# WEAKENING MEMORIES BY HALF-REMEMBERING THEM

Gregory Julius Detre

A DISSERTATION
PRESENTED TO THE FACULTY
OF PRINCETON UNIVERSITY
IN CANDIDACY FOR THE DEGREE
OF DOCTOR OF PHILOSOPHY

RECOMMENDED FOR ACCEPTANCE
BY THE DEPARTMENT OF PSYCHOLOGY

Advisor: Kenneth Norman

June 2010

© Copyright by Gregory Julius Detre, 2010. All rights reserved.

## Abstract

During retrieval, memories activate to different degrees and compete to be recalled. We suggest the following hypotheses: memories that activate highly win the competition and are strengthened; memories that activate moderately lose the competition and get weakened; the closer the competition, the greater the consequent strengthening and weakening; and memories that do not activate do not compete, and are unaffected.

In Chapter 1, we review the behavioral findings (Anderson and Spellman, 1995; Anderson and Green, 2001) that suggested such a nonmonotonic relationship between level of activation and subsequent accessibility, and three possible accounts of it (Anderson et al., 2004; Tomlinson et al., 2009; Norman et al., 2006).

In Chapter 2, we describe four behavioral experiments designed to control the degree to which a memory activates in order to cause it to be forgotten. Experiment B1 (no significant effect) used presentation duration in an RSVP task to attempt to control activation. Experiment B2 (no significant effect) introduced a novel 'watermark' suppression task with which we attempted to release the learning of new associations from proactive interference. Experiment B3 (no significant effect) tested whether competition-driven learning when forming new associations caused the old associations to be weakened. Experiment B4a (significant effect) replicated the Depue et al. (2007) think/no-think paradigm with emotional stimuli. Experiment B4b (significant effect) introduced a novel 'graduated exposure watermark task'.

In Chapter 3, we describe how the lessons learned from two fMRI pilots shaped the design of Experiment F7.

In Chapter 4, we describe Experiment F7 (significant effects), which applied MVPA and region-of-interest (ROI) methods to fMRI to provide a covert, neural measure of a memory's activation within the think/no-think paradigm, and thus to predict its subsequent accessibility.

In Chapter 5, we consider the lessons learned from this series of experiments, and propose future work.

# Contents

# 1 Introduction

## 1.1 Why do we forget?

Sometimes we forget because we didn't form much of a memory in the first place.

Sometimes we forget because we cannot produce a strongly-related cue for the memory we wish to recall.

Sometimes we forget because memories interfere with one another at retrieval. The prior existence of old memories can make it harder to recall newer memories (*proactive interference*), and the formation of new memories can make it harder to recall old memories (*retroactive interference*).

And sometimes it seems that we forget because individual memories have grown weaker.

These different causes of forgetting are not completely distinct. Indeed, we will argue that cuing, interference at retrieval and weakening may all be related by the way that memories activate and compete at retrieval.

During retrieval, memories activate to different degrees in response to the cue, and compete to be recalled. We hypothesize that:

1. Memories that activate highly win the competition and get strengthened.
2. Memories that activate moderately lose the competition and get weakened.
3. The closer the competition, the greater the consequent strengthening and weakening.
4. Memories that do not activate do not compete, and are unaffected.

In this introductory chapter, we will review two paradigms (*retrieval-induced forgetting* and *think/no-think*) that provide behavioral evidence to support this set of hypotheses. We will consider three prominent accounts of these behavioral findings (Section 1.3). Though they

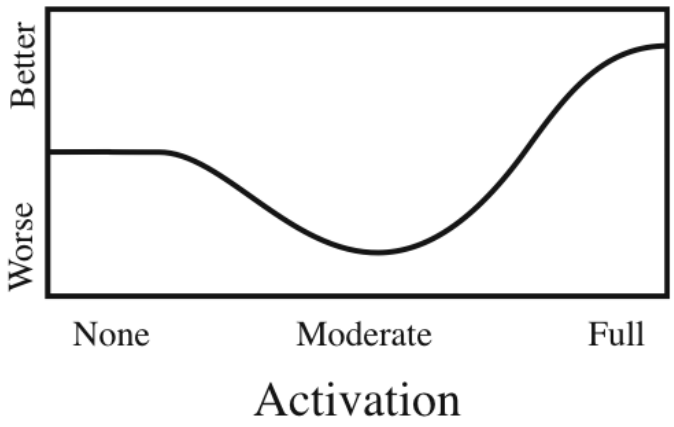


Figure 1: Nonmonotonic function relating the average activation of a representation to its subsequent accessibility (on the y-axis). Zero activation has no effect. Moderate activation will cause reduced accessibility. High activation will cause greater accessibility. From Newman (2008).

take very different form, there is considerable overlap in their broad predictions. Critically, all three make the counter-intuitive prediction that memories that are active enough to compete but not fully activated will be subsequently less accessible (although not all of them frame this in terms of weakening). Figure 1 shows this predicted nonmonotonic relationship between how much a memory activates and how easy it is to retrieve later.

To test and examine this nonmonotonic prediction, we ran a series of experiments designed to control and measure the amount of activation of and competition between memories, and to demonstrate and predict their consequently reduced accessibility at recall.

First, we will consider the retrieval-induced forgetting and think/no-think paradigms that motivated these predictions.

## 1.2 Two experimental paradigms

### 1.2.1 Retrieval-induced forgetting

In the best-known version of the *retrieval-induced forgetting* paradigm (Anderson and Spellman, 1995), subjects first learned a series of category-example paired associations (e.g.

*Fruit-Apple*, *Fruit-Pear*, *Fruit-Kiwi*, *Animal-Sheep*). Then, during the *retrieval practice phase*, they were cued with a subset of these pairs (e.g. *Fruit-Pe__*), and asked to retrieve the associates (e.g. *Pear*). Finally, their recall for all the pairs was tested.

We will consider four central findings:

1. *above-baseline facilitation* - memory for the practiced pairs (e.g. *Fruit-Pear*) was better than for the control pairs (e.g. *Animal-Sheep*).

2. *below-baseline suppression* - memory for the non-practiced related pairs (e.g. *Fruit-Apple*) was worse than for the control pairs.

3. memory for the weak category exemplars (e.g. *Fruit-Kiwi*) (*Fruit-Pe__*) was not impaired.

4. if the paired association was fully exposed (e.g. *Fruit-Pear*), instead of having to be retrieved from a partial cue (*Fruit-Pe__*), there was no competition between memories vying to be retrieved, and no change in accessibility of related memories (e.g. of *Apple*).

Facilitation of the practiced pairs relative to the control pairs is non-controversial - the act of being tested and correctly retrieving a memory has long been known to strengthen it (Karpicke and Roediger, 2008).

However, the suppression of the non-practiced related pairs is more interesting. Anderson and Spellman (1995) argue that the forgetting of the memory for *Fruit-Apple* results from competition at retrieval during practice. When cued with *Fruit-Pe__*, the representations for both *Pear* and *Apple* vied to be retrieved. *Pear* won the competition and was strengthened as a result, and *Apple* lost the competition and was weakened as a result.

This idea of competition-driven learning during retrieval is corroborated by the fact that there is no forgetting of weak category exemplars (*Kiwi*) that are hardly cued by the category

(*Fruit*), and no forgetting when the partial cue is replaced by a full exposure (requiring no retrieval). Likewise, the control pair (*Animal-Sheep*) remains unaffected by the retrieval practice phase, since *Fruit* does not cue *Sheep* at all, and so no competition or learning occurs.

**Cue-independent forgetting** Anderson makes a strong claim that this weakening effect acts on the *associate* item itself, rather than just severing the *association* between the pair of items. This assertion rests on the finding of *cue-independent forgetting* (Anderson and Spellman, 1995; Levy and Anderson, 2002) - even when cued with an *independent cue* such as *Red-A__*, the unpracticed associate *Apple* was less accessible. This would seem to suggest weakening of the associate representation (*Apple*) itself, rather than just the learned association (*Fruit -> Apple*) (however, see Camp et al. (2007) for problems with this approach). For the most part, we avoid discussion of whether the weakening affects the associate item, the association between the cue and the associate, or both. Our predictions are consistent with either possibility.

### 1.2.2 The think/no-think paradigm

The *think/no-think* paradigm (Anderson and Green, 2001) provides another approach for thinking about forgetting. There were three main phases:

1. In the *study phase*, subjects learned a series of paired associations.
2. In the *think/no-think* phase, they were repeatedly presented with the cues from these pairs.
   (a) For the *think pairs*, they were instructed to retrieve the associate.
   (b) For the *no-think pairs*, they were instructed to try not to let associate enter their consciousness.

The *baseline pairs* were not presented during the think/no-think phase at all.

3. In the *test* phase, subjects' recollection for the think, no-think and baseline pairs was tested.

Subjects' recall performance on the *think pairs*, (where they had tried to retrieve the associate) was better than for the baseline pairs that had not been practiced at all since being learned. This *above-baseline facilitation* is another example of correct retrieval causing a memory to be strengthened.

Subjects' recall performance on the *no-think pairs* (where they had deliberately refrained from retrieving the associate) was worse than for the baseline pairs which had not been practiced at all since being learned. This *below-baseline suppression* suggests that the act of deliberately suppressing the recently-learned associate to a cue from entering conscious awareness caused that associate to be forgotten. As in retrieval-induced forgetting, this reduced accessibility of the associate was also demonstrated using an independent cue, i.e. other than the cue learned as part of the paired association.

Levy and Anderson (2002) argue that these think/no-think results, in conjunction with those from retrieval-induced forgetting, provide strong evidence that memories can be weakened under a variety of circumstances, some under our deliberate control and some not.

### 1.2.3 What do retrieval-induced forgetting and think/no-think have in common?

We have introduced the *retrieval-induced forgetting* and *think/no-think* paradigms together because we think we can provide a common framework for understanding them.

Both paradigms involve studying paired associates, followed by a practice phase in which the cues trigger some amount of recollection, and then a final recall test of all the pairs.

Both the practiced pairs in retrieval-induced forgetting and the think pairs in think/no-think

were successfully retrieved and show an *above-baseline facilitation effect*. These memories were highly activated by the cues, won the competition at retrieval, and were consequently remembered better.

Both the unpracticed related pairs in retrieval-induced forgetting and the no-think pairs in think/no-think showed a *below-baseline suppression effect*. These memories were moderately activated by the cues, lost the competition at retrieval, and were consequently forgotten.

The fully exposed pairs and weak category exemplars in retrieval-induced forgetting did not elicit much competition at retrieval, and showed little consequent change in accessibility. Likewise, Levy (2008) showed that the more that subjects report intrusions during no-think trials (evidence of competition), the greater their subsequent below-baseline suppression effect.

### 1.2.4 The unreliability of the below-baseline suppression effect

The below-baseline suppression effect has been replicated a number of times with variants of both the retrieval-induced forgetting paradigm (e.g. Anderson et al. (1994); Anderson and Spellman (1995); Bauml (1996); Bauml (2002); Storm et al. (2007); Kuhl et al. (2007) - see survey by Levy and Anderson (2002)) and the think/no-think paradigm (e.g. Anderson et al. (2004); Depue et al. (2006); Depue et al. (2007); Levy and Anderson (2008)).

However, there have also been a number of published failures to replicate this below-baseline suppression effect (e.g. Hertel and Gerstle (2003); Algarabel and Martinez (2006); Butler et al. (2001); Bulevich et al. (2006); Bergström et al. (2007)). Notably, Bulevich et al. (2006) were unable to show a below-baseline suppression effect in a modified version of the Anderson and Green (2001) think/no-think paradigm, nor even when they attempted an exact replication of the original experiment with directions from Michael Anderson. Williams and Zacks (2001) failed to replicate the Anderson et al. (1994) finding of greater forgetting for the strong category exemplars (*Apple*) than the weak category examplars.

Indeed, most demonstrations of below-baseline suppression show only about a 10% drop in recall performance, with considerable variability both within and between subjects. *Why is this below-baseline suppression such a noisy and subtle effect?*

In short, we suggest it is because there has to be a competition at retrieval in which the to-be-forgotten representation activates moderately but below the threshold of strengthening. Calibrating this level of activation experimentally is very hard to do, since too much or too little will undermine the effect.

We expand upon this explanation below (Section 1.3).

## 1.3 Three accounts

We will consider three accounts of these retrieval-induced forgetting and think/no-think effects:

1. Top-down, targeted inhibition suppresses the competing representations (Anderson et al., 2004; Levy, 2008) - Section 1.3.1
2. After the memory trace is located during the first *sampling* stage of retrieval, its contents are replaced by a new interfering memory during the *recovery* stage (Tomlinson et al., 2009) - Section 1.3.2
3. Local competition between representations drives learning, explained in terms of the *oscillating learning algorithm* (Norman et al., 2006) - Section 1.3.3

With a little reframing, all three predict that the relationship between activation and learning is nonmonotonic: low activation has no effect, moderate activation causes forgetting, but a lot of activation improves recollection.

It is easy to think about why this nonmonotonicity might be a beneficial strategy for the brain to adopt. If multiple memories compete vigorously in response to cue, that indicates

that they are poorly differentiated, that the target memory needs to be strengthened, and/or that some of them need to be weakened. It makes sense then to strengthen those that activate the most, to ensure that they activate more next time and win more readily. Conversely, it makes sense to weaken the competing memories that activated partially, so that they do not get in the way so much in future. The cumulative effect of many such strengthenings and weakenings is to rebalance the memories and to differentiate them from one another. Indeed, Norman et al. (2007) showed that this differentiation as a result of competition is the reason that nonmonotonic algorithms (such as the oscillating learning algorithm) are extremely effective at learning many, many correlated patterns.

**Why competition-driven learning is hard to demonstrate experimentally** While this nonmonotonic plasticity curve conveys clear benefits to the learner, it makes it difficult to demonstrate forgetting of memories experimentally. This is because nonmonotonic models predict that the range of activation within which forgetting occurs is bounded on both sides - too little and the memory will be unaffected, too much and the memory will be remembered better. If this is true, then these below-baseline suppression effects are fickle because it is very difficult to reliably elicit just the right level of activation of a representation.

Put another way, striking the right level of competition between memories to cause maximal forgetting of the losing competitors is hard, because the closer the margin between the winning and losing memories, the greater the consequent change in accessibility. If the competing memories gain the upper hand even briefly, we can expect them to be remembered considerably better, canceling out any forgetting effect.

We will now consider the three nonmonotonic accounts in turn.

### 1.3.1 Top-down targeted inhibition - Levy & Anderson (2002)

Michael Anderson interprets the results from retrieval-induced forgetting and think/no-think in terms of targeted inhibition from executive control areas (Logan, 1994; Knight et al., 1999). As we have discussed (Section 1.2.3), we can consider retrieval practice in retrieval-induced forgetting and no-think trials in think/no-think as engendering a competition between memories vying to be retrieved in response to the cue. Like the teacher in a school yard, the executive control areas step in to resolve the competition, ensuring that the target representation wins and the competitor representations lose. They suggest that this intervention of executive control processes causes strengthening of the winning target representation and weakening of the losing competitor representations.

Levy and Anderson (2002) make an overt analogy between the control processes involved in overriding prepotent motor responses (as in the go/no-go task: Sakagami and Niki, 1994; de Zubicaray et al., 2000) and the control processes involved in suppressing memories (as in think/no-think) (see also Levy, 2008). More strongly, they suggest that control in both the motor and the memory domains may actually be regulated by overlapping neural processes. Plausibly, the competition at retrieval activates the anterior cingulate cortex (ACC) as a kind of conflict-detection alarm bell. The ACC, in turn, recruits the lateral prefrontal cortex (PFC) to down-regulate the unwanted competing memories. Depue et al. (2007) and Kuhl et al. (2007) provide further evidence that these frontal mechanisms are heavily involved during the no-think trials.

However, as Levy (2008) recognizes, this executive control need not necessarily be inhibitory - much the same result can be achieved by excitation of the target representation combined with local lateral inhibition (Miller and Cohen, 2001; Norman et al., 2007). By providing supplementary excitatory input to one of the representations vying to be activated, the prefrontal cortex could bias the competition in favor of the target representation, overriding the prepotent response. Local competition-driven learning would then cause the over-ridden representation to be weakened, without any top-down inhibition required.

We discuss the arguments in favor of targeted top-down inhibition (rather than top-down excitation combined with local lateral inhibition) further in Section 1.3.3.

The centrality of executive control in this account of memory weakening led Levy and Anderson (2008) to propose the *executive deficit hypothesis* - individual differences in executive control underly the individual differences in suppression of unwanted memories. They suggest that these differences in executive control account for a large proportion of the between-subjects variance in the laboratory paradigms described above, and also in different people's ability to suppress traumatic memories.

### 1.3.2 The two-stage theory of interference - Tomlinson et al (2009)

Models such as SAM (Search of Associative Memory; Raaijmakers and Shiffrin, 1981) and REM (Retrieving Effectively from Memory; Shiffrin and Steyvers, 1997) have been very successful at explaining many forgetting effects solely in terms of interference between memories at retrieval - in other words, even when we forget, the memory is still there, but is occluded by other memories. They do not incorporate any mechanism for structural weakening of memory traces.

So it seemed that the retrieval-induced forgetting and think/no-think below-baseline suppression effects (especially cue-independent forgetting - Section 1.2.1) posed a problem for such models, since it did not seem to be possible to accommodate these effects without positing some kind of structural weakening of memories.

Recently though, Tomlinson et al. (2009) have proposed an interesting reframing of the no-think cue-independent forgetting effect in terms of the 2-stage sampling and recovery model of recall featured in models like SAM:

1. The memory trace is located in the *sampling stage*
2. The details of the memory are retrieved in the *recovery stage*

Tomlinson et al. (2009) proposed that interference in the recovery stage might account for cue-independent forgetting, without requiring any weakening of the memory itself. Participants may sometimes have sampled the partial memory trace during a no-think trial, despite their efforts not to. Having sampled it, they then associated this memory trace with the no-think response (of sitting quietly). In other words, they activated the memory partially during the no-think trial, and in doing so, created a new 'sitting quietly' memory on top of it. During the final recall phase, the independent cue sampled the same location for the memory trace, but now recovered the newly-learned no-think 'sitting quietly' response, resulting in impaired recall performance without the actual memory for the associate being weakened.

To test this, they ran a modified version of the think/no-think paradigm that included an extra 'press enter' condition in which subjects were instructed to simply press the 'enter' key as quickly as possible in response to the cue. By their account, this would create an interfering 'press enter' memory occasionally whenever the associate memory trace was sampled, and that this 'press enter' memory trace would then interfere with the associate memory trace during recovery in the final recall phase. This task should not recruit any executive control memory suppression mechanisms of the kind described by Levy and Anderson (2002) (see Section 1.3.1). In line with Tomlinson et al. (2009)'s predictions, performance in this 'press enter' condition was significantly lower than in the baseline condition, and not significantly different from performance in the no-think task.

It seems reasonable to think this account might be extended to the retrieval-induced forgetting paradigm too. Perhaps the *Apple* representation is sampled at retrieval practice sometimes when *Fruit-Pe__* is the cue, but ultimately *Pear* is recovered. As a result, when cued later with either *Fruit-A* or *Red-A* (an independent cue), *Pear* interferes, so *Apple* is less accessible, and appears to have been weakened.

**Nonmonotonicity within the interference-based account** To our knowledge, the proponents of this interference-based account have not made any explicit prediction of a nonmonotonic relationship between activation and learning. However, it seems straightforward to accommodate this prediction within their framework:

1. At low levels of activation, neither sampling nor recovery of the associate occur, and so recall performance should be unaffected.

2. At moderate levels of activation, the association memory trace is being partially activated during the sampling stage, but not recovered, and the new interfering 'sitting quietly' memory trace is laid down. This is simply a restatement of Tomlinson et al. (2009)'s explanation given above for the below-baseline suppression effect.

3. At high levels of activation, perhaps subjects are actually recovering the original association memory trace fully and, in the process, strengthening that association.

This account proposes a very different mechanism to explain the below-baseline suppression effect in comparison with the structural weakening involved in the top-down targeted inhibition and oscillating learning algorithm accounts. However, as we can see, it generates a broadly similar set of nonmonotonic predictions. For the most part, we will focus on testing the predictions that these three accounts have in common, though we will briefly discuss ways to disambiguate the predictions made by this 'pure interference' account from the other 'structural weakening accounts' in Experiment B2 and in Section 5.3.2.

### 1.3.3 The oscillating learning algorithm - Norman et al (2006)

The oscillating learning algorithm (Norman et al., 2006; Norman et al., 2007; Norman et al., 2005) is a neural network learning rule that emphasizes the role of local competition-driven learning in its account of the retrieval-induced forgetting and think/no-think findings. We

will focus first on applying it to the retrieval-induced forgetting paradigm to illustrate how it works.

Each pair is modeled as a distributed representation (i.e. a pattern of activity over a set of units) in a neural network containing recurrent weights. The cues (e.g. *Fruit-Pe__*) are modeled by providing excitatory input to a subset of the units involved in the *Fruit-Pear* representation. The network has already learned weights that connect these units together strongly, and so it can retrieve the associate (*Pear*) to the cue (*Fruit-Pe__*) by allowing activity to spread from the units that are receiving external excitatory input to the remainder.

The interesting aspect of the network's learning occurs after the partial cue has been provided, where activity is spreading between units and the network is settling into an attractor.

**Inhibition, strengthening and weakening in the oscillating learning algorithm** In order to understand the network dynamics better, we need to consider how inhibition is implemented in the model.

In the brain, inhibitory inter-neurons regulate the overall level of activity by measuring the output from excitatory neurons and providing a dampening, inhibitory effect as this level of excitatory activity rises - just as the air conditioning kicks in when a room gets too hot. Rather than simulating these inhibitory inter-neurons directly, we make use of the k-winners-take-all (kWTA) algorithm (Minai and Levy, 1994; O'Reilly and Munakata, 2000) as a simplifying shortcut, which mimics the effect of inhibitory inter-neurons by adjusting the overall level of inhibition directly to maintain some setpoint of activity.

In the oscillating learning algorithm model, inhibition plays a more proactive role by oscillating on top of this kWTA setpoint sinusoidally.

1. At the activation setpoint, the active units tend to be part of the *Pear* representation, since they are receiving more excitatory input from the *Fruit-Pe__* cue than the *Apple*

units, and thus supporting one another more with spreading activation.

2. When we reduce the inhibition, more units will become active. Any units whose activity increases as the inhibition reaches its trough are part of competitor representations (*Apple*) that are finally free to activate, and so the oscillating learning algorithm decrements their weights.

3. When we increase the inhibition, fewer units will remain active. Any units whose activity decreases as the inhibition reaches its peak are part of the target representation (*Pear*) that are struggling to stay active, and so the oscillating learning algorithm increments their weights. [1]

   For a more detailed description of the algorithm, see Norman et al. (2006) and Norman et al. (2007).

In this way, the low inhibition serves to flush out and identify the units of competing representations so that they can be weakened, and the high inhibition serves as a stress-test that identifies the units of the target representation so that they can be strengthened.

**Competition in the no-think trials** It is less clear how to think about competition in the context of the no-think trials. We can consider the 'losing' memory to be the associate that is supposed to be suppressed, but subjects are not provided with an explicit 'target' memory that should trump this to-be-suppressed associate. We suggest that there are many possible representations that play the role of winners in this competition. For instance, in informal post-experimental debriefings, our subjects reported using a variety of different strategies during no-think trials, such as: focusing on the letters of the cue word; thinking about pre-experimental associations to the cue word; singing a song to themselves; and creating new substitute associations of their own. Each of these would provide target

[1] For clarity of exposition, we have simplified the description of how the sign of the learning algorithm changes as a function of the phase of the oscillation - actually, learning occurs at all four phases of the oscillation (Norman et al., 2006).

representations that could successfully compete with the to-be-suppressed representation during retrieval. Indeed, Hertel and Calcaterra (2005) reported that they only found below-baseline suppression for no-think trials when subjects used a substitute association. [2] Certainly, we could consider the 'press enter' instruction used by Tomlinson et al. (2009) as a kind of competitor, and indeed they treat 'sitting quietly' as a potentially interfering (and therefore competing) memory.

However, there is a case to be made that it is top-down inhibition, rather than substitution or local competition, that drives the below-baseline suppression effect. According to Levy (2008), around 10% of subjects reported "letting their mind go blank" as a strategy, and still showed a below-baseline suppression effect, and Bergström et al. (2007) only found cue-independent forgetting when subjects were instructed to use a 'thought suppression' rather than 'thought substitution' strategy. From the imaging literature, Depue et al. (2007) did not find increased activity in sensory cortical areas during no-think trials, which they took as evidence that subjects were not generating new substitute associations. And Levy (2008) points to the hippocampal down-regulation during no-think trials (Anderson et al., 2004; Depue et al., 2007; also Section 3.1.1) as further evidence of inhibition - however, this might be explained by biased competition and lateral inhibition, or (more speculatively) by a more diffuse spatial profile produced by multiple competing memories that elicits a lower overall BOLD response across the region of interest.

On balance then, it seems reasonable to frame the no-think trials in terms of competition, much like the retrieval practice trials.

In the next section, we emphasize that our central hypotheses concern the nonmonotonicity of the relationship between activation and learning, rather than competition *per se*.

[2] However, because Hertel and Calcaterra (2005) did not include an independent cue test, their below-baseline suppression when using substitutions could be explained purely in terms of interference effects, undermining any strong conclusion that we might want to draw from this experiment.

