A Neural Network Model of Retrieval-Induced Forgetting

Kenneth A. Norman, Ehren Newman, & Greg Detre

Department of Psychology
Princeton University
Green Hall
Princeton, NJ 08544

{korman, enewman, gdetre}@princeton.edu

Center for the Study of Brain, Mind, & Behavior Technical Report 06-01

Submitted to Psychological Review

3rd January 2006
Abstract

Retrieval-induced forgetting (RIF) refers to the finding that retrieving a memory can impair subsequent recall of similar memories. Here, we present a new model of how the brain gives rise to RIF. The core of the model is a recently developed neural network learning algorithm (based on neural theta oscillations) that leverages regular oscillations in feedback inhibition to strengthen weak parts of target memories and to weaken competing memories. We use the model to address several puzzling findings from the RIF literature, including: why retrieval practice leads to more forgetting than simply presenting the target item; how RIF is affected by the strength of competing memories and the strength of the target (to-be-retrieved) memory; and why RIF sometimes generalizes to “independent cues”, and sometimes does not. We also use the model to address non-monotonic effects of retrieval practice, whereby repeated practice first helps, then hurts recall of competing memories. For all of these questions, we show that the model can account for existing results, and we generate novel predictions regarding boundary conditions on these results.
## Contents

**Introduction: The puzzle of retrieval-induced forgetting**  
- The scope of the paper  
- RIF basics  
- RIF as competitor punishment  
- Other theories of forgetting  
- Finding RIF in the brain  

**Competitor punishment through oscillating inhibition**  
- Network architecture and the role of inhibition  
- Precis of the learning algorithm  
- Algorithm details  
- Theta oscillations: A possible neural substrate for the oscillating learning algorithm  
- General simulation methods  

**Simulations of retrieval-induced forgetting**  
- RIF simulation methods  
- Key parameters  
- Simulation 1: Basic RIF and retrieval-dependence  
  - Background  
  - Methods  
  - Results  
- Simulation 2: Effects of competitor strength and target strength  
  - Background  
  - Methods  
  - Results: Simulation of Anderson et al. (1994)  
  - Boundary conditions  
  - Summary and discussion of Simulation 2  
- Simulation 3: Comparing different types of independent cues  
  - Background  
  - Methods  
  - Results  
  - Discussion  
  - Boundary Conditions  
- Simulation 4: Nonmonotonic retrieval practice effects and prefrontal cortex  

**General discussion**  
- Theoretical implications  
- How competitive dynamics drive learning  
- Generalized forgetting via weakening of attractor states  
- The role of PFC and cognitive control  
- Comparison to other neural network models  
- Comparison to abstract computational models of memory  

**Conclusions**  

**Acknowledgments**  

**Appendix A: Details of the oscillating learning algorithm**  
- Basic Network Parameters  
- Initializing the Network  

**Appendix B: Effects of partial practice on network weights**  

**References**
Introduction: The puzzle of retrieval-induced forgetting

Over the past decade, several researchers (most prominently, Michael Anderson) have argued that competitors are punished during memory retrieval. According to this view, if a memory receives input from the retrieval cue, but the memory is not ultimately retrieved, then the memory is weakened. Anderson has argued that this retrieval-induced forgetting (RIF) effect is cue-independent (i.e., it generalizes to cues other than the previously utilized retrieval cue) and that it is competition-dependent (i.e., the amount that a memory is punished is proportional to how strongly it competes; see Anderson, 2003 for more discussion of these claims). Anderson and others have marshaled an impressive array of evidence for these principles, although not all studies have obtained results consistent with these claims (e.g., Perfect, Stark, Tree, Moulin, Ahmed, & Hutter, 2004).

The scope of the paper

In this paper, we present a new theory (implemented in neural network form) of how the brain gives rise to RIF effects. The introduction to the paper consists of four parts: In the RIF basics section, we describe the RIF paradigm, and we review evidence for cue-independent forgetting. Next, in the RIF as competitor punishment section, we describe Anderson’s theory of RIF, and review evidence supporting Anderson’s claim that the degree of forgetting is proportional to the degree of competition. In the Other theories of forgetting section, we briefly review Anderson’s arguments regarding why extant RIF results are problematic for other theories of forgetting (in particular, blocking and associative unlearning). Finally, in the Finding RIF in the brain section, we discuss possible neural mechanisms for RIF.

After providing an overview of existing findings and theories, we present our account of RIF. In the Competitor punishment through oscillating inhibition section, we describe a neural network learning algorithm (previously developed by Norman, Newman, Detre, & Polyn, in press) that leverages regular oscillations in neural feedback inhibition to strengthen weak target memories, and to weaken other (non-target) memories. The Norman et al. (in press) paper focused on the functional properties of the oscillating algorithm (how many patterns can it store, etc.). The present manuscript focuses on the psychological implications of the oscillating algorithm.

In the Simulations of retrieval-induced forgetting section, we show that the oscillating algorithm can account for detailed patterns of RIF data (e.g., RIF using independent cues; more RIF in high-competition vs. low-competition situations). More importantly, we show that the model provides a clear account of the boundary conditions on these findings. As such, the model can account for data that are (apparently) inconsistent with Anderson’s account of RIF, as well as data that are consistent with this account. It can also be used to generate novel, testable predictions about factors that will modulate the size of RIF effects.

In the General discussion, we describe how our theory of RIF relates to other theories of forgetting; we provide a summary list of predictions; we describe key challenges for theory; and we discuss how the model can be applied to other domains (besides RIF).

RIF basics

In this section, we describe the basic RIF paradigm and provide a brief overview of evidence for RIF (for a more thorough overview, see Anderson, 2003). In one commonly-used variant of the RIF paradigm (see, e.g., Anderson, Green, & McCulloch, 2000b), participants are given a list of category-exemplar pairs (e.g. Fruit–Apple and Fruit-Pear) one at a time and are told to memorize the pairs. Immediately after viewing the pairs, participants are given a practice phase where they practice retrieving a subset of the items on the list (e.g. they are given Fruit-Pear and must say Pear). After a delay (e.g., 20 minutes), participants’ memory for all of the items on the study list is tested. The paradigm is illustrated in Figure 1.

There are several notable results:

- Memory for practiced items (e.g., Fruit-Pear) is better than memory for control items that were not practiced and have no resemblance to practiced items (e.g., Animal-Sheep).
- Memory for non-practiced items that are similar to practiced items (e.g., Fruit-Apple) is worse than memory for control items.
- Forgetting of items like Fruit-Apple is not limited to situations where Fruit is used as a retrieval cue; forgetting also occurs when memory is tested with semantically related cues that were not presented at practice (e.g., using Red to cue for Apple). Anderson calls this property cue-independent forgetting, although (as discussed in Simulation 3) some types of test cues are more effective at eliciting RIF than others.

This basic pattern (facilitated recall of the practiced item, and cue-independent forgetting of similar, non-practiced items) has been observed when category-plus-one-letter-stem cues (like those depicted in Figure 1) are
used at test (Anderson et al., 2000b), and also when category cues (alone) are used at test (Anderson & Spellman, 1995; Camp, Pecher, & Schmidt, in press). Forgetting has been observed when the “independent cue” is a related extralist item (e.g., study Fruit-Pear, Fruit-Apple; practice Fruit-Pe; cue with “tell me a studied word that is related to Red and starts with A’; Anderson et al., 2000b; see also Carter, 2004). Forgetting has also been observed when the “independent cue” is a related word that was paired with the competitor at study, but not presented at practice (e.g., study Fruit-Pear, Red-Apple; practice Fruit-Pe; cue with Red-A; Anderson & Spellman, 1995). Cue-independent forgetting has also been observed using materials other than category-exemplar pairs (e.g., Anderson & Bell, 2001 used novel sentence stimuli). Finally, Carter (2004) demonstrated cue-independent forgetting in a semantic memory paradigm where the competitor was not itself presented at study (e.g., “hold” is an associate of both “carry” and “keep”; practicing retrieval of carry-drop impairs recall of “hold” given “keep” as a retrieval cue). The above examples are meant to provide a general sense of the kinds of studies that have found cue-independent forgetting; they are not meant to provide an exhaustive list (for other, recent examples of cue-independent forgetting, see, e.g., Veling & van Knippenberg, 2004; Johnson & Anderson, 2004; Shivde & Anderson, 2001).

In light of the aforementioned successes, it is also worth noting one published failure to obtain the basic RIF effect (regardless of cue-independence): Butler, Williams, and Zacks (2001) did not find RIF in a study that used dependent category-plus-two-letter-stem cues (e.g., study Fruit-Apple, practice Fruit-Pe; test with Fruit-Ap). Although it is not completely clear why this study failed to obtain RIF, it may relate to differences in how items were selected (in particular, Butler et al., 2001 may have taken fewer steps to minimize associations between category exemplars; we re-visit the issue of how associations between category exemplars can affect RIF in the General Discussion section).

Also, a recent study by Perfect et al. (2004) failed to find RIF when it used a different kind of independent cue: Instead of using a cue that was semantically related to the competitor itself (e.g., cuing for Apple using Red), they paired the competitor with an semantically unrelated word (e.g., Zinc-Apple) prior to the RIF experiment, and used this “external associate” to cue memory. We discuss reasons why RIF might be smaller for external, semantically unrelated cues (vs. semantically related cues) in Simulation 3.

**RIF as competitor punishment**

Anderson and colleagues explain the basic RIF finding in the following manner: When the practice cue is presented (Fruit-Pe), the Pear memory trace receives the most input (because it matches Fruit and “Pe”) but the Apple trace also receives some input (because it matches Fruit). Because Pear receives more input than Apple, Pear wins the competition to be retrieved, and consequently is strengthened. Because Apple receives a substantial amount of input but loses the competition to be retrieved, it is weakened. Furthermore, Anderson posits that the weakening accrues to the Apple representation itself (not just the connection between Fruit and Apple), thereby explaining why forgetting of Apple generalizes to cues other than Fruit.

The most important prediction of the competition-based account is that reducing the extent to which Apple competes with Pear should reduce forgetting of Apple. Anderson tested this by changing the practice phase
such that, instead of giving participants partial practice cues and asking them to complete the cues (Fruit-Pe), participants were given additional presentations of studied items (Fruit-Pear). We will refer to this latter condition as the full practice condition. The intuition here is that the relative match between the cue and Pear (vs. Apple) is larger in the full practice condition than in the partial practice condition, so there should be less competition between Apple and Pear in the full practice condition. According to the competition-based view, this implies that Apple should be hurt less in the full (vs. partial) practice condition. This was confirmed by Anderson and Shivde (in preparation) (see Blaxton & Neely, 1983; Ciranni & Shimamura, 1999; Anderson, Bjork, & Bjork, 2000a; Shivde & Anderson, 2001; Bauml, 1996, 2002 for related findings). We address the “retrieval-dependence” of RIF in Simulation 1, below.

Another way that Anderson has tested the competition-based account is by manipulating the association strength of the competing category-exemplar pairs. For example, participants might study Fruit-Apple, Fruit-Kiwi, and Fruit-Pear, then practice Fruit-Pe. In this example, strong associates of Fruit (Apple) should compete more than weak associates of Fruit (Kiwi), so strong associates should show more RIF than weak associates. This result was obtained by Anderson, Bjork, and Bjork (1994) (but see Williams & Zacks, 2001 for a failure to replicate the result). We address the effects of competitor strength on RIF in Simulation 2.

Other theories of forgetting

In describing his “competitor punishment” account of RIF, Anderson has been careful to distinguish this account from other theories of RIF, most prominently:

- **Blocking** theories, which posit that impaired recall of Apple is an indirect consequence of strengthening Pear, and that no actual weakening of Apple takes place (e.g., McGeoch, 1936). According to these theories, strengthening Pear at practice hurts subsequent recall of Apple, by increasing the odds that “Pear” will come to mind and block recall of Apple. Some theories of this type are referred to as *ratio rule* theories, because — according to these theories — the probability of recalling a memory is a function of the ratio of the strength of the sought-after memory, compared to other memories; as such, increasing the strength of Pear can impair recall of Apple, even if the actual strength of Apple is unchanged (see Mensink & Raaijmakers, 1988 for discussion of the wide range of phenomena that can be explained by ratio rule theories; see also Rundus, 1973; Anderson, 1983; Raaijmakers & Shiffrin, 1981; Gillund & Shiffrin, 1984).

- **Associative unlearning** theories, which posit that learning at practice involves weakening of the connection between Fruit and Apple (and strengthening of the connection between Fruit and Pear), but the Apple and Pear representations themselves are unaffected (e.g., Melton & Irwin, 1940).

See Anderson (2003) and Anderson and Bjork (1994) for a much more detailed overview of these (and other) theories of RIF. While blocking and associative unlearning theories can account for certain aspects of the RIF data space (e.g., the basic finding that practicing Fruit-Pe hurts participants’ ability to subsequently recall Apple using the cue Fruit-A), other aspects of the RIF data space are more problematic for blocking and associative unlearning theories.

With regard to blocking theories: The key claim of these theories is that forgetting of the competitor (Apple) is a consequence of strengthening of the practiced item (Pear). As such, a given manipulation should boost RIF if and only if that manipulation also boosts target strengthening. Several findings from the RIF literature contradict this prediction. For example, the Ciranni and Shimamura (1999) study (mentioned earlier) that found more RIF after partial vs. full practice also found equivalent levels of target strengthening after partial vs. full practice (for similar results, see, e.g., Anderson et al., 2000a and Anderson & Shivde, in preparation).

With regard to associative unlearning theories: The main prediction of these theories (illustrated in Figure 2) is that forgetting of Apple should be limited to the cue “Fruit” — other cues like Red-A should be able to bypass both the weakened Fruit-Apple association (and the strengthened Fruit-Pear association) and access the intact Apple memory. However, this prediction contradicts the finding (discussed earlier) that forgetting generalizes to cues other than “Fruit” (e.g., Anderson & Spellman, 1995).

In summary: These results suggest that blocking and simple forms of associative unlearning can not fully account for the RIF data space. However, as discussed later, we think that a more sophisticated version of associative unlearning (that operates on “micro-features” of distributed representations, as opposed to word-level concepts) plays an important role in RIF, and we also show that blocking can contribute to RIF in certain circumstances. We re-visit the issue of how our theory relates to Anderson’s theory, blocking, and associative unlearning in the General discussion.
Finding RIF in the brain

The results reviewed above suggest that brain mechanisms responsible for RIF need to be able to weaken memories according to the degree that they compete. Recently, Levy and Anderson (2002) and Anderson (2003) have focused on the possible role of prefrontal cortex (PFC) in mediating competitor punishment. There is a large body of research (see, e.g., Miller & Cohen, 2001) suggesting that PFC plays a role in guiding the on-line dynamics of competition, by providing extra activation to the contextually appropriate response (thereby ensuring that the correct response wins and other responses lose the competition; see Simulation 4 for more discussion of this issue). However, this “biased competition” idea does not address the most salient aspect of RIF: Namely, that losing the competition at time \( n \) affects the accessibility of that memory at future time points. Put another way: The central puzzle of RIF is not why one memory loses the competition (and other one wins). Rather, the puzzle is why losing the competition has lasting effects on the accessibility of that memory. Although there is some debate over exactly how long RIF effects last (e.g., MacLeod & Macrae, 2001), there is widespread agreement that RIF can last for at least 20 minutes (Anderson, 2003).

