic response delay, trials with an onset from 1 to 12 s before the onset of a volume with excessive head motion were modelled as events of no interest. To reduce the effect of signal drifts and of physiological fluctuations, the average signals for white matter and for CSF were included as two additional regressors of no interest. Lastly, a session specific mean regressor was added to the design matrix. Linear trends, temporal autocorrelations and low frequency signal fluctuations were handled with whitening and high-pass filtering, as standard implemented in the SPM5 software.

First level single participant analyses integrated data collected during three runs with the same task. Each run was modelled as described above. After voxel-wise estimation of the weights associated with each of the components in the model, contrasts were defined to estimate the strength of the BOLD response in each trial type relative to baseline. The resulting weight maps were used to compute whole brain percent signal change maps, which were used in the follow-up region of interest based group level analysis described below.

A whole brain group analysis integrated data collected during three runs of the same task from each of the participants. This fixed effects approach was chosen because the sample size was too small to yield sufficient statistical power in a random effects analysis. Due to the fixed effects approach the conclusions drawn from this group analysis have to be limited to the sample of participants studied and cannot be generalized to the population from which the sample was drawn. However, this does not affect the conclusions that can be drawn from the study, since the purpose of the group analyses was merely to provide a frame of reference for interpreting the multivariate pattern analysis results, which are single-subject results. After estimating the beta weights associated with each of the variables modelled, contrasts of each condition of interest relative to baseline were computed, as well as specific contrasts between these conditions. All these contrasts ignored the time and dispersion derivative regressors associated with the different conditions. The resulting statistical *t*-maps were evaluated at the single voxel significance level of 0.05, corrected for multiple comparisons based on Gaussian field theory (i.e., family-wise error correction) (Friston et al., 1991; Worsley et al., 1992).

Group-level region of interest (ROI) analyses were performed on the participant and condition specific percent signal change data that resulted from the first level analyses described above. These analyses were conducted as post-hoc follow-up analyses intended to reveal the specific nature of the BOLD signal modulations at selected activation sites. ROIs were defined by local maxima in the statistical maps created in the integrated group-level analyses, that were described above. A position of a ROI was defined by the MNI coordinates of the local maximum and comprised all the voxels within a 6 mm range of these coordinates. Eleven ROIs in the left hemisphere were examined. For each participant the average percent signal change was computed for the 20% voxels within the ROI with the strongest percent signal change (positive or negative) in a condition. This voxel selection procedure leads to an optimization of the ROI to the specific activations of each participant. The analysis was focused on identifying ROIs whose power for discriminating long term from immediate overt recall trials did not replicate in the monitoring task, which statistically translates into a significant interaction between task and trial type. Despite the small number of participants, a random effects analysis revealed six regions for which this interaction was significant. We therefore choose to report these random effects results.

Multivariate pattern analysis (MVP): recall task

Single trial BOLD response examples were obtained from normalized and smoothed single subject data as the average signal in volumes 3–4 (4–8 s) after cue onset of correct recall trials. The average number of correct recall trials was 38.2 (range 36–39) for the immediate condition, 34.4 (range 27–39) for the delayed condition and 36.8 (range 35–38) for the long term condition. Response strength values were voxel-wise rescaled to the 0–1 range. Rescaling parameters were computed on all training examples and applied to training and test examples, setting values exceeding the range to the range limits. Including only training examples for scaling parameter estimation ensured that the training results were independent from the specific test examples used.

Voxels were selected as features for the MVP analysis that had a grey matter tissue density in the individual's grey matter map of 0.5 or higher. Furthermore, voxels were excluded from MVP analysis that were not part of the GLM analysis specific brain mask, which identifies voxels with a clear T2* brain signal (SPM5 default is 0.8 times the mean intensity of voxel with signal exceeding 1/8 of total volume mean).

The spider linear support vector machine (SVM) algorithm (version 1.71; www.kyb.tuebingen.mpg.de/bs/people/spider) was used for MVP analyses. To compare the classification performance for delayed recall trials to that of long term memory and immediate recall trials we first created run-wise folds of the data, and cross validated by training on two runs and testing on the remaining run. This run-wise cross-validation is more conservative than trial-wise cross-validation. For each fold a separate recursive feature elimination analysis with an initial feature reduction step was performed. The initial feature reduction step comprised of selecting the 5000 voxels that discriminated most between the training classes. This selection was based on the unsigned t-statistic computed for each feature separately. Importantly, this selection step only included trials belonging to the training set at the particular fold. The training set within a fold was further split 10 times, each time leaving out another 1/10th of the available trials to avoid overfitting. The SVM classifier was trained on each split and the average absolute discriminative weights (|w|) of the features over the ten splits were computed. These averaged weights were used in the recursive feature elimination step to discard the 10% least discriminating features. The recursive feature elimination step was performed fifty times, each time discarding 10% of the features. At each iteration the average accuracy (and classifier certainty) over the 10 splits of the training set were computed. As said above, this entire procedure was repeated three times (once for each fold) changing the training and testing runs. Lastly, for each recursive feature elimination level the classification accuracy (and classifier certainty) averaged over the three folds at that level was obtained. The best recursive feature elimination level is the level with the highest average classification accuracy.

When testing the generalizability of training to discriminate immediate from long term retrieval examples to the delayed retrieval trials, only delayed trials were included in the test set that occurred in the run not used for training. This ensured the same level of independency between training and test data for both the reference categories (immediate and long term retrieval) and the target category (delayed retrieval). In the standard MVP generalization approach, the best iteration in the RFE approach is selected based on the accuracy obtained on the test data. Consequently, when testing on the delayed retrieval target examples, the iteration is chosen that gives the highest classification of these target trials. However, to maximize the rigour of testing the reverse inference hypothesis, we also report classification accuracies for target examples at the iteration that was most optimal for discriminating the reference examples, i.e., the iteration that gave the highest accuracy in classifying immediate versus long term retrieval examples.

Statistical significance was based on 250 re-analyses of the data following the exact same steps as described above (initial reduction, RFE and best iteration selection), but with random redistribution of the training labels. The label re-assignment mimics the null-hypothesis that there is no systematic association between feature values and classes, so that the class labels are interchangeable. Random accuracies were on average higher than 50.0% (47.0 to 56.2%) due to selection of the best of all iteration. The individual significance level was 0.05. The significance of individual participant results was the centile position of the observed accuracy amongst the 250 accuracies obtained with the repeated randomization procedure. Significance at the group level was established with the cumulative binomial coefficient, to take into account the accumulation of chance.

Parameter dependency of classification performance

Because there are many parameters to select in the entire MVP procedure, the influence of several of these parameter settings on the result was investigated. With respect to the initial univariate feature reduction step we investigated the effect of the number of features selected (All, 1000, 5000 or 2500 voxels) and the score used for the selection. De Martino et al. (2008) obtained the best results in a simple auditory discrimination tasks with a score highlighting the most active voxels within each of the training classes. Although we did not expect in our more complex and effortful cognitive task that the most active voxels would also be the most important voxels to distinguish between the two types of memory retrieval, we nonetheless investigated the effect of using this selection score. A second potentially influential factor is the algorithm used. While most of the analysis were done with the spider implementation of the SVM because of its speed, we verified the results obtained also with the least-squares SVM (Suykens et al., 2002; De Martino et al., 2008) and the Keerthi implementation (Keerthi and DeCoste, 2005). All three algorithms used linear kernels. Thirdly, we investigated the dependency of the results on the number of voxels retained at each recursive feature elimination level. Lastly, we examined the influence of the method used to obtain single trial response estimates. The following voxel-wise BOLD response strength estimations were examined: signal strength in the third scan (4–6 s) after the onset of the recall cue, the mean signal over the third and fourth scan (4–8 s) following the cue onset, response modelling with a single trial single regressor general linear model (GLM) estimation (referred to as the 'glm' method), and lastly with a single trial multiple regressors GLM estimation (referred to as the 'epoch' method). For the GLM methods the single trial analysis window was from −2 s to +10 s relative to the cue onset time. In both cases the parameters of the hemodynamic response function were varied recursively and the beta weight of best fitting result was retained as the estimate of the response strength. The parameters varied were the peak amplitude of the function (4 levels), the onset delay of the peak wave (4 levels) and its dispersion (4 levels). The difference between the 'glm' method and the 'epoch' method is that in the first the estimated model existed of only this parameterized hemodynamic response functions, whereas in the 'epoch' method all the other regressors that pertain to this single trial time window (except the ones that model the condition of interest) were borrowed from the first level GLM design matrix (for details see section Univariate analysis of fMRI task effects).