**Nonmonotonic relationship between average activation and consequent learning** In the account given above, competition at retrieval is driving the oscillating learning algorithm's weight changes. But at a coarser level, we can look at these weight changes as a function of representations' *average activity over the course of the oscillation*. The moderately active competing *Apple* representation is only present during the low-inhibition phase of the oscillation (*moderate average activation*) and gets weakened. The strongly active target *Pear* representation only falters during the high-inhibition phase of the oscillation (*high average activation*)and gets strengthened. Finally, the inactive control Sheep representation does not activate at all (*low average activation*) during the oscillation, and so is neither strengthened nor weakened. In this way, the oscillating learning algorithm predicts the nonmonotonic relationship between the average activity of a representation and its consequent weakening/strengthening shown in Figure 1.

**Extremely high levels of activation** Strictly speaking, the oscillating learning algorithm makes the further prediction that no learning will occur for extremely high levels of activation (beyond the right-hand end of the X axis of Figure 1), since very strong representations would not falter even during the high-inhibition stress-test portion of the oscillation - and without a decrease in activation as inhibition peaks, no weight changes will occur. However, we consider it unlikely that any of the paired associates learned in the laboratory experiments described below will create memories strong enough to activate in this upper range. For this reason, we will focus on the predictions relating to the lower three-quarters of the activation range (following Newman and Norman, 2010).

**Nonmonotonic plasticity curves in the literature** This notion of a nonmonotonic neural plasticity curve relating activation to learning is not new or unique to the oscillating learning algorithm (Bienenstock et al., 1982; Senn and Fusi, 2005; see Newman and Norman, 2010 for a wider survey, also discussed in Section 3.1.1).

The oscillating learning algorithm was originally designed to account for the rich and

varied findings in the human behavioral retrieval-induced forgetting domain, only briefly surveyed in Section 1.2.1 (see Norman et al. (2007) for a much more detailed survey). But we can also look to neurophysiology for evidence of nonmonotonic plasticity curves - the function relating concentration of Ca2+ ions (indicative of excitatory input) to learning (long-term depression and long-term potentiation) appears to have the same nonmonotonic shape (Hansel et al., 1996).

More generally, these nonmonotonic learning curves also have a number of desirable functional properties, including high and robust associative memory storage capacity, especially for correlated patterns (Norman et al., 2006).

Indeed, many of the same nonmonotonic predictions would be made by this entire family of nonmonotonic learning algorithms. However, direct evidence of nonmonotonic plasticity is relatively sparse - only Newman and Norman (2010) have set out to show this relationship in humans, and no one has yet looked for it in the various suppression paradigms described here.

## 1.4 Aims: controlling and measuring

Broadly, we will take two approaches to testing our prediction of a nonmonotonic relationship between memory activation and accessibility:

1. In our behavioral experiments, we will attempt to *finely control the degree of activation* of to-be-forgotten representations by engineering tasks that will moderately activate them, which should cause those representations to be forgotten. The hard part is to ensure that these representations do not accidentally over-activate, since that would cause them to be counter-productively strengthened. See Chapter 2.

2. Given that such careful titration of memory activation is very difficult to achieve, we will attempt to devise a *covert, neural measure of memory activation*. With this measure, we should be able to predict which trials will be remembered worse and which

better, based on the oscillating learning algorithm's hypothesized nonmonotonic relationship between activation and learning. See Chapters 3 and 4.

## 1.5 Overview of experiments

We will discuss a total of 7 experiments, the first 4 of which were behavioral, and the last 3 of which used fMRI:

1. Experiment B1 used a rapid serial visual presentation (RSVP) task to try and elicit partial activation and forgetting of associations by briefly cuing them. Although the results looked promising, the overall non-parametric test for the complete binning analysis was not significant - see Section 2.2.

2. Experiment B2 failed to demonstrate the suppression of A-B pairs (using our 'watermark' task) through a release from proactive interference when learning A-C pairs in an AB-AC paradigm - see Section 2.3.

3. Experiment B3 failed to show suppression in an *AB-AC* paradigm, designed to engender competition between the old and new associations in order to weaken the *A-B* associations - see Section 2.4.

4. Experiment B4a was an attempted replication of a think/no-think paradigm with emotional stimuli (Depue et al., 2006). Experiment B4b successfully adapted this paradigm to produce a below-baseline suppression effect with our 'graduated exposure watermark' task - see Sections 2.5 and 2.6.

5. Pilot Experiment F5 was an early fMRI pilot, testing how well we could classify successful from unsuccessful recalls - see Section 3.2.

6. Pilot Experiment F6 was our first attempt to apply MVPA to the think/no-think paradigm - see Section 3.3.

7. Experiment F7 was our culminating think/no-think fMRI experiment, demonstrating that a covert neural measure of memory activation shows a nonmonotonic relationship with subsequent accessibility - see Chapter 4.

# 2 Behavioral experiments

## 2.1 Introduction - controlling the degree of activation of the to-be-forgotten representations

We think of the basic no-think instructions (Anderson and Green, 2001) ("don't let the associated item enter your consciousness") as a means of eliciting a little but not too much activation of a representation. The cue presentation causes the associate to activate a little, but the subject's cognitive control mechanisms step in to ensure that the associate representation does not fully activate (see Section 1.3.1). Together, these two mechanisms counteract one another, keeping the activation of the to-be-forgotten associate's representation within the moderate-level forgetting band.

However, this procedure could easily fail - if the inhibitory cognitive control mechanisms were to respond insufficiently or too slowly, the representation would activate too much and become more accessible. On the other hand, if subjects were to become too adept at barring the associate from consciousness, it might hardly be activated at all, and hardly be forgotten.

In the experiments we propose, we aim to activate representations moderately while minimizing the number of 'intrusions' where the representation over-activates by a small margin and gets consequently strengthened. In Section 1.3, we suggested that the below-baseline suppression effect is so unreliable because even a few of these just-above-threshold intrusions produce occasional, large bursts of strengthening that cancel out the small forgetting effects.

## 2.2 Experiment B1 - rapid serial visual presentation (RSVP)

### 2.2.1 Introduction

In this experiment, we used a straightforward rapid serial visual presentation (RSVP) task to partially activate the to-be-forgotten associations by cuing them very briefly. The cue presentation duration serves as a proxy for the activation of its associate, predicting a nonmonotonic relationship between a cue's duration and the recall probability for its associate (as shown in Figure 1). In other words: final recall for cues presented very quickly should not be affected; final recall for cues presented at a medium rate should be impaired; and final recall for cues presented slowly should be improved.

### 2.2.2 Methods

**Participants** 31 subjects participated in this experiment, either for course credit or for a $12 payment. All of the subjects were drawn from the Princeton community.

**Stimuli** All of the cue-associate pairs in this experiment consisted of:

1. A location cue, e.g. 'Underwater', 'Gymnasium' or 'Helicopter'
2. A famous person associate, e.g. 'Britney Spears', 'Tom Cruise' or 'George W Bush'

Both a picture and a name were provided to the subject - see Figure 2.

In order to minimize primacy and recency effects, 3 extra pairs were presented at the beginning of each study run, and 3 extra pairs were presented at the end of each study run - these filler pairs were not included in any of the following analyses.

In order to minimize encoding variability, we tried to ensure that each subject would only be studying celebrities who were familiar to them. Prior to the experiment, each subject


Figure 2: Example location/celebrity stimuli pair used in Experiments B1, B2 and B3.

was presented with a large pool of celebrity stimuli from which they were asked to exclude any that were unfamiliar.

**Study phase** Each of the 80 location-celebrity pairs was presented twice (with a break between lists). Each presentation trial lasted 3750ms, with a 250ms inter-trial interval.

**Rapid serial visual presentation (RSVP) phase** The 80 pairs were divided evenly into an experimental RSVP group, and a baseline group.

The baseline pairs did not appear during this RSVP phase at all.

Subjects were presented with the location cues for the 40 RSVP pairs in a randomized order in rapid sequence, repeating each cue 12 times in each of 4 runs. They were instructed to press the spacebar whenever they noticed an interspersed 'oddball' animal image.

Each RSVP run consisted of roughly 500 trials (varying slightly depending on the number of oddball images inserted). Each RSVP pair was assigned a unique presentation duration, ranging from 30ms to 498ms in increments of 12ms. Each RSVP location cue stimulus was padded with a mask for a duration of at least 25ms to ensure a constant inter-stimulus

interval across trials. [3]

The oddball images were presented for a jittered duration of 180ms +/-150ms. Oddball images were randomly inserted into the RSVSP stream separated by at least 8 and at most 15 location trials. The oddball animal type (polar bear, tiger, dog, eagle) changed on each of the four runs. At the beginning of each run, subjects were familiarized with all of the oddball images for that animal type.

**Final recall phase** In this final recall phase, subjects' recollection of all of the associations (both RSVP and baseline groups) was tested, in randomized order. They were presented with the location cue and asked to type in the full name of the associated celebrity within 10,000ms. These typed responses were marked as correct if exactly or nearly the same (within three missed, translated or mistyped keystrokes) as the full celebrity name. For instance, a response of 'Jac Nihcolson' would have been marked as correct if the right answer were 'Jack Nicholson'.

### 2.2.3 Results

**Comparing baseline and RSVP pairs** The mean performance in the final recall phase across subjects for the baseline pairs (80.1%, SEM = 0.02) was higher than for the RSVP pairs (79.4%, SEM = 0.03), although this difference was not significant ($t(30) = 0.40$, $p > 0.05$) - see Figure 3.

**Binning RSVP pairs by cue duration** To examine the relationship between cue duration and recall, we grouped sets of pairs together into N bins based on their RSVP presentation duration. [4] For instance, if we were to divide the pairs into 10 bins, then the first bin

[3] The total inter-stimulus interval, including the mask and two screen refreshes, added up to 550ms per trials. This should be large enough to avoid complications from attentional blink effects (Raymond et al., 1992), which are most acute around 200-300ms (Nieuwenhuis et al., 2005), and have mostly fallen off by 500ms (Chun and Potter, 1995).

[4] As described in Section 2.2.2, the range of presentation durations was the same for each subject.

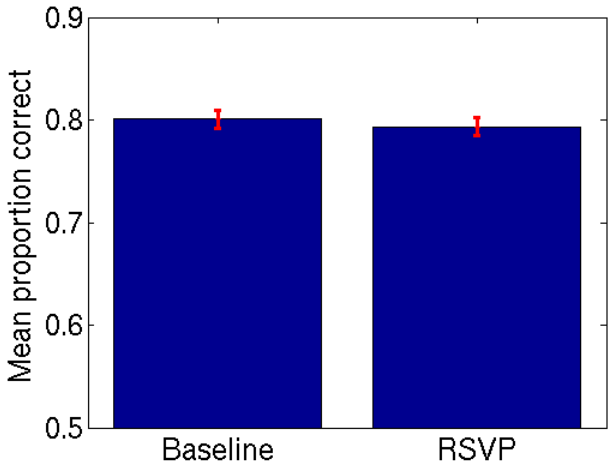


Figure 3: Comparison between the final recall performance for the baseline and RSVP pairs.

would contain the RSVP pairs with durations of 30ms, 42ms, 54ms or 66ms. For each subject, for each bin, we computed the mean cue duration and the proportion of final recall responses answered correctly. We could then compare the proportion correctly recalled across subjects for one bin with another bin (or against baseline) with a paired samples t-test.

This analysis is similar in spirit to the *binning analysis* applied to classifier activations in Experiment F7 (Section 4.8), and many of the same issues apply. As we discuss in Section 4.8.7, it is hard to estimate in advance how many bins to divide the pairs into, and there are tradeoffs to having too many or too few, so we ran the analysis for 3, 4, 5, 6, 7, 8, 9 and 10 bins - see Figures 4, 5, 6 and 7. We will refer to each of these analyses, run with a different number of bins, as a *bin-set* - in other words, there were 8 total bin-sets (3-10).

To assess whether moderately active memories were being forgotten, we compared the recall performance for all of the middle (i.e. excluding the first and last) bins with the recall performance for the first and last bins, and also for the baseline pairs.

We will focus initially on the 8 bin-set analysis. As can be seen in Figure 6b, the recall performance for the pairs in the 3rd bin was significantly below the recall performance

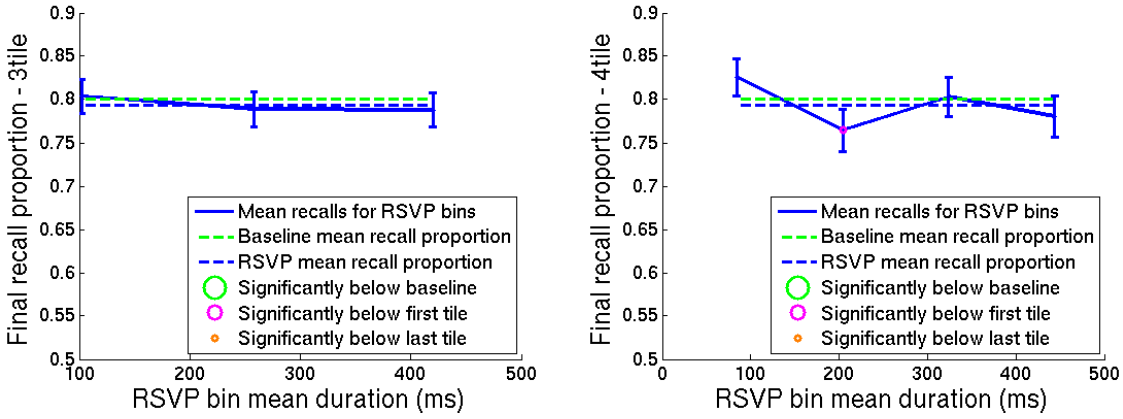


Figure 4: RSVP binning analysis - varying the number of bins: (a) 3 bins (b) 4 bins. N.B. The errorbars displayed are between-subjects, though the actual t-tests comparing each of the middle to the outer bins were paired samples.

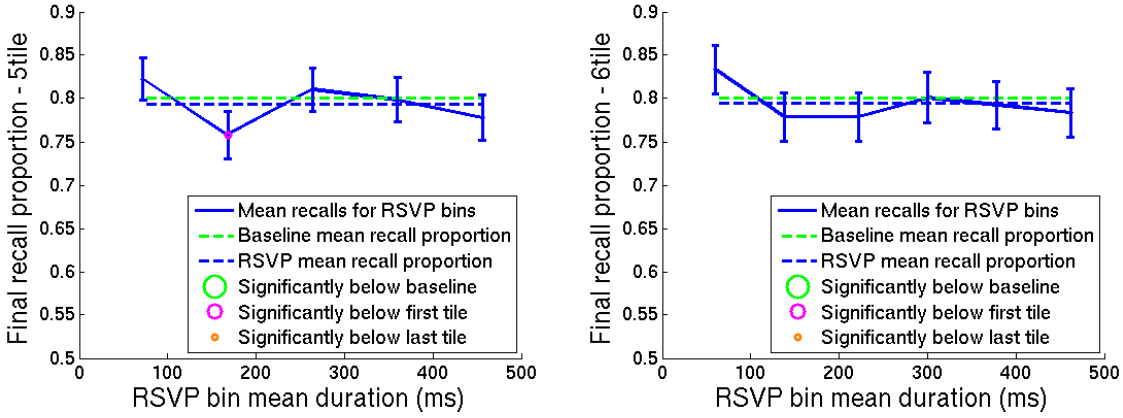


Figure 5: Binning analysis - varying the number of bins: (a) 5 bins (b) 6 bins. As described above, these are between-subjects errorbars.

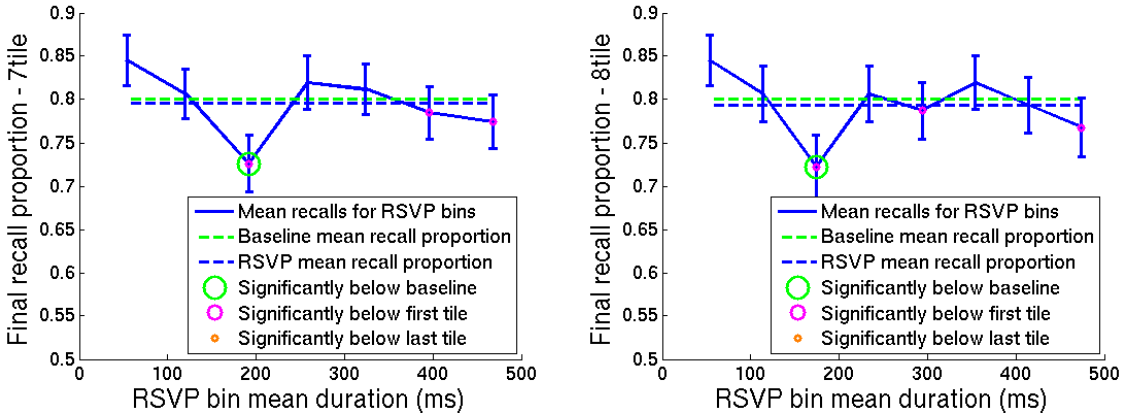


Figure 6: RSVP binning analysis - varying the number of bins: (a) 7 bins (b) 8 bins.

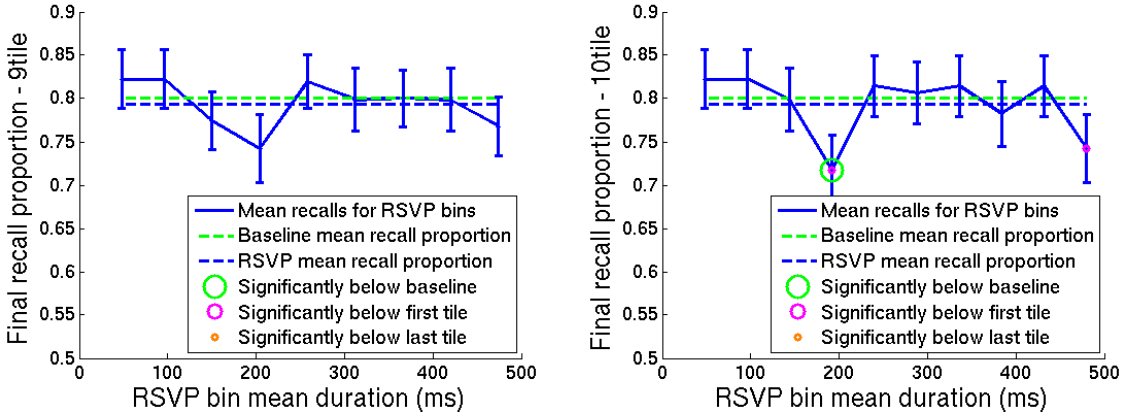


Figure 7: Binning analysis - varying the number of bins: (a) 9 bins (b) 10 bins.

for the baseline pairs ($t(30) = -1.91$, $p < 0.05$) (large green circle), and below the recall performance for the pairs in 1st bin ($t(30) = -2.98$, $p < 0.05$) (small magenta circle). However, there was no significant difference between the 3rd and the 8th (last) bin ($t(30) = -1.07$, $p > 0.05$). Performance in the 5th and 8th bins was also significantly below the 1st bin (small magenta circles). Unexpectedly, performance on the 1st bin was actually significantly *above* baseline ($t(30) = 1.98$, $p < 0.05$) (not marked on Figure 6b).

Performance in the 3rd bin was also significantly below baseline in the 7-tile and the 10-tile analysis (marked with green circles). In multiple cases, performance on middle bins was significantly below the that of the 1st bins, but since these were somewhat elevated, we do will not emphasize this.

None of the middle bins in the 8 bin-set analysis showed performance significantly below that of the last bin.

**Correcting for multiple comparisons with non-parametric statistics** In the preliminary statistics described above, we did not correct for multiple comparisons. Firstly, we ran our t-test analyses separately on each of the middle 6 bins. Secondly, we ran the complete bin-set analysis multiple times, with varying numbers of bins (3-10).

For both these reasons, we introduced a non-parametric analysis (permutation test; Nichols and Holmes, 2001) to correct for these multiple comparisons. This would determine whether the complete set of analyses across all the bin-sets showed a significant nonmonotonic effect. In other words, we wanted a single overall p-value that would reflect how often we should expect to see middle bins dipping this far below the first and last bins within their bin-set by chance.

1. We started by running the complete set of individual t-tests on each of the middle bins, for each of the (3-10) bin-sets, exactly as described in the previous section.

2. For each middle bin, we were running two t-tests: one to compare it with the first bin,

and one to compare it with the last bin (see description above). We now kept only the t-stat for the least significant of these two comparisons, for each of the middle bins. This yielded a single t-stat for each of the middle bins, for each of the (3-10) bin-sets.

3. For each complete bin-set, we picked the middle bin with the best t-stat. This yielded a single t-stat for each bin-set. For instance, in the case of the 8-bin analysis described above, we would have picked the t-stat for the 3rd bin compared against the last bin.

4. Finally, we took the mean across bin-sets of these t-stats (one per bin-set), yielding a single mean t-stat.

5. To create the null distribution for the permutation test, we scrambled the binary labels attached to each RSVP pair denoting whether that pair was recalled or forgotten, within subjects. After scrambling, each subject would still have the same proportion of pairs marked as recalled, but which were recalled and which forgotten would vary. Subjects who recalled all or none of their RSVP pairs were excluded from the analysis. All of the above steps were followed for each permutation, yielding a single mean t-stat each time.

6. We calculated the rank of the mean t-stat for the unscrambled (real) data relative to the null distribution of t-stats from the scrambled versions of the data. To compute the (one-tailed) p-value, we divided this rank by the number of permutations. [5]

This overall non-parametric test across all the middle bins of all the (3-10) bin-sets did not yield a significant effect (1000 permutations, $p = 0.16$).

[5] To be clear, the rank is computed as (1 + the number of null values above or equal to the real value). This way, if the rank of the mean t-tstat for the real data was higher than all of the mean t-stats for the scrambled data, the numerator would be set to 1, ensuring that the p-value can approach but never reach 0, even with many permutations. The denominator was set to (1 + the number of scrambled permutations) in order to include the real data in the count. This procedure follows the prescriptions in Nichols and Holmes (2001).

### 2.2.4 Discussion

When we compared the overall performance on RSVP and baseline pairs, we did not see a significant difference.

But when we binned the RSVP pairs by cue presentation duration, we started to see small but significant (uncorrected for multiple comparisons) below-baseline effects, especially for the bins in the lower-middle portion of the duration range. This is consistent with the nonmonotonic prediction that moderately activated memories would show reduced accessibility. The non-parametric permutation test that was run across all the middle bins in all the bin-sets was not quite significant though.

We had not anticipated the small but significant (uncorrected for multiple comparisons) above-baseline level of recall for the very 1st bin in the 8-tile analysis (and others). If this effect were robust, it would suggest that very very fast cue presentations (< 100ms) somehow increase the accessibility of the associate memories.

We did not show the predicted above-baseline effect for slower (higher duration) presentations. However, this could easily be because the activations elicited by these RSVP durations lie within the bottom two-thirds of the activation range of the nonmonotonic prediction shown in Figure 1. In order to test this, we plan to include some even slower cue presentations (above 500ms).

In this experiment, we sampled the range of cue presentation durations uniformly between 30ms and 498ms. Given that forgetting seems to occur for durations around 200ms, in future work we will sample this range more finely, and sample the upper range more coarsely.

Indeed, it would be interesting to directly compare the suppression efficacy of the standard no-think task with the this RSVP task (using a duration of around 200ms for all the pairs). One nice property of RSVP is that the presentations can be finely tweaked, and perhaps even calibrated for individual subjects depending on their base rate of intrusions.

In this analysis, we did not exclude any subjects based on their behavioral performance on the oddball task. However, it might be worth excluding outliers (as described in Section 4.7) as a screen for inattentiveness.

## 2.3 Experiment B2 - release from proactive interference

### 2.3.1 Introduction

In this experiment, we attempted to show a release from proactive interference by weakening the interfering memories. Our paradigm was based on the standard '*A-B A-C*' paired associate learning task (Barnes and Underwood, 1959), but with an additional 'weakening' phase inserted between the *A-B* and *A-C* study phases.

In a sense, this might seem like a circuitous dependent measure to use. We are trying to demonstrate that a memory has been weakened by measuring how much it proactively interferes with other memories later on. This approach was designed to pull apart the predictions of the pure interference-theory account from the other two structural weakening accounts.

This experiment employed a novel 'watermark task' (described fully in Section 2.3.2), where subjects were presented with the location cue, while being asked to look for a number of small, superimposed household object images to distract them from thinking about the associated celebrity. We hoped that the presented location cue would cause the associated celebrity's representation to activate slightly, but the watermark counting task would keep subjects engaged enough to avoid full recollection of the associate, activating it moderately as a result.

Subjects will have been presented with the cue for the *A-B* pairs multiple times in the watermark task. These presentations should, if anything, create *more* interfering memories and more interference at recall. But if subjects *were* to show a release from proactive interference for the watermark *A-B* pairs, that might be evidence that these pairs had in

fact been structurally weakened - otherwise, how else could more presentations of them be reducing the proactive interference they cause?

### 2.3.2 Methods

**Participants** 57 subjects participated in this experiment, either for course credit or for a $12 payment. All of the subjects were drawn from the Princeton community. The data were collected over two separate periods, in Spring 2008 and Spring 2010.

**Stimuli** The location-celebrity paired association images with labels used in this experiment are described fully in Section 2.2.2.