In this section, we present the core of our theory of RIF: A neural network learning algorithm that specifies how local synaptic modification mechanisms can implement selective weakening of strong competitors, and selective strengthening of weak parts of the to-be-learned (target) memory. In previous work, Norman et al. (in press) mapped out the algorithm’s capacity for storing patterns, and showed that the algorithm’s ability to punish competitors greatly improves its ability to memorize and recall overlapping input patterns (relative to similar algorithms that do not incorporate competitor punishment; this point is discussed in more detail in the General Discussion). While the development of the algorithm was inspired by behavioral data indicating competitor punishment, Norman et al. (in press) did not address the algorithm’s ability to account for this behavioral data. The goal of the present paper is to evaluate how well this algorithm works as a psychological theory, by exploring its ability to account for detailed patterns of RIF data.

The learning algorithm depends critically on oscillations in the strength of neural feedback inhibition. By way of background, we describe the basic structure of the model, and the role of inhibition in regulating excitatory activity in the model. Then, we provide an overview of how the learning algorithm leverages changes in the strength of inhibition to “flush out” strong competitors.
(so they can be punished), and to identify weak parts of target memories (so they can be strengthened). Finally, we provide a more detailed account of how synaptic weights are updated in the model, and we briefly discuss how the algorithm may be implemented in the brain by theta oscillations.

**Network architecture and the role of inhibition**

The simulations described here use a simple, one-layer network with rate-coded units, where all of the units are connected to all other units via modifiable synaptic weights. The simulations all involve training the network on a set of patterns (presented one at a time) and then testing the network’s ability to recall missing pieces of these patterns.

Recurrently connected networks like this one need some way of controlling excitatory activity, such that activity does not spread across the entire network (causing a seizure). In the brain, this problem is solved by inhibitory interneurons; these interneurons enforce a set point on the amount of excitatory activity within a localized region, by sampling the amount of excitatory activity in that region, and sending back a commensurate amount of inhibition (O’Reilly & Munakata, 2000). In our model, we capture this set point dynamic using a \( k \)-winners-take-all (kWTA) inhibition rule, which adjusts inhibition such that the \( k \) units in each layer that receive the most excitatory input are strongly active, and all other units are at most weakly active (O’Reilly & Munakata, 2000; Minai & Levy, 1994). We set \( k \) equal to the number of units in each studied pattern, such that (when kWTA is applied to the network) the best-fitting memory — and only that memory — is active.

**Precis of the learning algorithm**

The learning algorithm is based on the premise that memory retrieval is a competitive process. Thus, in order to optimize retrieval, weakening competing memories is just as important as strengthening target memories. Another key premise is that synaptic modification should be as frugal as possible: While there is a clear overall benefit to weakening competing memories, excessive weakening can have harmful consequences, if it ever becomes necessary to recall those competitors later. Thus, memory weakening should only be applied to non-target memories that are threatening to displace the target memory. Likewise, there is no benefit to strengthening a memory trace if that trace is already strong enough to support robust recall. Thus, strengthening should be limited to weak parts of the target memory (the parts that are most likely to be displaced by competitors).

In order to selectively strengthen weak target units, the model needs a way of identifying which parts of the target memory trace are weak (vs. strong). Likewise, in order to selectively punish strong competitors, the model needs a way of identifying which memories are strong competitors (vs. weak competitors). The learning algorithm achieves these goals by oscillating inhibition above and below its baseline level, and learning based on the resulting changes in activation. The major components of the algorithm are summarized here, and depicted graphically in Figure 3:

- **First**, the target pattern is presented to the network, by applying an external input to each of the units in the target pattern (this input is held constant throughout the entire trial). Given strong external input, the total amount of excitatory input will be larger for target units than non-target units. In this situation, the kWTA rule will set inhibition such that the target units are active, and other (non-target) units are inactive.

- **Second**, the algorithm identifies weak parts of target memories by raising inhibition above the baseline level of inhibition (set by kWTA). This acts as a “stress test” on the target memory. If a target unit is receiving relatively little support from other target units, such that its net input is just above threshold, raising inhibition will trigger a decrease in the activation of that unit. However, if a target unit is receiving strong support from other target units, such that its net input is far above threshold, it will be relatively unaffected by this manipulation.

- **Third**, the algorithm strengthens units that turn off when inhibition is raised (i.e., weak target units) by increasing weights from other active units. By doing this, the learning algorithm ensures that a target unit that drops out on a given trial will receive more input the next time that cue is presented. If the same pattern is presented repeatedly, eventually the input to that unit will increase to the point where it no longer drops out in the high inhibition condition. At this point, the unit should be well-connected to the rest of the target representation (making it possible for the network to “pattern complete” that unit) and no further strengthening will occur.

- **Fourth**, the algorithm identifies competitors by lowering inhibition below the baseline level of inhibition. Effectively, lowering inhibition lowers the threshold amount of excitation needed for a unit to become active. If a non-target unit is just below threshold (i.e., it is receiving strong input, but not quite enough to become active) lowering inhibition

\[ k \text{-winners-take-all (kWTA)} \]
Figure 3: High-level summary of the learning algorithm. For all sub-parts of the figure, target units (labeled with a T) and competitor units (labeled with a C) are ordered according to the amount of excitatory net input they are receiving. White units are active (excitation > inhibition) and black units are inactive (inhibition > excitation). Step 1 depicts what happens when the target pattern is presented to the network. Assuming that the external input (applied to the target units) is strong enough, the total amount of excitatory input will be higher for target units than for competitor units. In this situation, if $k$ = the number of target units, the $k$-winners-take-all rule sets inhibition such that the $k$ target units are above threshold, and competitor units are below threshold. Steps 2 and 3: Raising inhibition causes target units that are just above threshold to turn off; the learning algorithm then acts to strengthen these units. Steps 4 and 5: Lowering inhibition causes competitor units that are just below threshold to become active; the learning algorithm then acts to weaken these units. See text for additional details.
will cause that unit to become active. If a non-target unit is far below threshold (i.e., it is not receiving strong input), it will be relatively unaffected by this manipulation.

- Fifth, the algorithm weakens units that turn on when inhibition is lowered (i.e., strong competitors) by reducing weights from other active units. By doing this, the learning algorithm ensures that a unit that competes on one trial will receive less input the next time that cue is presented. If the same cue is presented repeatedly, eventually the input to that unit will diminish to the point where it no longer activates in the low inhibition condition. At this point, the unit is no longer a competitor, so no further punishment occurs.

**Algorithm details**

The Norman et al. (in press) learning algorithm adjusts connection strengths using the Contrastive Hebbian Learning (CHL) equation (Ackley, Hinton, & Sejnowski, 1985; Hinton & Sejnowski, 1986; Hinton, 1989; Movellan, 1990). CHL involves contrasting a more desirable state of network activity (sometimes called the plus state) with a less desirable state of network activity (sometimes called the minus state). The CHL equation adjusts network weights to strengthen the more desirable state of network activity (so it is more likely to occur in the future) and weaken the less desirable state of network activity (so it is less likely to occur in the future).

\[
dW_{ij} = \text{lrate} \left( X_i^+ Y_j^+ - X_i^- Y_j^- \right)
\]  

(1)

In the above equation, \( X_i \) is the activation of the presynaptic (sending) unit, \( Y_j \) is the activation of the postsynaptic (receiving) unit. The + and - superscripts refer to plus-state and minus-state activity, respectively. \( dW_{ij} \) is the change in weight between the sending and receiving units, and \( \text{lrate} \) is the learning rate parameter.

The description of the oscillating algorithm in Figure 3 shows inhibition changing in discrete jumps (between normal, high, and low inhibition). In the actual model, we implement the learning dynamics shown in Figure 3 by varying inhibition in a continuous, sinusoidal fashion, over the course of multiple time steps. At the outset of each trial, we set inhibition to its normal level (i.e., the level set by kWTA), such that — assuming the target units receive sufficient external input — all of the target units (and only those units) are active; this is the maximally correct state of network activity. Next, we distort the pattern of network activity by continuously oscillating inhibition from its normal level to higher-than-normal, then to lower-than-normal, then back to normal. Network weights are adjusted by applying the CHL equation to successive time steps of network activity. At each point in the inhibitory oscillation, inhibition is either moving toward its normal level (the “maximally correct” state), or it is moving away from this state. If inhibition is moving toward its normal level, then the activity pattern at time \( t + 1 \) will be more correct than the activity pattern at time \( t \). In this situation, we will use the CHL equation to adapt weights to make the pattern of activity at time \( t + 1 \) more like the pattern at time \( t \). However, if inhibition is moving away from its normal level, then the activity pattern at time \( t + 1 \) will be less correct than the activity pattern at time \( t \) (it will either contain too much or too little activity, relative to the target pattern). In this situation, we will use the CHL equation to adapt weights to make the pattern of activity at time \( t + 1 \) more like the pattern at time \( t \). These rules are formalized in Equation 2 and Equation 3.

If inhibition is returning to its normal value:

\[
dW_{ij} = \text{lrate} \left( (X_i(t+1) Y_j(t+1)) - (X_i(t) Y_j(t)) \right)
\]  

(2)

If inhibition is moving away from its normal value:

\[
dW_{ij} = \text{lrate} \left( (X_i(t) Y_j(t)) - (X_i(t+1) Y_j(t+1)) \right)
\]  

(3)

Figure 4 summarizes how the learning algorithm affects target and competitor representations. The algorithm strengthens the connections between target units that drop out (during the high inhibition phase) and other target units. Also, it weakens the connections between competitor units that pop up (during the low inhibition phase) and other units that are active during the low inhibition phase. The net effect of these weight changes is to increase the average degree of interconnectivity between the units in the target pattern, and to decrease the average degree of interconnectivity between the units in the competitor pattern (for evidence that these changes occur in our RIF simulations, see Appendix B).\(^1\)

The increased interconnectivity of the target pattern makes it a stronger attractor in the network: Because target units all send mutual support to one another, it is easier to activate the target pattern (i.e., it is a more “attractive” state of network activity), regardless of the cue. Likewise, the decreased interconnectivity of the

---

\(^1\)This target-strengthening and competitor-weakening is contingent on the assumption that target units are active given normal inhibition (and competitor units are not). If the external input coming into a target unit is weak, it is possible for a competitor unit to displace that target unit. If this occurs, connections coming in to the competitor unit will be strengthened and connections coming in to the target unit will be weakened. Some ramifications of this fact are discussed in Simulation 2 and Simulation 4.
Figure 4: Summary of the oscillating learning algorithm (Norman et al., in press). Part A shows how target and competitor activation change during different phases of the oscillation, and also how the sign of the learning rate varies as a function of oscillatory phase. The high-inhibition part of the oscillation causes some target units to drop out and then reappear; the low-inhibition part of the oscillation causes some competitor units to activate and then disappear. Part B shows how the activation changes in part A affect network weights. To a first approximation, the change in synaptic weights (coming in to a given unit) is a function of the change in that unit’s activation, multiplied by the current learning rate. Applying this heuristic to all four quadrants of the oscillation, the net effect of learning is to increase weights coming into target units, and to reduce weights coming into competitor units. More specifically, target units that dropped out during the high-inhibition phase become better linked to other target units; and competitor units that popped up during the low-inhibition phase are cut off from the target representation (and from each other).
competitor pattern makes it a weaker attractor in the network: Because competitor units do not send strong support to one another, it is easy for the network to slip out of the competitor activity pattern, and into some other pattern. This should hurt the network’s ability to retrieve the competitor pattern, regardless of which cue is used at test (cue-independent forgetting, in Anderson’s parlance). This basic principle (that RIF effects can extend to cues other than the cue used at practice) is demonstrated in all four of the simulations described below. One last point is that, even though the model shows some degree of generalized forgetting, it is also capable of showing different degrees of RIF for different cues (this is addressed in Simulation 3).

Theta oscillations: A possible neural substrate for the oscillating learning algorithm

As discussed in Norman et al. (in press), several findings suggest that theta oscillations (rhythmic changes in local field potential at a frequency of approximately 4 to 8 Hz in humans) could serve as the neural substrate for the oscillating algorithm:

- Theta oscillations depend critically on changes in the firing of inhibitory interneurons (Buzsaki, 2002; Toth, Freund, & Miles, 1997).

- Theta oscillations have been observed in humans in the two structures that are most important for semantic and episodic memory: cortex (e.g., Kahana, Seelig, & Madsen, 2001) and hippocampus (e.g., Ekstrom, Caplan, Ho, Shattuck, Fried, & Kahana, 2005).

- Theta oscillations are fast enough to support several complete oscillations per stimulus presentation, and slow enough to allow competitors to activate when inhibition is lowered.

- Most importantly, theta oscillations have been linked to learning, in both animal and human studies (e.g. Seager, Johnson, Chabot, Asaka, & Berry, 2002; Sederberg, Kahana, Howard, Donner, & Madsen, 2003). Several studies have found that the direction of potentiation (LTP vs. LTD) depends on the phase of theta (peak vs. trough; Huerta & Lisman, 1996; Holscher, Anwyl, & Rowan, 1997; Hyman, Wyble, Goyal, Rossi, & Hasselmo, 2003). This result mirrors the property of our model whereby the high-inhibition phase of the oscillation is primarily concerned with strengthening target memories (LTP) and the low-inhibition phase of the oscillation is primarily concerned with weakening competitors (LTD).

At this point, the linkage to theta is only suggestive. However, if we take the linkage seriously, it leads to several predictions that should (in principle) be testable using human electrophysiology. These predictions are described in the Neurophysiological predictions section at the end of the paper.

General simulation methods

All of our simulations were conducted in a one-layer network with 80 units (except for Simulation 3, which used 120 units).

Inhibition was oscillated by adding an oscillating component (at each time step) to the value of inhibition computed by k-winners-take-all. There was one full oscillation (from normal to high to normal to low to normal inhibition) per trial. The $k$ parameter for k-winners-take-all was set to $k = 8$, to match the fact that input patterns have 8 active units. For additional details regarding how the learning algorithm was implemented in our simulations, see Appendix A.

In graphs of simulation results, error bars indicate the standard error of the mean, computed across simulated participants. Most simulations used on the order of 1000 simulated participants. When error bars are not visible, this is because they are too small relative to the size of the symbols on the graph (and thus are covered by the symbols).