Time-varying multivariate pattern analysis

To further explore the quality and distribution over time of patterns that discriminate immediate from delayed and long term recall we performed additional MVP analyses at systematic time points relative to trial onset. The training set in this analysis always consisted of maps of the mean BOLD response strength over the 3rd and 4th scans relative to recall trial onset (4–8 s), for the immediate recall trials (class 1) and the long term recall trials (class 2). As test examples we used BOLD-signal strength measures in each trial at −1 to +5 TRs relative to stimulus onset. The resulting single volume single trial BOLD strength maps for all three conditions of interest (immediate, delayed and long term recall) were used as test examples for the MVP analysis in which the classifier was trained on all training examples (no training examples left out for training). All other parameters were as in the basic analysis described above (initial feature reduction to 5000 voxels, RFE with 50 iterations discarding 10% of the features, 10 splits on the training set). This analysis allowed investigating at what point in peri-stimulus time the distinguishing pattern emerged. The classifier's class estimate for individual test examples, rescaled to 0–1 range, was used as classifier evidence in favour of long term retrieval.

Voxel importance map

A voxel importance map was created by simply adding the weight maps of individual participants. MVP Parameters were as in the basic leave-one-run-out analysis. For each subject the weight map at the 12th iteration was used, because the highest accuracy in distinguishing between immediate and long term recall examples averaged across all participants was obtain at that iteration. At this iteration, after initial feature reduction to 5000 voxels, 1566 voxels were still included in the analysis of each participant. Of course, these voxels could be different for different participants. Hence, the importance attributed to voxels can reflect both the weight of the voxel in an individual participant's analysis and the number of participants in which the voxel contributed to the classification at that stage in the analysis.

Classification generalization over participants

MVP analysis was performed on examples derived from all participants but one, which was left out for testing. To allow sampling of voxels across participants, new examples were created that included the same voxels in all participants. The voxel selection mask used to create examples was derived from the SPM5 grey matter prior by thresholding it at a density > 0.33. Examples quantified the mean signal in scans 3–4 after the recall cue onset. One fold of the data consisted of all immediate and long term recall trials of five out of six participants. Testing data were examples from immediate and long term recall trails of the sixth participant. This process was repeated until data from each participant were used once as test data set. The MVP analysis at each fold consisted of an initial feature reduction to 5000 voxels, after which training with the spider SVM was iteratively performed 50 times, eliminating the 10% least weighted voxels. At each fold 10 splits of the training set were created and training was repeated 10 times, discarding every time a different split from the training set. The accuracy at a particular iteration of a particular fold was the average accuracy across the 10 splits at that iteration. The final accuracy reported was the maximum accuracy obtained in the 50 iterations, after completing and averaging results over all folds with a different participant left out as test set. The statistical significance was established in 250 randomizations of the training labels with an otherwise identical procedure (all leave-one-subject-out folds, and within fold splits and iterations), as described above.

Multivariate pattern generalization from recall to monitoring task

The same single subject multivariate pattern analysis procedures as described for the recall task were also followed for the monitoring task. However, data from monitoring trials were used only as test examples. Hence, training was always performed on immediate recall trials as class 1 and long term recall trials as class 2, and testing was performed on the immediate, delayed and long term monitoring trials. In the basic analysis leave-one-run-out cross-validation was performed, with the recall examples for training derived from two runs and the monitoring examples coming from the third run. Examples were created as the mean signal in the 3rd and 4th volumes after trial onset. In the follow-up analysis also monitoring examples based on the mean of the 5th and 6th volumes were used. For the time-varying classification, the training was on the same recall examples derived from the 3rd and 4th volumes, but no leave-out procedure was used. Hence, training was performed on all immediate and long term recall examples that were available for a particular participant, keeping the 10 splits, to avoid over-fitting of the data. The test examples were voxel-wise signal strengths measured in individual volumes recorded during the monitoring task, from −1 to +5 TRs relative to monitoring trial onset.

Repeated measures statistical analyses

The statistical significance of differences between the various task conditions across participants, for behavioural measures, fMRI percent signal change data, and classifier evidence, was evaluated with repeated measures Analysis of Variance (ANOVA). The small number of participants in our study did not yield enough degrees of freedom to use multivariate analysis of variance for multilevel within subject designs, which avoids the sphericity assumption. To correct for possible violations of the sphericity assumption the Greenhouse–Geisser correction was applied when estimating significance. To further explore the nature of significant effects revealed by the ANOVA results, pair-wise comparisons were made using paired samples *t*-tests with Tukey's studentized range correction for multiple comparisons (Howell, 2002). The significance level was set to 0.05. All reported p-values are two-tailed unless specified otherwise.

Visual presentation of whole brain results

Figures present data from the left hemisphere. Similar but less extensive results were observed in the right hemisphere. Brain maps are visualized in Caret v5.61 overlaid on the PALS-B12 template.

Results

Memory retrieval underlying delayed recall

Behavioural performance

Consistent with the theory that in immediate recall trials the second word is still readily available in working memory, the error rate in immediate recall trials (0.9 ± 1.3%) was significantly lower than in delayed (14.0 ± 12.3%) and long term recall trials (8.1 ± 6.6%; $F(1.4,6.9) = 5.6$, $p = 0.044$). Similarly, reaction times in correct recalled trials were faster in immediate (1.4 ± 0.6 s), than in delayed (2.4 ± 0.9 s) and long term recall trials (2.6 ± 1.1 s) ($F(1.2,4.8) = 23.1$, $p = 0.005$).

Whole brain univariate analysis

The results of the univariate group analysis of the fMRI data are summarized in Fig. 2. Overall, the extent of activation and deactivation was very similar for the encoding and recall phase (Fig. 2A,B), with some noticeable differences. Recall specific activity was observed in the lower motor-premotor regions and the audition-related superior temporal cortex, possibly reflecting the verbal response requirement in the recall, but not the encoding phase. In contrast, encoding induced more activation in intraparietal sulcus and lateral occipital–temporal cortex, presumably due to higher visual-attentional demand following our instruction to use visual encoding strategies.

The immediate and long term recall conditions were contrasted directly to investigate the univariate activity pattern that maximally distinguishes between working memory and long term memory retrieval (Fig. 2C). The working memory specific activation pattern included a large inferior parietal cluster (red in Fig. 2C; 1583 voxels left and 910 right), that only partially overlapped with typical task-deactivation regions in angular gyrus (Mars et al., 2011; Stiers et al., 2010) and extended anterior into task-activation cortex of the supramarginal gyrus (Mars et al., 2011). Smaller clusters were found in the task-deactivated cortex of superior temporal sulcus, dorsal and anterior medial prefrontal cortex, and anterior and posterior middle frontal gyrus. The opposite contrast revealed significant preference for long term recall in a large cluster (blue colour in Fig. 2C; 4493 voxels bilaterally) over posterior medial and retrosplenial cortex, in addition to several smaller clusters in task-activation regions. These results show that the task was effective in engaging the brain and that unique patterns of neural activity were associated with each of the two cognitive processes of interest.

Multivariate pattern classification

To investigate the feasibility of decoding the type of retrieval process underlying delayed recall, a support vector machine learning algorithm was trained on single trial brain activity maps from correctly recalled immediate and long term trials (i.e., the reference examples), and tested on similar maps from immediate and long term recall trials (i.e., the target examples) belonging to a different run of the same participant. In order to delineate the most optimal voxels for classification a recursive feature elimination (RFE) procedure was used (De Martino et al., 2008), which repeated the analysis 50 times each time eliminating the 10% least contributing voxels. The highest accuracy over all iterations for classifying the target examples is reported. In addition, we also report the accuracy for classifying target examples obtained at the iteration