**First study phase** In the first study phase, subjects studied 20 of *A-B* pairs to criterion. Each of the pairs was shown once, and thereafter subjects' recall of the associate to a given cue was tested repeatedly in blocks until they had responded correctly twice for each pair. This study-to-criterion procedure was designed to enable the formation of strong associations and to minimize the encoding variability between pairs.

**Watermark task** After all the *A-B* pairs had been learned to criterion, subjects entered the 'weakening' phase, where we attempted to weaken half of the *A-B* associations using the 'watermark' task. In each trial, subjects were presented with the location *A* cue from one of the to-be-weakened pairs with a number of line drawings of household items superimposed (see Figure 3). Subjects viewed these images for 3500ms, with each of the to-be-weakened location *A* images appearing eight times. Subjects were given the following instructions:

1. Pick out and count as many of these superimposed household items as you can.
2. At the same time, try not to think of the person *B* associates that you previously learned for the location *A* background images.

3. If the person *B* associate does come to mind, press the space-bar.

Every space-bar press caused the image's presentation to be extended by another 3500ms. There was no upper limit on the number of times an image's presentation could be extended in this way.

Our aim with this weakening phase was for subjects to partially process the location *A* image in the background. By asking them to search for the superimposed household objects, we hoped to preclude too much processing of the location image. However, from pilot studies, we fully expected the previously-learned person *B* associate to intrude into awareness occasionally, and so we used the space-bar pressing as a kind of self-report, so that we might at least identify these trials.

We will refer to those pairs whose *A* cue appeared in this way in the weakening phase as 'wiped' pairs, and the remaining half of the pairs as 'unwiped'. This terminology is to distinguish them from the 'baseline' *D-E* pairs, introduced for the first time in the second study phase, for which no proactive interference existed.

**Second study phase** After the weakening phase, subjects began the second study phase. The procedure for this was more or less identical to the first study phase. However, although the same location *A* cues were used as before, they were now paired with new famous person *C* associates. We predicted that both learning and recall of these *A-C* associations should be impaired by the existing, proactively interfering *A-B* associations, though less so for the wiped pairs. Furthermore, a number of *D-E* baseline pairs were introduced. These consisted of new locations paired with new famous people. Because these locations only appeared in the second study phase, they provided a baseline level of performance with no proactive interference from previous associations at all.

### 2.3.3 Results

**Proactive interference** Subjects required significantly more trials to learn the *A-C* pairs for the unwiped than the baseline pairs ($t(56) = 3.74$, $p < 0.05$). In other words, learning the earlier *A-B* pairs proactively interfered with learning the later *A-C* pairs, relative to the baseline pairs for which no existing *A-C* association had been learned.

**Comparing wiped and unwiped pairs** If the weakening phase was having its intended effect of weakening the *A-B* associations, then it should be easier to learn the *A-C* associations for the wiped than the unwiped pairs. We considered three metrics for this:

1. Most simply, the number of trials were required to learn the *A-C* associations to criterion can be compared for the wiped and unwiped pairs.

   The number of trials required to learn the associations to criterion in the second study phase for the wiped *A-C* pairs (mean 2.65, SEM 0.05) was highest, followed by the unwiped *A-C* pairs (mean 2.64, SEM 0.06) and the *D-E* baseline (mean 2.51, SEM 0.05) pairs.

   In opposition to our predictions, the wiped *A-C* pairs took slightly longer to learn than the unwiped pairs, though this difference was not significant ($t(56 = 0.44$, $p > 0.05$).

2. We also looked at the proportion of trials where subjects answered correctly on the first round of *A-C* tests, i.e. how often they learned the *A-C* associations after a single presentation as a different way of assessing the degree of proactive interference from the existing *A-B* associations. There was no significant difference between the wiped and unwiped pairs on this first *A-C* trial ($t(56) = 0.78$, $p > 0.05$).

3. Finally, we looked at how many trials required to learn each *A-C* association to criterion subtracting the number of trials required to learn the corresponding *A-B*

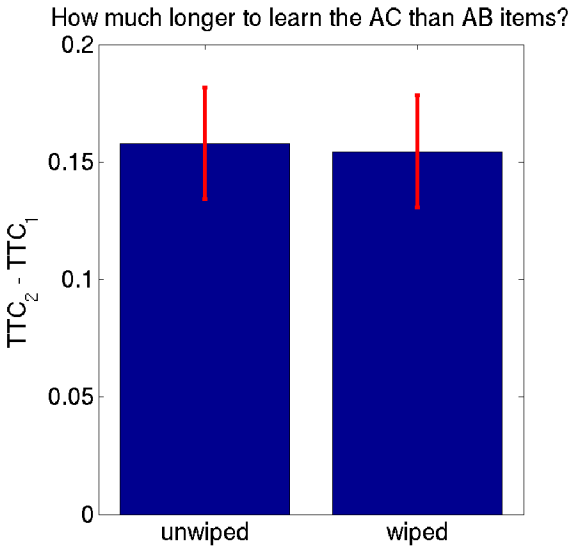


Figure 8: Comparison between the wiped and unwiped pairs, using the third metric, subtracting the number of trials required to learn the *A-B* association from the number of trials required for the corresponding *A-C* association to criterion. Not significant.

association to criterion. There was no significant difference between the wiped and the unwiped pairs ($t(56) = 0.07$, $p > 0.05$). See Figure 8.

### 2.3.4 Discussion

In the earlier proposal for this dissertation, we reported preliminary data from just 25 subjects in this experiment. These showed the basic proactive interference effect, and a promising but non-significant trend where the wiped *A-C* pairs appeared to be learned more quickly than the unwiped.

We proposed then to collect more data to see whether this difference between wiped and unwiped pairs would be significant for a larger subject pool. As described in Section 2.3.3, this difference is no longer close to statistical significance with the larger pool of subjects.

In Section 2.7 we consider a variety of reasons for why we might fail to show a suppression effect, all of which potentially apply here.

This is the only experiment we discuss that, if it were to show forgetting, might be hard to account for within a pure interference-theory account. We discuss this issue in more detail in Section 5.3.2.

## 2.4 Experiment B3 - competition-driven learning

### 2.4.1 Introduction

In this experiment, we set out to show that we could create competition during retrieval practice between an old association and a new association, and that this competition would cause weakening of the old association.

As in Experiment B2, this experiment is based on the *A-B A-C* (Barnes and Underwood, 1959) paradigm, but the dependent measure is quite different. In Experiment B2, we sought to first weaken the *A-B* pairs with the watermark task, and then to show a release from proactive interference when learning the *A-C* pairs. However, in this experiment, we tried to use the competition involved in learning the *A-C* pairs to drive learning, and then we measured recall for the *A-B* pairs directly. In our weakening retrieval practice condition, we tried to maximize the amount of competition from the to-be-weakened *A-B* pairs, while trying to keep their activation below the threshold of strengthening. This should maximize the amount of weakening. In contrast, learning new associations with full exposure should not elicit competition from the previous *A-B* associations at all, and so should not cause them to be weakened.

In some sense, this is an attempt to conceptually replicate Anderson et al. (2000). In a retrieval-induced forgetting experiment, they showed that retrieval practice for *Fr__-Orange* did not cause forgetting. *Fruit* was the only possible response, and so there was no competition at retrieval. In contrast (as discussed in Section 1.2.1), retrieval practice for *Fruit-Pe__* produced forgetting of related associates like *Apple* and *Orange* because they competed with (and lost to) *Pear* at retrieval.

### 2.4.2 Methods

**Participants** 37 participants (18 female) from the Princeton community participated in this experiment for payment.

**Design overview** The experiment was divided broadly into three phases:

1. *study A-B*
2. *study A-C - full exposure* vs *retrieval practice*
3. *final A-B recall*

We are interested in the competition during the *study A-C* phase between the old *B* associations and the new *C* associations being learned. The key manipulation is between the *full exposure* and *retrieval practice* conditions (see below). This is designed to modulate the competition during the *study A-C* phase:

1. In the *full exposure* group, the new *A-C* associations are presented, with no testing.
2. In the *retrieval practice* group, the new *A-C* associations are learned by testing - subjects had to practice retrieving the *C* association in response to the *A* cue.

We hypothesized that there would be much more competition in the retrieval practice than the full exposure group. This competition at retrieval should cause more weakening of the losing, competing *B* association. As a result, recall of the retrieval practice *A-B* pairs during the final recall phase should be much worse than recall for the full exposure pairs.

See Figure 9 to see a schematic of this design.

**Stimuli** For the *A-B* pairs, we used locations and celebrity images with text labels, as used in previous experiments (Section 2.3.2). As in Experiment B1, we asked subjects to filter out unfamiliar celebrities before the experiment began (Section 2.2.2).

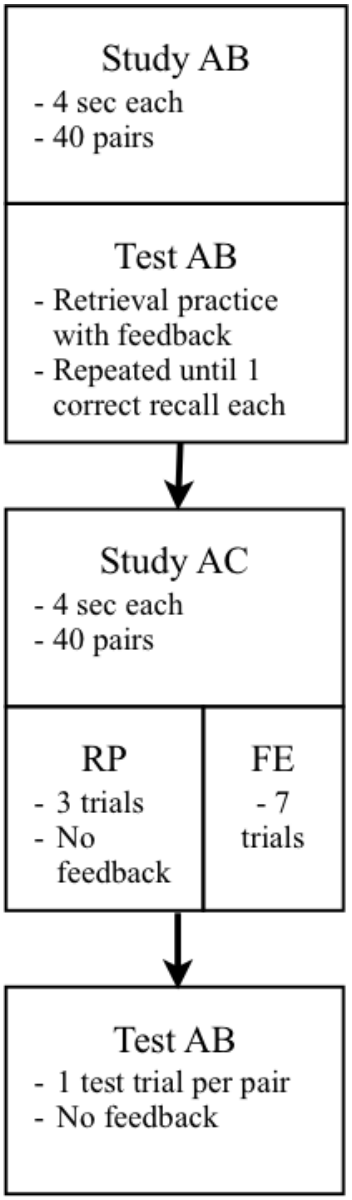


Figure 9: Schematic of the design of Experiment B3. From Carroll (2009).

For the *A*-*C* pairs, we used locations and miscellaneous objects. These miscellaneous *C* objects were drawn from a mix of different categories (e.g. animals, body parts, household objects, tools etc.).

**Study *A*-*B* phase** Subjects were presented with all 40 *A*-*B* pairs once for 4000ms, then tested on each of the pairs until they had recalled each successfully once.

**Study *A*-*C* phase** As described in Section 2.4.2, the 40 pairs were divided evently into *full exposure* and *retrieval practice* groups.

For the full exposure pairs, subjects were presented with both the *A* cue and *Z* associate together for 4000ms. Each *A*-*Z* pair was presented 7 times, once per run.

For the *retrieval practice* pairs, subjects were presented with each of the *A*-*C* pairs once. Thereafter, they were presented with just the cue, and required to type in *Z* associate. Each *A*-*Z* pair was tested 3 times in this way, spread over multiple runs. In order to maximize the amount of competition from the *A*-*B* association, we subliminally flashed a novel image of the *B* associated celebrity (as rapidly as the 60Hz computer screen could update) just prior to asking subjects to recall the *Z* associate on the third retrieval practice trial. When questioned afterwards, very few subjects reported seeing these images.

This phase was divided into 7 runs.

**Final *A*-*B* recall phase** Just as in Experiment B2 (Section 2.3), we hope to measure this weakening in terms of reduced proactive interference when learning the new A-C associations. The substitution procedure would provide an even more convincing demonstration of weakening, since adding new associations to the 'A' cue should only cause the amount of proactive interference to *increase*.

**Pilot experiments to ensure calibrate the full exposure and retrieval practice conditions**
Our hypothesis was that the competition during retrieval practice trials would cause weakening of the *A-B* associations, and so impair recall performance for the *A-B* retrieval practice pairs relative to full exposure.

However, it is well-established that the act of recollection in test-driven learning is a more effective way to learn than purely by presentation (Karpicke and Roediger, 2008). Lower performance for retrieval practice than full exposure pairs could then be confoundingly explained in terms of increased retroactive interference from more strongly-interfering retrieval practice *A-C* associations.

To compensate for this, we ran a number of pilot experiments (not described) to calibrate the learning efficacy of the full exposure and retrieval practice procedures. It was determined that 7 full exposure trials would create memories at least as strong as 3 retrieval practice trials. Thus, if we were to find lower *A-B* final recall performance for the retrieval practice pairs, it could not be explained in terms of greater retroactive interference.

### 2.4.3 Results

We compared the number of correctly recalled celebrity *B* associates recalled in the final recall phase for the full exposure group (mean 18.2, SEM = 0.32) and the retrieval practice group (mean 18.1, SEM = 0.30). A paired samples one-tailed t-test did not find this difference to be significant ($t(36) = 0.48$, $p > 0.05$) - see Figure 10.

### 2.4.4 Discussion

Our hypothesis was that there would be competition in the retrieval practice trials when learning the new *A-C* associations. The *A-C* pairs would win this competition, and the competing *A-B* associations would lose, impairing subsequent recall of the *A-B* associations in the final recall phase.

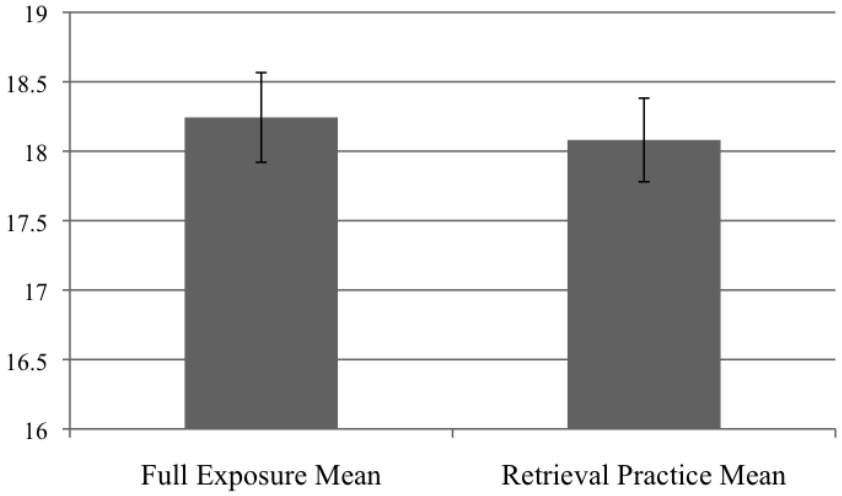


Figure 10: Mean number of correctly recalled associates for the full exposure and retrieval practice conditions. [From Carroll (2009)]

In contrast, there would be little or no competition at retrieval from the *A-B* pairs in the full exposure trials, since the direct presentation of the *A-C* would require no recollection and provide little opportunity for the *B* associates to activate.

However, we did not see this difference (see Section 2.4.3). Carroll (2009) describes a number of post-hoc analyses we ran to try to understand this null effect. We excluded subjects performing at ceiling using the criteria described by Depue et al. (2006). We also attempted to determine whether we might have over-compensated in the full exposure condition by having subjects learn *too* well. Finally, we examined whether individual differences in sleep or stress level might be a significant source of variation in between-subjects recall performance. None of these modified analyses substantively affected the results.

We had planned a number of more ambitious follow-up experiments combining ideas from Experiment B2 and this one. However, after failing to show a greater suppression effect here and in Experiment B2, we decided on a different, more careful approach. In Experiment B4a, we first attempt an exact replication of a well-established think/no-think

design, and then work within it to test our ideas about what makes for a good suppression task.

## 2.5 Experiment B4a - TNT with emotion stimuli replicating Depue et al (2006)

### 2.5.1 Introduction

After the previous two failures to produce a below-baseline suppression effect, we decided to take a try working within a well-established paradigm. While Experiment B4b involves a novel forgetting task that builds on it, this experiment is just a preliminary sanity check to make sure that we can replicate the published results.

Depue et al. (2006)'s behavioral think/no-think experiment compared emotional and non-emotional stimuli to test whether the cognitive control processes involved in emotional and non-emotional memories differ, and whether emotional memories might actually be easier to suppress. Indeed, they found greater below-baseline suppression of emotional than non-emotional memories after repeated no-think trials.

We sought to replicate this below-baseline suppression effect with emotional stimuli[6] so that we might then compare the standard no-think instructions with an alternative suppression task of our own in Experiment B4b (Section 2.6).

Ultimately, such suppression paradigms might be of significant clinical value to sufferers of disorders characterized by over-strong memories and associations - see Section 2.6.1.

All of the stimuli, instructions and methods were drawn directly from Depue et al. (2006), as described in greater detail in Fenstemaker (2009).

[6]Depue et al. (2006) also varied the number of repetitions (5 and 10). In this experiment, we only included the 10-repetitions condition, since they found the facilitation and suppression effects for this condition to be strongest.

### 2.5.2 Methods

**Participants** 23 Princeton undergraduates (19 female, mean age 21) participated in this experiment for payment. Data from 2 participants were omitted for failing to comply with instructions ($n = 1$) or performing at ceiling ($n = 1$) [7], leaving 21 subjects.

The experiment was divided into study, think/no-think and final recall phases.

**Software** All of the experiments described throughout were programmed using the Python Experiment-Programming Library (PyEPL) (Geller et al., 2007), and analyzed using a combination of custom Python and Matlab (Mathworks, Natick MA) code.

**Stimuli** 40 paired associates were generated by randomly pairing:

1. An anonymous male or female face image. Each participant was shown either all male or all female faces. See Figure 11 for example stimuli.
2. A negatively emotional scene image, drawn from the IAPS corpus. See Figure 12 for example stimuli.

**Study phase** Subjects were presented with 40 randomly-associated face-scene pairings (faces on the left, scenes on the right), grouped into two blocks. Subjects were then tested to criterion in blocks with 2-alternative forced-choice trials until they had responded correctly to 39 of the 40 pairings. [8]

[7] We used the same criteria applied by Depue et al. (2006) - subjects with 100% or 0% accuracy in more than half (i.e., 2 or more) of the conditions were excluded.

[8] The actual study procedure used was more complex than this - it was intended to exactly replicate Depue et al. (2006), with further details helpfully provided by Depue (B. Depue, personal communication, December 17, 2008), and described fully in (Fenstemaker, 2009).

Figure 11: Sample neutral anonymous male and female face cue images used in Experiments B4a and B4b.

Figure 12: Negatively emotional scene stimuli, similar to those used in Experiments B4a and B4b. N.B. The terms of use for the IAPS corpus prohibit reproduction of the actual stimuli, so these comparable images are provided for illustrative purposes. From Fenstemaker (2009).

**Think/no-think phase** 12 of the 40 pairs were used as *baseline* pairs and excluded entirely from this phase. 14 of the pairs were used as *think* pairs, and 14 were used as *no-think* pairs. Each pair appeared once for each of the 10 runs.

Both think and no-think trials consisted of a colored fixation cross presented for 1500ms, followed by a face for 4000ms, and then a 500ms inter-trial interval.

Think trials were indicated by green fixation crosses, for which subjects had been instructed to think of the scene previously associated with the face. No-think trials were indicated by red fixation crosses, for which subjects had been instructed to try not to let the previously associated scene enter their consciousness.

**Final recall phase** Each final recall trial consisted of a presentation of a face cue, and the instruction to type in a 3 or 4 word description of the scene that was paired with it. All 40 faces were tested like this, in a randomized order.

These scene descriptions were scored as correct or incorrect by a coder blind to their conditions.

### 2.5.3 Results

As reported by Depue et al. (2006), recall performance on think pairs was the highest of the three conditions (mean 75.2%, SEM = 0.03), followed by the baseline pairs (70.6%, SEM = 0.04), with performance on the no-think pairs being the lowest (68.7%, SEM = 0.03) - see Figure 13.

However, the above-baseline facilitation difference between the think and baseline conditions was not significant, ($t(20) = 1.43$, $p > 0.05$), nor was the below-baseline suppression difference between the no-think and baseline conditions ($t(20) = -0.47$, $p > 0.05$). Only the think vs no-think comparison was significant ($t(20) = 1.70$, $p = 0.05$).

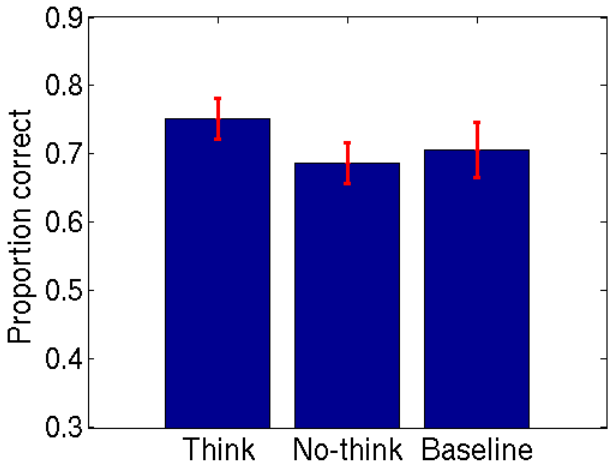


Figure 13: Recall performance by condition for Experiment B4a.

### 2.5.4 Discussion

While the comparisons between the experimental conditions and baseline did not reach significance, the broad pattern of results matched published findings.

Importantly though, our replication included only half as many subjects as Depue et al. (2006). We considered expanding our subject pool to confirm that these effects would reach significance, but we felt confident enough in the closeness of our replication and our attainment of the broad pattern of results to move on to Experiment B4b, where we adopted a different forgetting task in place of the standard no-think task - see Section 2.6.

## 2.6 Experiment B4b - based on Experiment B4a, with graduated exposure watermark task

### 2.6.1 Background

**Graduated exposure** *Graduated exposure* (also known as *systematic desensitization*) is a therapeutic technique used to treat conditions such as phobia, anxiety disorder and post-traumatic stress disorder (PTSD) (Wolpe et al., 1973). The aim is to reduce the negative

emotional response to some cue(s). In the case of PTSD, these cues might be evocative and emotionally loaded stimuli like wartime images, or even something innocuous like a loud noise. In the case of phobic patients, the cue is the phobic item, e.g. a spider.

When using the graduated exposure approach, subjects are exposed to the aversive stimuli gradually, starting with a very weak and attenuated cue. At each exposure level, subjects are taught to relax and keep their negative emotional response under control. As subjects are able to master their emotional response to low-level exposures, the strength of the cue is ramped up adaptively.

We can understand why graduated exposure might be a fruitful therapeutic approach for memory suppression in terms of the oscillating learning algorithm (see Section 1.3.3). In the case of PTSD, these negative associations have become so strong that they are too easily cued, even by innocuous stimuli such as loud noises. These memories lie at the very upper end of the activation range, so it is very difficult to activate them partially, in order to bring them into the range where they would be forgotten. As a means of avoiding these too-strong memories from blossoming into full recollections and becoming counter-productively strengthened, the graduated exposure approach combines very weak cues and relaxation techniques. [9]

### 2.6.2 Introduction

In this experiment, we reproduced almost all of the methods from Experiment B4a, except for swapping in a novel 'graduated exposure watermark' task in place of the standard no-think instructions.

[9] An alternative means of partially activating these too-strong memories might be to introduce a competing representation to provide a laterally-inhibiting dampening effect. Indeed, this would account for the success of approaches like Eye-Movement Desensitization and Reprocessing (EMDR) (Maxfield, 1999) - subjects are presented with cues to their negative associations while following a stimulus back and forth with their eyes. This secondary eye tracking task may provide a competing distraction that prevents the to-be-forgotten memory from activating fully.

**The watermark task** The watermark task was designed with the aim of causing greater below-baseline suppression than the no-think task, in the following ways:

1. The standard no-think instructions (Levy, 2008) encourage subjects to try and avoid letting the associate memory enter their consciousness, but they do not specifically provide a strategy for *how* to do this (though they do constrain subjects' behavior by urging them to fixate and attend to the cue stimulus throughout the no-think period). Subjects employ a variety of strategies during no-think trials, e.g. allowing their mind to wander, imagining further details about the cue or substituting a new association. In contrast, the object-counting aspect of the watermark task provided a concrete and specific task for subjects to focus on, rather than leaving it up to their discretion. In this way, we hoped to minimize the number of accidental intrusions and so reduce the amount of both within- and between-subject variability.

2. The task required subjects to perform a simple visual search with the face cues as a background so that subjects would be unwittingly and unavoidably processing the cue image to some degree, triggering partial recollection of the association.

3. Since we expected that the associate representation would be strongest for the early repetitions, we used the partially-visible early face cue images to try and avoid activating the associate representation too much, since that might cause it to be counter-productively strengthened.

4. After multiple watermark trials, we expected that the associate representation would have been at least somewhat forgotten (Kuhl et al., 2007), and so we gradually increased the visibility of the face cues. If we left the face cues at low-visibility throughout, we were concerned that only a small amount of forgetting might occur at the very beginning, after which they would cease to activate the associate representation sufficiently to cause appreciable forgetting.

### 2.6.3 Methods

**Participants** 46 Princeton University undergraduates (30 women, mean age = 20) participated in this experiment for payment. Data from 4 participants were omitted for failing to comply with instructions ($n$ = 3) or performing at ceiling ($n$ = 1), leaving 42 subjects.

**Graduated exposure watermark task** For this experiment, we replaced the standard no-think task used in Experiment B4a with the 'graduated exposure' version of the watermark task from Experiment B2.

Just as in Experiment B4a, the pairs were divided into either the baseline, think or watermark groups, and each the think and watermark trials were interspersed. Besides this change of forgetting task, all other aspects of the experiment remained the same.