The overall goal of this modeling work is to account for key empirical regularities in the RIF data space, and the boundary conditions on these regularities. As such, the modeling work described below focuses more on qualitative fits to general properties of the RIF data space, rather than quantitative fits to results from specific studies. We have also tried to minimize changes in parameters from simulation to simulation. The parameters that govern the underlying operation of the oscillating learning algorithm (e.g., oscillation size) were held constant across all of the simulations described here. All of the parameter changes that we did enact are described in the Key parameters section below.

Simulations of retrieval-induced forgetting

In this section, we explore how well the model can address several key findings from the retrieval-induced forgetting literature.

- In Simulation 1, we demonstrate the basic cue-independent forgetting result described in the RIF basics section above, and show that competitors are punished more after retrieval practice than after repeated study presentations. This latter result occurs
because the degree of competition between the target and the competitor is higher given partial (i.e., incompletely specified) retrieval cues, vs. when the full target item is presented. We also discuss effects of retrieval practice vs. repeated study on the practiced item itself.

- In Simulation 2, we explore how competitor strength and target strength interact with competitor punishment. In keeping with Anderson et al. (1994), we show that strong competitors tend to be punished more than weak competitors (because they are more prone to “pop up” at practice). The effects of target strength on competitor punishment are more complex: When the target item is weak, then boosting target strength increases competitor punishment, by increasing the odds that the target will win and the competitor will lose. However, once the target item is strong enough to win cleanly over the competitor at practice, further increases in target strength reduce competitor punishment (by increasing the “margin of victory” for the target, and consequently reducing competition). We also present simulations exploring how blocking effects are modulated by competitor strength and the structure of the test cue; and we present simulations showing that the strength of competitors relative to each other affects competitor punishment.

- In Simulation 3, we show that some “independent cues” are more likely to reveal competitor punishment effects than others. Specifically: RIF effects are smaller when the competitor is cued at test with an “external associate” that was linked to the competitor prior to the study phase, vs. when memory is cued using a feature of the competitor itself (Perfect et al., 2004). This difference is attributed to contextual focusing during the practice phase: Cuing with the study context prevents features that were encountered outside of the study context (e.g., the external associate) from activating during the low-inhibition phase. Because the external associate does not pop up at practice, connections involving the external associate are not weakened (so it retains its efficacy as a retrieval cue).

- In Simulation 4, we explore RIF in situations where the competitor is (initially) strong enough to displace the target memory. We show that — with the intervention of a “prefrontal system” that biases recall toward the target — practicing the target item can have nonmonotonic effects on competitor recall: Competitor recall first increases, then decreases as a function of target practice. We also demonstrate boundary conditions on this effect, as a function of overlap between the target item and the competitor, and as a function of how strongly target features are represented in the retrieval cue.

**RIF simulation methods**

Our basic RIF simulation procedure was structured to match the three phases of the retrieval-induced forgetting paradigm: A study phase, where the network learns about some patterns; a practice phase, where one of the studied patterns (but not others) is presented again, either in its entirety or in partial form, and a test phase, which measures the network’s ability to complete partial versions of studied patterns. These phases are described in more detail below. This simulation procedure was not meant to precisely mirror any particular RIF study. Rather, the goal was to set up the simplest simulation that could capture key RIF effects.

During the study phase, the network was trained on a small set of patterns. All of the patterns had eight out of a total of 80 possible units active (with activation = 1.0; other units’ activation was set to 0.0). The patterns were:

- A target pattern. This pattern was presented at study and also during the practice phase. This is analogous to Fruit-Pear in Figure 1.

- One or more competitor patterns. These patterns have 25% feature overlap with the target pattern. More specifically, the competitor patterns consist of two shared units (that are common to the target and all competitors) and 6 units that are unique to each pattern. In the “fruit” example, the 2 shared units can be seen as representing “fruit-ness”. Competitor patterns were presented at study but not at practice. This is analogous to Fruit-Apple in Figure 1.

- A set of control patterns (one control pattern for the target, and one control pattern for each competitor). The control patterns had 25% overlap with each other (via two features that were shared across all control patterns) but zero overlap with the target and competitor patterns. These items are analogous to Animal-Cow and Animal-Sheep in Figure 1.

During the study phase, the network was given multiple epochs of training on the patterns described above (the exact number of epochs varied from simulation to simulation). The presentation order of the items was randomly permuted for each epoch.

During the practice phase, the target item was presented once. As with the study phase, the oscillating learning algorithm was applied to the network. Most (but not all) of the simulations below used a partial practice cue, whereby external input strength was set to its maximal value (1.0) for the two units that were shared by the
target and competitor(s), and external input strength was set to a much lower value (.3) for the 6 units unique to the target. The partial practice condition is analogous to presenting Fruit-Pe___ at practice in Figure 1. Unless noted otherwise, target recall during the practice phase was at ceiling in all of the simulations reported in this paper. To ensure that we would be able to detect the effects of practice, the learning rate was set to a larger value for the practice phase than for the study phase (study phase learning rate = .02; practice phase learning rate = .24).

During the test phase, memory for each of the studied patterns was tested by presenting a single unit from each of the studied patterns (external input strength = 1.0 for the cue unit). Learning was turned off at test, and we did not oscillate inhibition, given that the purpose of the test phase was to read out the network’s knowledge (not to change weights). The network was given ample time to complete each pattern. For each pattern, we made sure that the unit that we used to cue memory was unique to the pattern being tested. The purpose of using minimal (1 unit) cues at test was to ensure that recall performance would be below ceiling at test (otherwise, there would be no way to examine target strengthening effects). Note that the units that we used to cue for competitors were independent cues for those items, in the sense that the units were not part of the cue used during the practice phase.

Figure 5 illustrates the patterns that were used during the study, practice, and test phases of our simulations. In RIF experiments, the test phase measures recall of unique (item-specific) properties of studied items (e.g., their orthography), rather than properties shared by practiced and non-practiced items (e.g., the fact that they are all fruits). To capture this fact in our model, we operationalized recall performance (for a given test item) by computing the average activity of the unique features of that item (excluding features that were part of the retrieval cue). We call this measure percent correct recall. To measure the effects of retrieval practice on targets and competitors, we computed the difference between recall of the item from the practiced category (e.g., the target, or the competitor) and recall of the corresponding control item.

Key parameters

As mentioned above, the vast majority of parameters were held constant across all of the simulations described here. This section describes the small number of parameters that were allowed to vary across simulations (and across conditions, within a simulation).

Variations in memory strength One of the key determinants of RIF in the model is the strength of the competing memories. In the simulations reported below, differences in memory strength were operationalized by varying the number of times that each pattern was presented during the initial study phase. For simplicity’s sake, when manipulating memory strength, we deliberately blurred together the effects of pre-experimental exposure to stimuli and the effects of encountering stimuli during the experiment itself. For example, Simulation 2 addresses data from Anderson et al. (1994) on how RIF interacts with the taxonomic strength of category exemplars (e.g., taxonomically strong fruits like “apple” vs. taxonomically weaker fruits like “kiwi”). To simulate differences in the strength of category exemplars, we simply varied the number of times that the “apple” and “kiwi” patterns were presented during the initial study phase. Thus, the “study” phase can be viewed as an amalgamation of pre-experimental and experimental exposures. As discussed in the Multiple levels of representation section in the General discussion, properly simulating differential effects of semantic vs. episodic memory strength will require a more differentiated network architecture (with separate networks supporting semantic and episodic learning). For the time being, none of the predictions addressed in this paper hinge on differential effects of semantic vs. episodic learning. Another limitation of the current network architecture is that, to avoid catastrophic interference, patterns need to be trained in an interleaved fashion with a low learning rate. Because of the need for a low learning rate, we had to present patterns on the order of 15 times (or more) during the study phase in order to get good recall of those patterns.

Variations in cue structure: Practice phase Another critical determinant of RIF in the model is how well the practice-phase retrieval cue specifies the target memory. If target units are not receiving enough input from the practice-phase cue, competitor units might be retrieved in place of target units, thereby reducing RIF (see Simulation 2 and Simulation 4). Conversely, if target units are receiving very strong input from the retrieval cue, this reduces the amount of competition and — through this — reduces RIF (see Simulation 1). Thus, it is useful to retain precise control of cue strength when simulating RIF effects.2 Practice-phase cue strength was manipulated in Simulation 1 (to mimic the effects of full vs. partial practice) and in Simulation 4 (to mimic the effects of PFC intervention in the retrieval process).

Variations in cue structure: Test phase In addition to varying the structure of the retrieval cue during the

---

2We considered two ways to construct a partial practice cue: One alternative was to provide a weak external input (of uniform magnitude) to all of the unique target units. The other alternative was to provide maximal external input to some unique units, and zero external input to others. We chose the former alternative because it allowed us to adjust the strength of the cue in a continuous fashion (by changing the strength of the external input applied to the unique units).
Figure 5: Figure illustrating the input patterns used during the study, practice, and test phases of our simulations. Note that the target has 25% overlap with the competitor, and the target control has 25% overlap with the competitor control, but neither control pattern has any overlap with the target or the competitor. In the partial practice cue, external input strength was set to 1.0 for features shared by the target and the competitor, and 0.3 for features unique to the target. The cues used to probe recall at test are shown on the bottom row. Note that the cue used to test competitor recall has no overlap with the partial practice cue, and thus qualifies as an “independent cue”.
practice phase, we also varied the structure of the retrieval cues that are used during the test phase. Specifically, in Simulation 2, to study blocking effects, we varied the relative proportion of shared vs. unique features that were included in retrieval cues at test. Also, in Simulation 3, we varied whether the competitor was cued at test using a feature of the competitor itself, or whether it was cued with an “external associate” that was paired with the competitor earlier. Apart from these manipulations, all of our simulations used a single unique feature to cue for items at test.

Variations in study pattern composition and network size Most of the simulations used input patterns where targets and competitors shared two out of 8 features. The two exceptions to this rule were Simulation 3, where we used 12-unit patterns (in order to accommodate “external cues” and context tags that were appended to the patterns), and Simulation 4, where we used 8-unit patterns but manipulated the level of overlap between the target pattern and the competitor.

Variations in practice-phase learning rate All of the simulations in the paper used the .02 study-phase learning rate mentioned in the RIF simulation methods section above, and most of them used the .24 practice-phase learning rate. The two exceptions were Simulation 3, where we also explored the effects of a smaller practice-phase learning rate, and Simulation 4 where we were interested in the incremental effects of multiple practice trials (and thus used a smaller per-practice-trial learning rate of .02).

Apart from the parameters mentioned above, all of the other parameters of the model were held constant across all of our simulations.

Simulation 1: Basic RIF and retrieval-dependence

Background The goal of this simulation is to demonstrate the basic “cue-independent” RIF effect described in the RIF basics section above, and also to explore the effects of using partial retrieval cues at practice (“retrieval practice”) vs. full retrieval cues (“extra study”). While (as discussed above) there have been numerous studies that have used the “fruit-apple” paradigm to examine cue-independent forgetting, and there have been numerous studies comparing full vs. partial practice, the only study to combine these two manipulations (to our knowledge) in the “fruit-apple” paradigm is Anderson and Shivde (in preparation). Results from that study are shown in Figure 6. The left-hand panel shows that both full practice and partial practice improved target recall in this study; the finding that partial practice did not improve target recall more than full practice is consistent with other RIF studies (e.g., Ciranni & Shimamura, 1999). The right-hand panel shows that partial practice affected competitor recall but full practice did not. Below, we explore whether the model can generate this pattern of results.

Methods During the initial study phase, the network was trained for 15 epochs on the target pattern, a single competitor pattern, and two corresponding control patterns. The target was presented once per epoch and the competitor was presented three times per epoch (making for 15 study trials for the target and 45 study trials for the competitor; the corresponding control items were presented 15 and 45 times, respectively).3

The methods for the partial practice condition were exactly as described in the RIF simulation methods section, above. The only difference between the full practice condition and the partial practice condition was the amount of external input that was applied to unique target units at practice: In the partial practice condition, external input strength was set to 0.3 for unique target units, whereas in the full practice condition external input strength was set to its maximal value (1.0).

Results Activation dynamics at study In our first analysis, we set out to characterize the dynamics of target vs. competitor activation during the study phase, and how these dynamics change with training. Figure 7 depicts how target activation and competitor activation fluctuate over the course of a single study trial (during which the target pattern was presented as input, and inhibition was oscillated as per the description in the General simulation methods section above). The dotted line shows the time-course of the inhibitory oscillation. The solid lines depict target and competitor activation dynamics early in the training process (when the target pattern is being presented for the second time), whereas the dashed lines depict target and competitor activation dynamics late in the training process (when the target pattern is being presented for the fifteenth time). We measured target activation and competitor activation at each time step by computing the average activation of the units that were unique to the target and competitor patterns, respectively.

Early in training, the net input difference between the target and competitor (when the target is presented) is still relatively small, such that target units are just above threshold and competitor units are just below threshold. As such, target activation decreases sharply when inhibi-

---

3We used a strong competitor to make the retrieval competition effects more salient in Figure 7 and Figure 8 below. The same qualitative pattern of effects is observed when the competitor and target are of equal strength, but (quantitatively) the effects are less salient.
Effects of Full vs. Partial Practice on Target Recall and Competitor Recall
Data from Anderson & Shivde (in preparation)

Figure 6: Data from Anderson and Shivde (in preparation), showing the effects of full vs. partial practice on targets and competitors. This experiment used independent cues for competitors and dependent cues for targets. The left-hand figure shows that practice boosts target recall in both the full and the partial practice conditions (slightly more so in the full practice condition). The right-hand figure shows that practice hurts competitor recall in the partial practice condition, but not the full practice condition.
Figure 7: Graphs illustrating activation dynamics when the target item is presented, early in training (during the second target presentation; solid lines) and late in training (during the fifteenth target presentation; dashed lines). The black and gray lines indicate the activation of unique features of the target and competitor patterns, respectively. The dotted line tracks the inhibitory oscillation (i.e., the amount of inhibition that is added to the inhibition value computed by kWTA on each time step). Early in training, increased inhibition causes a substantial decrease in target activation. Also, decreased inhibition leads to an increase in competitor activation. The learning algorithm updates weights to decrease the likelihood that these activation changes will be observed on future study trials; see text for discussion. Additional training reduces (but does not totally eliminate) the decrease in target activation during the high inhibition phase, and it almost completely eliminates the increase in competitor activation during the low inhibition phase. Thus, the learning algorithm has the effect of “ironing out the bumps” observed in the solid-line graph.
hibition is increased, and competitor activation increases when inhibition is reduced. The decrease in target activation during the high inhibition phase triggers strengthening of weights to weak target features, making it less likely that target units will show decreased activation (in response to increased inhibition) on subsequent study trials. Also, the increase in competitor activation during the low inhibition phase triggers weakening of weights to competing units, making it less likely that competing units will show increased activation (in response to decreased inhibition) on subsequent trials.

The dashed-line part of Figure 7 illustrates how the weight changes that occur early in training affect activation dynamics later in training. By the fifteenth epoch of training, the net input gap between the target and competitor is much larger, such that target units are relatively far above threshold and competitor units are relatively far below threshold. As such, the target representation shows less of an activation decrease during the high-inhibition phase, and competitor activation hardly increases at all during the low-inhibition phase.