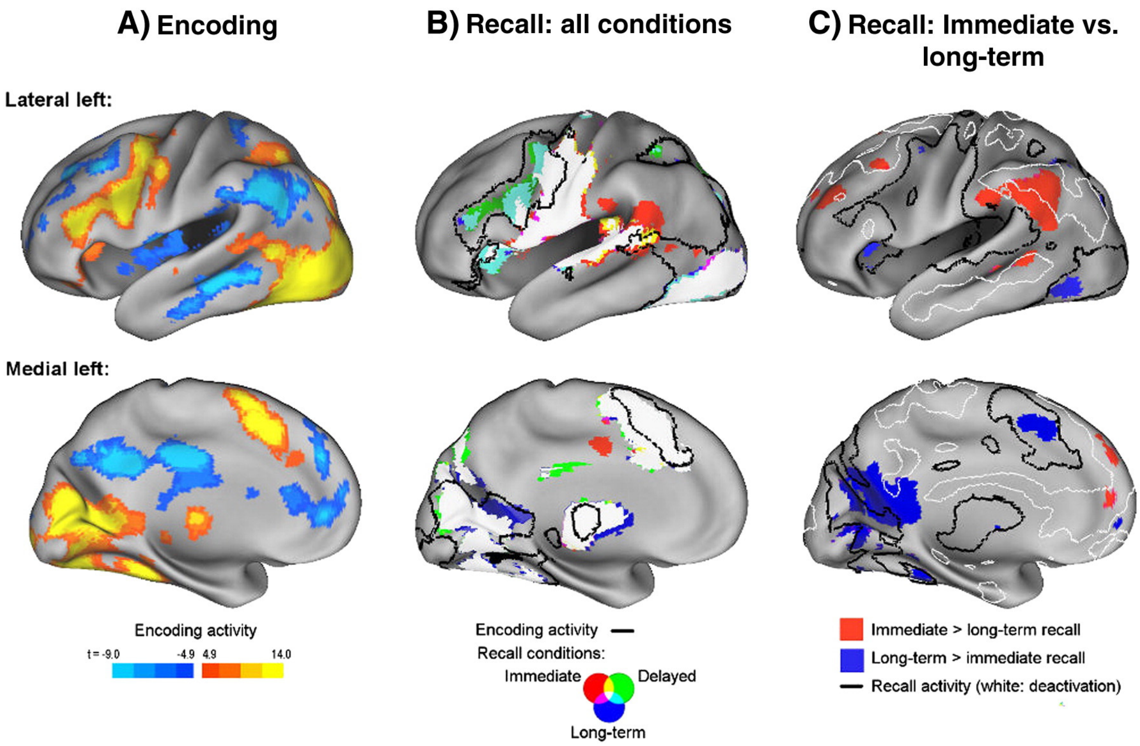
A) Encoding
B) Recall: all conditions
C) Recall: Immediate vs. long-term
Lateral left:
Medial left:
Encoding activity
t = -9.0
-4.9
4.9
14.0
Encoding activity
Recall conditions:
Immediate
Delayed
Long-term
🟥 Immediate > long-term recall
🟦 Long-term > immediate recall
Recall activity (white: deactivation)

Fig. 2. Univariate brain activation patterns during the word pair overt recall task. A) Encoding versus null statistical t-map, thresholded at $t \geq 4.98$ ($p = 0.05$, corrected). B) Immediate, delayed, and long term recall activation (versus null; *t*-maps thresholded as above). Black lines delineate encoding activation. C) Immediate versus long-term recall *t*-maps thresholded as above (red: immediate > long term; blue: inverse).

that was best for classifying the reference examples. The classification of long-term recall trials was significantly above chance level in all six participants (mean highest accuracy 77.4 ± 5.0%, $p < 0.0001$; mean accuracy at iteration with highest reference classification accuracy 73.4 ± 4.4%, $p < .00001$), whereas immediate recall trials were identified above chance in four participants at the iteration most optimal for target classification (76.9 ± 10.6%, $p < 0.0001$) and in all six participants at the iteration most optimal for classifying reference examples (74.7 ± 4.6%, $p < 0.0001$) (Fig. 3 and see Inline Supplementary Table S1). To further demonstrate the selectivity as well as the specificity of this classification result, the receiver operating characteristic (ROC) curves were computed for individual participants (Fig. 3D). This analysis shows that there was a strong precedence of hits (true positives) over false positives, with 77 to 100% of hits at the point where the first false positive appeared. The average area under the curve (0.995 ± 0.006) was significantly higher than in a random classification — i.e., at the time of stimulus presentation, when no process information is yet available (0.443 ± 0.053, $t(5) = 27.8$, $p < .0001$).

With respect to the delayed judgement of learning effect, in all 6 participants the delayed recall trials (from a different run than the training examples) were significantly classified as long term memory recall trials (75.9 ± 5.1%, $p < 0.0001$). Testing delayed recall trials at the optimal RFE iteration for classifying reference trials reduced the group average accuracy to 71.5 ± 6.4%, but classification was still above chance in each of the 6 participants ($p < 0.0001$). In fact, the accuracy was stable over a wide range of RFE iterations (Fig. 3B), indicating that the critical information is not confined to a limited set of voxels. The results were also not critically dependent on other analysis parameters, such as the algorithm, the amount or score used for initial feature reduction, or the method for single trial response assessment (see Inline Supplementary Figure S1). These result warrant the conclusion that activation patterns invoked by delayed overt recall trials match the characteristic activation pattern of long term retrieval.

Time-varying classification

As a further control analysis we investigated whether ongoing neural activity directly preceding the recall phase contributed to recall classification. This could for instance be the case if participants engage in continued working memory maintenance, which is directly beneficial to the recall task in immediate trials, but irrelevant to performance in delayed and long term recall trials. New classification tests with the classifier trained on the reference examples described above but tested on examples generated at each time point from −2 to +10 s relative to recall cue onset revealed that the discriminative pattern emerged only from the 3rd volume after cue onset onward (all post-hoc pairwise p-values < 0.001) (Fig. 3C). Similarly, when the classifier was trained and tested on each volume relative to recall cue onset, there was no discernable difference between immediate and long term recall trials just before (−2 to 0 s; $p = 1.0$) or after (0 to 2 s; $p > 0.25$) the cue (see Inline Supplementary Figure S2).

Classifier evidence in relation to behavioural performance

Several studies found that higher classifier accuracy during recall (Rissman et al., 2010) or encoding (Kuhl et al., 2012; Xue et al., 2010) is associated with better recall performance. This relation is assumed to reflect the strength of (re-)activation of the memory contents. In the present study, aimed at predicting the recall process rather than its content, this relationship is not expected. Rather, we would predict, if anything, that the classifier evidence reflects the effort and length of retrieval, which is higher in difficult to retrieve items. In line with this, the classifier long term retrieval evidence for not recalled delayed/long term word pairs (0.77 ± 0.38: number of trials = 7.6 ± 5.3) was somewhat, but not significantly, higher than for correctly recalled pairs (0.51 ± 0.18; n trials = 69.2 ± 7.7; $t(4) = 1.2$, $p = 0.29$). This is even more noteworthy since we only used successful recall trials for training. Moreover, for correct recalled items shorter reaction times