Our graduated exposure watermark task was designed to partially activate subjects' memory for the scene associates by careful presentation of the face images as cues. Specifically, we used the following devices to attempt to control the degree of mnemonic activation of the associated scene representations:

1. We superimposed roughly 12 dark and light household object watermarks over the face cue background. Subjects were familiarized with these objects at the beginning of the think/watermark phase. During each watermark trial, subjects were instructed to focus on counting as many of these watermark objects as they could find. This task was intended to provide something besides the scene associate to occupy their mind.

2. As in the standard no-think instructions, subjects were also asked to prevent the scene associated with the background face from coming into their minds.

3. By analogy with the 'graduated exposure' approach described in Section 2.6.1, we slowly ramped up the visibility of the background face images over the course of the think/no-think phase - see this visibility progression in Figure 14.

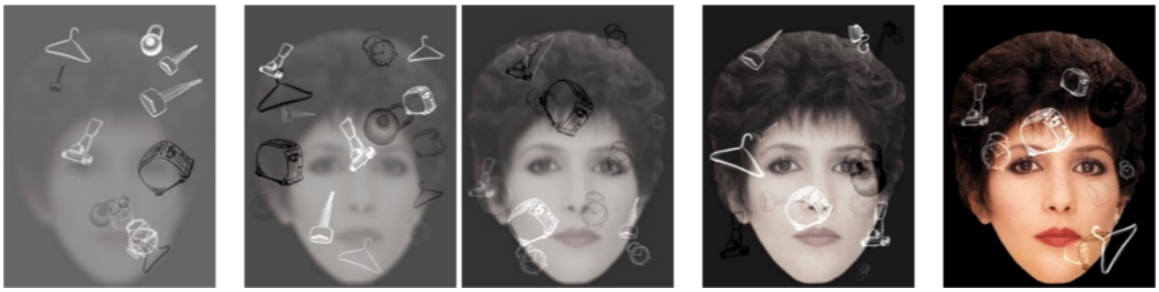

Figure 14: Sample stimuli from the forgetting task employed in Experiment B4b. Subjects were instructed to count the superimposed household object watermarks in each image. Over the course of multiple no-think trial repetitions, the visibility of the background face images was increased - the five visibility levels are shown here.

For the first presentations, the face cue images shown were almost unrecognizably blurred, darkened and grayscaled. After every two repetitions, we reduced the degree of Gaussian smoothing, increased the brightness and re-saturated the colors, making the faces more distinct and recognizable each time. By the final, 10th repetition, the face images had been restored to their original levels of sharpness, brightness and color saturation.

### 2.6.4 Results

As predicted, and in line with the results from Experiment B4b (Section 2.5.3), recall performance on think pairs was the highest of the three conditions (mean 70.2%, SEM = 0.03), followed by the baseline pairs (68.5%, SEM = 0.03), with performance on the no-think pairs being the lowest (64.1%, SEM = 0.03).

The above-baseline facilitation difference between the think and baseline conditions was not significant, ($t(41) = 0.67$, $p > 0.05$). However, unlike in Experiment B4a, the below-baseline suppression difference between the no-think and baseline conditions ($t(41) = -1.89$, $p < 0.05$) *was* significant. As before, the difference between think and no-think items was also significant ($t(41) = 2.22$, $p < 0.05$).

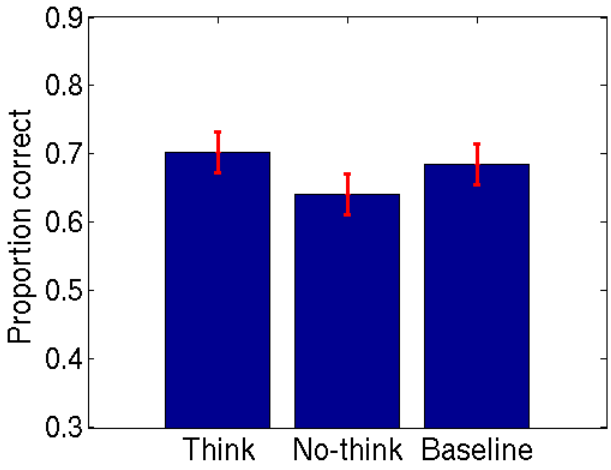


Figure 15: Recall performance by condition for Experiment B4b.

### 2.6.5 Discussion

**The watermark task** We attempted to incorporate ideas from graduated exposure (see Section 2.6.1) into the design of the watermark task, by steadily increasing the visibility of the face cue. In this experiment, the schedule for increasing the visibility was held constant, changing slightly every two repetitions. In other words, the graduated exposure was not adaptively driven by the strength of the representation we sought to suppress. So, for instance, if the associate memory were to accidentally intrude halfway through the watermark phase, it might get strengthened, after which it would become even more difficult to avoid an intrusion again, since the face cues would be becoming more and more visible. For this reason, it might be beneficial to adapt the cue visibility in response to some dependent measure of associate memory strength. One could employ the behavioral trial-by-trial intrusion responses used by Levy (2008), or attempt something more sophisticated but noisy based on a covert neural measure (as described in Chapters 3 and 4).

Since the watermark pairs were remembered significantly worse than the baseline pairs, we can consider the basic goal of the watermark task to have been met. However, further work is needed to determine whether the counting task, the graduated exposure, or some

combination of them both, were necessary to cause this below-baseline suppression. In Experiment B2 (see Section 2.3), we incorporated a variant of this watermark task without the graduated exposure - in this case, the baseline-suppression effect was not significant. This might suggest that the graduated exposure played a significant role in the below-baseline suppression effect here. However, the design, instructions, stimuli and dependent measure also differed in Experiment B2 so we cannot say this with any confidence.

## 2.7 General discussion

### 2.7.1 Summary of behavioral results

Our aim with these behavioral forgetting procedures was to find a way to activate representations in a moderate and controlled manner.

1. *Experiment B1* showed a small but significant below-baseline effect for one of the bins in the RSVP task.

2. *Experiment B2* failed to show a release from proactive interference when learning *A-C* pairs after applying the basic (non-graduated exposure) watermark task to the *A-B* pairs in an AB-AC paradigm, despite promising preliminary results.

3. *Experiment B3* failed to show suppression of old associations by eliciting competition from new associations in an AB-AC paradigm.

4. In *Experiment 4a*, we decided to re-orient by first replicating an experiment design and stimuli used successfully twice in the literature (Depue et al., 2006; Depue et al., 2007). We reproduced the overall pattern of results although our subject pool was too small for them to be significant.

5. In *Experiment B4b* (based on Experiment 4a), we replaced the standard no-think procedure with a novel *graduated exposure watermark* procedure, which combined a

distracting visual search task, instructions to suppress the associate, and graduated exposure of the cue over time. This graduated exposure task produced a significant suppression effect.

### 2.7.2 Failing to control the activation of the associate memory

There are a number of possible explanations for the null suppression effects in Experiments B2 and B3. Our central hypothesis relates the degree to which the association memory is activating during the weakening phase to the degree of strengthening or weakening of that memory. In order for the memory to be weakened, it must fall consistently in the middle of the range of activation (see Figure 1). This theory about the shape of the nonmonotonic curve could be entirely correct, but if the weakening task fails to reliably keep the associate memory within this middle activation range, we would not see the predicted weakening effect. There are a number of possibilities for why we might not be reliably activating memories to the right degree:

1. *Too strong*

   If the weakening task were to activate the associated memory too strongly (towards the right hand side of the nonmonotonic curve), then the 'weakening' trials would actually be counter-productively strengthening the association.

   Indeed, this is why we designed the weakening task in Experiment B2 to extend the duration of a trial after an intrusion. We hoped that this might give subjects an extra opportunity to bring the activation back down after an intrusion, and break this vicious cycle of intrusions begetting more intrusions. However, this approach of extending trials after intrusions could just as easily backfire (see the discussion on 'too much variability' below).

2. *Too weak*

   On the other hand, if the weakening task were to fail to activate the associated

memory enough (towards the left hand side of the nonmonotonic curve), then the 'weakening' trials would have very little effect. If a memory fails to activate at all, it will be neither strengthened nor weakened.

3. *Too much variability in weakening trials*

   We have considered that our weakening procedure might be activating associations too strongly, or too weakly. Perhaps the most likely possibility is that there was considerable variability in the degree to which the associations were activating from subject to subject, from pair to pair, and from trial to trial. This variability would cause the memories to be weakened sometimes, strengthened sometimes, and other times not affected at all. From introspection, it certainly seemed as though sometimes the background location cue image in Experiment B2 hardly registered at all, while other times it happened to catch one's eye and strongly trigger the association.

   More complicated still, we should take into account the cumulative, sequential effect of the weakening trials. An early strong intrusion could cause a memory to be strengthened, causing it to intrude again and again, getting strengthened each time. Likewise, a moderate activation early in the weakening phase might cause a memory to be weakened enough to stay within that moderate range for the rest of the weakening phase. It might be possible to build a model to capture the cumulative effect of intrusions in this way, but it would probably need a more refined measure of memory activation.

### 2.7.3 Could the stimuli be the problem?

It may be noteworthy that two of the experiments (B2 and B3) that failed to show the below-baseline suppression effect used the location-celebrity image stimuli. In contrast, Experiments B4b and B4a used the face-IAPS scene pairs, and showed significant and nearly-significant below-baseline suppression effects. Likewise, Experiment F7 used different stimuli and also showed nearly significant below-baseline suppression. The only

exception to this is Experiment B1, which used the location-celebrity pairs and showed below-baseline suppression for some of the bins.

Might the location-celebrity stimuli be unsuitable for demonstrating a suppression effect? These celebrity and location images were highly familiar and semantically rich. We know that the brain has specialized and highly differentiated representations for people and places. Furthermore, subjects were encouraged to form elaborative encodings between the location and celebrity associations - indeed, visualizing a person in a location is easy to do, and is widely used as a highly-effective technique for forming strong and distinguishable memories (Yates, 1966). They pictured the celebrities performing silly or memorable actions, creating strong, vivid, episodic memories.

Taken together, these facts suggest a number of possible explanations for why these stimuli might (in retrospect) be ill-suited to experiments on memory suppression :

1. There could be considerable variability in encoding of the different pairs. Some celebrities and locations simply make for memorable associations. For instance, remembering Michael Jordan in the basketball court is likely to be easy. This encoding variability is almost impossible to avoid with these stimuli.

   Since the pairings and condition groupings were all randomized separately for each subject, there's no reason to think that this encoding variability would systematically affect our results. However, if the encoding variability were to account for a much greater portion of the variance than the effect of interest, it might simply obscure it.

   We did attempt to minimize the encoding variability as best we could, by:

   (a) Introducing primacy and recency filler pairs.

   (b) Pre-filtering the stimuli separately for each subject to make sure that all the celebrities were familiar in later experiments.

   (c) Checking the across-subject recall performance for individual stimuli to ensure there were no particular items that were much easier or harder than others to

remember.

(d) Using a learn-to-criterion procedure during the study phase. Once a pair had been learned, it was not studied any more, to try and ensure that each pair was learned to the same fixed criterion level.

But these measures may not have been sufficient.

2. Because these location-celebrity pairings form such strong memories, we may be facing a ceiling effect. In other words, it may be that our forgetting tasks are weakening the memories somewhat - but if they are very strong to begin with, the weakening may not be sufficient to pull them below the threshold for successful recollection very often. As a result, the suppression effect may be there, but very small - this, combined with a large degree of encoding variability, would make the suppression effect almost impossible to detect.

   Note that this is a quite separate point from the one raised in Section 2.7.2 regarding the activation during recollection being too strong. In that case, we were suggesting that the suppression task might actually be causing strengthening if the activation of the memory lay at the upper end of the nonmonotonic range. In other words, both of these issues (the ceiling effect, and the possibility of counter-productive strengthening) are related to the strength of the location-celebrity memories, but they are conceptually distinct.

3. Finally, there may be an issue about the nature of the representation of these elaborative, episodic memories. Since both the locations and the celebrities are so semantically rich and highly differentiated, it may be that even after being weakened, the degraded representations were still easily distinguishable. In contrast, less differentiated noun-noun stimuli might become more confusable after weakening, creating competition at retrieval, and highlighting the effect of the suppression.

Given all of these concerns, it is unsurprising that many of the previous successful RIF and

TNT experiments have used noun-noun word pair stimuli (instead of images), since they may make for less rich, weaker and perhaps more suppressible memories.

### 2.7.4 Future work

Given the discussion about the potential issues with location-celebrity stimuli, the most obvious next step would be to re-run the failed experiments using different stimuli. The RSVP task in Experiment B1 shows the most promise in this regard. Indeed, the success of Experiment B4b using the graduated watermark task and the anonymous face cue stimuli provides some cause for optimism in this regard.

Secondly, it would be valuable to determine which aspect(s) of the watermark task (visual search as a distraction task, the instruction to avoid thinking of the associate, or graduated exposure of the cue) are critical to its success.

### 2.7.5 If we cannot control intrusions, perhaps we can at least track them?

Our ability to control the degree of activation of the to-be-forgotten representations is limited. Both between-subjects variables (e.g. sleep, trauma experience) (Levy and Anderson, 2008) and within-subjects variables (e.g. attentiveness, familiarity) conspire against us. Instead, we might be better off accepting that the to-be-forgotten representation is sometimes going to activate too little, sometimes too much, and sometimes just the right amount to be maximally forgotten. If we could measure this activation level, and show that it predicts the degree of subsequent behavioral forgetting, that would provide powerful support for the theory.

The simplest way to measure intrusions would be to ask subjects after each trial how much the to-be-forgotten representation had 'entered their consciousness'. Indeed, this is the procedure adopted by Levy (2008) and in Experiments B2 and F5. Unfortunately, we expect this self-reporting to be imperfect, and more worryingly still, might actually make

it *more* likely that subjects' minds will inadvertently stray towards that which they are seeking to avoid recollecting, in a kind of 'ironic control' failure to suppress the intrusions (Wegner, 1994).

Better still, we would like to have a neural measure of the degree to which the to-be-forgotten representation is active. In the remaining chapters, we discuss our attempts to use fMRI as just such a covert, neural measure of the activity of the to-be-forgotten representation, so that we might relate this to, and predict, the degree of subsequent forgetting.

# 3 Early attempts to use fMRI as a covert measure of memory activation

## 3.1 Introduction - measuring the degree of activation of the to-be-forgotten representations

The behavioral experiments described in Chapter 2 were designed to produce a below-baseline suppression effect by controlling the degree to which the associate memories activated. We varied the suppression task, the design, the stimuli, reporting of intrusions and the dependent measure of forgetting. Despite our efforts to find an experimental design that would activate the memory reliably at just the right level to cause forgetting, we remained concerned that the to-be-forgotten associate memory was activating too strongly on some trials, and too weakly on others. We believe this variability in activation during no-think trials might be a big part of the reason that it is so hard to reliably and robustly elicit the suppression effect (see Section 2.7.2).

In response, we attempted to address this concern about variability in memory activation with fMRI. Introducing imaging obviously will not reduce the variability, but we hoped that it might allow us at least to *measure* it.

We ran six pilot experiments to optimize the timing, stimuli, design and analysis parameters for this purpose. In this chapter, we consider just two of these pilots (Experiments F5 and F6), and how the lessons learned from them shaped the design of our final experiment (F7).

Finally, in the next two chapters, we will discuss this final, most evolved experiment (F7) for using fMRI to provide a covert, neural measure of a memory's activation with the think/no-think paradigm.

In the next section, we will consider prior neuroimaging work that has motivated the designs and analyses of these experiments, before introducing the MVPA approach we adopted to tackle these questions.

### 3.1.1 Previous work

**Newman & Norman (2010)** Newman and Norman (2010) set out to test the oscillating learning algorithm's (Section 1.3.3) nonmonotonic predictions about how competition drives learning in negative priming using MVPA methods applied to EEG. In their paradigm, subjects were presented with two stimuli at once, and asked to make a judgment about one, while ignoring the other. They predicted that, as a result of this competition, the ignored competing representation would be weakened. As a result, subjects should be slower to make judgments about this previously ignored item than an unseen control item. This basic behavioral negative priming effect has been shown before, though it is small and unreliable (see Section 1.2.4).

Newman and Norman (2010) used classifiers to measure the neural activation of the ignored stimulus. They predicted that the activation level of the competing, ignored stimulus should predict the size of the subsequent negative priming effect - specifically, a moderate level of excitation should show the most negative priming, while a high level of excitation should show the least. Indeed, this is exactly what they found, even showing a (non-significant) hint of positive priming for the ignored stimuli that activated most, as though subjects had accidentally attended to them, allowing them to win the competition. This success motivated our attempt to measure the activation of associated memories in a think/no-think task, and thus predict whether they will be remembered or forgotten (Section 1.4).

**Anderson et al (2004)** Anderson et al. (2004) were the first to run a think/no-think experiment with fMRI, following the paradigm in Anderson and Green (2001) fairly closely. They argued that forgetting involves an active control process, since a variety of cognitive control areas were found to be more active during the no-think than the think trials - these areas included bilateral dorsolateral prefrontal cortex, bilateral ventrolateral prefrontal cortex, BA 45 and BA 46 and the ACC.

The only area significantly less active for no-think than think trials was the hippocampus, perhaps because hippocampal recollection is suppressed by these frontal control processes. In keeping with this idea of hippocampal activity tracking subsequent memory, they found greater hippocampal activity during remembered than forgotten think trials. However, the hippocampus' role seems to be more complex than this. They also found an interaction - while the hippocampus was less active for no-think trials overall, it was more active for the forgotten than remembered no-think trials [10]. This is the opposite of what might be expected if hippocampal activity straightforwardly predicted subsequent recollection.

Anderson et al. (2004) suggest that this greater hippocampal activity for forgotten than remembered no-think trials may reflect momentary intrusions of the suppressed forgotten items during suppression, strongly triggering a control response which in turn dramatically suppressed the no-think memory. Corroborating this, they found that activity in the right dorsolateral prefrontal cortex negatively correlated with hippocampal activity.

**Kuhl et al (2007)** Kuhl et al. (2007) adapted the retrieval-induced forgetting paradigm for fMRI to test hypotheses about the effects of competition-driven learning over time. In their experiment, subjects practiced cue-associate pairs, where each cue had been studied with multiple associates (e.g. ATTIC-dust, ATTIC-junk).

They suggested that during the retrieval practice period, when retrieving 'ATTIC-dust' (the practiced pair), there would be competition from 'ATTIC-junk' (the unpracticed pair). This unpracticed pair would compete at retrieval, lose, and be suppressed. After multiple such retrieval practice trials, there should be less competition from the (now weaker) unpracticed pair.

If conflict detection and resolution processes are recruited during retrieval practice competition, we would expect them to be highly active early in the retrieval practice phase, but less and less active as the competition diminishes after multiple retrieval practice trials.

[10] Furthermore, this effect was bigger for subjects who showed a large below-baseline suppression effect for no-think trials.

Indeed, they found that each subject's overall below-baseline suppression score (similar to the one described in Section 4.3) correlated with their reduction in ACC and right vlPFC over the course of the retrieval practice phase. They also showed that hippocampal activation early in the retrieval practice phase correlated with initial engagement of the ACC and later competitor forgetting. Taken together, this would fit with the idea that early competition at retrieval triggers conflict detection and control responses, which in turn lead to suppression over multiple trials of the intrusive, competing recollection, eventually leading to less conflict detected and control exerted.

However, this does not conclusively establish that these frontal processes are inhibitory. As discussed in Sections 1.3.1 and 1.3.3, prefrontal cortex's cognitive control role may be to provide targeted excitation, rather than inhibition (Miller and Cohen, 2001).

### 3.1.2 Multi-voxel pattern analysis (MVPA)

In all of our fMRI experiments, we make heavy use of the multi-voxel pattern analysis (MVPA) approach for analyzing fMRI data (Norman et al., 2006). Until recently, the conventional approach for analyzing fMRI datasets was to run a mass-univariate set of statistical tests to generate brain maps, based on the general linear model (Worsley and Friston, 1995), with each statistical test being run separately on each voxel. In contrast, MVPA analyses incorporate information from multiple voxels simultaneously, usually through the use of machine learning algorithms - we will focus here on the use of pattern classifiers.

**What is a pattern classifier?** Classifiers are algorithms that learn to discriminate between different classes of patterns - they can be used to ask 'how can we recognize a pattern of brain activity as being associated with one cognitive state versus another?' by learning which patterns of voxel activity are predictive of one class of cognitive states vs another. Once trained, they can be used to guess which cognitive state is associated with a given

pattern of brain activity.

Classifiers are trained with a series of labeled observations. For most of these analyses, an observation consists of a subset of voxel activations from a single brain image as the input, and a single output unit per cognitive state category. During training, the output unit activation values are specified by the experimenter, and the classifier learns the set of weights that best map from the input values to these desired output values (Norman et al., 2006; Polyn et al., 2005; Haynes and Rees, 2006). During testing, the classifier is provided with the input values alone, and it generates its 'guessed' activation values for each of its output units.

To make a guess about which class (i.e. cognitive state) is associated with a given brainstate (i.e. pattern of voxel activities in a single brain image), we can simply pick the classifier output unit with the highest activation. However, for many of the analyses described in this and the following chapter, we will make use of these continuous-valued output unit activities directly, rather than simply asking which is the highest.

**Too many voxels, not enough observations** In principle, the classifier will learn larger 'weights' from input voxels that contain useful information for the discrimination being learned, and will learn to ignore (by setting the weights to be smaller) voxels whose activity does not discriminate between the classes. However, we have many more voxels than observations, and so this problem is under-determined. Regularized classifiers deal better with this issue by incorporating a 'penalty' term - this effectively incurs a cost for non-zero weights, inducing the classifier to pick just a small subset of voxels in its solution. Even with regularization though, we have found it to be helpful to use feature selection algorithms to whittle down the number of voxels ('features') that the classifier sees, by excluding voxels that are found to be least significant by a mass-univariate GLM contrast.

**Using MVPA and mass-univariate approaches in complementary ways** Norman et al. (2006) discuss some of the advantages of the MVPA approach. Most critically for our purposes, the classifier provides a sensitive estimate of the degree to which different cognitive states, processes and representations are present (i.e. active) in the brain on an image-by-image or trial-by-trial basis. This is the means by which we will attempt to read out a covert, neural measure of recollections and intrusions during think and no-think trials.

We will still rely on the standard mass-univariate GLM approach for feature selection, defining regions of interest, and for visualization of the regions whose activity differs between cognitive states of interest.

**Peeking** One important caveat needs to be mentioned - when conducting MVPA analyses, it is critical that none of the same observations used to test the classifier are present during training. Otherwise, such 'peeking' spuriously elevates the classifier's apparent ability to generalize (just as students who sneak in the answers to an exam get more answers right without having learned the material), and would invalidate any further conclusions drawn based on the classifier's activation levels.

Likewise, observations that will be used to test the classifier must be hidden from the feature selection algorithm. Otherwise, the same spurious improvement in generalization would be observed, just as a student told exactly which pages of the textbook to read would be at an advantage in an exam, even if they were not told exact questions will be asked.

**Cross-validation** In many of our analyses, we will be training the classifier on one phase of the experiment, and testing it on a different phase. In this case, it is easy to avoid peeking.

However, sometimes, we will want to train and test the classifier on observations from the same phase. In this case, we will carefully separate the observations into distinct training and testing subsets. Following the standard cross-validation procedure (Polyn et al., 2005;

Norman et al., 2006), we will run this procedure multiple times, each time with a fresh classifier, holding out a different subset of the observations as the testing set. In this way, each observation takes a turn at being part of the testing set, to provide a fair estimate of the classifier's ability to generalize.

**Balancing conditions within the training and testing sets** The classifier is seeking statistical regularities in the training set that it can exploit to help it to generalize to the testing set. If, for instance, more of the training observations were drawn from one class than the other, the classifier would learn to be biased *a priori* to guess the more numerous class. For our purposes, this complicates and confounds our ability to interpret the classifier's activations. There are multiple approaches for correcting and adjusting for this imbalance, but we will adopt the simplest and least problematic - we will randomly exclude ('under-sample') observations from the more numerous class(es) to ensure that the classes are balanced, both within the training and testing sets.

## 3.2 Pilot experiment F5 - attempting classification of recall success

### 3.2.1 Introduction

In this pilot experiment, we aimed to calibrate behavioral, scanning and classification parameters to optimize the classification of successful from unsuccessful recalls. In this way, the classifier's activation would provide a measure of memory retrieval strength. We could then apply this measure of retrieval strength to the no-think trials, as an index of the activation of the associate memory. However, this pilot experiment did not include any no-think trials.

### 3.2.2 Data collection methods

**Participants** 7 subjects participated in a paid fMRI experiment.

**Overview** The scanning portion of this experiment consisted of two main phases: a brief study phase, followed by a longer cued recall phase. Participants were scanned during both the presentation and cued recall phases.

**Study phase** Participants were rapidly presented (around 2500ms per trial [11]) with a sequence of 160 location-celebrity associations (e.g. Fountain - Jack Nicholson). Each stimulus was presented just once, and consisted of both an image and a text label, just as in Experiment B1 (Section 2.2.2).

**Context shift phase** In between the presentation and cued recall phase, we introduced a 2-minute imagination/autobiographical memory task (picturing the floor plan of your parents' home) that was intended to introduce a context shift, and slightly reduce recollection performance in the cued recall phase.

We ran the anatomical scan during this context shift phase.