The asymmetry evident in Figure 7, whereby the residual increase in activation during the low-inhibition phase (the “competitor bump”) is much smaller than the residual decrease in activation during the high-inhibition phase (the “target dip”) is a straightforward consequence of the way that inhibition was parameterized in the model. As discussed above, the kWTA algorithm places the inhibitory threshold between the unit receiving the \( k^\text{th} \) most net input, and the unit receiving the \( k + 1 \text{st} \) most net input. However, given this constraint, the algorithm has some leeway in exactly where it sets inhibition: The “normal” level of inhibition can be set relatively high (so the inhibitory threshold is just below the \( k^\text{th} \) unit), or it can be set relatively low (so the inhibitory threshold is just above the \( k + 1 \text{st} \) unit). This is controlled by a parameter called \( i_{k^\text{th} \text{st} \text{kt} \text{pt}} \). By tuning this parameter, it is possible to adjust the relative amount of target strengthening vs. competitor punishment in the model: Moving the inhibitory threshold closer to the target distribution increases the extent to which target units drop out during the high inhibition phase (and are strengthened) and reduces the extent to which competitors activate during the low inhibition phase (and are punished).

In choosing a value of \( i_{k^\text{th} \text{st} \text{kt} \text{pt}} \), we were guided by data showing that full practice helps target recall, but does not hurt competitor recall (see, e.g., data from Anderson & Shivde, in preparation, as shown in Figure 6; see also Ciranni & Shimamura, 1999; Anderson et al., 2000a). To fit this data, we selected a value of \( i_{k^\text{th} \text{st} \text{kt} \text{pt}} \) (.325) that places the inhibitory threshold closer to the target distribution than the lure distribution. This parameter setting ensures that target units will be more affected by the inhibitory oscillation than competitors (as shown in Figure 7), and thus targets will show larger learning effects than competitors in the full practice condition (see Figure 10 and Figure 11 below). The value of \( i_{k^\text{th} \text{st} \text{kt} \text{pt}} \) was held constant across all of the simulations described in this paper.

**Activation dynamics during practice** The practice phase consists of one trial during which the network is presented with either a full or partial version of the target pattern. We can compare what happens during full vs. partial practice using the same kind of activation dynamics graph that we used in the preceding section. Figure 8 illustrates how target and competitor activation fluctuate over the course of a representative partial practice trial (dashed line) and a representative full practice trial (solid line). As in Figure 7, we measured “target activation” and “competitor activation” at each time step by computing the average activation of unique target units and unique competitor units, respectively.

The input to the network in the full practice condition exactly matches the input that was used when the target item was presented at study, thus the dynamics in this condition (depicted in the solid-line part of Figure 8) are identical to the dynamics observed at the end of the study phase (depicted in the dashed-line part of Figure 8). As a result of the network’s extensive experience with the full target pattern, the dip in target activation during the high-inhibition phase is relatively small, and the increase in competitor activation during the low-inhibition phase is even smaller. The small dip in activation during the high inhibition phase should lead to some target strengthening. However, since the “competitor bump” during the low-inhibition phase is virtually nonexistent, we would expect very little (if any) competitor punishment in this condition.

Retrieval dynamics in the partial practice condition (depicted in the dashed-line part of Figure 8) differ strongly from dynamics in the full practice condition: The target shows a large dip in activation when inhibition is raised above its normal level, and the competitor shows a large increase in activation when inhibition is lowered below its normal level. Because target activation decreases during the high inhibition phase, and competitor activation increases during the low inhibition phase, we should find both strengthening of the target representation and weakening of competitors in this condition.

The difference in retrieval dynamics between full vs. partial practice can be explained by considering how the distributions of net input values associated with targets and lures differ in the two practice conditions.

The upper part of Figure 9 illustrates the distribution of net input values for target and competitor units during full practice. Because all of the target units are receiving
Figure 8: The figure plots activation dynamics during full practice (solid line) and partial practice (dashed line). The black and gray lines indicate the activation of unique features of the target and competitor patterns, respectively. The dotted line tracks the inhibitory oscillation (i.e., the amount of inhibition that is added to the inhibition value computed by kWTA on each time step). The inhibitory oscillation has a larger effect on target and competitor activation in the partial practice condition than in the full practice condition.
strong external input (as well as input from each other) the target distribution is located far above the competitor distribution. Given the wide separation between the distributions, the inhibitory threshold is not very close to either distribution. However, because of the particular \(i_{K_WTA} \) setting used in these simulations (see previous section), the inhibitory threshold is placed slightly closer to the studied-item distribution, thereby explaining the asymmetry (mentioned earlier) whereby raising inhibition slightly lowers target activity, but lowering inhibition does not raise competitor activity in this simulation.

The lower part of Figure 9 illustrates the distribution of net input scores for target and competitor units during partial practice. As discussed above, the partial practice cue is structured such that target units that are \( \text{shared} \) with the competitor receive a strong external input \( (1.0) \), whereas \( \text{unique} \) target units receive a smaller external input \( (0.3) \). This results in a situation where shared target units receive far more net input than competitors, but unique target units receive only slightly more net input than competitors. Given this distribution of net inputs, the kWTA algorithm places the inhibitory threshold in the (very small) gap between the weakest target unit and the strongest competitor. Because unique target units are just above threshold, raising inhibition results in a strong decrease in the activation of these units. Likewise, because strong competitor units are just below threshold, lowering inhibition results in a strong increase in the activation of these units.

**Effects of full vs. partial practice on target recall**

The left-hand plot in Figure 10 shows the effects of full vs. partial practice on recall of the target item. Target strengthening effects were obtained (relative to control items) after both full and partial practice. However, in keeping with the finding that raising inhibition has a larger effect on target activity in the partial (vs. full) practice condition (see Figure 8 above), target strengthening effects were larger in the partial practice condition. This simulation result is consistent with studies showing a generation effect on subsequent recall, whereby memory is better after generating an item (as is done in the partial practice condition) vs. simply reading it (as is done in the full practice condition; see Slamecka & Graf, 1978 for more information on the generation effect).

As stated in the **Precis of the Learning Algorithm** section, we believe that target strengthening effects are attributable primarily to learning that occurs during the high-inhibition phase of the inhibitory oscillation. To test this hypothesis, we ran additional simulations where we restricted learning at practice to the high-inhibition phase of the inhibitory oscillation (note that learning at study used both phases). The results of these simulations are shown in the right-hand plot in Figure 10. The same basic qualitative pattern of results was present in these simulations, thereby confirming that differential target strengthening during the high-inhibition phase is sufficient to cause the generation effect observed here.

With regard to the generation effect: As mentioned earlier, several studies have found equivalent levels of target strengthening after partial vs. full practice, as opposed to the pattern observed here (see, e.g., Ciranni & Shimamura, 1999; see also Anderson et al., 2000 for a related finding). The Anderson and Shivde (in preparation) results shown in Figure 6 show a numerical trend in the opposite direction: more target strengthening after full vs. partial practice. Given the large number of studies (outside of the RIF domain) that have found a generation effect, it seems reasonable to ask why RIF studies typically do not obtain this effect. The key to explaining this discrepancy may be the fact that, in RIF studies, participants do not always successfully retrieve the target word on partial practice trials. In **Simulation 2**, we show that imperfections in target retrieval during partial practice can reduce target strengthening in the partial practice condition and — through this — eliminate the generation effect.

**Effects of full vs. partial practice on competitor recall**

The left-hand plot in Figure 11 shows the effects of full vs. partial practice on competitor recall. RIF effects (relative to control items) were obtained in the partial practice condition but not the full practice condition; this is consistent with the finding (see Figure 8 above) that substantial competitor pop-up occurs the low-inhibition phase of the inhibitory oscillation during partial practice (but not during full practice).

As stated in the **Precis of the Learning Algorithm** section, we believe that RIF effects are attributable (at least in part) to learning that occurs during the low-inhibition phase of the inhibitory oscillation. To test this hypothesis, we ran additional simulations where we restricted learning at practice to the low-inhibition phase of the inhibitory oscillation (note that learning at study used both phases). The results of these simulations are shown in the right-hand plot in Figure 11. The same basic qualitative pattern of results was present in these simulations, thereby confirming that learning that occurs during the low-inhibition phase is sufficient to cause RIF in the par-
Figure 9: The figure illustrates the distribution of net input scores for target units (marked with a T) and competitor units (marked with a C) during full practice (upper bar) and partial practice (lower bar), when inhibition is set to its normal (baseline) level. The punishment zone marks the range of net input values (below the inhibitory threshold) that would be pushed above-threshold when inhibition is lowered, thereby leading to competitor punishment. The strengthening zone marks the range of net input values (above the inhibitory threshold) that would be pushed below-threshold when inhibition is raised, thereby leading to target strengthening. The gap between the lowest target unit and the highest competitor unit is smaller in the partial practice condition, so more target units fall into the strengthening zone, and more competitor units fall into the punishment zone.

Figure 10: Graph of the effect of full vs. partial practice on target recall. The left-hand figure plots the effects of a complete oscillation at practice (high inhibition and low inhibition), and the right-hand figure plots practice effects when learning is restricted to the high-inhibition phase at practice. In both plots, there is significant target strengthening (relative to control items) in both the full-practice and partial-practice conditions, but strengthening effects are larger in the partial-practice condition (a generation effect).
22 A Neural Network Model of Retrieval-Induced Forgetting

Figure 11: Graph of the effect of full vs. partial practice on competitor recall. The left-hand figure plots the effects of a complete oscillation at practice (high inhibition and low inhibition), and the right-hand figure plots practice effects when learning is restricted to the low-inhibition phase at practice. In both plots, there is significant RIF (relative to control items) in the partial-practice condition but not in the full-practice condition.

Discussion

In this simulation, we showed that the model captures several key aspects of the RIF data space:

- Both full and partial practice boost retrieval of the target item, as evidenced by better recall of this item vs. control items.

- Partial practice leads to competitor punishment (as evidenced by worse recall of the competitor than control items) but full practice does not lead to competitor punishment.

- Given that we used an independent cue to probe for the competitor, our results confirm the finding that competitor-punishment can be obtained even when there is no overlap between the cue used to probe for the competitor at test (e.g., Red-A) and the cue that was used to probe for the target at practice (e.g., Fruit-Pe).

As discussed earlier, we believe that RIF effects are — at least in part — attributable to an overall decrease in the strength of connections between competitor units (making the competitor a weaker attractor). Furthermore, we believe that beneficial effects of practice on the target item are attributable to an overall increase in the strength of connections between target units (making the target a stronger attractor). To address these hypotheses, we ran a detailed analysis of how partial practice affects network weights in this simulation; the results of this analysis are presented in Appendix B. In keeping with the idea that partial practice weakens the competitor attractor and strengthens the target attractor, the weight analysis showed that partial practice reduces the overall degree of interconnectivity of units in the competitor representation, and that it boosts the overall degree of interconnectivity of units in the target representation.

Boundary conditions

This simulation provides a mechanistic account of why competitor punishment is larger after partial practice (“retrieval practice”) vs. full practice (“extra study”). In the model, there is nothing special about partial practice per se. Rather, the key determinant of both target strengthening and competitor punishment is the gap in net input between target units and competitor units. Competitor punishment was smaller in the full (vs. partial) practice condition because the gap in net input between targets and lures was larger in this condition (see Figure 9). This view implies that

---

5To be clear, we do not believe that competitor weakening during the low-inhibition phase is the only cause of forgetting in the model; in Simulation 2, we discuss how target strengthening during the high-inhibition phase can sometimes lead to blocking effects.
it should be possible to get competitor punishment effects in the full practice condition, if the target memory is extremely weak and the competitor is extremely strong (such that target units are receiving only slightly more net input than competitor units). A recent study conducted by Shivde and Anderson (2001) provides some data relevant to this issue: Shivde and Anderson (2001) explored the effects of full practice in a study with weak targets (subordinate meanings of homographs) and strong competitors (dominant meanings of those homographs) and failed to show a cue-independent RIF effect. At first glance, this appears to be evidence against our prediction. However, it is important to keep in mind that even the “weak” meaning of a homograph (e.g., interpreting “arms” as “weapons”) has been encountered hundreds of times in the person’s lifetime, and thus might not be weak enough to allow competitors to come to mind during full practice. A better way to test this prediction might be to use novel targets the have been studied under relatively impoverished conditions (and, as such, should have truly weak memory traces). For example: Participants could be given brief training on novel nonwords that sound like familiar words, and we could then explore how full practice of the (weak) nonwords affects subsequent recall of the similar-sounding familiar words.

**Simulation 2: Effects of competitor strength and target strength**

**Background**

In this simulation, we explore how competitor strength and target strength interact with RIF. The first behavioral experiment to explore these factors in detail was Anderson et al. (1994). With regard to competitor strength: Anderson et al. (1994) found more RIF for strong vs. weak competitors (where competitor strength was operationalized in terms of taxonomic frequency; but see Williams & Zacks, 2001 for a failure to replicate this result). Bauml (1998) obtained a similar result, using an output interference paradigm: Retrieving moderate-frequency items at test causes output interference for subsequently-tested strong items but not subsequently-tested weak items. With regard to target strength: In the same study where they manipulated competitor strength, Anderson et al. (1994) also (orthogonally) manipulated target strength, and found no effect of target strength on competitor punishment (i.e., competitor punishment was just was large given strong vs. weak targets). The data from Anderson et al. (1994), Experiment 3 (showing the pattern described above) are shown in Figure 12.

We set out to determine whether our model can generate the pattern of results observed by Anderson et al. (1994) (a competitor strength effect but no target strength effect). The finding of more punishment for strong vs. weak competitors is highly compatible with the explanatory framework outlined earlier (in the Precis of the Learning Algorithm section). Figure 13 schematically illustrates the amount of net input received by target units, units belonging to strong competitors, and units belonging to weak competitors. We expect that units belonging to strong competitors will receive more net input than units belonging to weak competitors, insofar as the components of strong attractors are more densely interconnected than the components of weak attractors. Because units belonging to strong competitors are closer to threshold than units belonging to weak competitors, units belonging to strong competitors are more likely to activate (and be punished) when inhibition is lowered.

While the competitor strength effect observed by Anderson et al. (1994) appears to be compatible with our explanatory framework, the same explanatory framework also implies (contrary to what was observed by Anderson et al., 1994) that competitor punishment should be lower given strong vs. weak targets. More specifically:

- In the model, strengthening a target pattern amounts to strengthening the connections between the units in that pattern. As such, units participating in strong target patterns receive more net input (from each other) than units participating in weak target patterns.

- The k-winners-take-all rule places the inhibitory threshold a fixed proportion of the distance between the top k units (ranked in terms of net input) and other units. Thus, boosting the amount of net input received by target units has the effect of boosting the inhibitory threshold (pulling it away from competitors).