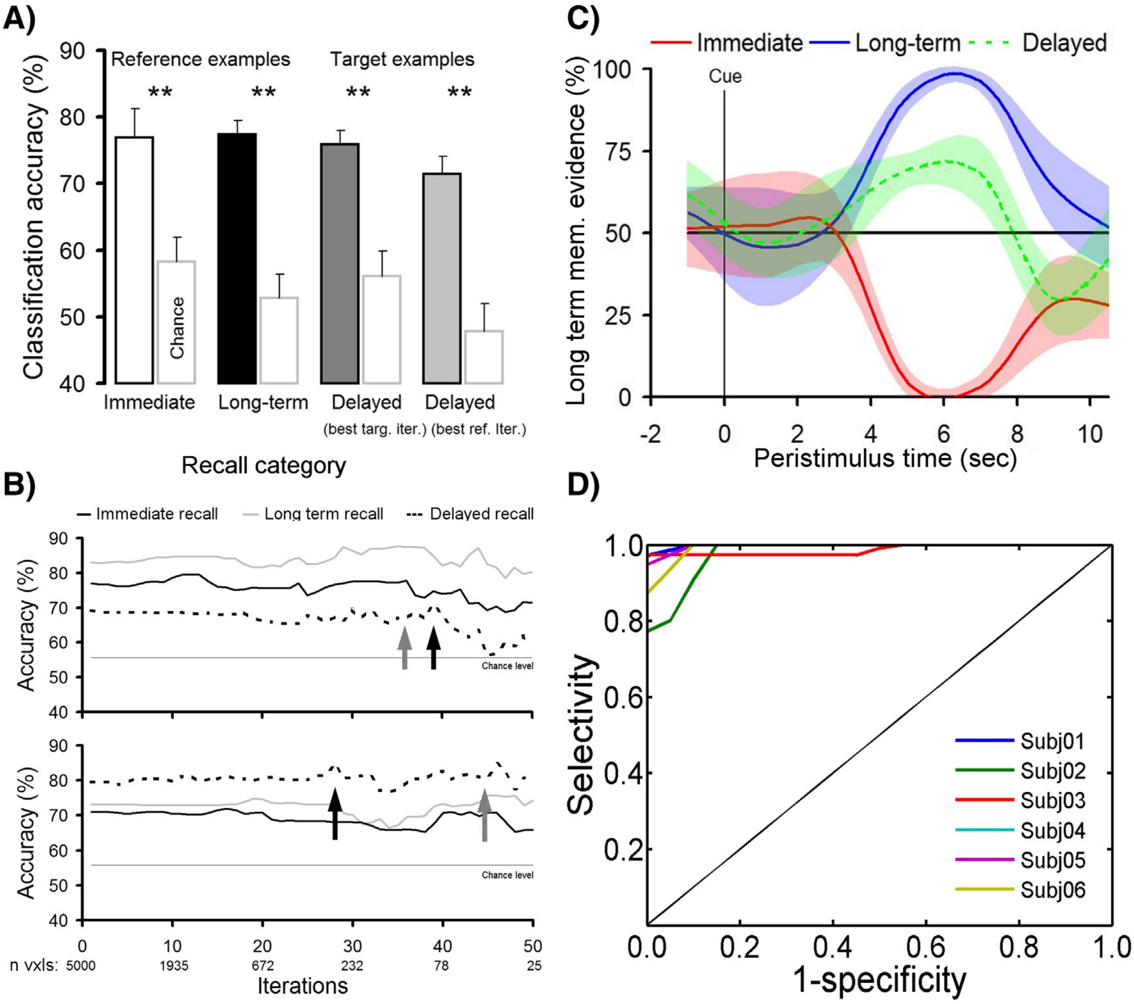

Fig. 3. Overt recall classification accuracy for recall time lag conditions. A) Best test accuracies per condition, compared to chance classification. Chance level is based on 250 randomizations of the training labels (see Inline Supplementary Table S1). Light grey bar represents accuracy for classifying delayed trials at the iteration optimal for classifying the reference examples. B) Accuracy as a function of recursive feature elimination iteration for two representative participants. Black arrows mark best iteration for classifying delayed recall trials (dark grey bar in A); Grey arrows indicate delayed recall accuracy at best iteration for classifying reference examples (light grey bar in A). See Figure S1 for effect of other parameter settings on classification result. C) Emergence of discriminative activation pattern over time. The classifier was trained on averaged signal at TRs 3 and 4, and tested on signal strength at successive TRs throughout peristimulus time (−1 TR to +5 TR from recall cue). Filled areas represent ±1 SD. D) Individual ROC curves for identifying long-term versus immediate recall trials. Selectivity is the proportion of hits; 1-specificity is the proportion of false positives.

were associated with lower classifier evidence, in delayed (mean $r = 0.283 \pm 0.111$; $t\ (4) = 5.7$, $p = 0.005$) and long term recall trials (mean $r = 0.218 \pm 0.172$; $t\ (4) = 2.8$, $p = 0.047$). This correlation was not different from zero in the immediate recall condition (mean $r = 0.169 \pm 0.222$, $t\ (4) = 1.6$, $p = 0.165$).

Intersubject consistency of discriminative patterns

To get an impression of the consistency of the discriminative patterns across participants, the weight maps obtained for each participant at RFE iteration 12, which yielded the highest mean accuracy over participants, were averaged (Fig. 4 left). The map emphasizes voxels with a stronger contribution in more participants. Contributing voxels were located throughout the brain and did not single out specific structures as specifically important. Also, voxels in the cerebellum and subcortical structures (colliculus superior, striatum, thalamus, hippocampus, and amygdala) contributed to the classification in at least some participants.

As a more objective measure of spatial consistency, we performed intersubject classification using leave-one-participant-out cross validation with examples from immediate and long term recall trials, and initial feature reduction to 5000 voxels, in combination with recursive feature elimination. The best accuracy of 65.3% was significantly above chance (52.9 ± 2.9%, $p = 0.0039$). Fig. 4 (right) shows the voxel weights at iteration 12 (mean accuracy = 64.7%, $p = 0.0040$).

Classifying judgement of learning trials

To reversely infer that memory retrieval processes underlie monitoring of learning it needs to be shown that the successful classification of overt recall trials generalizes to brain activity maps obtained during the monitoring task. The same retrieval processes are assumed to be executed in the monitoring task, but now covertly and embedded in a complex metacognitive task that invokes additional cognitive processes (e.g., monitoring, scaling of internal information, dosed responding, reading the response scale). This task was in all respects similar to the overt recall task, including identical time lag conditions, but with the cued recall response replaced by a judgement of learning response (Fig. 1B). Instead of overtly retrieving the second word, participants had to judge verbally on a 6-point scale their confidence that they would recall the second word at a later test. This task was performed by five participants.[1]

[1] Due to technical failure data recorded from one of the six participants during the monitoring task was lost.

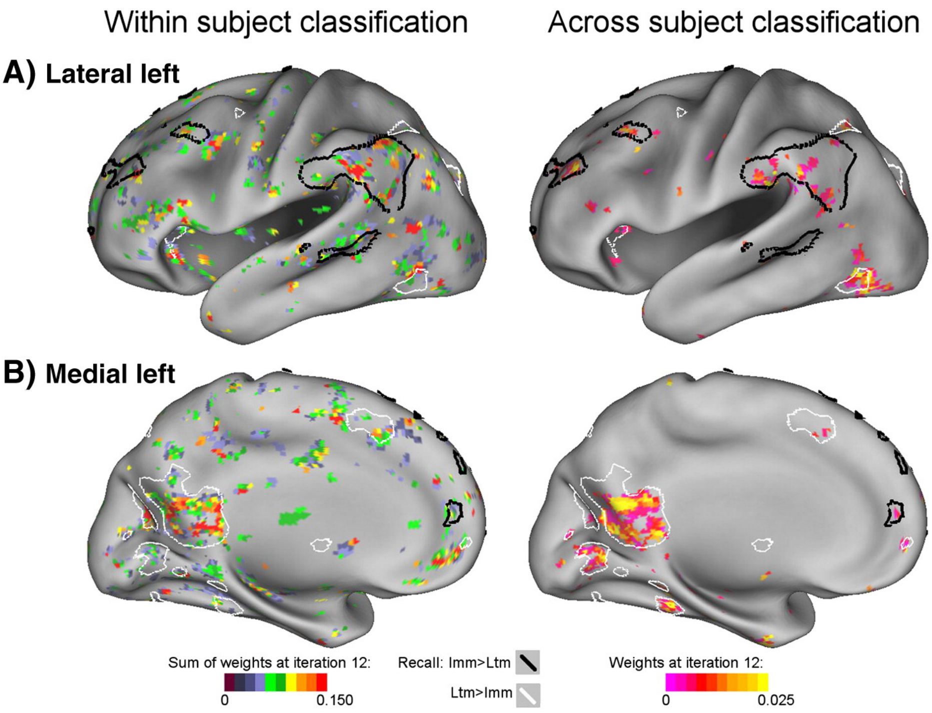
Within subject classification
Across subject classification
A) Lateral left
B) Medial left
Sum of weights at iteration 12:
0
0.150
Recall: Imm>Ltm
Ltm>Imm
Weights at iteration 12:
0
0.025

Fig. 4. Voxel importance maps for classifying immediate versus long-term recall trials. Left. Voxel weights summed over participants at RFE iteration 12 of within subject classification with leave-one-run out cross validation. Outlines mark voxel clusters with significantly higher univariate BOLD response during immediate than long-term recall (black lines) and the inverse (white lines) contrast (Fig. 2C). Right. Voxel weights for between subject classification at the 12th RFE iteration.

Behavioural performance

The response time for immediate monitoring trials (2.2 ± 0.4 s) did not differ from long term (2.5 ± 0.6 s) or delayed monitoring trials (2.7 ± 0.7 s; $F(1.5, 5.9) = 4.9$, $p = 0.062$), but the time delay of monitoring did influence the confidence scores of participants that they would later recall the word pair (test recall x judgement time lag interaction: $F\,(2,8) = 13.0$, $p = 0.009$). While confidence scores for immediate monitoring did not differ for correct (4.2 ± 0.9) and incorrect (4.7 ± 1.3) recalled word pairs ($p > 1.0$), confidence scores for delayed/long term trials were significantly higher if the word was recalled at the later test (5.1 ± 0.6 vs. 3.8 ± 1.0), $p < 0.05$). This confirms that the relative accuracy of monitoring, commonly expressed as the gamma correlation[2] between monitoring scores and (in)correct recall, is higher at a delay after encoding (Nelson and Dunlosky, 1991; Thiede et al., 2003).