**Cued recall phase** Each cued recall trial consisted of:

1. A 4000ms presentation of the cue (e.g. Fountain - ???), during which subjects attempted to form a vivid recollection of the associate.
2. A 4000ms period during which they pressed a button to indicate: 1) successful recollection; 2) lack of recollection, or 3) a false recollection. They were presented with the correct associate at the same time as this response so that they were able to make the judgment. The ordering of the buttons was randomized on each trial.

Each cued recall trial was separated by a 6-8s fixation interval.

[11] This was varied from subject to subject - see Section 3.2.2.

**50% recall performance** We sought to bring subjects' behavioral recall performance close to 50% - this would provide the classifier with many examples of both successful and unsuccessful recalls. To this end, we ran an earlier behavioral practice experiment with separate stimuli to estimate each subject's overall recall performance - we used these results to calibrate the study presentation timing individually for each subject by hand, setting a faster or slower presentation rate to adjust their recall performance during the scanned cued recalls.

**2-back task to familiarize locations** In order to minimize the variability between stimuli due to familiarity, we ran participants in a simple 2-back task before the study phase, to familiarize them with the location images.

We also asked each subject to filter out any celebrities from the pool that they were unfamiliar with.

### 3.2.3 Analysis methods

We discuss our scanning parameters, preprocessing and classification procedures in more detail in Section 4.4 and 4.5. A summary description follows.

To select our features, we picked 1000 voxels from across the whole brain whose variability was best explained by a GLM modelling successful and unsuccessful recollections as separate regressors.

We trained a regularized logistic regression classifier [12] to distinguish successful vs unsuccessful recollections (ignoring false recollections), using just the cue presentation volumes at the beginning of each trial, i.e. while participants were striving to recollect the associate.

In order to equate the number of observations in the successful and unsuccessful recall classes (see Section 3.1.2), we used a leave-one-trial-out cross-validation procedure. Specif-

[12] The MVPA Toolbox (Detre et al., 2006) *train_logreg.m* function, with a penalty of 50.

ically, we first threw away surplus observations from the over-represented class to create balanced training sets, keeping aside one observation from each class for testing. We defined a 20-second 'moat' either side of each testing observation that we excluded from both training and testing in order to avoid spurious generalization resulting from the haemodynamic lag. We ran this entire leave-one-trial-out procedure 50 times, each time leaving out a randomly chosen pair of trials (one per class), re-running the GLM to pick the input voxels, and training a fresh classifier.

#### 3.2.4 Results

Across 7 subjects, we were able to classify successful vs unsuccessful recollection with around 70% accuracy (where chance performance would be 50%). Although we tried a large number of different parameterizations of the basic classifier setup (e.g. varying the number of voxels, classifier type, GLM details), none of these alternative versions provided any improvement over our default parameters.

#### 3.2.5 Discussion: issues with Pilot Experiment F5 that prompted design decisions in Pilot Experiment F6

Based on the results from Pilot Experiment F5, we made a number of improvements to our experimental design.

**Switching from classifying recall success to multiple association categories** We were able to classify successful vs unsuccessful recollections significantly above chance (70%), but nonetheless, we had reservations about this approach.

Primarily, we were concerned about our ability to keep behavioral recall performance reliably close to 50%. Our efforts to calibrate each subject's presentation timing to affect behavioral performance proved to be labor-intensive and error-prone, and we worried that

we might sometimes end up with very few training trials (after balancing our two classes) if a subject's performance varied too much towards floor or ceiling.

Secondly, we were concerned that the feature selection and classification might select voxels that would work during training on think trials for distinguishing successful from unsuccessful recall, but that some of these voxels might not generalize well to no-think trials. For instance, the hippocampus has been shown to activate more for successful than unsuccessful recalls in think trials, which would make it seem to be an ideal region of interest for training a classifier. However, its role in no-think trials is more complicated - Anderson et al. (2004) reported that it responded *more* to forgotten than remembered no-think trials (see Section 3.1.1). Finally, the hippocampal response changes over the course of the no-think phase as a function of the amount of suppression (Levy, 2008).

Given these concerns about having enough successful and unsuccessful recall trials, about generalizing from think to no-think, and about the activation profile changing over the course of the no-think phase, we elected to try a different strategy for using classifiers to read out the activation of associate memories.

Until now, we had used celebrity images as our sole association category. In the next experiment, we introduced two new association categories (animals and tools). From now on, instead of trying to train the classifier on successful vs unsuccessful recalls, we would be training the classifier to discriminate between these association categories. Then we could read out the activity of the relevant classifier output unit for the associate in each trial as a measure of its recollection activation, following the approach used by Newman and Norman (2010).

This decision simplified things considerably. Firstly, it meant that we could dispense with the behavioral performance calibration step. Instead of trying to push behavioral performance down to 50%, we could now afford to push it as high as possible, since we'd only be training the classifier to discriminate the association categories on correct think trials.

Secondly, it meant that we could do much more to minimize the encoding variability between pairs. In Pilot Experiment F5, we rapidly presented the pairs at study a single time. We could not test whether subjects had successfully encoded any of them since that act of recollection would have cemented the association (Karpicke and Roediger, 2008). It was impossible then to determine which pairs had been successfully encoded during the study phase without inadvertently boosting recall performance at the same time. Having decided to distinguish between image categories on correct trials, we could afford to allow behavioral performance to improve somewhat. As a result, we could now ensure that subjects studied each pair to criterion during the study phase, minimizing the encoding variability between pairs.

**Maximizing data** We made a number of smaller design decisions to try and maximize the amount of data that we would have to train and test the classifier on.

We kept the inter-trial intervals between study trials small, but grouped trials with the same image association category together into 'miniblocks'. The haemodynamic response would smear the volumes within a miniblock together, but since they would be labeled the same, this wouldn't present a serious problem. This allowed us to shorten the duration of the study phase.

We also tightened up the timing of the wipe trials, so that we might pack more trials into the wipe phase.

Finally, we scanned each subject twice in two separate sessions, in the hope that we might be able to combine the data from the two sessions.

**Lots of analysis options** We wanted to give ourselves a number of options for data analysis. With this design, we could train on the study presentation phase, the study test phase, or the think trials. We could choose to label our trials based on 1) successful and unsuccessful recalls, 2) the three different image association categories, or even 3)

behavioral responses during the think and no-think trials.

## 3.3 Pilot Experiment F6 - first attempt at fMRI think/no-think experiment

### 3.3.1 Introduction

This was our first attempt to run a full think/no-think experiment in the scanner. As described in the previous section, we had learned a number of lessons from our early classification pilots, and so this experiment was designed to maximize the number of think and no-think trials on which to train, to include multiple associate categories, and to allow for multiple classification approaches.

Notably, following Levy (2008), we also asked subjects to report how much the associate memories were intruding during each no-think trial.

### 3.3.2 Data collection methods

**Participants** 5 subjects participated in a paid fMRI experiment spanning 2 sessions.

**Study presentation phase I** Subjects were presented with 60 pairs. Each pair consisted of a location cue, associated with either an animal, a celebrity or a tool. Both stimuli in the pair consisted of an image and a text label. Pairs were grouped into 'miniblocks' of the same associate category - for instance, we might present two tool pairs, followed by two celebrity pairs, followed by two animal pairs, and so on. Each pair was presented for 3750ms with a 250ms inter-trial interval.

**Study test phase** Subjects were tested on each pair once. Each test trial consisted of three parts:

1. A 4000ms cue-only presentation during which subjects were asked to recreate a vivid

and accurate mental image of the animal/celebrity/tool associated with that location cue.

2. A 2000ms response period in which subjects were then shown the correct association, and asked to indicate whether they had correctly recalled it:

   (a) YES = "you could have accurately stated the animal/celebrity/tool's name/label aloud before the correct answer was displayed"

   (b) NO = "if you drew a mental blank"

   (c) MIS = "if you had recollected the wrong thing, or only partially recollected the right answer".

   The orderings of these three button options were randomized on each trial.

3. A 6000ms inter-trial interval during which subjects performed a simple brightness-change fixation task (see Section 4.2.2).

**Think/no-think phase** The 60 pairs were divided into 36 'think' pairs, 12 'no-think' pairs and 12 'baseline' pairs:

Each think trial consisted of three parts:

1. A 4000ms presentation of the location cue alone, outlined in a bright *green* rectangle. This indicated to subjects that they should recreate a vivid and accurate mental image of the animal/celebrity/tool associated with that location cue, just as in the study test trials.

2. A 2000ms response period in which subjects were asked to indicate whether they had correctly recalled the association, much like the study test trials (YES, NO, MIS). Unlike the study test trials, subjects were not shown the correct association during these think trial response periods, and so we did not have any independent verification of the accuracy of their judgments.

3. A 6000ms inter-trial interval during which subjects performed the same fixation task as before.

Each no-think trial consisted of three parts, mirroring the timing and structure of the think trials:

1. A 4000ms presentation of the location cue alone, outlined in a bright *red* rectangle. This indicated to subjects that they should avoid thinking about the associated animal, celebrity or tool. Subjects were asked to perform a simpler version of the 'watermark' task (first described in Section 2.3) during these no-think trials that would simultaneously force them to attend to the location cue while distracting them from recollecting the associate - they were told that some of the location cue images would contain zero, one or two small watermark images of a sun. Their job was to scan over the location image, and press a button every time they noticed one of these watermarks.
2. A 2000ms response period in which subjects were asked to indicate whether they had experienced an intrusion of the associated animal, celebrity or tool.
   (a) NONE = no intrusive recollection of the associated image
   (b) SOME = some recollection of the associated image
   (c) LOTS = strong recollection of the associated image
3. A 6000ms inter-trial interval during which subjects performed the same fixation task as before.

The baseline pairs did not appear at all during the think/no-think phase.

**Recall phase** In this phase, subjects' recollections of the associations were tested, much as in the think trials.

Subjects remained in the scanner for this phase while we ran the anatomical scan. As a result, we do not have functional data for these recalls, and could afford to dispense with the inter-trial interval, leaving just the 4000ms location-only cue and the 2000ms recall response.

**Study presentation phase II** Finally, we re-represented all of the pairs once more, just as in study presentation phase I. We included this extra study presentation run because in the hope that these data would make good training data for the classifier. However, in pilots, we had found that including two study presentation runs as well as a study test run raised the level of behavioral recall performance during the think/no-think phase too high. By including this study presentation phase at the end, we benefited from the extra training data while avoiding the boost in behavioral performance.

Each subject was scanned in this way twice, on separate days. These two sessions were identical in structure, but employed distinct stimuli.

### 3.3.3 Methods - classification

There are many options for training and testing the classifier in this experiment. The main analysis we ran involved training the classifier on the study-presentation and study-test phases, and testing it on the no-think trials. The classifier was trained to discriminate between the three associate categories.

### 3.3.4 Results - classification

Cross-validation performance on the study presentation phase for the three association categories was 51%.

Generalization performance to the think trials when classifying the three association categories was 50%.

Generalization performance to the no-think trials when classifying the three association categories was 37%.

Chance performance would be 33%.

We ran many more analyses than this, but for clarity of exposition, we will not describe them.

### 3.3.5 Discussion: issues with Pilot Experiment F6 that prompted design decisions in Experiment F7

**Behavioral performance was too high, not enough below-baseline suppression** Despite our attempts to keep behavioral performance at least a little below ceiling, we found that subjects performed very well on the final recall phase, and showed very little suppression - in most cases, a correct recollection during the study test phase almost always led to a correct recollection in the final recall. For our final experimental design, we made further modifications to deal with this.

We split data collection across days. Subjects studied the associations in a behavioral phase on day 1, and then were scanned during the think/no-think phase on day 2.

We changed the location/celebrity stimuli to try and make them less vivid, sticky and verbalizable.

We also modified the instructions - instead of urging subjects to use elaborative encoding to form a vivid mental image that associated the associate with the cue, we simply instructed subjects to "form a connection between the noun and the photo so that when you are given the word, you can recall the photo".

**Removing the no-think intrusions judgment** Based on Levy (2008), we expected that we could ask subjects to report whether they had experienced an intrusive recollection during the no-think trials without affecting the below-baseline behavioral suppression effect for

the no-think pairs. However, given the fragility of this below-baseline suppression effect, we removed this intrusion report from the no-think trials, just in case. This also freed up some scanning time, allowing us to include more no-think trials.

**Maximizing think and no-think data** We wanted to maximize the number of think and no-think trials for training and testing the classifier. To this end, we chose not to scan during the study portion of Experiment F7, and devoted the entirety of our scanning time to the think/no-think phase.

We also speeded up the timing of the think and no-think trials. We shortened the inter-trial interval from 6000ms to 4000ms - this involves a difficult tradeoff between being able to include more trials, but potentially contaminating each trial with the haemodynamic lag from the previous trial. Given that we had determined from Pilot Experiments F5 and F6 that the second volume of the 4000ms cue-only presentation provided the highest signal, this would leave at least 6000ms between volumes, which we hoped would be enough. As described above, we also removed the intrusion reports during the no-think trials, but retained the responses for the think trials (though modified to 4-category forced choice judgments).

**Heterogeneity of image categories** We had chosen animals, celebrities and tools as our image categories since we had good reason to believe that these categories are represented saliently by the visual system, and should therefore be distinguishable for the classifier. However, the individual exemplars within each category were unavoidably heterogenous and richly semantically differentiable. Even though George Bush and Britney Spears (say) are both classed as celebrities, the difference in their mental representations might be nearly as large as the difference between representations of George Bush and (say) a fountain pen. In other words, the within-category hetereogeneity of our stimuli might have made it hard for the classifier to define clear boundaries separating the representations of the different categories. To remedy this, we used different stimuli, chosen to have fewer

pre-experimental semantic associations and greater homogeneity within-category: cars, anonymous male faces, bedroom scenes and shoes. As described in Section 4.2.2, we weighted the frequencies of these categories to emphasize the face and scene categories that we expected to have most success classifying.

We also hoped that the introduction of scenes as an associate category would provide another strongly classifiable category besides faces. This would alleviate our concern that classification performance in this experiment was being driven by the identifiability of the face category.

**No-think task** The simple watermark-detection task that we used for the no-think trials was designed to provide subjects with a task that would distract them from recollecting the association, while still forcing them to attend to and process the cue. As in Experiments B2 and B4b, we had hope that this kind of task would control and direct subjects' mental processes more closely than the rather open-ended instructions often used in think/no-think experiments, e.g. "try to avoid allowing the associate to enter your consciousness". Based on preliminary pilot behavioral results with a similar watermark task, we had reason to believe that this might be an effective no-think task - however, as described in Experiment B3, collecting more subjects in the same behavioral paradigm subsequently undermined this result.

So, in order to more faithfully replicate succcessful think/no-think experiments in the literature, we adopted the more straightforward and standard no-think instructions for the next version of the experiment (B. Levy, personal communication, September 6th, 2009).

**Functional localizer run** Since we now planned for the study phase to occur on the previous day (to bring down behavioral performance), this freed up some extra scanning time.

We added a short block-design 1-back task as a 'functional localizer' at the end of the exper-

iment. The data from this task could be used to demarcate functional regions in individual subjects that responded vigorously to the new face and scene associate categories. The data could also be used for feature selection and classifier training.

## 3.4 Discussion

We have described two fMRI pilot experiments that laid the path for our final, most evolved attempt to use fMRI to provide a covert, neural measure of a memory's activation within the think/no-think paradigm.

As a result of this process of evolution, we attempted to incorporate the best aspects of all the experiments so far in Experiment F7. As described in the following chapter, Experiment F7 showed a nearly-significant behavioral below-baseline suppression effect, above-chance classification of association category on no-think trials, and we were able to use the MVPA readout of memory activation for each no-think trials to show the predicted nonmonotonic relationship with subsequent recall.

# 4 Experiment F7 - main fMRI think/no-think experiment

## 4.1 Introduction

In this experiment, we sought to directly test our hypothesis relating memory activation to its subsequent accessibility, as motivated by the previous chapter (Section 3.1). We hoped to measure how much a memory was activating during a think/no-think paradigm, and to predict the likelihood of that memory being correctly recalled in a later cued-recall test. To recap, in Section 1.3.3, we hypothesized that:

1. trials on which the associated memory activated *least* would show *middling* recall performance
2. trials on which the associated memory activated *moderately* would show the *worst* recall performance
3. trials on which the associated memory activated *most* would show the *best* recall performance

In this chapter, we describe the analyses for Experiment F7, designed to test this set of predictions.

Broadly, we tried two distinct approaches in parallel: one based on MVPA pattern classification, and one using regions of interest (ROIs; defined anatomically by hand in conjunction with a functional GLM contrast).

A number of decisions had to be made along the way, each of which could materially affect the results. To determine which of these decisions were consequential, we reran this analysis a few different ways, described below.

## 4.2 Methods - data collection

**Participants** 31 subjects (19 female, aged 18-35) participated in a paid experiment spanning 2 days, advertised as an experiment on 'attention and mental imagery'. All of the subjects were native English speakers, and were drawn from the Princeton community. One subject was excluded for falling fast asleep during the scanning, leaving 30 subjects.

Our paradigm was based on the think/no-think paradigm, first described by Anderson and Green (2001), and later adapted for fMRI by a number of researchers, including Anderson et al. (2004), Depue et al. (2007) and Levy (2008). We discuss the think/no-think paradigm in more detail in Section 1.2.2, and also in Pilot Experiment F6. In essence, subjects learned paired associations, practiced recalling some of them and suppressing others, and then were finally tested on their recall for the associates.

Subjects received printed instructions before each phase of the experiment. They were asked to read these, before discussing them carefully with the experimenter before each phase, to ensure their understanding and compliance.

### 4.2.1 Study phase (day 1, outside the scanner)

On the first day, subjects learned a set of paired associations between words and images. The words were common, imageable nouns (e.g. 'arrow', 'fountain' and 'steamboat'), and the images were photographs drawn from four categories (1/3 faces, 1/3 bedroom scenes, 1/6 cars and 1/6 shoes). See Section 4.2.5 for further details on stimuli choices and generation.

**Initial presentations** Each of the pairs was presented once initially. In each presentation trial, the cue word appeared alone for 1500ms (to ensure that subjects attended to it), and then both the cue word and associate image were presented together for 4000ms - see Figure 16.

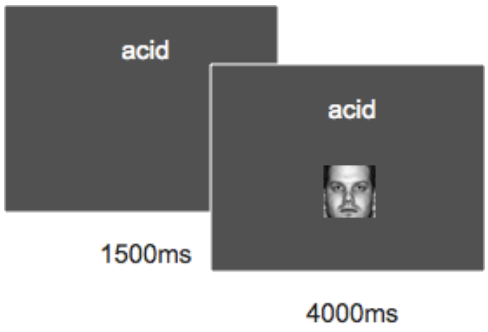


Figure 16: Study phase - initial presentations

**Testing with feedback** For the rest of the study phase, subjects' recollection of each of the paired associates was tested using cued recall in a randomized order. For each pair, they were shown the cue word for 4000ms, then asked to make a 4-alternative forced choice for the category of the associated image (2000ms time limit). If they were correct, they were then asked to make a 4-alternative forced choice between 4 individual, familiar exemplars from that category (2500ms time limit). Both these 4-alternative forced choice tests used button presses and randomized orderings. After each button press, subjects received either a red *'X'* or a green *':)'* as brief feedback (750ms). If their responses on either of these forced-choice cued recall tests were wrong (or too slow), the cue and image paired association were re-presented together for 4000ms. See Figure 17.

In order to minimize encoding variability due to primacy and recency effects, two filler pairs (1 car associate and 1 scene associate) were inserted at the beginning and two more at the end of the presentation run and each testing run - these pairs did not appear at all in the rest of the experiment.

Each time a subject answered both the category and exemplar tests for a pair correctly, that pair was marked as 'correct', and was dropped from further testing. In other words, every pair was tested (with re-presentation for wrong responses) until it had been answered correctly once. This study-to-criterion procedure was designed to enable the formation of reasonably strong associations and to minimize the encoding variability between pairs.

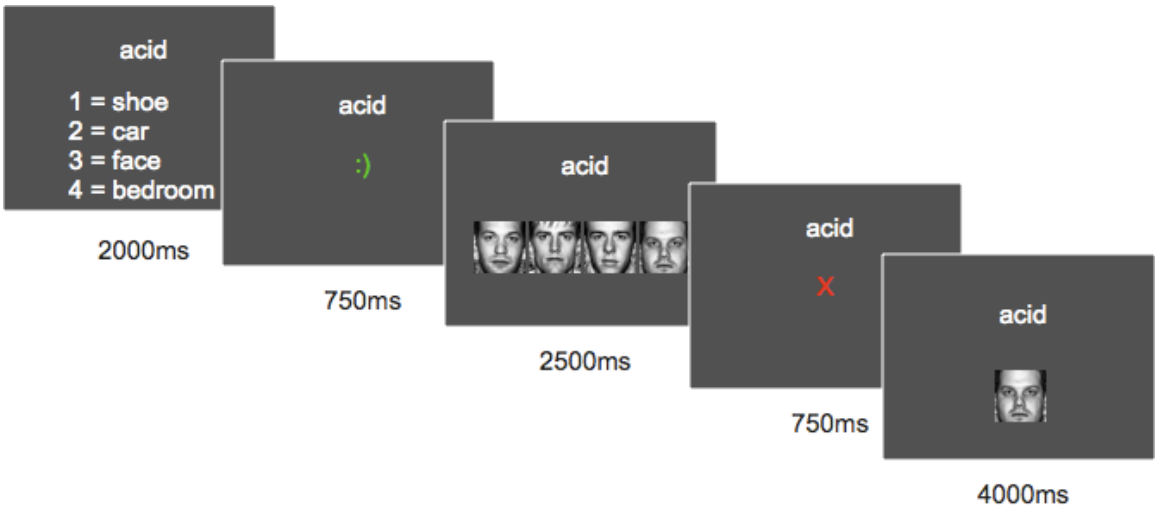


Figure 17: Study phase - testing with feedback. In this case, the subject answered the category correctly but the exemplar incorrectly, and so received feedback.

Following previous think/no-think experiments (B. Levy, personal communication, September 6th, 2009), we aimed for subjects to achieve a behavioral performance of around 70% on the final recall phase (on the next day). Through pilot testing (not described), we adjusted the timing, study-to-criterion and exemplar distinguishability to achieve roughly this level of performance.

### 4.2.2 Think/no-think phase (day 2, inside the scanner)

During the think/no-think phase, the 54 pairs were randomly assigned to either the *think* (36), *no-think* (8) or *baseline* (10) groups. For the think pairs, subjects practiced recalling the associates. For the no-think pairs, they practiced suppressing recollection of the associates. The baseline pairs did not appear at all during this phase. This think/no-think/baseline grouping was the central experimental manipulation, and provided the core data to which the classifier was applied.

The think/no-think phase was divided into 6 runs. Each think pair appeared once per run, and each no-think pair appeared twice, for a total of 6 repetitions per think pair, and 12 repetitions per no-think pair. The associations for the think trials were divided into 1/3

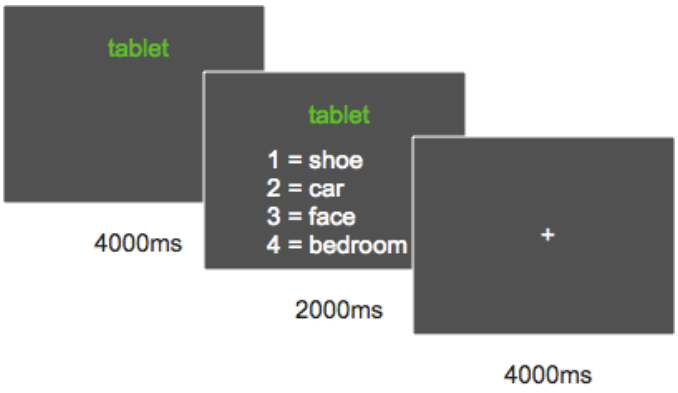


Figure 18: Think trial.

faces, 1/3 scenes, 1/6 cars and 1/6 shoes. The associations for the no-think and baseline trials were divided evenly into 1/2 faces and 1/2 scenes. The ordering of the trials was randomized.

Each think trial consisted of a word-only cue presentation (4000ms), a cued recall test (2000ms), and then a fixation task (4000ms). During the word-only cue presentation, subjects were cued with the word for that pair in green ink and asked to form a vivid and detailed mental image of its associate for as long as the word is on the screen. Then, for the cued recall, they responded to a 4-alternative forced choice with the category of the associate. For the fixation task, subjects were asked to fixate on a small "+" in the center of the screen, and to count silently how many times it changed brightness for as long as the cross remained on the screen. See Figure 18.

Each no-think trial consisted of a word-only cue presentation (4000ms) and then a fixation task (4000ms). During the word-only cue presentation, subjects were cued with the word for that pair in red ink and asked to try as hard as possible to avoid thinking about the associated photograph - to keep it from entering consciousness. Subjects were told that they could accomplish this goal in any way they saw fit, as long as they kept paying attention to and looking at the red word throughout the presentation period. The fixation task was the same as for think trials. See Figure 19.

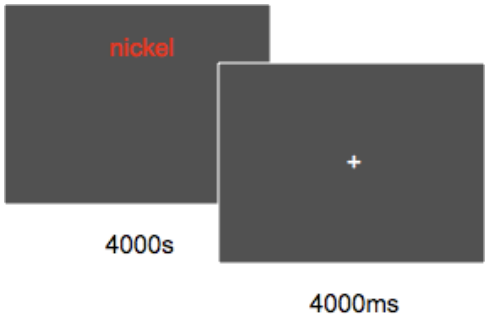


Figure 19: No-think trial.