- Because competitors are far below the inhibitory threshold in the strong-target condition, they do not activate when inhibition is lowered, so they are not punished.

Figure 14 illustrates hypothetical net input distributions given a strong target vs. a weak target.

In summary, based on Figure 13 and Figure 14, we would expect more punishment for strong vs. weak competitors, and less competitor punishment given strong vs. weak targets (contrary to the Anderson et al., 1994 finding of a competitor strength effect but no target strength effect). In the simulations below, we show that (as expected) strong competitors are punished more than weak competitors. With regard to target strength effects: In situations where the target wins cleanly during practice, increasing target strength reduces competitor punishment (as per the logic in Figure 14, and contrary to Anderson et al., 1994). However, when the target does not win
Figure 12: Graph of results from Anderson et al. (1994), Experiment 3, showing how competitor strength and target strength interact with competitor punishment. There is more punishment for the strong competitor than the weak competitor (in both the Weak Target and Strong Target conditions). Competitor punishment effects (collapsing across the strong and weak competitor) are of similar size in the Strong Target condition and the Weak Target condition.

Figure 13: Illustration of the distribution of net input scores for target units (marked with a T), units belonging to strong competitors (marked with an S), and units belonging to weak competitors (marked with a W). Units belonging to strong competitors are closer to the inhibitory threshold, which in turn should lead to greater punishment for strong vs. weak competitors.
Figure 14: The figure schematically illustrates the distribution of net input scores for target units (marked with a T) and competitor units (marked with a C) during partial practice for weak targets (upper bar) and strong targets (lower bar). Competitors are closer to the inhibitory threshold in the weak target condition than the strong target condition, so they are more likely to be punished in the weak target condition.

cleanly (i.e., some competitor units receive more input than some target units), the opposite pattern can occur: Increasing target strength can increase competitor punishment (by increasing the odds that the competitor will lose and the target will win). This latter property of the model makes it possible to explain the findings from Anderson et al. (1994).

Methods
To simulate the effects of competitor strength, we made some minor modifications to the simulation methods described in the previous section (described below). Otherwise, all of the simulation parameters were exactly the same as the parameters used in Simulation 1.

First, instead of just using one competitor, we used two competitors:

- A strong competitor item that was presented five times per epoch at study (75 times total), but was not presented at practice (corresponding to Fruit-Apple)
- A weak competitor item that was presented once per epoch at study (15 times total), but was not presented at practice (corresponding to Fruit-Kiwi)

We manipulated target strength by varying the number of target-item presentations during the study phase from 15 to 135 (in steps of 15). In this simulation, there were three control items (corresponding to each of the 3 items from the practiced category: the target, the strong competitor, and the weak competitor). The number of study presentations for each control item was identical to the number of study presentations for the corresponding item from the practiced category.

Results: Simulation of Anderson et al. (1994)
To simulate the Anderson et al. (1994) experiment (which orthogonally manipulated the strength of targets and competitors), we can compare the condition where the target was presented 15X (the Weak Target condition) to the condition where the target was presented 75X (the Strong Target condition). This arrangement has the desirable property that (as in the Anderson et al., 1994 study) the strength value of the strong target is the same as the strength value of the strong competitor (75X), and the strength value of the weak target is the same as the strength value of the weak competitor (15X).

Results from these conditions are shown in Figure 15. Overall, the results from this simulation line up well with the results from Anderson et al. (1994): Increasing competitor strength led to a large increase in RIF, but increasing target strength by the same amount did not have a large effect on RIF (numerically, RIF was slightly larger given strong vs. weak targets).

The finding of greater RIF for strong vs. weak competitors (in the model) can be explained in terms of the principles expressed in Figure 13. Strong competitors are closer to the inhibitory threshold, so they show a larger increase in activation when inhibition is lowered (and thus more punishment). We measured the activation of strong and weak competitors at the trough of the inhibitory oscillation (i.e., when inhibition was lowest): Collapsing across the Strong Target and Weak Target conditions, the average activation of the strong competitor (out of 1.0) was .824 (SEM = .005), and the average activation of the weak competitor was .003 (SEM = .002). One small discrepancy between these results and the results of Anderson et al. (1994) is that we obtained a small (but above-zero) RIF effect for weak competitors, whereas the RIF effect was nonsignificant for weak competitors in Anderson et al. (1994). The RIF effect for weak competitors in the simulation can not be attributed to direct punishment of these items by the oscillating learning algorithm (since, as mentioned above, weak competitors did not “pop up” during the low-inhibition phase). The only remaining explanation is that the RIF effect observed for weak competitors is a blocking effect.
A Neural Network Model of Retrieval-Induced Forgetting

Figure 15: Graph of how competitor strength and target strength affect competitor punishment in the model. The qualitative pattern of results mirrors what was observed in Anderson et al. (1994), Experiment 3 (see Figure 12): There is more punishment for the strong competitor than the weak competitor (in both the Weak Target and Strong Target conditions). Competitor punishment effects (collapsing across the strong and weak competitor) are slightly larger in the Strong Target condition than the Weak Target condition.
(i.e., an indirect effect of target strengthening on competitor recall). The issue of blocking effects for weak competitors is addressed in much more detail in the Interactions between competitor strength and blocking effects section, below.

With regard to target strength effects: Earlier, we had argued that increasing target strength should reduce competitor punishment (by increasing the “margin of victory” for the target; less competition leads to less RIF). However, contrary to this principle, we found in this simulation that increasing target strength led to a small increase in competitor punishment. The reason for this discrepancy is simple: Our previous discussion of target strength effects assumed that the target would win cleanly over the competitor in both the Weak Target and Strong Target conditions. However, in this simulation, the target memory was not strong enough to win cleanly over the strong competitor in the Weak Target condition. In situations like this, increasing target strength helps to boost competitor punishment, by ensuring that target activation stays on top of competitor activation throughout the trial. Figure 16 shows the activation dynamics of the target and strong competitor during the practice trial, in the Weak Target condition: During the early part of the trial, the target is more active than the competitor, but during the latter part of the low-inhibition phase, the competitor overtakes the target.\(^6\) This relatively small imperfection in target retrieval dynamics is nonetheless sufficient to reduce competitor punishment. Going from the Weak Target condition to the Strong Condition helps to clean up this imperfection in retrieval dynamics (so the competitor does not overtake the target), which — in turn — increases competitor punishment.

Effects of imperfect practice dynamics on target strengthening The Weak Target condition of this simulation also gives us an opportunity to explore how imperfect target retrieval dynamics (as shown in Figure 16) affect target strengthening. The fact that the target representation comes on initially and dips down during the high-inhibition phase should lead to strengthening of the target representation; however, the fact that the target starts to drop out at the end of the low-inhibition phase (just like the competitor) should lead to some amount of weakening, which counteracts the initial strengthening.

The deleterious effects of imperfect target retrieval (during partial practice) on target strengthening can be quantified by comparing the effects of partial practice to the effects of full practice.\(^7\) In Simulation 1 above, we discussed how (when the target is cleanly retrieved during partial practice) partial practice typically leads to more strengthening than full practice — a generation effect. We also raised the possibility that imperfect retrieval of the target during partial practice could have the effect of reducing (and possibly reversing) the generation effect, by reducing the amount of target strengthening in the partial practice condition. Figure 17 provides a concrete illustration of this point, using results from the Weak Target and Strong Target conditions of our simulation. There is a “reverse generation effect” (more strengthening after full vs. partial practice) in the Weak Target condition (15X target study), which (as discussed above) suffers from imperfect recall dynamics at practice, and a normal generation effect in the Strong Target (75X target study) condition (where target retrieval dynamics are more regular). The finding that subtle imperfections in retrieval dynamics can sabotage the generation effect, coupled with the observation that — in real-world RIF experiments — target recall at practice is far from perfect, provides an explanation of why RIF studies typically do not obtain a generation effect (e.g., Ciranni & Shimamura, 1999; Anderson et al., 2000a; Anderson & Shivde, in preparation).

Boundary conditions

Having demonstrated that the model can simulate the two key results from Anderson et al. (1994) (i.e., increasing competitor strength boosts RIF, but increasing target strength does not reduce RIF) we now explore boundary conditions on these findings. First, in the Boundary conditions on the null target strength effect section, we show that increasing target strength does reduce RIF if we use a more powerful target strength manipulation. Next, in the Interactions between competitor strength and blocking effects section, we show that weak competitors (vs. strong competitors) are differentially susceptible to blocking effects. Finally, in the Effects of relative competitor strength section, we show that RIF is affected by the strength of competitors relative to each other, in addition to the strength of competitors relative to targets.

Boundary conditions on the null target strength effect

In our simulation of Anderson et al. (1994) above, we examined two levels of target strength (15X and 75X) and showed that competitor punishment was slightly higher in the Strong Target (75X) condition. Here, we explore a wider range of target strength values and show that the overall pattern of target strength effects is more complex than was evident in the previous simulation.

Figure 18 plots the effect of target strength on competitor punishment (collapsing across weak and strong competitors), with target strength values ranging from 15X to 135X (in 30X increments). The figure shows a nonmonotonic effect of target strength on competitor punishment: Increasing target strength from 15X to 45X
Figure 16: Practice-phase activation dynamics in the Weak Target condition. The target item is more active during the high-inhibition phase, but the competitor overtakes the target at the end of the low-inhibition phase.
results in more competitor punishment, but additional increases in target strength reduce competitor punishment. By the time that target strength reaches 135, the competitor punishment effect has been almost completely eliminated. The nonmonotonic pattern observed here can be explained in terms of two contrasting effects of target strength: First, as discussed in the simulation of Anderson et al. (1994) above, increasing target strength from low values to moderate values boosts competitor punishment by helping to ensure that the target wins (and the competitor loses). Once the target is strong enough to win cleanly over the competitor, further increases in target strength work to reduce competitor punishment, by increasing the margin of victory for the target and — through this — reducing the extent to which competitors activate during the low inhibition phase (see Figure 14).

Interactions between competitor strength and blocking effects In this section, we show that weak competitors are especially susceptible to blocking effects. More specifically: raising target strength indirectly hurts recall of weak competitors (more so than strong competitors), even in the absence of actual synaptic weakening. In our simulation of Anderson et al. (1994) above, we argued that the small (but nonzero) level of RIF observed for weak competitors was attributable to blocking. Here, we show that manipulating the structure of retrieval cues used during the test phase can lead to a large and selective increase in blocking effects for weak competitors, to the point where weak competitors and strong competitors show similar levels of RIF.

The most straightforward way to measure blocking effects is to manipulate when learning occurs during the practice phase. If we limit learning to the low-inhibition (competitor punishment) phase at practice, so no target strengthening occurs, RIF effects that are due to blocking should disappear, but RIF effects that are actually due to competitor weakening should persist. It is also diagnostic to explore what happens when we limit learning to the high-inhibition (target strengthening) phase at practice. In this case, any RIF that is observed has to be due to blocking (since no synaptic weakening is taking place). In summary: RIF effects that are due to blocking should be present in the high-inhibition-only condition but not the low-inhibition-only condition; RIF effects that are due to actual synaptic weakening should be present in the low-inhibition-only condition but not the high-inhibition-only condition.

Figure 19 shows the results from the Strong Target condition of the previous simulation (strong competitor = 75X study, weak competitor = 15X study, target = 75X study), when learning at practice is limited to the low-inhibition phase and the high-inhibition phase, respectively. The weak competitor shows an RIF effect in the high-inhibition-only (target strengthening) condition but not the low-inhibition-only (competitor punishment) condition, suggesting that the weak-competitor RIF effect observed in Figure 15 is entirely attributable
Figure 18: Graph of how target strength affects competitor punishment. The gray bars indicate competitor recall (averaged across the strong competitor and weak competitor conditions). The black bars indicate recall of the corresponding control items. The effect of target strength is nonmonotonic: Going from target strength 15 to target strength 45, competitor punishment increases. However, further increases in target strength beyond this point reduce competitor punishment. Increasing target strength to 135 almost completely eliminates competitor punishment.
Competitor Punishment as a Function of Competitor Strength

Low Inhibition Only

High Inhibition Only

Percent Correct Recall

Strong Weak Strong Weak

Control Item Competitor

Figure 19: Plot of competitor punishment effects in the Strong Target condition, as a function of the type of learning used at practice. The left-hand figure shows practice effects when learning is turned off during the high-inhibition (target strengthening) phase at practice. Turning off target strengthening eliminates RIF for the weak competitor, suggesting that these RIF effects were attributable to blocking. The right-hand figure shows practice effects when learning is turned off during the low-inhibition (competitor punishment) phase at practice. Even though synaptic weakening is not taking place, RIF effects are nonetheless obtained for the weak competitor. This provides additional converging evidence that RIF effects for the weak competitor are due to blocking as opposed to actual weakening of the competitor's representation.
to blocking (no actual competitor punishment occurs, because in this simulation the weak competitor does not activate during the low-inhibition phase at practice). In contrast, the strong competitor shows an RIF effect in the low-inhibition-only condition but not in the high-inhibition-only condition, suggesting that the strong-competitor RIF effect observed in Figure 15 is primarily attributable to actual synaptic weakening (as opposed to blocking).

Figure 20 illustrates why weak memories are more susceptible to blocking effects than strong memories. Each memory (attractor state) stored in the network can be viewed as an “energy well”. Strong memories correspond to deep energy wells, and weak memories correspond to shallow energy wells (Part A). In this scheme, the process of activation spreading through the network can be construed as a marble that rolls around in the energy landscape. The marble is drawn (“attracted”) toward the energy wells, just as the neural network is drawn toward activation states corresponding to stored patterns. Now we can consider the effect of strengthening one memory on recall of other memories. In a network with multiple energy wells, deepening one energy well (i.e., strengthening one memory) distorts the rest of the attractor landscape. If the system is settled into a deep energy well (corresponding to a strong stored memory), then deepening another energy well probably will not dislodge the marble. However, if the system is settled into a shallow energy well (corresponding to a weak stored memory), then deepening another energy well could cause the marble to spring loose and roll into the newly deepened energy well. This example illustrates how strengthening one memory is more likely to disrupt recall of other, weak memories than other, strong memories.

The fact that weak memories are differentially susceptible to blocking implies that it should be possible to selectively manipulate RIF for weak items via factors that affect the amount of blocking. One way to manipulate the amount of blocking that occurs is to manipulate the composition of the retrieval cues used during the test phase of the experiment (i.e., the third and final phase of the experiment, after the practice phase has occurred). Intuitively, if the test cue that is used to probe for the competitor is comprised of features that are *unique* to the competitor, this will limit the spread of activation to the strengthened target item, and blocking effects will be relatively small. In contrast, if the test cue contains features that are *shared* by the target item and the competitor, this will promote activation of the target and (through this) lead to blocking effects.