Whole brain univariate analysis

The univariate group analysis provided little unequivocal evidence that the activation patterns characteristic of immediate and long-term memory retrieval patterns re-emerged during the monitoring task (Fig. 5B). Searching for voxels that were more active during immediate monitoring of word pair learning yielded two clusters, neither of which overlapped with the immediate recall preferring clusters of the overt cued recall task. In the opposite direction, clusters showing significantly higher signal during long term than immediate monitoring only partially overlapped with the long term recall preferring clusters during the overt recall task. In fact, only 17.4% of all the voxels that differentiated between immediate and long term trials either in the overt recall or in the monitoring task, were shared by the two tasks. To further detail the difference in univariate activation between the monitoring and the overt recall task, we investigated the response characteristics of clusters showing immediate or long-term preference in the overt recall but not in the monitoring task in a ROI-based group analysis (clusters out-lined in Fig. 5B). Five clusters preferring immediate recall (numbered dots in black out-lines in Fig. 5B) and one cluster preferring long-term over immediate recall (dot number 5 in Fig. 5B), showed a significant response profile difference over the three time delay conditions during monitoring compared to recall (time lag by task interaction, all ps < 0.05). In all these ROIs this was because they were equally engaged during the three types of monitoring, but differently engaged during immediate compared to delayed and long-term overt recall (Fig. 5C).

Classification performance

In contrast to what the univariate results suggested, the classifier could recognized delayed monitoring trials significantly above chance as long term retrieval examples in all five participants examined (Fig. 6A and see Inline Supplementary Table S2). Mean classification accuracy was 79.1 ± 4.4% ($p < 0.0001$) at the optimal iteration for classifying delayed monitoring and 74.1 ± 6.8% ($p < 0.0001$) at the optimal iteration for decoding immediate and long term overt recall trials. Similarly, mean accuracy for long term monitoring trials was 81.0 ± 9.6% (significant in 4 out of 5 participants, $p < 0.0001$). In contrast with this, the accuracy for classifying immediate monitoring trials (51.9 ± 11.1%) did not differ from chance (see Inline Supplementary Table S2). Similar accuracies were observed at the optimal iteration for decoding immediate and long term overt recall trials: long term monitoring 79.1 ± 7.5% (significant in 5 of 5 participants, $p < 0.0001$); immediate monitoring 47.9 ± 10.9% (ns). Despite the reduced performance for immediate monitoring trials, the ROC curves were still significantly different from chance (Fig. 6D) as the average area under the curves (0.728 ± 0.054) was significantly higher than for classification at the time of trial start — i.e., when no process information was yet available (0.488 ± 0.072, $t\,(4) = 6.1$, $p = 0.002$).

An indication of why immediate monitoring trials were not recognized as immediate recall events was provided by follow-up analyses in which the trained classifier was tested on examples created at different time points relative to trial onset. As Fig. 6B shows, immediate

[2] Unfortunately, we could not use the gamma correlation in our data due to the very low number of recall errors (only 3.2 ± 1.3 errors for delayed and 2.6 ± 3.2 for long-term items), which makes the non-parametric gamma coefficient highly unreliable (Siegel & Castellan, 1988).

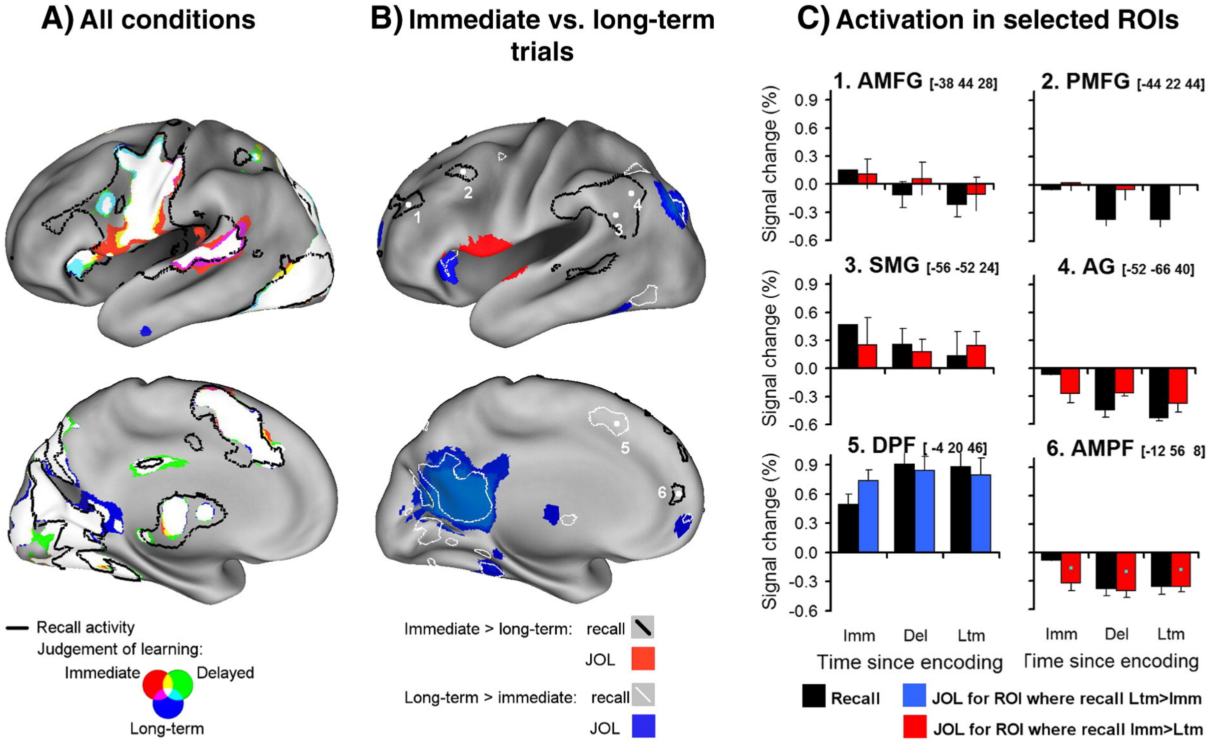

Fig. 5. Univariate brain activation patterns during monitoring of learning. A) Immediate, delayed, and long term monitoring activation (versus null; t-maps thresholded at $t \geq 4.98$, p = 0.05, corrected). Black lines delineate activation during similar conditions in the overt recall task (see Fig. 2B). B) Immediate versus long-term monitoring (t-maps thresholded as above; red: immediate > long term; blue: inverse contrast). Black/white lines mark activation/deactivation for the same contrasts in the recall task. C) Activity modulation in regions showing an immediate vs long term activity difference in the recall task (Fig. 2C), that also showed a significant task (overt recall vs monitoring) by time (immediate, delayed, long term) interaction in a follow-up region-based group analysis. Red bars: Immediate > long term monitoring; blue bars: opposite contrast; black bars refer to the recall task. Imm = immediate; Del = delayed; Ltm = long-term; JOL = judgement of learning; AMFG = anterior medial frontal gyrus; PMFG = posterior medial frontal gyrus; AG = angular gyrus; SMG = supramarginal gyrus; DPF = dorsal medial prefrontal; AMPF = anterior medial prefrontal.

monitoring trials were identified, but at a later time point than long term and delayed trials. To confirm this the classifier, trained on overt recall trials at 4–8 s after trial onset, was tested on immediate monitoring examples reflecting signal strength at 8–12 s after trial onset (see Inline Supplementary Table S3). The classifier could identify these immediate monitoring examples as immediate recall trials significantly above chance level in three of the five participants (p = 0.00116). Moreover, when comparing the timing of the accuracy peak for classifying monitoring versus recall trials of the same encoding history type, we found a significant delay in peak time only for the immediate trials (t(4) = 6.5, p = 0.009, Bonferroni corrected) (Fig. 6C). The variability across participants in decoding immediate monitoring trials was considerable, however (see Inline Supplementary Tables S2 and S3). In one participant immediate monitoring trials were significantly classified in the 4–8 s time window. In two participants the classification reached significance only at the 8–12 s window, while in the last two participants it never rose above chance level. It seems therefore, that in the last four participants the second word was not retrieved in order to make the judgement of learning.

Discussion

Several previous studies demonstrated the potential of multivariate pattern analysis for decoding the contents of cognitive states in the sensory (Haxby et al., 2001; Kamitani and Tong, 2005; Lewis-Peacock and Postle, 2008) and the cognitive domains (Haynes et al., 2007; Poldrack et al., 2009; Woolgar et al., 2011). In these demonstrations, content specific activation in lower tier sensory and integrative association cortices are important drivers of the classification. Here we showed that cognitive processes themselves, instead of content-dependent categories, can also be decoded from single trial brain activation patterns. This was evident from the above chance identification of the long term retrieval pattern even when participants were not overtly recalling the target words, but instead engaged in a metacognitive monitoring of their state of learning these words. Moreover, the additional demonstration that delayed recall and delayed monitoring trials were significantly above chance recognized as long term retrieval examples confirms the re-activation hypothesis regarding the delayed judgement of learning effect. These results open a new perspective on “mind reading” (Norman et al., 2006) in that they show the feasibility of decoding not just the content to which a person orients his/her attention, but also the cognitive process s/he is employing implicitly to accomplish a certain task.