The think and no-think trials were fairly similar in structure and timing. Both started with a word-only cue period (though with different tasks), and ended with the fixation task. However, the think trials also included the cued recall test in the middle.

Note that there were no image presentations or feedback given during any part of the think/no-think phase.

The detailed instructions for the no-think trials attempted to forestall potential misunderstandings and undesirable strategies, drawing on experience from previous experiments and in consultation with Benjamin Levy (B. Levy, personal communication, September 6th, 2009). Notably, subjects were discouraged from deliberately thinking about the no-think associates at any point during the think/no-think phase and from averting their gaze during the word-only cue period of no-think trials. They were also questioned about their strategies after the experiment to confirm that the instructions had been followed.

### 4.2.3 Functional localizer (day 2, inside the scanner)

In the final functional scanning run, subjects performed a 7-minute 1-back task on images of cars, faces, scenes and shoes. Our aim here was to generate a clean, robust neural signal in response to viewed images that we could use to localize low-level and medium-level posterior visual areas (such as the FFA and PPA), and that could also be used to train the classifier. We had good reason to think that these same areas would also be activated during

mental imagery of the same categories (Yi et al., 2008; Johnson and Johnson, 2009), so that a classifier trained on this functional localizer phase might generalize to the think/no-think phase.

Each image was presented for 1s as part of a 16-image block. Subjects were asked to respond on each trial by button press indicating whether the current image matched the previous. Each block comprised a single category of images, e.g. solely faces. There were 18 blocks in total (6 face, 6 scene, 3 car, 3 shoe). We created three between-subjects counterbalanced 1-back designs, in each case ensuring there were 10 matches in each block, that each exemplar appeared the same number of times as every other in that category, and that every category block followed and was followed by every other roughly the same number of times. Each block was separated by a 10s fixation period to allow the haemodynamic response to subside.

Although the functional localizer stimuli were generated in the same manner and belonged to the same four categories as the association images previously studied, all of the exemplars were novel.

Subjects were instructed to respond on each trial with a button press to indicate whether the current image exactly matched the previous image. These trial-by-trial responses provided a straightforward indication of alertness that helped us pick out inattentive subjects - see Section 4.7.

#### 4.2.4 Behavioral final recall phase (day 2, immediately after the scanning session)

Subjects' recollection of all the pairs was tested in this final phase of the experiment, conducted after all the scanning had been completed. On each trial, subjects were first presented with a word-only cue, in black ink (4000ms). They were then presented with a 4-alternative forced choice for the category of the associated image (2000ms), followed by a 4-alternative forced choice for the individual exemplar (2500ms). No feedback was given.

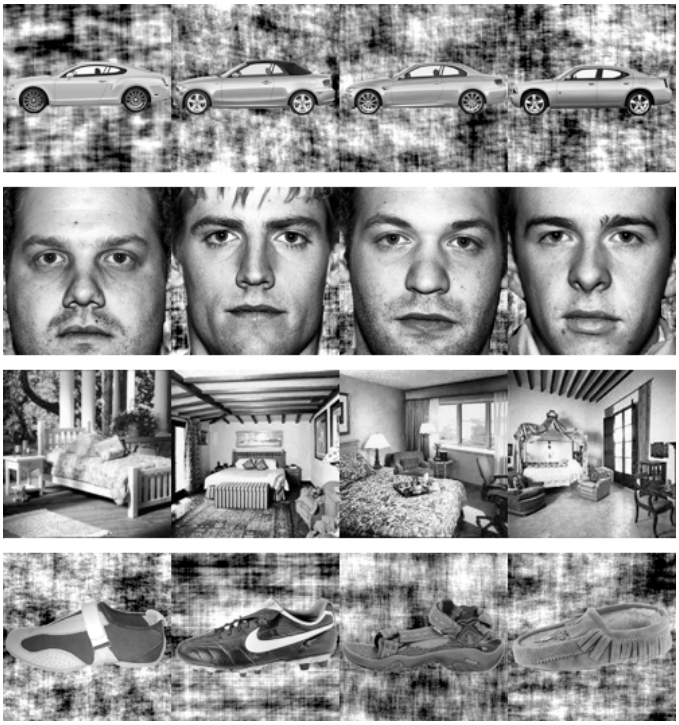

Figure 20: Examples of the car, face, scene and shoe stimuli used in Experiment F7.

A lack of response was marked as incorrect. Unlike the study phase, subjects were always presented with both the category and the exemplar forced choices. Subjects were asked to do their best to recall the associates, even if they had previously been presented in red as no-think pairs, or excluded from the think/no-think phase altogether as baseline pairs.

### 4.2.5 Stimuli

In total, there were 54 pairs in the main experiment: 18 face, 18 scene, 9 car and 9 shoe associates. There were two car pairs and two shoe pairs set aside for use as filler stimuli during the study phase. There were also 10 associations per category for the functional localizer images (reused in each block). All of the association images were black and white photographs. See Figure 20.

The word cues were drawn from the Toronto Word Pool, filtered to include only short, common, imageable nouns, and to exclude nouns that were judged to be semantically re-

lated to any of the image categories (to minimize encoding variability between word/image pairs), leaving a pool of 612 to draw from randomly.

**Faces** The faces were all anonymous and unfamiliar square-cropped male faces.

**Scenes** For our place images, we considered a variety of different kinds of images. Newman and Norman (2010) had shown that images of houses could be classified with EEG and Polyn et al. (2005) had been successful classifying photographs of famous locations with fMRI. However, the cropped house images floated against a blank background, which might activate some of the object-sensitive areas, and we wanted to avoid the heterogeneity and semantic associations of the famous locations. We picked indoor bedroom scenes in an attempt to maximize the response from the PPA (N. Turk-Browne, personal communication, July 20th, 2009). We hoped that the indoor bedroom scenes would maximize the degree and specificity of scene-related neural processing.

**Issues with 2-category classification** However, there is a serious disadvantage to running a classification with just two categories. We wanted to obtain an independent readout for both categories during the no-think trials, so that we could distinguish between the activation of the 'relevant' (corresponding to the image associate's category) and the 'irrelevant' classifier outputs. In a 2-category classification, the independence of the two outputs cannot be assured - if category-A was more identifiable than category-B, the classifier might learn to pick out the category-A trials and treat the category-B trials as *not*-category-A. In such a case, the two classifier outputs are no more independent from one another than the seats on a see-saw. In a sense, the classifier is doing exactly what it is supposed to - it is finding a reliable regularity in its training data which is being reproduced in the statistics of its outputs. After all, if all the trials in the training data belong exclusively to one or the other category, then the weights for the two categories will be strongly negatively correlated.

We considered a number of ways in which we could identify and perhaps compensate for this non-independence of classifier outputs in the 2-category case. For instance, we might include a number of rest (i.e. fixation) trials, where the classifier's outputs would both be set to zeros, i.e. 'none of the above'. But even this might not adequately decorrelate the categories, since the input values in the rest trials would look so different from the presentation trials that the classifier might find a way to suppress its outputs during rest trials without affecting its representations for the two categories. Furthermore, it might require longer breaks between trials to allow the haemodynamic response to more fully subside towards baseline.

Ultimately, the best solution seemed to be to introduce further image categories. By forcing the classifier to find new boundaries in its input space to discriminate between our main two categories as well as these extra categories, we hoped to force it to find features that picked out both face and scene categories. To put this another way, to reduce its training error with multiple categories simultaneously, the classifer could not fall back on the category-A vs not-category-A strategy, but would have to learn to discriminate category-A vs categories-B-C-and-D as well as category-B vs categories-A-C-and-D (not to mention picking out categories C and D too). This should increase the independence of the category-A and category-B outputs.

So, adding further categories should help produce an independent readout for each category. However, for a fixed amount of training data, adding categories would mean fewer observations per category, which would significantly impact generalization performance. Furthermore, these extra categories needed to meet the same criteria as faces and places: well-established classifiability; clearly-defined functional localization; similar activation during visual presentation and mental imagery; and comparable levels of behavioral within-category discriminability. We considered cars, shoes, chairs, body parts, flowers, animals, tools and a number of other categories, but none of them met all these criteria.

In response, we decided to add two new categories (cars and shoes). The cars and shoes

would appear only during the think and functional localizer trials (which we intended to use primarily to train the classifier), and not at all during the no-think trials (whose trials we intended to use to test the classifier). We would only include two classifier output units (for faces and places), using the car/shoe trials as a kind of 'rest'.

Within-category behavioral discriminability for the categories should be roughly the same, for much the same reasons that we wanted to avoid introducing systematic discrepancies in their low-level visual characteristics. If the individual exemplars of one category were particularly hard to remember and distinguish, this might evince a reliably salient neural response during training trials (perhaps involving greater attention, effort or frustration) that might not be present during mental imagery. To equalize this behavioral difficulty across categories, we ran behavioral pilots using the same two-day paradigm as the actual experiment, and adjusted within-category stimulus similarity based on pilot subjects' behavioral performance.

**Controlling the low-level visual characteristics of the image categories** We sought to minimize the systematic discrepancies in low-level visual characteristics between the stimuli in our four image categories. For instance, we describe below how we ensured that the images in all four categories were the same size, overall shape and luminance. Consider what might otherwise happen if, say, all the face images were small and dark while our scenes were large and bright. We might expect that a classifier trained to discriminate the brainstates elicited by visual presentations of faces and scenes would opportunistically pick up on these differences, since features such as size and luminance should be clearly visible in the gross V1 representations for the two categories. However, top-down visualization is likely to be driven by higher-level size- and luminance-invariant representations. As a result, low-level properties such as size and luminance might not be reliably reproduced during mental imagery. Our concern was that such systematic discrepancies in low-level visual characteristics between image categories would then affect how well a classifier trained on visual presentations would generalize to mental imagery.

**Boundary shape and size** The scene photographs were rectangular, yet the cars, faces and shoes all had irregular boundaries and took up differently-sized areas on the screen. To compensate for this, we generated a number of noisy background images by scrambling the Fourier components of the scenes, and placed each car, face and shoe image onto one, making them the same rectangular size and shape as the scenes.

**Overall luminance** Inevitably, the various photographs differed in their luminance profile. In an effort to reduce this, we utilized Matlab's *imadjust* and *adapthiseq* functions to readjust the contrast, normalize the luminance within each 'tile' of the image and then smooth the boundaries between tiles.

To combine the separate boundary shape/size and luminance compensation procedures described above, we first equalized the scene images, generated the scrambled backgrounds, superimposed the other categories on top of the backgrounds, and then ran the luminance equalization for these compound images.

### 4.2.6 Scanning details

The fMRI data were acquired on a Siemens Allegra 3-Tesla scanner at the Center for the Study of Brain, Mind, and Behavior at Princeton University. Anatomical brain images were acquired with a fast (5-minute) MP-RAGE sequence containing 160 sagittally-oriented slices covering the whole brain, with a voxel size of 1.0 x 1.0 x 1.0mm, and a 256mm field of view. Functional images were acquired with an EPI sequence, containing 34 axial slices covering almost the whole brain, collected with a TR of 2000ms, a voxel size of 3.0 x 3.0 x 3.96mm and a 192mm field of view.

The first six runs were for the think/no-think phase (253 volumes each). The 7th run was for the functional localizer phase (238 volumes). The final run was for the anatomical scan.

Each run began with a 10s blank period to allow the scanner signal to stabilize, and ended

with an 8s blank period to allow for the time lag of the haemodynamic response.

In total, we collected 253 volumes for each of the 6 think/no-think functional runs, followed by 238 volumes for the functional localizer run, totaling 1756 functional volumes. Combined with the 5-minute anatomical scan, this amounted to a little over an hour of scanning, excluding breaks between runs and the brief localizer scout and EPI test runs beforehand.

## 4.3 Methods - behavioral analysis

For each pair, we had two 4-alternative forced-choice measures of recollection success during the final recall phase - one for the category of the association, and one for the individual exemplar within that category. There are various criteria that might be used for marking a pair as 'recalled':

1. *category* - if the category response was correct (ignoring the exemplar)
2. *exemplar* - if the exemplar response was correct (ignoring the category)
3. *both category and exemplar* - if both the category *and* the exemplar responses were correct.

For most of the analyses in this chapter, we will consider a pair to have been recalled correctly only if *both* the category and the exemplar responses were correct.

### 4.3.1 Functional localizer

Subjects responded match vs non-match on every single trial of the 1-back task during the functional localizer. We calculated the proportion of times each subject responded correctly, primarily as a means of determining that subjects were still paying attention by the end of the long functional scan.

In Section 4.7, we discuss how we use this functional localizer behavioral data to exclude subjects who were not following the instructions.

## 4.4 Methods - preprocessing and brain maps

### 4.4.1 fMRI preprocessing

The functional data were preprocessed using the AFNI software package (Cox, 1996). Differences in slice timing were corrected by interpolation to align each slice to the same temporal origin. Every functional volume was motion-corrected by registering it to a base volume near the end of the functional localizer (7th) run, which directly preceded the anatomical scan (Cox and Jesmanowicz, 1999). Signal spikes were then smoothed away on a voxel-by-voxel basis. Each voxel's timecourse was normalized into a percentage signal change by subtracting and dividing by its mean (separately for each run), truncating outlier values at 2. The data were smoothed using a Gaussian blur with a full-width half-maximum of 4mm. [13] Baseline, linear and quadratic trends were removed from each voxel's timecourse (separately for each run). The functional data were then imported into Matlab (Mathworks, Natick MA) using the Princeton MVPA toolbox (Detre et al., 2006). In Matlab, each voxel's timecourse was finally z-scored (separately for each run).

A brain-only mask was created (dilated by 2 voxels to ensure no cortex was accidentally excluded).

Each subject's anatomical scan was warped into Talairach space using AFNI's automated *@auto_tlrc* procedure. These rigid-body warp parameters were stored and used later in the group analyses.

[13] As described in 4.10.1, the face vs scene group analysis mask was created using a FWHM of 8mm instead of 4mm, to maximize overlap between subjects.

### 4.4.2 Functional localizer GLM masks

In order to determine which areas' activations differed significantly between faces and scenes, we ran a GLM group analysis on the functional localizer run alone. First, we ran an individual-subjects GLM on the 8mm-smoothed, percent signal change normalized data, prior to detrending and z-scoring. We scoped the analysis to the inside of the brain, adding 'regressors of no interest' for constant baseline, linear trends and quadratic trends. We did not include the motion correction parameters in the GLM, since this rarely helped in previous experiments.

The category labels of the functional localizer blocks were the regressors of interest. These were convolved with SPM's standard model of the haemodynamic response function before being passed into the GLM.

Each of these individual-subjects GLMs were warped into Talairach space (using the warp parameters defined for their anatomical). A two-factor ANOVA was run on each voxel (in Talairach space), with face vs scene as the fixed effect, and the individual subjects as random effects, to determine which areas differed significantly for faces and scenes, across subjects.

### 4.4.3 Intersecting the individual-subjects and group analysis GLMs

Picking the right set of input features for the classifier makes a big difference to its ability to generalize, especially when there are only a small number of noisy training observations. Although the scene-selective regions tended to be fairly consistently located across subjects, the location of the face-selective regions varied more widely.

We wanted to scope our analysis to regions that we had *a priori* reason to believe would contain reliable, generalizable signal, but whittled down to the relevant subset for each individual subject. To do this, we intersected (a) the voxels that passed the face vs scene group analysis using a very liberal threshold with (b) the top 2000 voxels in the individual-

subjects face vs scene GLM.

The number of voxels that remained after these intersections varied widely between subjects. Visual inspection confirmed that they tended to be located in the posterior areas of the cortex that we had expected.

## 4.5 Methods - MVPA-based approach

### 4.5.1 Introduction

All MVPA classification analyses were performed using a combination of AFNI (Cox, 1996) and the Princeton MVPA toolbox in Matlab (Detre et al., 2006). As described in Section 3.1.2, the classifier was trained with labeled example brainstates, and instructed as to which association category (e.g. faces or scenes) was associated with each brainstate. Then, we could ask it to generalize to new brainstates where we did not have an *a priori* sense of how much the representations of the associate categories were activating, such as during the no-think trials.

During the previous pilot experiments, we had experimented with a wide range of preprocessing steps, feature selection methods, classifier algorithms and parameters. For the most part, we found that the details of the classification algorithms and parameters made only a small difference to classification performance. In contrast, the quality and quantity of the training data, the feature selection, and the similarity between the training and testing data, all make a large difference. We will focus our description on our default subset of parameters and approaches that reliably worked well.

### 4.5.2 Preprocessing for classification

All our classification analyses started with the AFNI preprocessing steps described in Section 4.4.1. While the GLM brain maps benefited from 8mm smoothing, the classification

performance was consistently highest when using 4mm smoothing (as reported by Polyn et al., 2005, and consistent with analyses in our pilot experiments). Secondly, we used the percent signal change normalized, detrended, z-scored data for classification - those steps were unnecessary for the GLMs since such artifacts could be modeled away as regressors of no interest.

### 4.5.3 Peeking

Our primary analyses used the functional localizer phase for training data, and the no-think trials within the think/no-think phase as testing data. In supplementary analyses, we tried including the think trials as training data, or (separately), testing on the think trials. We took care to avoid ever training the classifier with the same data on which it would be tested (to avoid the problem of 'peeking' described in Section 3.1.2).

### 4.5.4 Cross-validation

In the few analyses where we wanted to train and test the classifier on different subsets of the same phase, we used the leave-one-out cross-validation procedure described in Section 3.1.2. For instance, when attempting to determine the amount of signal in our functional localizer run, we would train the classifier on 12 of the 18 blocks, and test it on the remaining 6, then repeat this twice more, holding out a different 6 blocks each time.

### 4.5.5 Feature selection

To pick the set of voxels to use as input features, we intersected the individual-subject and group analysis GLM maps (as described in Section 4.4.3) to create a mask for each subject. For the same reason that we kept the training and testing data carefully segregated, this voxel selection step never included any of the testing data (see Section 3.1.2).

### 4.5.6 Ridge regression

Previous experiments had indicated that a regularized logistic regression classifier would perform reliably well on this kind of fMRI data. However, the sigmoidal output of the logistic regression would be binarized, which might skew or obscure the middle of the activation range critical to our hypothesis. Since we wanted to be able to use the output of the classifier as a continuous-valued readout of the activation of the associate category, we used ridge regression in place of the logistic regression (following Newman and Norman, 2010), since it does not apply a sigmoid or other nonlinearity to its output. Usefully, it incorporates an (L2) regularization term (much like the regularized logistic regression), which helps ensure that only the most useful input features get heavily weighted. We benchmarked our basic classification analyses using both logistic and ridge regression to confirm that ridge could perform comparably, before attempting to apply it to our later analyses.

Strictly speaking, ridge regression is not a 'classifier', but rather a regression algorithm. As described, this suited our purposes when reading out continuous-valued output activations for our binning and slope analyses (Section 4.8). When we wanted to be able to treat it as a classifier, we trained a separate ridge regression model for each class, and then applied a straightforward performance metric - for each testing observation, we picked the ridge regression model with the larger output value as the 'classification guess'. We will term the output from these separate ridge regression models as 'output units' for simplicity.

### 4.5.7 Labels

For most of our primary analyses, the classifier was being trained to discriminate between the brainstates elicited by the association categories. In other words, we had one output unit for faces, and another for scenes.

As described in Section 4.2.5, we intended to use the extra car and shoe categories as a kind

of 'rest'. To this end, we included the car and shoe trials as part of the training data, but they did not have their own output units. As a result, the car and shoe trials were labeled with as just having zeros for both the face and scene output units. The car and shoe trials never appeared as testing observations.

### 4.5.8 Timepoints

Both the think and no-think trials started with a word-only cue presentation during which subjects either attempted to recall or suppress the associate image. This was the critical part of the trial on which we wanted to test our classifier, since this is where we expected the signal for the recollective and mental imagery processes to be highest. Based on previous experiments, we picked out just the period 2-4s after cue presentation onset (i.e. the second of the two cue presentation timepoints) to focus our analysis upon. In pilot experiments, this timepoint had consistently yielded the best classification results - we can speculate that the cognitive processes of interest (mental imagery in the think trials, and suppression in the no-think trials) had been fully engaged by this point. For our classifier training labels, we created boxcar regressors spanning just this second volume from each trial, convolved these with the gamma-variate model of the haemodynamic response, and then thresholded them (setting the threshold at half the maximum value in the convolved timecourse). This had a very similar effect to simply shifting the regressors forward by three timepoints (as per Polyn et al., 2005).

## 4.6 Methods - region-of-interest (ROI)-based approach

As an alternative to looking at the timecourse of classifier activations during the no-think phase, we wanted to be able to look directly at the timecourse of the BOLD response in the high-level visual areas that are known to respond to face and scene stimuli presentations and mental imagery. In other words, we wanted to devise a neural readout of associate

activation without using MVPA methods, just by averaging over voxels within carefully defined regions of interest.

First, we ran an individual-subjects GLM on the functional localizer phase, contrasting the 'face' and 'scene' trials. We identified the approximate bilateral locations of the fusiform gyrus, parahippocampal gyrus and retrosplenial cortex based on anatomy.

We picked out the peak face-selective voxel in our *face* > *scene* contrast (in terms of the T-test statistical significance) that lay within the fusiform gyrus. Often, there appeared to be distinct anterior and posterior face-selective regions within the fusiform gyrus - since this dichotomy fit with previous findings (Yi et al., 2008; Johnson and Johnson, 2009), we kept both as regions of interest. Not all subjects showed both anterior and posterior FFA regions, nor did all subjects show bilateral FFA activity, and the location varied somewhat between subjects.

We then picked out the peak scene-selective voxel in our *scene* > *face* contrast that lay within the parahippocampal gyrus in each hemisphere, and the peak voxels in the retrosplenial cortex. These tended to be more consistently placed across subjects.

For each peak voxel, we created a spherical mask centered on that voxel (of radius 2mm) in anatomical resolution, and then downsampled this mask to functional resolution, yielding a small spherical mask containing approximately 7 voxels.

We loaded in the included voxels for the face-related regions of interest, and averaged across them in space. This yielded a single timecourse reflecting face-related activation. We then performed the same averaging for the scene-related regions of interest. These two category-specific timecourses could then be processed just like the two classifier activation timecourses in all of our subsequent analyses, providing a less processed neural signal as a comparison to the classifier output.

## 4.7 Methods - subject exclusion criterion

For our main analyses, we excluded subjects whose behavioral performance on the functional localizer task (Section 4.2.3) lay more than two standard deviations below the mean (mean 82.4%, S.D. 0.14). 3 (of 30) subjects failed to meet this criterion of 55% correct responses. On closer inspection, it became clear that these subjects had not been correctly following the instructions (that had been both written and spoken) for responding during the functional localizer task, and so for our main analyses, we considered only the remaining 27 subjects.

In Sections 4.9, 4.12.1 and 4.13.1, we show that the broad patterns of results were preserved when these 3 subjects were included.

## 4.8 Methods - binning analysis

### 4.8.1 Introduction

Our central prediction nonmonotonically relates the level of activation of a memory to its subsequent accessibility (Figure 1). In the following *binning analysis*, following Newman and Norman (2010), we hoped to read out the activation level of associate representations during no-think trials, and to predict whether those representations would be more or less accesible as a consequence.

In these binning analyses, we grouped together stimulus pairs that activated to roughly the same degree. Then, we could calculate the proportion correctly recalled for each bin. The nonmonotonic hypothesis suggests that the bins for trials in the lower-middle of the activation range should have the worst behavioral recall performance.

### 4.8.2 Overview of the binning analysis

This binning analysis can be broken down into three main stages:

1. *Defining the bin boundaries for each subject* - within each subject, we grouped the trials into bins according to the 'relevant' classifier output.

2. *Running a logistic regression on each subject individually* [14] - within each subject, we trained a logistic regression classifier to classify each pair as remembered or forgotten, based on which of the bins its trials were placed in. Specifically, the classifier inputs for each observation were counts of how many times that pair's trials appeared in each bin. The set of counts for each pair always summed to 12 (the number of repetitions per no-think pair).

3. *Aggregating the betas across subjects* - finally, the logistic regression beta weights per subject were averaged across subjects to create figures such as Figure 24. The more positive the beta weight for a bin, the more likely that pairs whose trials were placed in that bin would be recalled, and vice versa for smaller betas.

We will now describe these steps in more detail.

### 4.8.3 Pulling out the 'relevant' classifier activity for each trial

We tested the classifier on a single volume drawn from each no-think trial (see Section 4.5.8), and noted the activity level of the two (face and scene) output units. For each trial, we can consider the output unit for the category of the associate as the 'relevant' unit, and the other as the 'irrelevant' unit.

[14] Note that this use of a classifier to classify successful-vs-unsuccessful recalls for each pair (based on which bins its trial activations had been placed into), is entirely separate from the prior ridge regression step classifying faces and scenes.

### 4.8.4 Defining the bin boundaries for each subject

We first aggregated all the trials together across pairs but within subjects. Given that each subject was presented with 8 no-think pairs, each of which was repeated 12 times, this totaled 96 no-think trials for each subject.

For each subject, we split these 96 trials into a number of evenly-sized bins based on their 'relevant' classifier activity value.