The next simulation was identical to the Strong Target condition above, except we orthogonally manipulated the number of shared and unique features that were included in retrieval cues during the test phase. Specifically, some cues included zero (out of two) shared features and some cues included one shared feature. Crossed with that, some cues included two (out of six) unique features, and some cues included five (out of six) unique features. In Anderson’s parlance, competitor retrieval cues with zero shared features are “independent cues” (since they have no overlap with the cue that was used to retrieve the target during the practice phase), and competitor retrieval cues with one shared feature are “dependent cues”. In the discussion below, we use abbreviations like “1S2U” to refer to the various conditions of the simulation, where the number before “S” refers to the number of shared features in the cue, and the number before “U” refers to the number of unique features in the cue (so, “1S2U” refers to the “one shared feature, two unique features” condition). Note that, while the properties of the *test cues* (i.e., the cues used during the test phase of the simulation) were manipulated in this simulation, the properties of the *practice cue* were held constant (we used our standard “partial practice” cue).

With regard to competitor recall: We expected that blocking would be maximal in the 1S2U condition (low specificity dependent cues) and that blocking would be minimal in the 0S5U condition (high specificity independent cues). Consequently, we expected that forgetting of the weak competitor would be largest in the 1S2U condition and smallest in the 0S5U condition.

Figure 21 shows the results of the simulation, for all four cue types. The top row of the figure shows the results for (low-spectrum) cues with two unique units: Going from the 0S2U cue condition (an independent cue) to the 1S2U cue condition (a dependent cue) greatly increases blocking, as evidenced by a sharp and selective increase in RIF for the weak competitor. This increase effectively erases the competitor strength effect in the 1S2U condition (i.e., the weak and strong competitor show equivalent levels of RIF).

Interestingly, it appears that including a high number of unique units in the cue can protect against the deleterious effect of including shared units. The bottom row of the figure shows the results for (high-spectrum) cues with five unique units. Here, going from the independent cue condition (0S5U) to the dependent cue condition (1S5U) does not result in increased RIF for the weak competitor. As such, both conditions show a normal competitor strength effect (i.e., more RIF for the strong vs. weak competitor).

Overall, the results fit with our expectation that blocking effects (as manifested by weak-competitor punishment) would be numerically largest in the “low specificity, dependent cue” (1S2U) condition and smallest in
Figure 20: The selective vulnerability of weak memories to blocking can be explained using energy-well diagrams, where weak memories correspond to shallow energy wells, and strong memories correspond to deep energy wells. Spreading activation in the network can be construed as a marble rolling downhill in the energy landscape. Part A: Energy-well diagram illustrating a situation where memory for the competitor is strong and memory for the target is weak. Part B: How strengthening the target memory affects recall of the competitor. As in part A, the left-hand well corresponds to the competitor and the right-hand well corresponds to the target. To illustrate competitor recall, we start by placing the marble in the left-hand well. When memory for the competitor is strong (top row), strengthening the target memory (deepening the right-hand well) distorts the energy landscape, but this distortion is not sufficient to dislodge the marble from the left-hand well. When memory for the competitor is weak (bottom row), deepening the right-hand well causes the marble to roll from the left-hand well to the right-hand well. In our model, this “escaped marble” corresponds to intruding the (newly strengthened) target in place of a previously stored (but weak) competitor.
Figure 21: Plot of competitor punishment effects in the Strong Target condition, as a function of competitor strength and test cue type. Practice-phase learning occurred during both phases of the oscillation (high-inhibition and low-inhibition). Test cues were manipulated along two orthogonal dimensions: The number of features in the cue that were unique to the cued item (2 or 5), and the number of features in the cue that were shared with other items (0 or 1). Competitor retrieval cues with zero shared units are nominally independent cues, and competitor retrieval cues that share one unit with the target pattern are nominally dependent cues. Each cue condition is labeled with a tag like “1S2U” that specifies the number of shared and unique features in the cue (“1S2U” = 1 shared feature, 2 unique features). Switching from the 0S2U condition (low specificity, independent cue) to the 1S2U condition (low specificity, dependent cue) increases blocking, as evidenced by a large increase in RIF for the weak competitor. However, switching from the 0S5U condition (high specificity, independent cue) to the 1S5U condition (high specificity, dependent cue) does not affect RIF for the weak competitor. Taken together, these results indicate that (in the model) high cue specificity protects against blocking effects.
the “high specificity, independent cue” (0S5U) condition. The absolute size of the RIF effect for the weak competitor in “high cue specificity” conditions (0S5U and 1S5U) was very close to zero, consistent with the Anderson et al. (1994) finding of a null RIF effect for weak competitors. The finding that cue specificity selectively impacts RIF for weak competitors leads to several testable predictions, which are described in the Discussion section below.

Effects of relative competitor strength Our explanation of competitor strength effects (e.g., in Figure 13) has, up to this point, focused on the strength of competitors relative to targets as a key determinant of competitor punishment. Here, we show that (in addition to being affected by the strength of competitors relative to targets), competitor punishment also is affected by the strength of competitors relative to each other. Intuitively, this a consequence of the fact that inhibitory mechanisms monitor and respond to network activation, even during the low-inhibition phase. When inhibition is lowered, the strongest competitor activates first. This increase in activation triggers a compensatory increase in the strength of inhibition, which makes it more difficult for additional competitors to become active (thereby sparing them from punishment).

We can demonstrate this point about relative competitor strength in a two-competitor simulation, by holding the strength of one competitor constant and manipulating the strength of the other competitor. We ran a simulation with the first competitor (hereafter referred to as competitor one) was presented 45X at training and the target was presented 75X. We manipulated the strength of the second competitor (hereafter referred to as competitor two) from 15X to 75X. Our primary interest in this simulation is how forgetting of competitor one (whose strength, relative to the target, is held constant throughout the simulation) is affected by changes in the strength of competitor two.

Figure 22 shows the results of this simulation; the top part of the graph shows RIF for competitor one (whose strength is fixed at 45X) and the bottom part of the graph shows the peak activation of competitor one during the low inhibition phase at practice (i.e., how much the competitor “pops up”). As expected, increasing the strength of competitor two reduces the extent to which competitor one activates during the practice phase, and (consequently) reduces punishment for competitor one.

Summary and discussion of Simulation 2

Competitor strength In the simulations above, we showed that the model can generate the competitor strength effect (more punishment for strong competitors) observed by Anderson et al. (1994), but we also showed that there are circumstances where this pattern will not be observed. Specifically, the model predicts that manipulations that boost blocking (e.g., use of low-specificity, dependent cues) should selectively boost RIF for weak (vs. strong) competitors. If this blocking effect is large enough, it can wipe out the competitor strength effect.

Apart from blocking effects, the simulations also point to the importance of evaluating both the strength of the competitor relative to the target, and the strength of the competitor relative to other competitors, when predicting RIF effects. The Anderson et al. (1994) study held relative competitor strength constant (in a given condition, all of the competitors were strong, or all of the competitors were weak) and then manipulated the strength of these competitors relative to the targets. One clear prediction from Figure 22 is that, if we held target strength and competitor strength (for some competitors) constant, and increased the strength of other competitors, this should reduce the amount of RIF that we observe for the competitors whose strength is not being manipulated.

Target strength The target strength results presented here confirm a basic prediction of our modeling framework (less competitor punishment as the net-input gap between targets and lures increases), and add an important boundary condition on this effect: The effect of target strength on RIF can reverse in situations where target recall during practice is less than perfect. This latter fact may help explain the lack of a target strength effect in Anderson et al. (1994). Target recall at practice was below ceiling in that study (e.g., in Experiment 3, recall accuracy at practice was 82% for strong targets, and 67% for weak targets), thereby placing performance in exactly the parameter regime where we would expect null or reversed target strength effects.

This account leads to the following prediction: If it were possible to increase the probability of successful target retrieval in the weak target condition, without compromising the overall power of the target strength manipulation, we would expect to find a clear target strength effect (less competitor punishment for strong vs. weak targets). Put another way: The ideal conditions for obtaining a target strength effect involve comparing targets that are strong enough to be recalled successfully at practice (but just barely) with targets that are much
Punishment of Competitor One (Strength = 45) as a Function of the Strength of Competitor Two

Figure 22: Plot of how punishment of competitor one (studied 45X) varies as a function of the strength of the competitor two. The target was studied 75X, and competitor two’s strength was varied from 15X to 75X. The graph also plots the peak activation (“pop up”) of competitor one during the low-inhibition phase at practice. When competitor one is stronger than competitor two, it activates strongly at practice, and it shows a large competitor punishment effect. As the strength of competitor two rises, competitor one activates less and less at practice, leading to reduced punishment effects. These results clearly show that punishment is affected by the strength of the competitor relative to other competitors (in addition to the strength of the competitor relative to the target).
stronger.

One final point regarding target strength effects relates to the issue of blocking. Anderson et al. (1994) point out that target strength effects (less competitor punishment for strong targets) could arise for reasons other than competitor weakening per se. For example, if weak targets undergo more strengthening than strong targets at practice (due to ceiling effects or other factors), this will differentially increase weak targets’ ability to interfere with (block) competitor recall at test. This differential increase in blocking could, on its own, result in more RIF given weak vs. strong targets. While we agree that (logically) this is a possibility, we are sure that blocking is not solely responsible for our finding (shown in Figure 18) that, as target strength increases, competitor punishment asymptotically goes to zero. If this finding were attributable to indirect effects of target strengthening, it should go away when we turn off learning during the high-inhibition phase at practice (where target strengthening takes place). However, we ran additional control simulations (not shown here) and found that the same qualitative pattern of target strength results is obtained when we turn off learning during the high-inhibition phase.

**Simulation 3: Comparing different types of independent cues**

**Background**

As discussed above, Anderson has argued that RIF is cue-independent, meaning that subsequent retrieval of competitors is impaired no matter what cue is used at test. Extant studies provide a clear “existence proof” that RIF can be observed given independent cues (i.e., cues that were not used at practice; Anderson & Spellman, 1995). However, at this point, it is unclear whether RIF extends to all independent cues, or whether RIF is limited to specific subtypes of independent cues.

Recently, Perfect et al. (2004) challenged Anderson’s notion of cue-independence, by showing that some types of independent cues are (apparently) insensitive to RIF. Specifically: Perfect et al. (2004, Experiment 3) modified the standard RIF procedure by including a pretraining phase (prior to the study, practice, and test phases outlined in Figure 1) where each category exemplar was associated with a unique, unrelated word cue. For example, “Apple” might be associated with “Zinc”. Then, at test, Perfect et al. (2004) compared recall using two different types of cues:

- category-plus-fragment cues (e.g., study Fruit-Apple; practice Fruit-Pe; test with Fruit-
- p -)
- cues from the pretraining phase (e.g., study Fruit-Apple, test with Zinc).

Note that the second type of cue does not overlap with the practice cue (Fruit-Pe), so it is “independent” according to Anderson’s standard terminology, whereas the first type of cue is “dependent”. Perfect et al. (2004) found RIF using the category-plus-fragment cues but failed to find any RIF when they tested using cues from the pretraining phase (Zinc). Figure 23 shows the results from Perfect et al. (2004), Experiment 3; in the figure, “Standard” refers to the category-plus-fragment cue condition and “External” refers to the condition where participants were cued with novel associates from the pretraining phase.

The goal of this simulation is to explore why Perfect et al. (2004) did not obtain an RIF effect when they used cues from the pretraining phase. Given that (as discussed above) other studies have found RIF with independent cues, the use of independent cues per se can not be the cause of their failure to obtain an RIF effect. Furthermore, other studies have found RIF using novel associates as cues (e.g., Anderson & Bell, 2001; Ciranni & Shimamura, 1999; Radvansky, 1999; MacLeod & Macrae, 2001); thus, the use of novel associates as cues per se can not be used to explain the null RIF effect either. Having accounted for these factors, there is one highly salient difference between the Perfect et al. (2004) experiment and other studies that succeeded in finding RIF effects with novel-associate cues: In the studies that found RIF effects, the (novel) cue-target association was learned during the main study phase, whereas in Perfect et al. (2004) (Experiment 3) the novel cue-target association was learned outside of the study phase. As such, our simulation was specifically set up to address the role of contextual information in modulating RIF. In the simulation, we compared two test conditions:

- The standard independent cue condition matches the cue used in our previous simulations; this is meant to reflect the situation where participants study Fruit-Apple, practice Fruit-Pe, and are subsequently tested with Red-A.
- The external cue condition is meant to mirror the situation from Perfect et al. (2004) where participants pretrain with Zinc-Apple, study Fruit-Apple, practice Fruit-Pe, and are tested with Zinc.

Below, we show that — even though both test cue types are nominally “independent” (i.e., they do not overlap with the practice cue) — the model predicts more RIF for standard independent cues than external cues. We first provide a high-level account of why the model makes this prediction, and then we present detailed methods and results.
Competitor Punishment as a Function of Test Cue Type

Data from Perfect et al. (2004), Experiment 3

Figure 23: Graph of data from Perfect et al. (2004), Experiment 3, showing competitor punishment when memory for the competitor (Apple) is tested with a category-plus-fragment cue (the Standard condition) vs. when memory is tested with a semantically unrelated word (e.g., Zinc) that was associated with Apple during the pretraining phase (the External condition). RIF is present in the Standard cue condition but not the External cue condition. Data were taken from the analysis shown in Perfect et al. (2004), Table 4, where participants were selected to ensure matched recall of control items.
Overview of simulation  
To simulate the two conditions of interest, all of the items (the target, the competitor, and their controls) were presented along with a “study context tag”, reflecting the fact that these items were presented during the study phase of the experiment. This tag is activated during the practice phase also, and serves to focus retrieval on items from the study phase. In addition to being presented along with the study context tag, the competitor and its control were also associated with external cues (analogous to Zinc in the Perfect et al., 2004 study).

Figure 24 illustrates the sequence of events that occurs at practice. We assume that participants actively try to retrieve memories from the study phase (as opposed to the pretraining phase), by cuing with features of the study context. This active targeting of study-phase memories should result in greater “pop up” of features of the competitor (which were active during the study phase) as opposed to features of the external cue (which were only active at pretraining, not at study). As a result of this “pop up” of competitor features, two things happen:

- Overall, the links between the features of the competitor get weaker; this effectively weakens the competitor’s attractor and makes it harder to retrieve competitor features, regardless of the cue.
- As illustrated in Part D of Figure 24, a crucial property of the standard “independent cue” (Red) is that Red is a feature of the competitor (Apple). Since (in this example) Red is among the competitor features that popped up at practice, the connections between Red and other competitor features have been weakened. This situation — where the cue has been specifically “cut off” from the features that it is supposed to activate — results in an extra decrease in recall, beyond what you get merely from attractor weakening. Importantly, because the external cue (Zinc) does not activate at practice, the links between this cue and the features of the competitor retain their efficacy.

In summary: We expect that attractor weakening should adversely affect recall, regardless of what type of cue is used at test. However, we also expect that external cues should retain their efficacy more than standard independent cues, so less RIF should be observed for external cues than standard independent cues.