Multivariate pattern analysis and reverse inference

The immediate and long term overt recall and monitoring trials served as the external criterion for evaluating the validity of reverse inference from multivariate pattern analysis. The significant above chance classification of these trials convincingly demonstrates the feasibility of reverse inferring long-term memory retrieval within the task paradigm used. This result is obtained despite limited knowledge of how memory retrieval processes are neurobiologically implemented. In the multivariate approach the knowledge problem is by-passed by starting from a set of carefully created examples that constitute an ostensive definition of the processes of interest. After extracting the most distinctive features during the training of the classifier, a direct test examines how well new examples of brain activation patterns fit the definition. This provides a direct test of the validity of the extracted pattern for differentiating the target processes within the context of the task paradigm.

Multivariate pattern analysis warrants reverse inference because it goes a long way in meeting the two proposals made by Poldrack

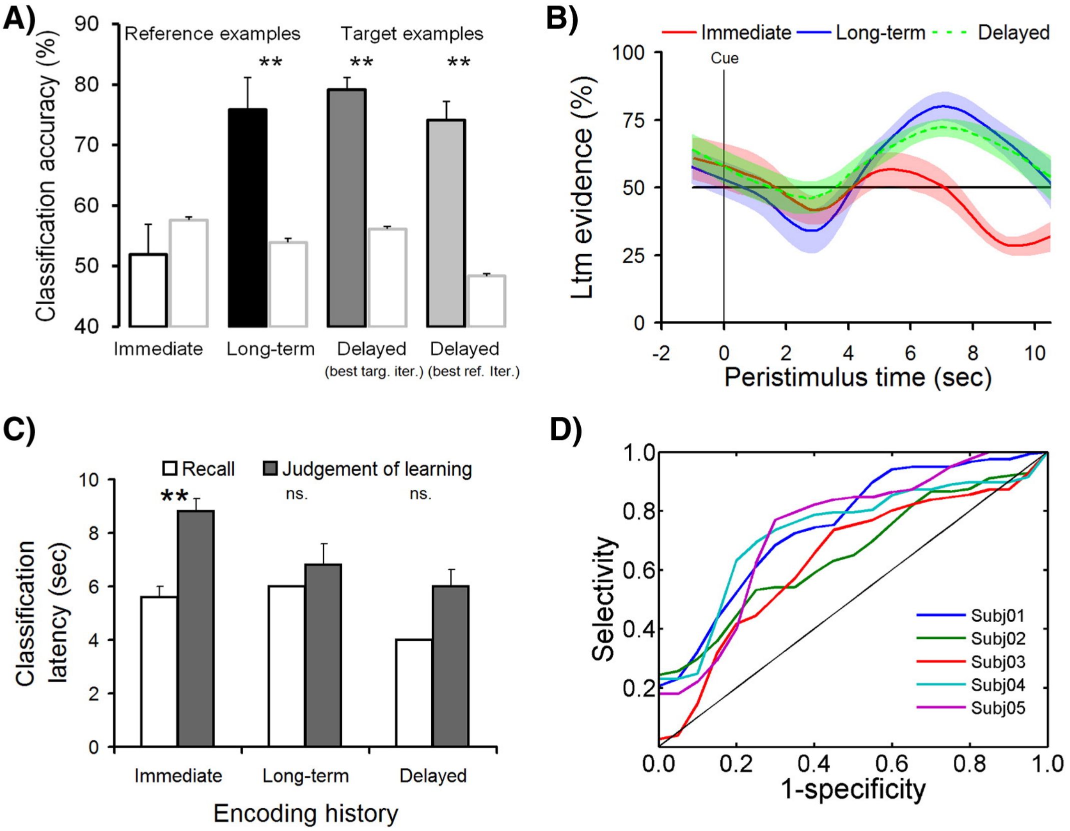

Fig. 6. Classification accuracy for monitoring trials after training on immediate and long term recall trials. A) Best test accuracies per monitoring condition, compared to chance level (250 randomizations of the training labels) (see Inline Supplementary Tables S2 and S3). Light grey bar represents accuracy for classifying delayed trials at the iteration optimal for classifying the reference examples. B) Discriminative activation pattern over time. The classifier, trained on mean signal at TRs 3–4 in the recall task, was tested on signal strength at successive TRs throughout peristimulus time (−1 TR to +5 TR from monitoring cue). Filled areas represent ±1 SD. C) Classifier evidence peaked significantly later in the monitoring compared to the recall task for immediate trials, but not for delayed and long term trials. D) Individual ROC curves for identifying long-term versus immediate monitoring trials; Selectivity is the proportion of hits; 1-specificity is the proportion of false positives.

(2006) to improve confidence in reverse inference. The first is to increase the selectivity of the activation pattern for the cognitive process of interest. Focusing on sets of smaller regions instead of a single large region is likely to improve selectivity, because specific processes most probably engage specific functional networks and subloops within them (Stiers et al., 2010). It is clear that this suggestion is maximally met in multivariate pattern analysis. The training identifies an optimal number of the smallest spatial units with a selective response towards the processes of interest, without the need for a priori region selection based on theoretical assumptions. Poldrack's second suggestion was to increase the prior probability that the cognitive process of interest is actually engaged by the cognitive task studied — i.e., if we are certain that the task engages the cognitive process, the prior probability is 1.0. This is under control of the experimenter. We designed the overt recall task used here as reference task in such a way that the immediate and long term trials only differed in the (hidden) neural source from where the target word was retrieved. As a consequence, we zoomed in on the processes of retrieval regardless of their content or outcome (being successful or not) and confirmed their operation in a thorough validation study.

An alternative interpretation of our results might be that the classification doesn't reflect retrieval specific differences, but merely a generalized difference between immediate and long term trials, such as for instance higher effort in the latter. Neither the univariate results, nor the multivariate results agree with this interpretation, however. In univariate analyses task-general effects are associated with a typical distribution of activations and de-activations in the attention/salience and default mode network, respectively (see for instance Fig. 2A). The distribution of voxels that differentiate between immediate and long term recall trials shows only a limited amount of overlap with this typical activity pattern (Fig. 2C). This suggests that retrieval specific effects are being observed. Moreover, some differentiating regions showed effects in the direction opposite to the direction of the presumed generalized effect (for instance, cluster 3 in Fig. 5B–C). In the multivariate results, on the other hand, voxels contributing to the classification fell in large part outside of the regions showing a univariate difference between immediate and long term recall trials (Fig. 4, left panel). Moreover, classification accuracies were high at all iterations of recursive feature elimination (Fig. 3B) and even without univariate feature reduction (Figure S1A), suggesting that the result is not critically dependent on any specific subset of regions. This was confirmed by a follow-up MVPA analysis in which all voxels showing a significant univariate effect ($p < 0.001$, uncorrected for multiple comparisons) had been eliminated from the feature matrix. This analysis yielded classification accuracies nearly identical to those in the original analysis (see Figure S3).

While multivariate pattern analysis was effective in decoding cognitive processes, the interpretation of the spatial map of contributing voxels is less straightforward than in a traditional univariate analysis (Jimura and Poldrack, 2012; Poldrack, 2011). First of all, far more voxels contribute to the classification than would be expected from a univariate approach. This is due to the higher sensitivity of the multivariate approach (e.g., De Martino et al., 2008; Haxby et al., 2001), but in our study it most likely also reflects individual differences in strategy. Secondly, the voxels identified as contributing to a classification are not all the voxels that carry relevant information. This it clear from the fact that in the recursive elimination process a wide range of iterations, each with a different subset of voxels, yielded high classification accuracies. Thirdly, that a voxel contributes to classification still doesn't tell us whether the underlying tissue is critical for the processes studied (Poldrack, 2011), only that they were differentially engaged. Finally, it is important to note that the specificity of a classifier for a particular cognitive process is also dependent on the cognitive taxonomy used to build the reference task conditions. Here we focussed on two types of retrieval, as proposed by delayed judgement of learning theory and in line with the two-stage model of working memory. According to this model, unique processes are involved in retrieval from the focus of attention, but there is no fundamental difference between retrieval of short and long delay items outside the current focus of attention. (Fiebach et al., 2006; Lewis-Peacock and Postle, 2008; Lewis-Peacock et al., 2012; Sakai et al., 2002). We could also have started from the three way division of working memory, that further distinguishes between a limited capacity buffer of direct access storage and long term memory (Nee and Jonides, 2011, 2013). Including examples from all three retrieval processes might have allowed us to determine whether immediate recall involves retrieval from the focus of attention or the direct access buffer, or else, whether delayed recall involves retrieval from the direct access buffer or from long term memory. Increasing the training classes will yield distinctive activity patterns that emphasize more unique aspects of a cognitive process. On the other hand, information on critical processes that are common to some of the processes studied will be lost.