### 4.8.5 Running a logistic regression on each subject individually

Next, we used the activations for a given pair to predict whether it would be remembered or forgotten. More specifically, we counted how often the 12 trials for a given pair appeared in each of the bins to create a kind of profile of how active that pair's representation had been during the no-think phase. We would predict that a pair whose trial activations tended to land in the middle bins, for example, would be remembered less well than a pair whose trial activations landed in the higher bins.

This stage of the analysis was also run separately for subject. For each subject, for each pair, we counted how often that pair occurred in each of the bins. For instance, the 12 trials for a given pair for a given subject might yield a vector of occurrence-counts such as this [1, 6, 3, 1, 1, 0] for the 6 bin-set. As stated above, we would predict that this pair would be forgotten, since its trial activations tend to lie in the lower-middle range.

For each subject, we trained a regularized logistic regression classifier to predict whether a pair would be forgotten or remembered. Each pair provided a labeled observation consisting of inputs for each bin (the occurrence counts per bin, each ranging from 0-12) and a binary recall-or-forgotten output value. The logistic regression classifier was not tested, and so no observations were withheld, providing a total of 8 observations (pairs) per subject. The value of this classifier lay in its beta weights, one per bin, learned for the remembered-vs-forgotten prediction.

### 4.8.6 Aggregating the betas across subjects

After training a separate logistic regression for each subject, we averaged the sets of betas across subjects (one set per subject), to yield a final set of mean beta weights. These mean beta weights are plotted as the Y values in Figure 24.

### 4.8.7 Why did we pick this binning procedure?

This three-step binning procedure deserves some justification.

**Why bin at all?** In short, we need to bin because our dependent measure (whether the pair was remembered or forgotten in the final recall phase) is binary - however, our prediction relates the level of activation to a (continuous-valued) probability of recall. We needed to aggregate groups of trials together to yield a continuous-valued empirical estimate of probability of recall for each activation range.

**Why did we pick 3-10 bins into which to divide our trials?** At the lower end, we clearly need at least three bins in order to demonstrate a non-monotonic effect. Of course, three bins is probably too few, since that might lump together trials from different portions of the nonmonotonic curve - this aliasing might very well smooth away the effect we are seeking. On the other hand, adding too many bins would dilute our statistical power, since there would be too few trials in each bin to observe reliable differences between them. For this reason, we expected that somewhere between 4 and 8 bins would offer the best compromise - for completeness, we show the results for 3-10 bin-sets.

### 4.8.8 Further details

Some subjects recalled all or none of the no-think pairs correctly. In this case, the logistic regression was unable to learn weights to discriminate the two classes, and these subjects

were excluded from this analysis.

## 4.9 Results - behavioral

### 4.9.1 Requiring both category and exemplar responses to be correct

Behavioral performance on the final recall phase was highest for think pairs (mean 63%, SEM = 0.02), then baseline (mean 58%, SEM = 0.02) and then no-think (mean 54%, SEM = 0.02). See Figure 21b. In this version fo the analysis, we treated a recall as correct only if both the category and exemplar responses were correct.

We wanted to measure whether suppressing recollection during no-think trials would cause forgetting of the no-think pairs. We compared the final recall behavioral performance for the no-think pairs with the baseline pairs (which did not appear at all during the think/no-think phase). The below-baseline suppression comparison between no-think and baseline pairs was clearly trending in the right direction, but not significant ($t(26) = 0.82$, $p > 0.05$).

The above-baseline facilitation effect predicted for think trials is the corollary to the below-baseline suppression of no-think items. The above-baseline facilitation comparison between think and baseline pairs was nearly but not quite significant ($t(26) = 1.48$, $p > 0.05$).

### 4.9.2 Requiring just the exemplar response to be correct

Figure 21a shows the effect of redefining what it means for a pair to be correct, focusing only on the exemplar responses and ignoring the category responses. The broad pattern of final recall phase behavioral results does not change, though the below-baseline suppression is still not significant.

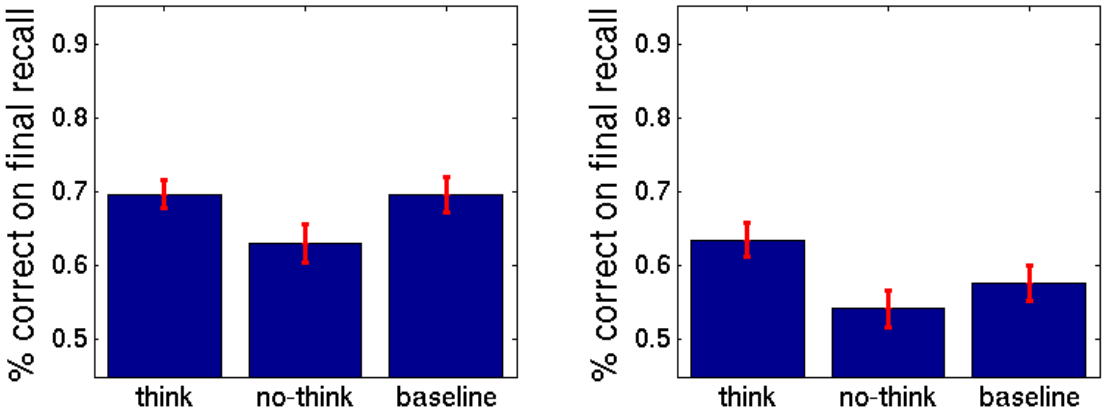


Figure 21: Behavioral performance on the final recall task, modifying the criterion for defining a final recall response as correct: (a) if the exemplar response was correct; (b) if both the category and the exemplar responses were correct.

### 4.9.3 Removing the subject exclusion criterion

When we removed our criterion for excluding subjects (described in Section 4.7, and included all 30 subjects, the pattern of behavioral results remained almost unchanged (see Figure 22). The above-baseline comparison was still not quite significant ($t(29) = 1.53$, $p > 0.05$). The below-baseline comparison moved closer to significance ($5(29) = 1.30$, $p > 0.05$).

## 4.10 Results - brain maps

### 4.10.1 Group analysis GLM

Figure 23 shows the map resulting from running a face vs scene contrast within the functional localizer phase, running a second-level ANOVA group analysis on multiple invidual subject maps (in Talairach space), each smoothed with a Gaussian kernel with a FWHM of 8mm.

As predicted, the group analysis picked out mostly posterior areas, including the fusiform gyrus, parahippocampal gyrus and retrosplenial cortex.

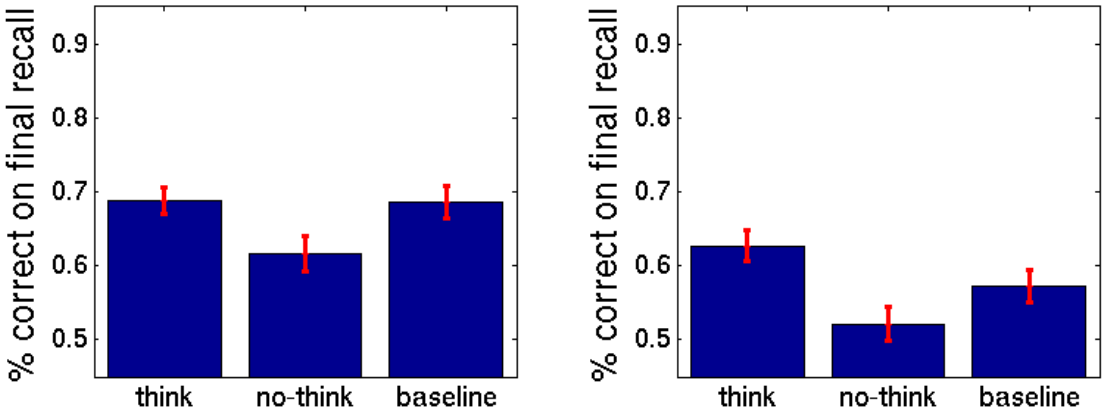


Figure 22: Behavioral performance on the final recall task, across all subjects (n = 30): (a) if the exemplar response was correct; (b) if both the category and the exemplar responses were correct.

## 4.11 Results - classification

### 4.11.1 Can we classify face vs scene during the functional localizer (cross-validation)?

Our ability to discriminate the brainstates for faces and scenes is central to all of our more complex analyses.

The functional localizer phase is the only part of the experiment where subjects are actually being presented with images while being scanned. As a result, this is the phase where we expect the signal to be cleanest, and neural discriminability to be highest. To confirm this, we ran a cross-validation classification within this functional localizer phase, training on 2/3 of the blocks and testing on the remaining 1/3.

Mean cross-validation classification performance for the functional localizer face and scene blocks was 89%, significantly above the 50% level of chance performance ($p < 0.05$).

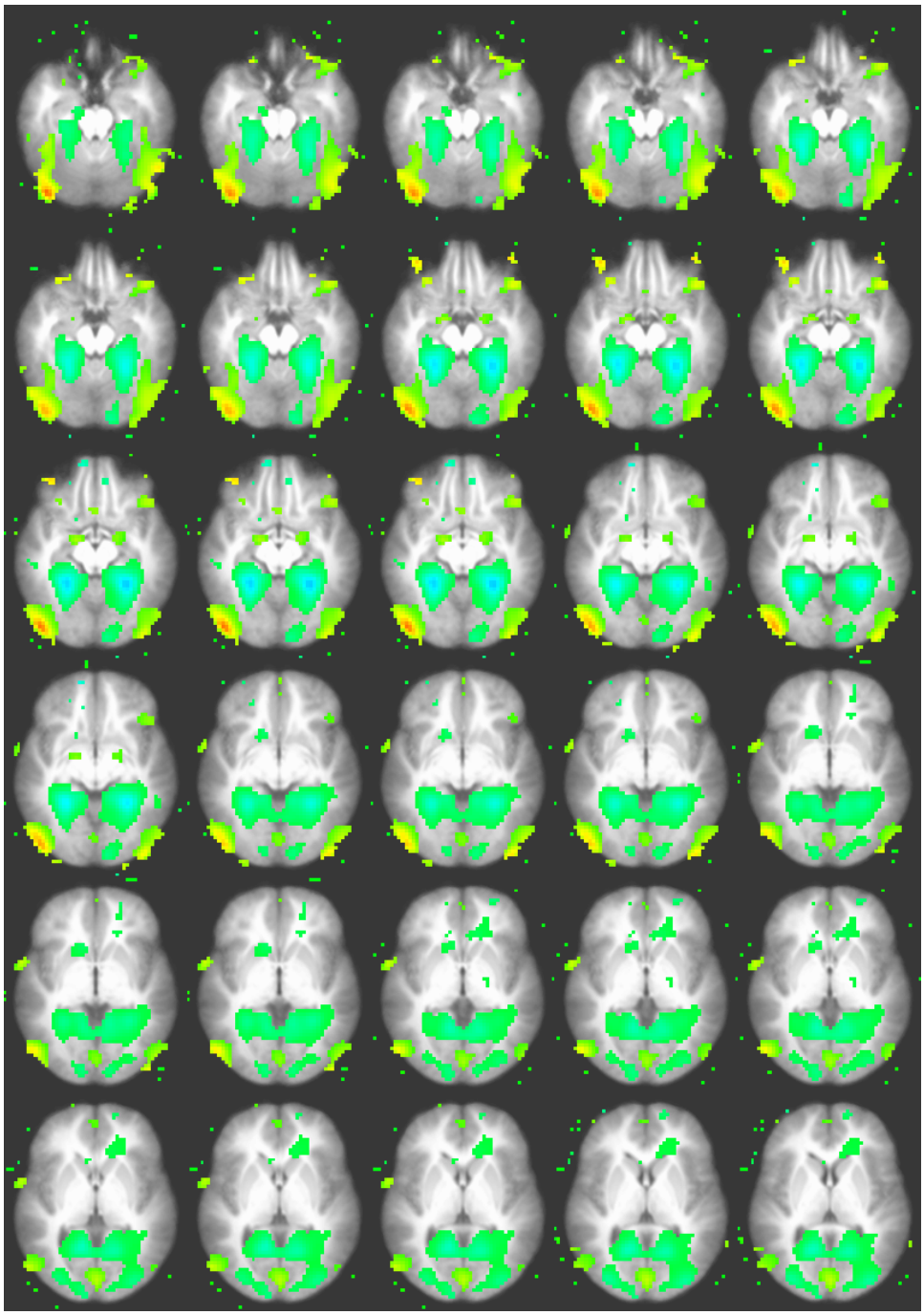

Figure 23: Group analysis with a face > scene contrast on just the functional localizer phase, using an (uncorrected) p threshold of 0.01.

### 4.11.2 Can we generalize from functional localizer phase to the think trials?

Having established a high ceiling for face vs scene classification when training on testing on visual presentations during the functional localizer phase, we next assessed whether a classifier trained on this functional localizer phase could generalize to the think trials. In this case, subjects were attempting to form a vivid and detailed mental image of the associate for the word cue being presented.

Mean classification performance for face vs scene on the think trials was 62%, significantly above the 50% level of chance performance ($p < 0.05$).

### 4.11.3 Can we generalize from the functional localizer phase to the no-think trials?

Finally, we assessed whether a face vs scene classifier trained on the functional localizer phase could generalize to the no-think trials. This is a more difficult discrimination than generalizing to the think trials since subjects are actively trying to suppress recollection of the associate whose category we are trying to determine.

Mean classification performance for face vs scene on the no-think trials was 53%, significantly above the 50% level of chance performance.

## 4.12 Results - binning analysis - MVPA-based approach

In this section, we discuss the results for the main MVPA-based approach (using the relevant classifier output, requiring both category and exemplar responses to be correct, and excluding 30 subjects).

Figures 24 and 25 show the results for the MVPA-based analysis, binning by between 3 and 10 bins. The bins are plotted at equal intervals along the X axis. The Y-axis shows the beta weights of the logistic regressions, averaged across subjects (as calculated in section 4.8.6).

To assess the significance of the non-monotonic U-shape for each middle bin in each bin-

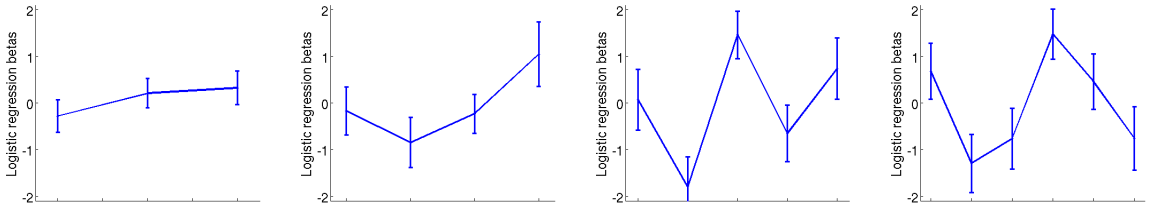


Figure 24: MVPA-based binning analysis, using the relevant classifier output, excluding 3 subjects, MVPA-based approach. Varying the number of bins: (a) 3 bins (b) 4 bins (c) 5 bins d) 6 bins.

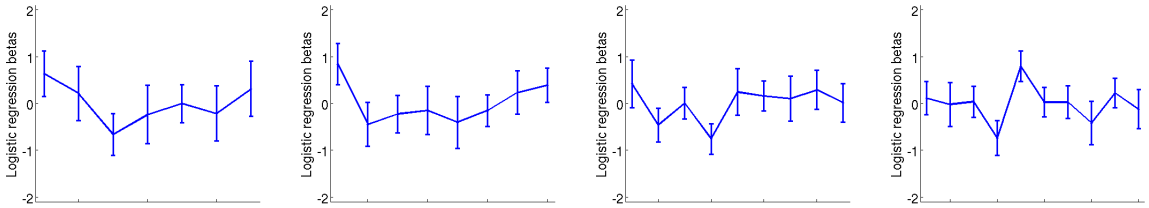


Figure 25: MVPA-based binning analysis, using the relevant classifier output, excluding 3 subjects, MVPA-based approach. Varying the number of bins: (a) 7 bins (b) 8 bins (c) 9 bins (d) 10 bins.

set, we ran paired t-tests comparing the beta weights for the middle (all but the first and last) bins with the beta weights for the outer bins (first and last). Specifically, for each of the middle bins, we ran a one-tailed t-test against both of the outer bins - if *any* of the middle bins were significantly smaller than *both* of the outer bins, that bin was considered significant, and marked with a red circle in Figures 24 and 25. None of the individual middle bins in any of the bin-sets were significantly lower than both of the outer bins.

The error bars shown are standard error bars. However, the across-subject t-tests comparing the betas for each of the middle bins with the betas in the first and last bins conducted were paired samples.

In Experiment B1, we ran a similar kind of binning analysis to group trials based on their cue presentation duration. In order to correct for the multiple comparisons involved in running t-tests on each of the middle bins in each of the bin-sets, we ran a non-parametric permutation test across all of the middle bins and all of the bin-sets, permuting the recalled-

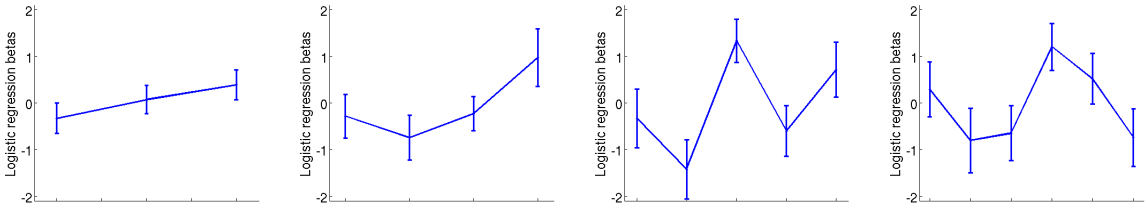


Figure 26: MVPA-based binning analysis - including all 30 subjects. Varying the number of bins: (a) 3 bins (b) 4 bins (c) 5 bins (d) 6 bins.

vs-forgotten values for each pair within subjects. This yielded an overall p-value, reflecting how often we should expect to see middle bins dipping this far below the first and last bins within their bin-set by chance. We applied the same procedure (described in full in section 2.2.3) to determine whether the nonmonotonic effects for this full analysis (run across all the middle bins and multiple bin-sets) were significant.

We found this non-parametric test for the main MVPA-based binning analysis to be non-significant, but very close (200 permutations, p = 0.06).

### 4.12.1 Removing the subject exclusion criterion from the MVPA-based binning analysis

We re-ran the main MVPA-based analysis described in Section 4.12, relaxing our subject exclusion criterion (described in Section 4.7) to include all 30 subjects - see Figures 26 and 27.

As before, none of the individual middle bins in any of the bin-sets were significant. We also ran the non-parametric permutation test procedure (see Section 4.12) to determine the significance of any nonmonotonic effects across all the middle bins in all the bin-sets. In this case, the inclusion of these extra 3 subjects nudged the nonmonotonic effect to become narrowly significant (200 permutations, p = 0.04).

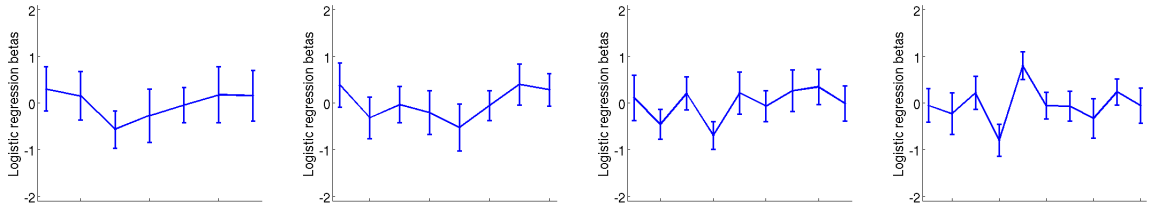


Figure 27: MVPA-based binning analysis - including all 30 subjects. Varying the number of bins: (a) 7 bins (b) 8 bins (c) 9 bins (d) 10 bins.

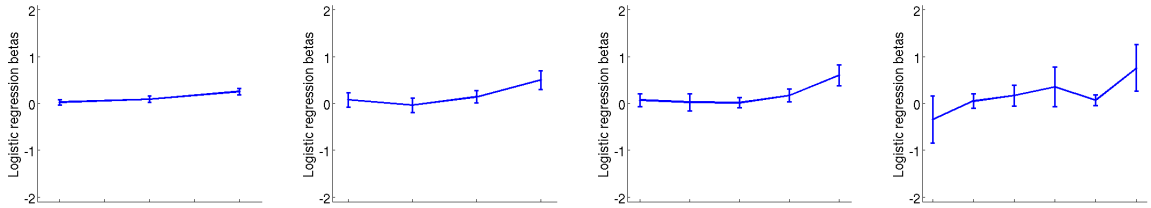


Figure 28: MVPA-based binning analysis - applied to the think rather than the no-think trials. Varying the number of bins: (a) 3 bins (b) 4 bins (c) 5 bins (d) 6 bins.

### 4.12.2 Applying the MVPA-based binning procedure to the think trials

Figures 28 and 29 show the results of testing the classifier on the think (instead of the no-think) trials, and running the binning analysis on these trials.

N.B. due to a technical glitch, two subjects had to be excluded from this particular analysis, leaving 25 instead of 27 subjects.

None of the individual parametric binning analyses were significant, nor was the non-

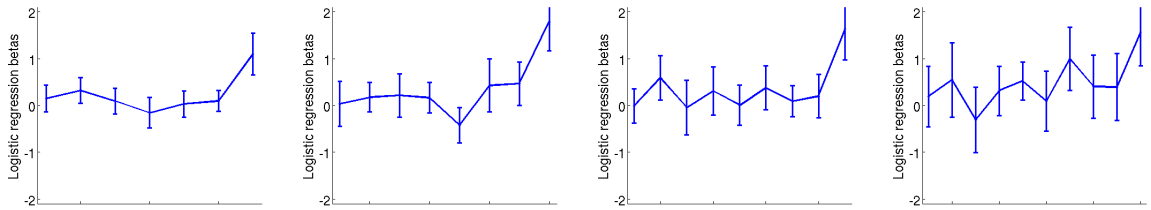


Figure 29: MVPA-based binning analysis - applied to the think rather than the no-think trials. Varying the number of bins: (a) 7 bins (b) 8 bins (c) 9 bins (d) 10 bins.

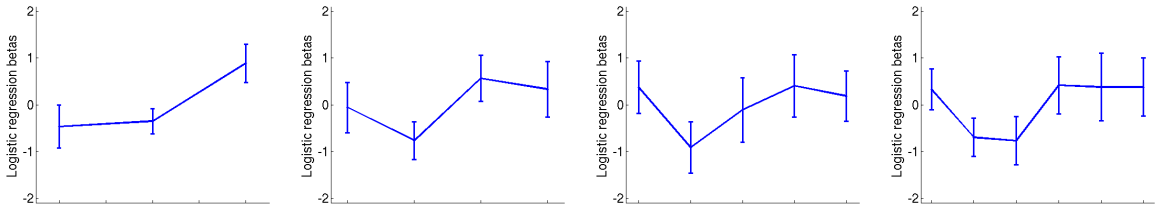


Figure 30: ROI-based binning analysis. Varying the number of bins: (a) 3 bins (b) 4 bins (c) 5 bins (d) 6 bins.

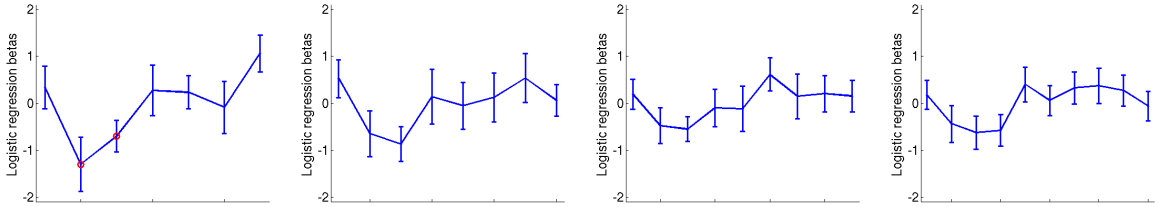


Figure 31: ROI-based binning analysis. Varying the number of bins: (a) 7 bins (b) 8 bins (c) 9 bins (d) 10 bins.

parametic permutation test (200 permutations, p = 0.74).

### 4.13 Results - binning analysis - ROI-based approach

As described in Section 4.6, we wanted to know whether the nomonotonic effect was visible even without the use of a classifier. We followed the main MVPA-based analysis where possible (excluding the same 3 subjects, using the output of the 'relevant' ROI, and requiring both the category and exemplar final recall responses to be correct). Figures 30 and 31 show the average activity within the regions of interest.

As before, we ran t-tests on each of the middle bins for each of the bin-sets. Two of the lower-middle bins in the 7-tile analysis were significantly below both the first and last of the 7-tile bins.

The non-parametric analysis (as described in Section 4.12, based on the RSVP procedure in Section 2.2.3) proved highly significant (200 permutations, $p < 0.01$).

### 4.13.1 Removing the subject exclusion criterion from the ROI-based binning analysis analysis

The non-parametric results for the ROI-based approach remained significant when we relaxed the subject exclusion criterion to include all 30 subjects (200 permutations, $p < 0.01$), though we have not included the figures here.