Methods  
This simulation used a 120-unit network (instead of the standard 80-unit network). We augmented our standard 8-unit patterns (target, competitor, and 2 controls) with a 4-unit “study context tag” that was shared by all 4 patterns, so the new patterns consisted of 12 active units in all. In addition to these 4 patterns (which were each studied 45X) we presented two new patterns, which were also studied 45X:

- The 8-unit competitor pattern was paired with a 4-unit “external cue” (which had no overlap with any of the other patterns or with the study context tag).
- The 8-unit “competitor control” pattern was paired with its own (distinct) external cue.

As in our prior simulations, all of the study patterns were trained in an intermixed fashion.  
The cue that we used at practice was the same as the partial practice cue from Simulation 1, except that it also included all 4 of the “study context tag” units (with external input strength for these units set to 1.0).

At test, recall of the competitor was cued using either:

- One of the features of the external cue (analogous to “Zinc-“)
- One of the unique features of the competitor (i.e., our standard independent cue, analogous to “Red-A“)

Results  
The results of the simulation are shown in Figure 25 (left-hand panel): As predicted, and in keeping with the results of Perfect et al. (2004), more RIF was observed given the standard cue vs. the external cue. However, contrary to Perfect et al. (2004), we found a large RIF effect for external cues, whereas Perfect et al. failed to find any RIF in the external-cue condition. We ran a follow-up simulation to address this discrepancy: While the model clearly predicts that there should be some RIF (due to attractor weakening) in the external-cue condition, we thought that it might be possible to reduce the absolute size of this effect in the model, thereby bringing our simulation results more into line with the results from Perfect et al. (2004). To accomplish this goal, we made two changes:

9Our simulation would have matched up better with the actual paradigm if we trained the two new patterns (corresponding to the “pretraining” patterns in the Perfect et al., 2004 study) prior to the 4 standard patterns. However, as discussed in the Key Parameters section earlier, the current network architecture requires interleaved training across multiple training epochs in order to acquire and retain patterns. The need for interleaved training in the model makes it impossible to mimic the multi-phase training procedure used in the Perfect et al. (2004) study. In the General Discussion section at the end of the paper, we discuss how implementing the oscillating learning algorithm in a hippocampal architecture would allow for rapid, successive memorization of patterns without catastrophic interference.
Figure 24: Illustration of competitor-punishment effects during the practice phase in Simulation 3. Apple and Pear are represented as distributed collections of interconnected features. F = Fruit features shared by Apple and Pear; R = Red (a feature unique to Apple); P = a feature unique to Pear. Because Apple and Pear were both presented during the study phase, all of the constituent features of Apple and Pear have been linked to a representation of the study context. In addition, all of the features of Apple have also been linked to a representation of Zinc (because Apple and Zinc were paired during the pre-training phase). Light colors represent active representations, and dark colors represent inactive representations. Part A: During the partial practice phase, the system is cued with Fruit, some features unique to Pear (“P”), and the study context representation. Part B: In response to this cue, activation spreads to the other features of Pear. Part C: When inhibition is lowered, activation spreads to other features of Apple (because these features are receiving input both from Fruit and also from the study context) but not to the representation of Zinc — Zinc features receive some input (by way of Apple features) but this input is not enough to push these features above threshold. Part D: The learning algorithm weakens the connections between Apple features that popped up during Part C and other active features. Note that, because the Zinc representation stayed inactive throughout the entire practice trial, the learning algorithm has no effect on the connection between Zinc and Apple.
Figure 25: Graph of competitor punishment in the model, when memory is tested using our standard “independent cue” vs. an external independent cue that is unlikely to pop up during practice. The left-hand panel shows competitor recall when practice $lrate = .24$ (our standard learning rate). RIF is observed in both conditions, but more RIF is observed for the standard cue vs. the external cue. The right-hand panel shows competitor recall when we use a lower practice $lrate (.16)$ and learning is limited to the low-inhibition (competitor punishment) phase at practice, to eliminate blocking. These changes reduce the overall amount of RIF, such that (in keeping with Perfect et al., 2004) the absolute magnitude of the RIF effect is close to zero for external cues. However, the basic qualitative pattern of results is unchanged: The RIF effect is still significant for both conditions, and more RIF is observed for the standard cue vs. the external cue.
• First, we lowered the practice-phase learning rate from .24 to .16.

• Second, we thought that some of the RIF observed in our simulation might be attributable to blocking. In order to explore the effects of attractor weakening on competitor recall, uncontaminated by blocking effects, we limited learning at practice to the low-inhibition (competitor punishment) phase.

The right-hand side of Figure 25 shows the results of simulations where the practice-phase rate was set to .16, and learning was limited to the low-inhibition phase at practice. As expected, RIF effects were smaller, overall, with these two changes in place.\(^\text{(10)}\) As a result of this overall decrease, RIF was numerically close to zero in the external-cue condition (thereby approximating the Perfect et al., 2004 finding of null RIF for external cues). However, apart from this general decrease in RIF, the qualitative pattern of results was basically unchanged: The RIF effect, though smaller, was still statistically reliable for both cue types, and it was still reliably larger for standard cues than external cues.

Discussion

The results of this simulation strongly confirm the claim made by Perfect et al. (2004) that different cues can elicit different degrees of RIF. Specifically, our simulation results match the Perfect et al. (2004) finding that “external cues” from the pretraining phase reveal less RIF than more standard types of cues. The model’s explanation for this finding is that, because of contextual targeting, external cues do not pop up at practice; since these cues do not activate, their connections to the competitor memory are not weakened, so they retain their efficacy at test. We ran additional analyses that confirm our key claim that the external cue did not pop up at practice: The average peak activation of the external cue (Zinc) during the low-inhibition phase at practice was .016 out of 1 (SEM = .002). In contrast, the average peak activation for the competitor itself (Apple) was .882 out of 1 (SEM = .001).

The main difference between our simulation results and the results from Perfect et al. (2004) is that we consistently obtained significant RIF effects for the external cue, whereas Perfect et al. (2004) did not. According to the model, RIF in the external-cue condition is primarily due to attractor weakening: Reducing connections between the constituent features of “apple” makes it harder to recall, regardless of which cue is used. However, we also showed that depending on parameter settings the size of this external-cue RIF effect can be quite small in the model (relative to the standard-cue effect, and on an absolute scale also). As such, Perfect et al.’s failure to observe an external-cue RIF effect may be attributable to their use of a design that is not powerful enough to detect a (possibly very small) effect.\(^\text{(11)}\)

Importantly, while this simulation describes one way in which context can modulate RIF, there are other possible ways that context can modulate RIF that are not addressed by this simulation. In the simulation above, we focused on indirect effects of contextual cuing on RIF: Contextual cues at practice influence which (non-contextual) features pop up, which in turn determines whether connections between these (non-contextual) features get weakened. However, in addition to these indirect effects, the learning algorithm also directly adjusts the connections between contextual and non-contextual features (e.g., Figure 24, part D, shows that the learning algorithm weakens connections between the “study context” tag and the features of Apple; the issue of how context-item associations can change during an RIF paradigm is also discussed by Perfect et al., 2004). In future simulation work, we will explore the ramifications of these direct changes in context-item association strength for RIF.\(^\text{(12)}\)

Boundary Conditions

As mentioned earlier, we think that a key difference between the Perfect et al. (2004) study (described above) and other studies that have found RIF effects using novel-associate cues (e.g., Anderson & Bell, 2001) is that Perfect et al. (2004) trained up these novel associations during a pretraining phase, whereas the other studies trained up the novel associations during the main study phase of the experiment. According to our “contextual targeting” account, the fact that the novel cue was encountered during the main study phase in Anderson and Bell (2001) makes it much more likely that this cue would pop up during retrieval practice, which — in turn — should boost the degree of RIF observed with that cue. It is

\(^\text{10}\)Follow-up simulations indicated that both of the changes contributed to the observed decrease in RIF.

\(^{11}\)Another possibility is that participants encode Apple using different semantic features in the presence of Zinc vs. the presence of Fruit (M. C. Anderson, personal communication). If the Apple features that are active at study (and pop up at practice) are different from the Apple features that are active during pre-training, this will reduce the extent to which weakening of Apple (during the practice phase) affects the Zinc-Apple memory.

\(^{12}\)For example, if participants routinely use the “study context” tag as a cue during the final test phase, and the learning algorithm weakens the connections between this tag and competitor features, this could result in some degree of RIF regardless of what other (non-contextual) information is included in the retrieval cue. In this simulation, it was not necessary to include contextual cuing during the final test phase in order to generate the key difference in RIF between standard and external cues (hence, for simplicity’s sake, we left it out). However, it seems possible that other findings will emerge that can only be explained by considering how context-item associations are adjusted at practice, and how these associations are utilized at test.
worth noting that the Perfect et al. (2004) paper also includes experiments where the external cue was presented during the main study phase (Experiments 1 and 2), and these studies still failed to find RIF for the external cue. However, crucially, these studies used faces as the “external cues” and words as the retrieval targets. Given that participants were trying to retrieve words (but not faces) at practice, it is unlikely that faces that were associated with competitors would have activated at practice, thus their efficacy as retrieval cues should be relatively preserved.

This explanation leads to a simple prediction: If Experiment 3 of Perfect et al. (2004) were modified such that the novel associations (e.g., Zinc-Apple) were mixed in with the main study phase, in a way that made it difficult for participants to filter out these associations based on context, this should boost the amount of RIF observed for novel-associate cues like Zinc.

**Simulation 4: Nonmonotonic retrieval practice effects and prefrontal cortex**

**Background** In this simulation, we explore the effects of top-down attentional modulation (mediated by PFC) on competitor punishment. According to biased competition theories of cognitive control, top-down attentional modulation is especially important in situations where the target memory is substantially weaker than competing memories. In this situation, PFC supports recall of the target memory by sending extra activation to the distinctive features of that memory (thereby “biasing the competition” toward the target; Miller & Cohen, 2001; Desimone & Duncan, 1995).

To explore interactions between PFC and competitive dynamics in the model, we need to decide on a way of operationalizing top-down PFC control. The simplest way to accomplish this in the model is to vary the strength of the cue. When the unique features of the target memory are receiving strong external input, this constrains the network to retrieve memories that are consistent with the cue. Conversely, when the unique features of the target memory are receiving relatively weak external input, the network may settle into an attractor state that is inconsistent with the cue.

To constrain our account of how PFC intervenes at retrieval, we set out to simulate data from Johnson and Anderson (2004) (Experiment 1) on how recall of subordinate word meanings affects subsequent recall of the dominant meaning. In this study, Johnson and Anderson (2004) selected homographs with a dominant noun meaning and a subordinate verb meaning; e.g., for the word “prune”, the dominant (noun) meaning is “a fruit derived by partially drying a plum” and the subordinate (verb) meaning is “to trim” (as in, “prune your hedge”).

This experiment did not have separate study and practice phases. Instead, participants were repeatedly induced to retrieve the subordinate (verb) meaning of the homograph by completing fragments consistent with that meaning (e.g., prune-t_·m). Each time that a target word like “prune” was presented, it was presented with a different verb associate (in fragment form). Johnson and Anderson (2004) measured the effects of verb retrieval practice on the accessibility of the dominant noun meaning by asking participants to recall the noun meaning using an independent cue. For example, after practicing recall of the verb meaning of “prune”, the accessibility of the dominant noun meaning (“fruit”) might be tested by asking participants to complete “Yogurt-F” with the first word that comes to mind.

Figure 26 shows the results of this study. The main finding is that verb retrieval practice has a nonmonotonic effect on the accessibility of the noun meaning: Initially, retrieving the verb meaning increases the accessibility of the noun meaning, but — with additional retrievals — the accessibility of the noun meaning starts to decrease.

Biased competition theories can explain the decreasing part of the curve by positing that PFC provides an extra “boost” to the verb meaning. Specifically: PFC increases the extent to which verb features are included in the cue, thereby ensuring that the verb meaning will win; when the noun meaning loses the competition, it gets weakened. However, this can not be the whole story: If PFC were always successful at getting the verb to win (and the noun to lose), we would expect a monotonic decrease in noun recall, instead of the nonmonotonic pattern observed here. To explain the observed pattern of results in terms of our model, we need to posit that, with a small number of practice trials, the noun meaning is winning the competition often enough to show a net increase in accessibility; but as the number of practice trials is increased, the verb meaning starts to take over. The challenge for our model is to provide a parsimonious account of why this shift occurs.

As we will show, the model can account for the shift, given the one key assumption that — during a verb retrieval trial — PFC does not provide its “boost” to the verb meaning right away. This delay in the onset of PFC control gives the network a chance to retrieve the noun meaning, thereby strengthening the noun meaning and incrementally weakening the verb meaning. Later in the trial, PFC kicks in and forces the network to retrieve the verb meaning (thereby strengthening the verb mean-
Effect of Verb Retrieval Practice: Experiment Data from Johnson & Anderson (2004, Exp. 1)

Figure 26: Graph of data from Johnson and Anderson (2004), Experiment 1. Repeated retrieval practice of the subordinate verb meaning of a homograph (e.g., PRUNE-trim) first boosts, then lowers recall of the dominant noun meaning (Fruit). Recall of the noun meaning was tested using an independent, extralist cue (Yogurt-F).
ing and weakening the noun meaning). In the model, the effects of strengthening an item (on subsequent recall of that item) tend to outweigh the effects of punishing that item, when it pops up as a competitor. Thus, the net effect of the dynamic described above (where, during the early part of the practice trial, the noun meaning wins the competition, then PFC gets the verb meaning to win) is that both the noun meaning and the verb meaning are strengthened. Eventually, due to ceiling effects on noun representation strength (which allow the verb to catch up in strength to the noun), the system starts to recall the verb meaning at the outset of the trial, instead of the noun. Once this shift occurs, no further noun strengthening takes place, and the model shows a steady decrease in accessibility of the noun meaning.

Methods For this simulation, we used two 8-unit patterns (corresponding to the noun and the verb meanings of a single homograph). The patterns had 50% feature overlap; the overlapping part can be construed as representing the shared lexical identity of the two words, and the unique parts of the two patterns can be viewed as representing the distinctive features of the noun and verb meanings (respectively). In the Boundary Conditions section below (see Figure 29), we show that high levels of pattern overlap are required in order for the model to show the nonmonotonic pattern of retrieval practice effects observed by Johnson and Anderson (2004).

In the first phase of the simulation, the noun and verb patterns were trained into the network using the oscillating learning rule. Both patterns were presented in their entirety; noun and verb trials were interleaved together at training. To establish a strength asymmetry between the noun and verb meanings, the noun pattern was presented for 225 study trials (in total) and the verb pattern was presented for 75 study trials in total.14 As with our prior simulations, the learning rate at study was set to .02.