This task-dependency of our result brings us back to the problem of the generality of the specificity-requirement, as discussed by Hutzler (2014). It is clear that the reverse inference demonstrated here is confined to the context of the reference task and the hypotheses formulated for the delayed judgement of learning effect. Reverse inference in its most general form – i.e., reading from ongoing BOLD measurements whether at any particular time a particular cognitive process or brain state emerged – requires the most general specificity of the biomarker for the process or state. To achieve this with multivariate pattern classification, the alternative process class cannot be restricted to one process (e.g., immediate recall), but needs to be as broad as possible. Future studies have to address the feasibility of this.

The delayed judgement of learning effect

Our multivariate pattern analysis showed that brain activation maps of delayed, but not immediate, monitoring trials are recognized as long term memory retrieval examples, thus confirming the re-activation interpretation of delayed judgement of learning (Dunlosky and Nelson, 1994; Nelson and Dunlosky, 1991; Thiede et al., 2005). Less convincing was the evidence that the information available during immediate monitoring is retrieved from working memory or the focus of attention, as predicted by the dual memories account of the delayed judgement effect (Dunlosky and Nelson, 1994; Rhodes and Tauber, 2011). Nonetheless, our implementation of immediate and delayed/long term judgement of learning was effective, because at the behaviour level the judgements made in delayed and long-term trials were relatively more accurate than in immediate trials. Participants gave significantly higher confidence scores for word pairs that they did recall in the post-test compared to word pairs that they did not recall, whereas there was no significant difference in immediate monitoring trials.

Rather than providing evidence that retrieval from working memory underlies immediate judgements of learning, our results suggest that in immediate monitoring participants more often do not retrieve the target word to make their judgement. Our variable decoding result in this condition is in accordance with behavioural studies reporting that immediate judgements are influenced by such trivial factors as cue word characteristics (Finn and Metcalfe, 2008; Koriat, 1997; Koriat and Bjork, 2006; Kornell and Bjork, 2009; Rhodes and Castel, 2008, 2009) and ease of encoding (Castel et al., 2007; Hertzog et al., 2003), which are irrelevant to evaluating the accessibility of the target word. On the other hand, immediate judgements are not influenced by manipulations that interfere with retrieval and that do affect delayed judgements of learning (Dunlosky and Nelson, 1992; Eakin and Hertzog, 2012).

Acknowledgments

This work was supported by The Netherlands Organisation for Scientific Research (Nederlandse Organisatie voor Wetenschappelijk Onderzoek — NWO), grant number PROO 411-07-153.

Conflict of interests

There is no conflict of interest for any of the authors.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.neuroimage.2016.02.008.

References

Aguirre, G.K., 2003. Functional imaging in behavioral neurology and cognitive neuropsychology. In: Feinberg, T.E., Farah, M.J. (Eds.), Behavioral Neurology and Cognitive Neuropsychology. New York, McGraw-Hill, pp. 85–96.

Cabeza, R., Ciaramelli, E., Moscovitch, M., 2012. Cognitive contributions of the ventral parietal cortex: an integrative theoretical account. TICS 12, 338–352.

Castel, A.D., McCabe, D.P., Roediger 3rd., H.L., 2007. Illusions of competence and overestimation of associative memory for identical items: evidence from judgments of learning. Psychon. Bull. Rev. 14, 107–111.

Chang, L.J., Yarkoni, T., Khaw, M.W., Sanfey, A.G., 2013. Decoding the role of the insula in human cognition: functional parcellation and large-scale reverse inference. Cereb. Cortex 23, 739–749.

Christoff, K., Owen, A.M., 2006. Improving reverse neuroimaging inference. Cognitive domain versus cognitive complexity. TICS 10, 352–353.

Chua, E.F., Schacter, D.L., Sperling, R.A., 2009. Neural correlates of metamemory: a comparison of feeling-of-knowing and retrospective confidence judgments. J. Cogn. Neurosci. 21, 1751–1765.

Cusack, R., Cumming, N., Bor, D., Norris, D., Lyzenga, J., 2005. Automated post-hoc noise cancellation tool for audio recordings acquired in an MRI scanner. Hum. Brain Mapp. 24, 299–304.

Daselaar, S.M., Prince, S.E., Dennis, N.A., Hayes, S.M., Kim, H., Cabeza, R., 2009. Posterior midline and ventral parietal activity is associated with retrieval success and encoding failure. Front. Hum. Neurosci. 3, e13.

De Martino, F., Valente, G., Staeren, N., Ashburner, J., Goebel, R., Formisano, E., 2008. Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns. NeuroImage 43, 44–58.

D'Esposito, M., Postle, B.R., 1999. The dependence of span and delayed-response performance on the prefrontal cortex. Neuropsychologia 37, 1303–1315.

D'Esposito, M., Ballard, D., Aguirre, G.K., Zarahn, E., 1998. Human prefrontal cortex is not specific for working memory: a functional MRI study. NeuroImage 8, 274–282.

Do Lam, A.T.A., Axmacher, N., Fell, J., Staresina, B.P., Gauggel, S., Wagner, T., et al., 2012. Monitoring the mind: the neurocognitive correlates of metamemory. PLoS One 7, e30009.

Dunlosky, J., Hertzog, C., 2001. Measuring strategy production during associative learning: the relative utility of concurrent versus retrospective reports. Mem. Cogn. 29, 247–253.

Dunlosky, J., Nelson, T.O., 1992. Importance of the kind of cue for judgments of learning (JOL) and the delayed-JOL effect. Mem. Cogn. 20, 374–380.

Dunlosky, J., Nelson, T.O., 1994. Does the sensitivity of judgments of learning (JOLs) to the effects of various study activities depend on when the JOLs occur? J. Mem. Lang. 33, 545–565.

Eakin, D.K., Hertzog, C., 2012. Immediate judgments of learning are insensitive to implicit interference effects at retrieval. Mem. Cogn. 40, 8–18.

Fiebach, C.J., Rissman, J., D'Esposito, D.E., 2006. Modulation of inferotemporal cortex activation during verbal working memory maintenance. Neuron 51, 251–261.

Finn, B., Metcalfe, J., 2008. Judgments of learning are influenced by memory for past test. J. Mem. Lang. 58, 19–34.

Formisano, E., De Martino, F., Bonte, M., Goebel, R., 2008. Who is saying what? Brain-based decoding of human voice and speech. Science 322, 970–973.

Fox, P.T., Friston, K.J., 2012. Distributed processing; distributed functions? NeuroImage 61, 407–426.

Friston, K.J., Frith, C.D., Liddle, P.F., Frackowiak, R.S.J., 1991. Comparing functional (PET) images: the assessment of significant change. J. Cereb. Blood Flow Metab. 11, 690–699.

Fuster, J., 1989. The Prefrontal Cortex: Anatomy, Physiology, and Neuropsychology of the Frontal Lobes. Raven Press, New York.

Fuster, J.M., Alexander, G.E., 1971. Neuron activity related to short-term memory. Science 173, 652–654.

Goldman-Rakic, P.S., 1987. Circuitry of the prefrontal cortex and the regulation of behavior by representational memory. Handbook of Physiology 5 pp. 373–417.

Haxby, J.V., Gobbini, M.I., Furey, M.L., Ishai, A., Schouten, J.L., Pietrini, P., 2001. Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science 293, 2425–2430.

Haynes, J.-D., Sakai, K., Rees, G., Gilbert, S., Frith, C., Passingham, R.E., 2007. Reading hidden intentions in the human brain. Curr. Biol. 17, 323–328.

Hertzog, C., Dunlosky, J., Robinson, E., Kidder, D., 2003. Encoding fluency is a cue used for judgments about learning. J. Exp. Psychol. Learn. Mem. Cogn. 29, 22–34.

Howell, D.C., 2002. Statistical Methods for Psychology. fifth ed. Duxbury Press, Belmont, CA.

Huijbers, W., Vannini, P., Sperling, R.A., Pennartz, P.M., Cabeza, R., Daselaar, S.M., 2012. Explaining the encoding-retrieval flip: memory-related deactivations and activations in the posteromedial cortex. Neuropsychologia 50, 3764–3774.

Hutzler, F., 2014. Reverse inference is not a fallacy per se: cognitive processes can be inferred from functional imaging data. NeuroImage 84, 1061–1069.

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

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

Kao, Y.C., Davis, E.S., Gabrieli, J.D., 2005. Neural correlates of actual and predicted memory formation. Nat. Neurosci. 8, 1776–1783.