## 4.14 Discussion

**Summary** In this experiment, we described two main binning analysis pathways (MVPA-based and ROI-based) for reading out a covert, neural measure of the activation of the associate memory during no-think trials. We binned the no-think trials based on this activation measure, then ran a logistic regression to predict whether a given pair would be recalled or forgotten based on which bins it occurred in. A non-parametric permutation test was run on all the middle bins in all of the bin-sets for the following binning analysis variants:

1. The main MVPA-based analysis (excluding 3 subjects, using the 'relevant' classifier output, and requiring both the category and exemplar final recall responses to be correct) was not quite significant.
2. However, when the subject exclusion criterion was relaxed, including all 30 subjects, this version of the MVPA-based analysis became just significant.
3. There was no significant nonmonotonic effect when the MVPA-based binning analysis was applied to the think rather than the no-think trials.
4. The main ROI-based analysis was highly significant.
5. The ROI-based analysis remained highly significant even after relaxing the subject exclusion criterion.

These results suggest that we can indeed use fMRI to provide a covert neural measure of associate activation and that this activation does appear to relate nonmonotonically to the subsequent accessibility of the associate memory.

### 4.14.1 Can we draw rely on the classifier's output when its generalization performance is so low?

In Section 4.11, we reported that generalization performance from the functional localizer phase to the no-think phase was around 53% - above the chance performance level of 50%, but only barely. It is reasonable to think that this might not provide enough signal to be able to make fine-grained discriminations between low, moderate and high activations of the associate.

However, we had good reasons to believe that we might be able to extract useful and graded signal from the classifier before running the binning analysis, even with this level of no-think generalization performance. To begin with, in preliminary analyses (not reported here), we could see clear differences in the classifier outputs for subsequently remembered and forgotten no-think pairs, indicating that the classifier was picking up on something meaningful. More generally, previous papers have successfully used classifiers to make fine-grained discriminations between levels of activity even when classifier performance is only slightly above chance (e.g. Newman and Norman, 2010; McDuff et al., 2009). Even though each individual classifier output value will be very noisy, across the nearly 3000 trials (for up to 30 subjects), we might expect much of this noise to cancel out.

### 4.14.2 Why did the ROI-based analysis work better than the MVPA-based analysis?

We had expected that the MVPA approach would provide a more sensitive readout of the associate activation, and thus be more likely to show any nonmonotonic effects. However, the ROI-based analysis turned out to yield a much more robust nonmonotonic effect, with

much cleaner-looking graphs (compare the ROI-based Figures 30 and 31 with the MVPA-based Figures 24 and 25).

There are a number of differences between the MVPA- and ROI-based analyses that could be giving rise to this difference:

1. The ROI-based analysis operates directly on the BOLD response. In contrast, the classifier first passes this BOLD response through its matrix of learned weights. It could be that some aspect of the classifier training on the functional localizer data is causing it to emphasize features that do not work reliably when applied to estimating the degree to which the no-think associate memories are being recollected.

2. There are dramatically fewer features in the ROI-based analysis. Preliminary MVPA analyses (not reported here) with comparable numbers of voxels tended to yield slightly inferior classification performance, though it may be worth revisiting these analyses in the light of the ROI results.

3. Beyond just the sheer quantity of voxels, the ROI-based analyses picked their features in a very different way. While the MVPA analyses relied entirely on intersecting GLM contrasts, the ROI analysis relied on picking the peak voxel within an anatomical region, defined by hand. It could be that the MVPA GLM group analyses excluded important areas such as the FFA (whose position tended to vary widely between subjects), or that incorporating *a priori* anatomical information helps pick out voxels with reliable signal.

We hope to learn more about why the ROI-based analyses are working so much better by applying classifiers to the ROI features in future analyses. If the ROI features are what makes the difference, then a classifier trained on these features should do at least as well as straight averaging in space.

# 5 General discussion

## 5.1 Summary

We set out to show how cuing, interference at retrieval and weakening might all be related by the way that memories activate and compete at retrieval. We reviewed two main behavioral paradigms, retrieval-induced forgetting and think/no-think. We described three very different accounts of these results: top-down targeted inhibition; the 2-phase interference theory; and the oscillating learning rule. We then argued that all three make a common set of predictions: memories that activate highly win the competition and get strengthened; memories that activate moderately lose the competition and get forgotten; the closer the competition, the greater the consequent strengthening and forgetting; and memories that do not activate do not compete, and are unaffected.

In our 4 behavioral experiments, we attempted to finely control the degree of activation of to-be-forgotten representations by engineering tasks that would moderately activate them. Experiment B4b was successful in showing a significant below-baseline effect for these moderately activated representations. In Experiment B1, the non-parametric permutation test results were promising but not significant.

Over the course of our 3 fMRI experiments, we developed multiple covert, neural measures of memory activation. In Experiment F7, we showed a significant nonmonotonic relationship between the activation of a memory (as measured either with a classifier, or directly from the BOLD activity within regions of interest) and its subsequent recall accessibility.

### 5.1.1 Experiment B1 - RSVP

We first set out to try and map this nonmonotonic activation/accessibility curve behaviorally using an RSVP task (Experiment B1), with cue presentation duration as a proxy for associate activity. The results from this were promising - while there was no overall difference

between the RSVP and the baseline pairs, some of the pairs whose cues were presented for around 200ms showed below-baseline recall. When we corrected fully for multiple comparisons with a non-parametric permutation test, the overall effect across bins and bin-sets was not significant, though fairly close.

More data would help determine whether this below-baseline effect for bins around 200ms is real. We would also like to test whether the unexpected above-baseline performance for very fast presentations is meaningful. Finally, we want to widen the range of presentation durations to see if we can map out the whole shape of the nonmonotonic curve, and show above-baseline performance for the very slow durations.

### 5.1.2 Experiment B4b - graduated exposure watermark task

We introduced the 'graduated exposure watermark' task, incorporating 3 devices in an attempt to control the degree of activation of the associate memories:

1. Counting the superimposed watermark household objects, to provide a distracting alternative to thinking about the scene associate.
2. As in the standard no-think instructions, subjects were also asked to prevent the scene associated with the background face from coming into their minds.
3. By analogy with the 'graduated exposure' approach described in Section 2.6.1, we slowly ramped up the visibility of the background face images over the course of the think/no-think phase.

The graduated exposure watermark task successfully produced a significant suppression effect. This suppression effect was not larger than the effect for the no-think task. However, unlike the no-think task, there are many ways in which the graduated exposure watermark task might be parameterized to make it more reliable and more effective in the future (Section 5.5.2).

### 5.1.3 Experiment F7 - main fMRI think/no-think

While Experiment B1 provided a continuous-valued *independent measure* for *controlling* activation, we can think of Experiment F7 as providing a corresponding continuous-valued *dependent measure* for *reading out* activation.

We aggregated groups of trials based on the activation of the associate (as measured with a classifier, or directly from the BOLD activity in regions of interest), and found a significant nonmonotonic relationship between these activations and the subsequent recall probability with our binning analyses. The nonmonotonic effect was strongest for the ROI-based analyses, but also significant for the MVPA-based analysis when we relaxed the subject exclusion criterion.

## 5.2 Do these results support the nonmonotonic learning hypothesis?

We have presented a number of experiments designed to test the hypothesis that there is a nonmonotonic relationship between the degree of activation and the subsequent accessibility of a representation.

Of the 4 behavioral experiments, only one (Experiment B4b) was significant, after correcting for multiple comparisons. On the face of it, these behavioral null effects cast doubt on the nonmonotonic hypothesis. However, as discussed in Section 2.7, there are a number of potential theoretical and practical obstacles that we think may explain why only 1 of the 4 experiments showed a significant effect. Indeed, it was the difficulty of controlling the trial-to-trial variability in activation that spurred development of the neuroimaging paradigm.

The fMRI results from Experiment F7 were more encouraging. We found significant nonmonotonic effects for the MVPA-based binning analysis when we included all 30 subjects, and for both variants of the ROI-based analysis. These results support the idea that the unreliability of the forgetting effects in the above behavioral experiments (and previously

in the literature - see Section 1.2.4) stems from our inability to control how much the associations activate. Using neuroimaging, we can at least measure this variability, and account for it.

## 5.3 Comparing the three theories that make a nonmonotonic prediction

### 5.3.1 Probing the top-down targeted inhibition account

According to the top-down targeted inhibition account (Levy and Anderson, 2002), cognitive control plays a central role in suppressing no-think responses, causing them to be weakened.

Our results were consistent with this hypothesis. The two experiments that worked best involved explicit instructions to avoid thinking about the associate. In Experiment B4b, the graduated exposure watermark task included this injunction as part of the instructions, along with the watermark counting task to provide a distracting goal to focus on. In Experiment F7, subjects' primary goal during the no-think trials was to avoid thinking about the associate or letting it enter their consciousness.

The earlier behavioral experiments made this suppression instruction less explicit. For instance, in the case of the RSVP task in Experiment B1, subjects were simply told to stay focused on looking for the oddball image, with no mention made of suppressing the associates. These Experiment B1 results were promising, but still not significant. If they had been, they might have suggested that explicit top-down targeted inhibition from cognitive control processes is not necessary for forgetting. As discussed in Section 2.7.4, we think that Experiment B1 could be modified to work more successfully in the future, potentially informing our view on the necessity of top-down targeted inhibition in forgetting.

### 5.3.2 Probing the interference-based account

According to the Tomlinson et al. (2009) interference account, subjects sometimes sampled the location of a memory trace during no-think trials, and then replaced the contents of that memory trace with a new 'sitting quietly' memory. Later, when their recall for this memory was tested, they would often recover the new, interfering 'sitting quietly' memory in place of the original association, and so exhibit below-baseline suppression.

We argued that despite the different mechanisms involved, even this interference-based theory predicted the same nonmonotonic relationship between activation and subsequent accesibility of a memory. Our primary aim has been to bolster the sparse evidence for this nonmonotonic prediction, rather than finding ways to disambiguate the three accounts that generate it (Section 1.3). However, in the next section, we do consider how some variant on the Experiment B2 paradigm might be used to separate out the predictions of the interference-based theory, if the experiment were to yield a significant below-baseline forgetting effect.

**Structural weakening of memory traces** The interference-based theory accounts for the findings of cue-independent forgetting without incorporating structural weakening of memory traces.

In Experiment B2, we attempted to design a paradigm whose findings could not be explained in this way by a pure-interference account, i.e. which could only be explained in terms of memory weakening. Our aim was to show that the 'watermark task' could weaken proactively interfering *A-B* associations, and so make it easier to then learn new *A-C* associations. However, we did not show this release from proactive interference.

We have discussed a number of reasons why our behavioral experiments might have failed to produce forgetting effects, even if the theory motivating them was right (Section 2.7). All we can say is that if a variant of this paradigm were to yield a significant release

from proactive interference, it might be hard to account for without positing structural weakening of memories. After all, creating *extra* watermark-counting memories of a cue should not *reduce* the proactive interference when learning new associations to those cues, unless the watermark task had weakened those associations. In this way, this potentially presents a means of pulling apart the predictions of the weakening accounts from the pure interference-based account.

### 5.3.3 Probing the oscillating learning algorithm account

In Section 1.3.3, we mentioned that the oscillating learning algorithm predicts that no learning will occur at extremely high levels of activation.

We did not expect to see evidence of this return to baseline at the right hand end of the nonmonotonic activity/accessibility curve in any of the experiments described here (and nor did we). In order for the the representations to remain intact, even at the peak of the oscillating inhibition, they have to be extremely strong - and it is difficult to create such strong memories in the laboratory. As a result, this is one of the least-tested predictions of the oscillating learning algorithm.

This prediction is specific to the oscillating learning rule, and is not made by the other members of the nonmonotonic learning rule family (Section 1.3.3), nor by the other two accounts. As a result, it would provide a way to distinguish the various accounts in future.

## 5.4 Future work

### 5.4.1 Stimuli in the behavioral experiments

In chapter 2, we presented 4 behavioral experiments, only two of which produced a significant forgetting effect. In Section 2.7.3, we consider the possibility that the location/celebrity stimuli might be a root cause of these null effects. It would be straightforward to attempt a

close replication of Experiments B2 or B3 using words or less rich, verbalizable and differentiated images (such as the word-face/scene pairs used in Experiment F7), to see whether this simple change makes a noticeable difference.

### 5.4.2 Determining what made the graduated exposure watermark task successful

As discussed, there are three components to the graduated exposure watermark task (counting the watermarks, avoiding recollection of the associate, and graduated exposure of the cue). It would be valuable to determine which of these is most important for producing below-baseline suppression. We would like to run variants of Experiment B4b with modified watermark tasks that only comprise one or two of the devices. Having established their relative contributions, it might then be possible to think about how to tweak them to make the graduated exposure task more reliable and effective. For instance, the rate at which the visibility of the cues was increased might be too slow or too fast, or it might be possible to devise a distractor task better titrates the degree to which subjects process the cue in the background.

### 5.4.3 Adaptive exposure - using real-time MVPA to modulate the graduated exposure

As discussed in 2.6.5, the graduated exposure schedule for increasing the visibility of the cue image was fixed in Experiment 4b, changing slightly every two repetitions. In other words, the graduated exposure was not adaptively driven by the strength of the representation we sought to suppress.

In Experiment F7, we demonstrated that we could use classifiers to read out the strength of the memory activation well enough to reveal the shape of the nonmonotonic relationship between activation and learning. If we could apply these same trial-by-trial measures to neuroimaging data in real-time (DeCharms et al., 2005), we might be able to adaptively change the visibility of the cue to maximize competition while minimizing intrusions.

From an experimental point of view, this real-time readout of memory activation would provide a much finer-grained control over memory activation, and could be used to modulate competition to more directly test the nonmonotonic predictions of competition-driven learning.

## 5.5 Concluding remarks

### 5.5.1 Applicability of the nonmonotonic prediction to other domains

The nonmonotonic activity/accessibility curve provides a framework for understanding how competition and level of activation can drive learning. This framework could apply outside the retrieval-induced forgetting and think/no-think paradigms, e.g. to domains such as task switching, cognitive dissonance reduction and metaphor comprehension (Norman et al., 2006). Indeed, we have already discussed Newman and Norman (2010)'s successful demonstration of a nonmonotonic effect in an EEG negative priming experiment.

### 5.5.2 Clinical applications

In Section 2.5, we discussed the possibility that the nonmonotonic predictions might have important clinical applications for therapeutic suppression for patients suffering from disorders such as phobias and post-traumatic stress disorder (Fenstemaker, 2009).

The competition-dependent learning framework places heavy emphasis on the importance of minimizing intrusions, since even a slight increase in the strength of a competing memory could push it above threshold and cause it to be counter-productively strengthened. This issue is even more critical for clinical applications, since the memories being weakened are so especially strong and intrusive.

The standard no-think instructions are very open-ended and hard to parameterize. In contrast, it should be possible to precisely calibrate both the RSVP task (by modifying the

cue presentation speed) and the graduated exposure watermark task (by modifying the visibility of the background cue images) to make them potentially suitable for patients with poor executive control and/or abnormally strong and intrusive emotional memories. Better still, we might be able to combine them with real-time neuroimaging to modulate these parameters adaptively, to help even more in producing approaches of real clinical value to patients.

### 5.5.3 Application to daily life

As discussed in Section 5.2, neuroimaging provides a way to measure the variability of activation on a trial-by-trial basis. By measuring it, we can account for this variability and predict when we should see subsequent forgetting, but not affect it.

However, if our aim is to apply our predictions to clinical situations or daily life, we will need to be able to more straightforwardly and reliably control this variability. New behavioral paradigms (e.g. based on our RSVP and graduated exposure watermark tasks) might provide greater control over memory activation, but if they require careful presentations of external stimuli they will be limited in their usefulness for everyday life. For now then, perhaps the most important take-home message is to try one's best not to let the to-be-forgotten memories activate strongly, otherwise they may well get strengthened.

## References


Algarabel SLJV, Martinez JL (2006) Inhibitory voluntary control of memory: Effect of stimulus onset asynchrony on reaction time to suppressed memories. *Psicologica* 27:57–77.

Anderson MC, Bjork RA, Bjork EL (1994) Remembering can cause forgetting: retrieval dynamics in long-term memory. *Journal of Experimental Psychology. Learning, memory, and cognition* 20:1063–87.

Anderson MC, Green C (2001) Suppressing unwanted memories by executive control. *Nature* 410:366–369.

Anderson MC, Spellman BA (1995) On the status of inhibitory mechanisms in cognition: Memory retrieval as a model case. *Psychological Review* 102:68–100.

Anderson MC, Bjork EL, Bjork RA (2000) Retrieval-induced forgetting: evidence for a recall-specific mechanis. *Psychonomic Bulletin and Review* 7:522–530.

Anderson MC, Ochsner KN, Kuhl B, Cooper J, Robertson E, Gabrieli SW, Glover GH, Gabrieli JDE (2004) Neural systems underlying the suppression of unwanted memories. *Science* 303:232–235.

Barnes JM, Underwood BJ (1959) "Fate" of first-list associations in transfer theory. *Journal of Experimental Psychology* 58, 2:97–105.

Bauml K (1996) Revisiting an old issue: Retroactive interference as a function of the degree of original and interpolated learning. *Psychonomic Bulletin and Review* 3:380–384.

Bauml KH (2002) Semantic Generation Can Cause Episodic Forgetting. *Psychological Science* 13:356–360.

Bergström ZM, Velmans M, de Fockert J, Richardson-Klavehn A (2007) ERP evidence for successful voluntary avoidance of conscious recollection. *Brain research* 1151:119–33.

Bienenstock E, Cooper L, Munro P (1982) Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex. *J. Neuroscience* 2:32–48.

Bulevich JB, Roediger HL, Balota DA, Butler AC (2006) Failures to find suppression of episodic memories in the think/no-think paradigm. *Memory & Cognition* 34:1569–1577.

Butler KM, Williams CC, Zacks RT, Maki RH (2001) A Limit on Retrieval-Induced Forgetting. *Cognition* 27:1314–1319.

Camp G, Pecher D, Schmidt HG (2007) No retrieval-induced forgetting using item-specific independent cues: evidence against a general inhibitory account. *Journal of Experimental Psychology. Learning, memory, and cognition* 33:950–8.

Carroll KT (2009) Investigating Competition-Dependent Learning: The Effect of Retrieval Practice on Memory for Unrelated Episodic Associations (Undergraduate thesis, Princeton University).

Chun MM, Potter MC (1995) A two-stage model for multiple target detection in rapid serial visual presentation. *Journal of experimental psychology. Human perception and performance* 21:109–27.

Cox RW, Jesmanowicz A (1999) Real-time 3D image registration for functional MRI. *Magnetic resonance in medicine* .

Cox RW (1996) AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research* 29:162–173.

de Zubicaray GI, Andrew C, Zelaya FO, Williams SC, Dumanoir C (2000) Motor response suppression and the prepotent tendency to respond: a parametric fMRI study. *Neuropsychologia* 38:1280–91.

DeCharms RC, Maeda F, Glover GH, Ludlow D, Pauly JM, Soneji D, Gabrieli JDE, Mackey SC (2005) Control over brain activation and pain learned by using real-time

functional MRI. *Proceedings of the National Academy of Sciences of the United States of America* 102:18626–31.

Depue BE, Banich MT, Curran T (2006) Suppression of emotional and nonemotional content in memory: effects of repetition on cognitive control. *Psychological Science* 17:441–447.

Depue BE, Curran T, Banich MT (2007) Prefrontal regions orchestrate suppression of emotional memories via a two-phase process. *Science (New York, N.Y.)* 317:215–9.

Detre GJ, Polyn SM, Moore CD, Natu VS, Singer BD, Cohen JD, Haxby JV, Norman KA (2006) The Multi-Voxel Pattern Analysis (MVPA) toolbox In *Organization of Human Brain Mapping (Florence, 2006)*.

Fenstemaker SC (2009) Cognitive control of valenced memories: intentional forgetting and its clinical implications (Undergraduate thesis, Princeton University).

Geller AS, Schlefer IK, Sederberg PB, Jacobs J, Kahana MJ (2007) PyEPL: a cross-platform experiment-programming library. *Behavior Research Methods* 39:950–8.

Hansel C, Artola A, Singer W (1996) Different threshold levels of postsynaptic [Ca2+]i have to be reached to induce LTP and LTD in neocortical pyramidal cells. *Journal of Physiology, Paris* 90:317–9.

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

Hertel PT, Calcaterra G (2005) Intentional forgetting benefits from thought substitution. *Psychonomic Bulletin and Review* 12:484–489.

Hertel PT, Gerstle M (2003) Depressive deficits in forgetting. *Psychological Science* 14:573–578.

Johnson MR, Johnson MK (2009) Top-down enhancement and suppression of activity in category-selective extrastriate cortex from an act of reflective attention. *Journal of Cognitive Neuroscience* 21:2320–7.

Karpicke JD, Roediger HL (2008) The critical importance of retrieval for learning. *Science* 319:966–8.

Knight RT, Staines WR, Swick D, Chao LL (1999) Prefrontal cortex regulates inhibition and excitation in distributed neural networks. *Acta psychologica* 101:159–78.

Kuhl BA, Dudukovic NM, Kahn I, Wagner AD (2007) Decreased demands on cognitive control reveal the neural processing benefits of forgetting. *Nature Neuroscience* 10:908–14.

Levy BJ, Anderson MC (2008) Individual differences in the suppression of unwanted memories: The executive deficit hypothesis. *Acta Psychologica* 32:474–488.

Levy B, Anderson M (2002) Inhibitory processes and the control of memory retrieval. *Trends in Cognitive Science* 6:299–305.

Levy BJ (2008) Controlling intrusive memories: behavioral and neural correlates of successful and failed memory suppression Ph.D. diss., University of Oregon.

Logan GD (1994) *On the ability to inhibit thought and action: A users' guide to the stop signal paradigm*, pp. 189–239 Academic Press, San Diego, CA, US.

Maxfield L (1999) Eye Movement Desensitization and Reprocessing: An Empirical Review of the Effectiveness of EMDR as a Treatment for PTSD. *Traumatology* 5:1–17.

McDuff SGR, Frankel HC, Norman KA (2009) Multivoxel pattern analysis reveals increased memory targeting and reduced use of retrieved details during single-agenda source monitoring. *The Journal of neuroscience : the official journal of the Society for Neuroscience* 29:508–16.

Miller EK, Cohen JD (2001) An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience* pp. 167–202.

Minai AA, Levy WB (1994) Setting the Activity Level in Sparse Random Networks. *Neural Computation* 6:85–99.

Newman EL (2008) Testing a model of competition-dependent weakening through pattern classification of EEG. *PhD thesis* .

Newman EL, Norman KA (2010) Moderate Excitation Leads to Weakening of Perceptual Representations. *Cerebral cortex* .

Nichols TE, Holmes AP (2001) Nonparametric Permutation Tests For Functional Neuroimaging : A Primer with Examples 25:1–25.

Nieuwenhuis S, Gilzenrat MS, Holmes BD, Cohen JD (2005) The role of the locus coeruleus in mediating the attentional blink: a neurocomputational theory. *Journal of experimental psychology. General* 134:291–307.

Norman KA, Newman EL, Detre GJ (2007) A neural network model of retrieval-induced forgetting. *Psychological Review* 114:887–953.

Norman KA, Newman EL, Detre GJ, Polyn SM (2006) How inhibitory oscillations can train neural networks and punish competitors. *Neural Computation* 18:1577–610.

Norman KA, Newman EL, Perotte AJ (2005) Methods for reducing interference in the Complementary Learning Systems model: oscillating inhibition and autonomous memory rehearsal. *Neural Networks* 18:1212–28.

Norman KA, Polyn SM, Detre GJ, Haxby JV (2006) Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences* 10:424–30.

O'Reilly RC, Munakata Y (2000) Computational Explorations in Cognitive Neuroscience: Understanding the Mind p. 504.

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

Raaijmakers JGW, Shiffrin RM (1981) Search of associative memory. *Psychological Review* 88.

Raymond JE, Shapiro KL, Arnell KM (1992) Temporary suppression of visual processing in an RSVP task: an attentional blink? . *Journal of experimental psychology. Human perception and performance* 18:849–60.

Sakagami M, Niki H (1994) Spatial selectivity of go/no-go neurons in the monkey prefrontal cortex. *Experimental Brain Research* pp. 165–169.

Senn W, Fusi S (2005) Learning only when necessary: better memories of correlated patterns in networks with bounded synapses. *Neural Computation* 2138:2106–2138.

Shiffrin RM, Steyvers M (1997) A model for recognition memory: REM - retrieving effectively from memory. *Psychonomic Bulletin and Review* 4:145–166.

Storm BC, Bjork EL, Bjork RA (2007) When intended remembering leads to unintended forgetting. *Quarterly Journal of Experimental Psychology* 60:909–915.

Tomlinson TD, Huber DE, Rieth CA, Davelaar EJ (2009) An interference account of cue-independent forgetting in the no-think paradigm. *Proceedings of the National Academy of Sciences* 106:15588–93.

Wegner DM (1994) Ironic processes of mental control. *Psychological Review* 101:32–54.

Williams CC, Zacks RT (2001) Is retrieval-induced forgetting an inhibitory process? *The American Journal of Psychology* 114:329–54.

Wolpe J, Brady JP, Serber M, Agras WS, Liberman RP (1973) The current status of systematic densitization. *The American Journal of Psychiatry* 130:961–965.

Worsley K, Friston K (1995) Analysis of fMRI time-series revisited-again. *NeuroImage* 2:173–181.

Yates F (1966) *The Art of Memory*.

Yi DJ, Turk-Browne NB, Chun MM, Johnson MK (2008) When a thought equals a look: refreshing enhances perceptual memory. *Journal of Cognitive Neuroscience* 20:1371–80.