Keerthi, S.S., DeCoste, D., 2005. A modified finite Newton method for fast solution of large scale linear SVMs. J. Mach. Learn. Res. 6, 341–361.

Keulers, E.H.H., Stiers, P., Jolles, J., 2011. Developmental changes between ages 13 and 21 years in the extent and magnitude of the BOLD response during decision making. NeuroImage 54, 1442–1454.

Kirwan, C.B., Stark, C.E., 2004. Medial temporal lobe activation during encoding and retrieval of novel face–name pairs. Hippocampus 14, 919–930.

Koriat, A., 1997. Monitoring one's own knowledge during study: a cue-utilization approach to judgments of learning. J. Exp. Psychol. Gen. 126, 349–370.

Koriat, A., Bjork, R.A., 2006. Mending metacognitive illusions: a comparison of mnemonic-based and theory-based procedures. J. Exp. Psychol. Learn. Mem. Cogn. 32, 1133–1145.

Koriat, A., Goldsmith, M., 1996. Monitoring and control processes in the strategic regulation of memory accuracy. Psychol. Rev. 103, 490–517.

Kornell, N., Bjork, R.A., 2009. A stability bias in human memory: overestimating remembering and underestimating learning. J. Exp. Psychol. Gen. 138, 449–498.

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

Lewis-Peacock, J.A., Postle, B.R., 2008. Temporary activation of long term memory supports working memory. J. Neurosci. 28, 8765–8771.

Lewis-Peacock, J.A., Drysdale, A.T., Oberauer, K., Postle, B.R., 2012. Neural evidence for a distinction between short-term memory and the focus of attention. J. Cogn. Neurosci. 24, 61–79.

Mars, R.B., Jbabdi, S., Sallet, J., O'Reilly, J.X., Croxson, P.L., Olivier, E., et al., 2011. Diffusion-weighted imaging tractography-based parcellation of the human parietal cortex and comparison with human and macaque resting-state functional connectivity. J. Neurosci. 31, 4087–4100.

Miller, E.K., Erickson, C.A., Desimone, R., 1996. Neural mechanisms of visual working memory in prefrontal cortex of the macaque. J. Neurosci. 16, 5154–5167.

Nee, D.E., Jonides, J., 2011. Dissociable contributions of prefrontal cortex and the hippocampus to short-term memory: evidence for a 3-state model of memory. NeuroImage 54, 1540–1548.

Nee, D.E., Jonides, J., 2013. Neural evidence for a 3-state model of visual short-term memory. NeuroImage 74, 1–11.

Nee, D.E., Berman, M.G., Moore, K.S., Jonides, J., 2008. Neuroscientific evidence about the distinction between short-and long term memory. Curr. Dir. Psychol. Sci. 17, 102–106.

Nelson, T.O., Dunlosky, J., 1991. When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: the "delayed-JOL effect". Psychol. Sci. 2, 267–270.

Norman, K.A., Polyn, S.M., Detre, G.J., Haxby, J.V., 2006. Beyond mind-reading: multi-voxel pattern analysis of fMRI data. TICS 10, 424–430.

Okada, K., Vilberg, K.L., Rugg, M.D., 2012. Comparison of the neural correlates of retrieval success in tests of cued recall and recognition memory. Hum. Brain Mapp. 33, 523–533.

O'Toole, A.J., Jiang, F., Abdi, H., Penard, N., Dunlop, J.P., Parent, M.A., 2007. Theoretical, statistical, and practical perspectives on pattern-based classification approaches to the analysis of functional neuroimaging data. J. Cogn. Neurosci. 19, 1735–1752.

Oztekin, I., McElree, B., Staresina, B.P., Davachi, L., 2009. Working memory retrieval: contributions of the left prefrontal cortex, the left posterior parietal cortex, and the hippocampus. J. Cogn. Neurosci. 21, 581–593.

Passingham, R.E., Toni, I., Rushworth, M.F.S., 2000. Specialisation within the prefrontal cortex. The ventral prefrontal cortex and associative learning. Exp. Brain. Res. 133, 103–113.

Pereira, F., Mitchell, T., Botvinick, M., 2009. Machine learning classifiers and fMRI: a tutorial overview. NeuroImage 45, S199–S209.

Poldrack, R.A., 2006. Can cognitive processes be inferred from neuroimaging data? TICS 10, 59–63.

Poldrack, R.A., 2011. Inferring mental states from neuroimaging data: from reverse inference to large-scale decoding. Neuron 72, 692–697.

Poldrack, R.A., Halchenko, Y.O., Hanson, S.J., 2009. Decoding the large-scale structure of brain function by classifying mental states across individuals. Psychol. Sci. 20, 1364–1372.

Postle, B.R., 2006. Working memory as an emergent property of the mind and brain. Neuroscience 139, 23–38.

Ptito, A., Crane, J., Leonard, G., Amsel, R., Caramanos, Z., 1995. Visual–spatial localization by patients with frontal-lobe lesions invading or sparing area 46. Neuroreport 6, 1781–1784.

Ranganath, C., Blumenfeld, R.S., 2005. Doubts about double dissociations between short-and long term memory. TICS 9, 374–380.

Rhodes, M.G., Castel, A.D., 2008. Memory predictions are influenced by perceptual information: evidence for metacognitive illusions. J. Exp. Psychol. Gen. 137, 615–625.

Rhodes, M.G., Castel, A.D., 2009. Metacognitive illusions for auditory information: effects on monitoring and control. Psychon. Bull. Rev. 16, 550–554.

Rhodes, M.G., Tauber, S.K., 2011. The influence of delaying judgments of learning on metacognitive accuracy: a meta-analytic review. Psychol. Bull. 137, 131–148.

Rissman, J., Greely, H.T., Wagner, A.D., 2010. Detecting individual memories through the neural decoding of memory states and past experience. PNAS 107, 9849–9854.

Sakai, K., Rowe, J.B., Passingham, R.E., 2002. Active maintenance in prefrontal area 46 creates distractor-resistant memory. Nat. Neurosci. 5, 479–484.

Siegel, S., Castellan Jr, N.J., 1988. Nonparametric statistics for the behavioral sciences. 2nd ed. McGraw-Hill, New York.

Stiers, P., Mennes, M., Sunaert, S., 2010. Distributed task coding throughout the multiple demand network of the human frontal–insular cortex. NeuroImage 52, 252–262.

Suykens, J.A.K., Van Gestel, T., De Barbanter, J., De Moor, B., Vanderwalle, J., 2002. Least Squares Support Vector Machines. World Scientific Publishing. http://www.esat.kuleuven.be/sista/lssvmlab/.

Talairach, J., Tournoux, P., 1988. Co-planar Stereotactic Atlas of the Human Brain. Beorg Thieme Verlag, Stuttgart.

Thiede, K.W., Anderson, M.C.M., 2003. Summarizing can improve metacomprehension accuracy. Contemp. Educ. Psychol. 28, 129–160.

Thiede, K.W., Dunlosky, J., 1999. Towards a general model of self-regulated study: an analysis of items for study and self-paced study time. J. Exp. Psychol. Learn. Mem. Cogn. 25, 1024–1037.

Thiede, K.W., Anderson, M.C.M., Therriault, D., 2003. Accuracy of metacognitive monitoring affects learning of texts. J. Educ. Psychol. 95, 66–73.

Thiede, K.W., Dunlosky, J., Griffin, T., Wiley, J., 2005. Understanding the delayed-keyword effect on metacomprehension accuracy. J. Exp. Psychol. Learn. Mem. Cogn. 31, 1267–1280.

Tourville, J.A., Reilly, K.J., Guenther, F.H., 2008. Neural mechanisms underlying auditory feedback control of speech. NeuroImage 39, 1429–1443.

Vannini, P., O'Brien, J., O'Keefe, K., Pihlajamäki, M., Laviolette, P., Sperling, R.A., 2011. What goes down must come up: role of the posteromedial cortices in encoding and retrieval. Cereb. Cortex 21, 22–34.

Woolgar, A., Thompson, R., Bor, D., Duncan, J., 2011. Multi-voxel coding of stimuli, rules, and responses in human frontoparietal cortex. NeuroImage 56, 744–752.

Worsley, K.J., Evans, A.C., Marrett, S., Neelin, P., 1992. A three-dimensional statistical analysis for rCBF activation studies in human brain. J. Cereb. Blood Flow Metab. 12, 900–918.

Xue, G., Dong, Q., Chen, C., Lu, Z., Mumford, J.A., Poldrack, R.A., 2010. Greater neural pattern similarity across repetitions is associated with better memory. Science 330, 97–101.