In the second phase of the simulation, the model was repeatedly cued to recall the verb meaning. On each practice trial, we strongly cued the 4 features that were shared across the two patterns (external input strength = 1.0) and weakly cued the unique features of the verb pattern (external input strength = .1). We then oscillated inhibition for one full cycle (with learning turned on). This is meant to simulate the part of the trial where recall occurs in the absence of PFC control. Next, to simulate the effects of PFC coming on later in the trial, we increased the external input to the unique features of the verb pattern (external input strength = .5) and oscillated inhibition for another full cycle (again, with learning turned on). Because we wanted to trace out the pattern of results observed in the Johnson and Anderson (2004) experiment (depicted in Figure 26).

The results of our simulations are presented in Figure 27. First, we will briefly consider the effects of repeated retrieval of the (dominant) noun meaning (depicted in the left-hand part of Figure 27). The results of this simulation mirror the results of our previous simulations: Noun wins the competition easily, resulting in strengthening of the noun meaning and weakening of the verb meaning. The effects of repeated retrieval of the (subordinate) verb meaning (depicted in the right-hand part of Figure 27) are more complex: Recall of the verb meaning shows a monotonic increase as a function of number of practice trials. In contrast, recall of the noun meaning shows a nonmonotonic pattern, whereby it increases, then decreases, as a function of the number of verb practice trials. The latter finding reproduces the pattern of results observed in the Johnson and Anderson (2004) experiment (depicted in Figure 26).

As discussed earlier, the nonmonotonic effect of verb practice on noun strength occurs because, for the first few practice trials, both the noun meaning and the verb meaning win the competition, at different points in the trial: Before PFC comes on, noun wins; and after PFC comes on, verb wins. While this is happening, both the noun meaning and the verb meaning are strengthened, but the verb meaning shows more strengthening (because of ceiling effects on noun strength). Once the verb meaning gets strong enough to win without PFC intervention, the strength of the noun meaning starts to decline mono-

---

14 This large strength asymmetry was necessary to ensure that the noun pattern would be cleanly retrieved in response to weak verb cues at the beginning of the practice phase.

15 In the noun practice simulation, the verb (competitor) test cue is an independent cue and the noun (target) test cue is a dependent cue.
Simulation Results: Effects of Verb and Noun Retrieval Practice

Figure 27: Simulation of the effects of verb practice (left-hand plot) and noun practice (right-hand plot). Repeated noun practice results in a monotonic increase in noun strength and a monotonic decrease in verb strength. Effect of verb practice are more complex: Repeated verb practice has a nonmonotonic effect on noun strength, whereby it first increases, then decreases, as a function of the number of practice trials.

Boundary conditions In addition to showing that the model can reproduce the nonmonotonic pattern of results from Johnson and Anderson (2004), we also ran simulations to characterize the boundary conditions on this effect. In one set of simulations, we varied noun-verb pattern overlap (i.e., how many units the noun and verb patterns had in common). The default value for this overlap parameter was 50% = 4/8 units in common. In a second set of simulations, we examined the effects of varying PFC cue strength (i.e., the strength of the external input applied to verb units during the “PFC active” part of the practice trial). The default value for this PFC cue strength parameter was .5.

The results of these simulations are shown in Figure 29. The overlap simulation results are shown in the left-hand panel, and the cue strength simulation results are shown in the right-hand panel.

With regard to overlap, increasing overlap between the noun and verb patterns leads to increased punishment of the noun representation. This can be explained as follows: Increasing overlap increases the extent to which the noun pattern pops up (as a competitor) on trials where the verb pattern wins; this extra competition leads to extra punishment. When overlap is very low, then the noun pattern is hardly punished at all once the verb takes over; as such, the noun is able to retain its initial gains in strength, even after 8 practice trials. Put another way: When overlap is high, strengthening one representation tends to lead to punishment of the other; whereas, when overlap is low, one representation can be strengthened without hurting the other.

With regard to cue strength, reducing verb cue strength (during the “PFC active” part of the practice trial) from .50 to .30 does not have a large effect: A cue

---

Note that the verb representation does not have to surpass the noun meaning in strength before it starts to win. In the absence of PFC intervention, the verb representation still receives a small amount of input from the cue (external input strength = .1). This weak cue provides an extra boost to the verb representation and allows it to win before it has fully caught up in strength to the noun.
Figure 28: Plot of competitive dynamics during (attempted) verb retrieval. The four graphs show the activation of unique noun features and unique verb features during the practice trial, as inhibition is oscillated. Importantly, these graphs depict the first half of each practice trial, prior to PFC intervention; during the second half of each practice trial (not shown here), PFC intervenes and provides additional input to the verb representation. The four graphs illustrate retrieval dynamics during the 1st, 2nd, 3rd, and 6th practice trials, respectively. During the first practice trial, in the absence of PFC input, the noun wins the competition to be retrieved, and the verb pops up as a competitor. During the second practice trial, the noun still wins, but the verb representation shows more activation overall. By the third practice trial, the verb has taken over, but the noun still shows considerable activity. By the sixth practice trial, the verb dominates over the noun, such that the noun representation barely pops up at all during the low inhibition phase.
A Neural Network Model of Retrieval-Induced Forgetting

Figure 29: Effect of verb practice on noun recall, as a function of the overlap of the verb & noun patterns (left-hand figure) and the strength-of-cuing for verb features during the “PFC active” part of the practice trial (right-hand figure). Varying overlap changes the overall amount of noun punishment: For high overlap values, verb practice asymptotically results in a sharp decrease in noun recall, whereas — for lower overlap values — verb practice has a net positive effect on noun recall. Varying cue strength also affects the overall amount of noun punishment: For moderate PFC cue strength values, noun recall shows the nonmonotonic curve observed in Figure 27, whereby noun recall first increases and then asymptotically decreases. In contrast, for lower PFC cue strength values, noun recall monotonically increases as a function of verb practice.
strength value of .30 is still sufficient to cause the verb to win. However, when verb cue strength is less than .30 during the “PFC active” part of the practice trial, the noun pattern ends up winning during both the “PFC active” and “PFC inactive” parts of the practice trial. This leads to monotonic strengthening of the noun pattern, and monotonic weakening of the verb pattern. Another way of parsing the results shown in the right-hand panel of Figure 29 is that the asymptotic level of RIF is higher given high (vs. low) PFC cue strength.

Taken together, the overlap and cue strength findings show that nonmonotonic effect of verb practice on noun recall (whereby noun recall after 8 practice trials is similar to noun recall before practice) is highly parameter-dependent. Depending on the exact amount of overlap, and the efficacy of PFC intervention in the recall process, it is possible to observe a monotonic increase in noun recall (e.g., with low PFC cue strength), and it is also possible to greatly increase the amount of noun punishment by increasing the overlap of the noun and verb patterns.

Discussion This simulation shows that the model can explain findings from paradigms where the competitor is (initially) strong enough to be recalled in place of the target representation. In order to generate the non-monotonic pattern observed by Johnson and Anderson (2004), we had to posit that PFC intervened in recall (forcing the system to recall the weak target pattern), and also that PFC intervention is not perfectly efficient (such that the strong competitor sometimes wins out over the weak target pattern). We also showed that the nonmonotonic effect of verb practice on noun recall is only present for certain parameter settings (e.g., 50% pattern overlap, but not 38% pattern overlap).

Importantly, this particular simulation is relevant, not only to the homograph paradigm used by Johnson and Anderson (2004), but also to the “think-no-think” (TNT) paradigm used by Anderson and Green (2001) and Anderson, Ochsner, Kuhl, Cooper, Robertson, Gabrieli, Glover, and Gabrieli (2004). In the TNT paradigm, participants study novel paired associates (e.g., “elephant-wrench”) until their memory for the pairing is very strong. Then, during a second phase, they are given cue words (e.g., “elephant”) and are asked to either think of the trained associate (the think condition) or to avoid thinking of the trained associate (the no-think condition). Later, memory for the associate (“wrench”) is probed using a category-cued recall test (“Name a tool that starts with w”). Anderson and Green (2001) found a significant RIF effect for items assigned to the no-think condition after 16 no-think trials, but the RIF effect was often not evident after a smaller number of no-think trials (e.g., 8 trials). Anderson and Green (2001) argue that, on no-think trials, participants focus on diversionary thoughts (basic semantic properties of elephants like “elephants are wrinkly”) in order to prevent themselves from thinking of the strong associate. Framed this way, the TNT paradigm is almost exactly isomorphic to the homograph paradigm: As in the homograph paradigm, the competitor (the trained associate) is much more accessible than the target (the participant’s pre-experimental semantic representation of elephant). In the TNT paradigm, this difference in accessibility is a function of the trained associate having been repeated several times recently in the study context. Also, just as PFC had to intervene in the homograph paradigm to ensure that the (weak) verb representation won the competition, PFC needs to intervene in TNT to ensure successful recall of (relatively weak) diversionary thoughts (Levy & Anderson, 2002).

Given this close mapping, all of the predictions derived in the context of the noun-verb simulation should apply here. For example, as per the “PFC cue strength” simulations shown in Figure 29, manipulations that increase the efficacy of PFC intervention should lead to increased (asymptotic) RIF for the no-think item. A recent TNT study conducted by Hertel and Calcaterra (2005) provides support for this view. Hertel and Calcaterra (2005) reasoned that participants would be able to use cognitive control more effectively (and thus show more RIF) if they were given a specific “substitute target” to think about, in place of the no-think item. In keeping with this prediction, participants in the substitute-target condition showed much more RIF than participants who were not given substitute targets. Another prediction is that, when PFC intervention is moderately effective (such that the no-think item occasionally pops into consciousness, but participants can suppress it most of the time) it should be possible to observe a nonmonotonic competitor punishment effect (as per Figure 27). In support of this view, Depue, Banich, and Curran (in press) ran a TNT study with emotionally valenced stimuli (which should be especially difficult to block out). They observed a robust non-monotonic pattern, where participants showed facilitation of competitor recall after five no-think trials and inhibition of competitor recall after 10 no-think trials.

General discussion

The research presented here shows how a small number of simple learning principles can be used to account for a wide range of RIF findings. Specifically, we described a learning algorithm incorporating the principles that:

- Lowering inhibition can be used to identify competing memories so they can be punished
• Raising inhibition can be used to identify weak parts of memories so they can be strengthened.

Using these principles, the model can simulate RIF results ranging from effects of competitor strength, to effects of retrieval practice vs. repeated presentation, to the nonmonotonic competitor punishment effects found by Johnson and Anderson (2004). Furthermore, the model leads to several novel predictions regarding boundary conditions on these effects.

The discussion section is divided into four parts.

• First, we discuss how our model relates to other theories of RIF. This section covers the role of competitive dynamics in driving learning; how “associative unlearning” theories of RIF can be reconciled with theories that posit weakening of the competitor itself; the role of PFC and top-down executive control in modulating RIF; how our model relates to other neural network models of learning and memory; and how our model relates to abstract computational models of memory.

• Second, we provide an overview of novel behavioral predictions generated by the model, regarding how various factors will modulate competitor punishment and target strengthening.

• Third, we discuss challenges for the model; specifically: how to provide a principled account of RIF effects occurring at different levels of the processing system (e.g., in episodic vs. semantic memory); how to account for data on the (possibly transient) time-course of RIF; how to account for the effects of target-competitor similarity in RIF; and developing a more detailed model of how PFC interacts with other structures during memory retrieval.

• Fourth, We discuss other applications of the model (besides modeling RIF data); specifically: we discuss our prior (Norman et al., in press) and ongoing attempts to characterize the functional properties of the oscillating learning algorithm (e.g., how many patterns can it store, compared to other algorithms); we also discuss other psychological domains involving competitor punishment that could be addressed by the model.

**Theoretical implications**

**How competitive dynamics drive learning**

One of the most important ideas presented here is that the amount of learning that occurs (on a given trial) is a function of the net input differential between the target memory and competing memories. Assuming that the target memory wins the competition (i.e., target units receive more net input than competitor units), then more learning occurs when the margin of victory for the target memory is small vs. when the margin of victory is large. In the paper, we showed that this simple framework can explain several important data points, including:

• The finding that more competitor punishment occurs given partial vs. full practice (Anderson et al., 2000a; see Figure 9 and Figure 11).
• The generation effect, whereby more target strengthening occurs given partial vs. full practice (Slamecka & Graf, 1978; see Figure 9 and Figure 10).
• The finding that strong competitors are punished more than weak competitors (Anderson et al., 1994; see Figure 13 and Figure 15).

We also discussed two data points that appear to be inconsistent with the simple competitive-learning framework outlined here:

• The finding from Anderson et al. (1994) that increasing target strength did not reduce competitor punishment.
• The finding that full and partial practice can lead to equivalent levels of target strengthening (i.e., no generation effect is observed; for an example, see Ciranni & Shimamura, 1999).

In both cases, we were able to reconcile these findings with the competitive-learning framework based on the observation that (in these studies) target recall was not perfect. We ran simulations showing that, when the target does not win cleanly over competing units, this mitigates both competitor punishment and target strengthening effects (see Simulation 2, Figure 17 and Figure 18). Perhaps the most important contribution of this competitive-learning framework is that it provides a straightforward way of characterizing boundary conditions on RIF effects; these predictions are reviewed in the Summary of predictions section below.

**Generalized forgetting via weakening of attractor states**

In this section, we explore the implications of this work for extant psychological theories of forgetting.

**Blocking vs. weakening**  As discussed in the Introduction, some theories of forgetting (such as Anderson’s) posit that forgetting is driven — at least in part — by actual weakening of stored memory traces. In contrast, ratio rule theories (also referred to in this paper as blocking theories) posit that impaired competitor recall is an indirect consequence of target strengthening, and that no
actual weakening of the competitor takes place. In accordance with Anderson’s theory, our model posits that weakening of stored memory traces contributes to RIF (although blocking can also contribute; see discussion below). To illustrate that RIF in our model is not entirely attributable to blocking, we ran simulations where we turned off learning during the “high inhibition” (target strengthening) part of the inhibitory oscillation at practice, and showed that RIF effects are still obtained (see *Simulation 1, 2, and 3*).

**Associative unlearning vs. inhibition** Within the realm of models that posit actual weakening, Anderson distinguishes between associative unlearning models and “truly inhibitory” models of weakening (see, e.g., Anderson, 2003 and Anderson & Bjork, 1994). As illustrated in Figure 2, associative unlearning involves decrementing the connection between the cue (Fruit) and the competitor (Apple). In contrast, “true inhibition” (using Anderson’s terminology) involves weakening the Apple representation itself. According to Anderson & Bjork, 1994, the key differential prediction between these theories is that, if the representation of Apple (itself) is inhibited, then impaired recall should be observed to some extent with all cues. In contrast, with associative unlearning, impaired recall should only be observed when the same cue (Fruit) is used at practice and at test.

While the latter prediction (i.e., that forgetting should be limited to the cue “Fruit”) appears to follow naturally from associative unlearning theory, we think this inference is incorrect. We believe that this mistaken inference has been fostered, at least in part, by certain misleading aspects of the diagram used to depict associative unlearning theory in Figure 2:

- First, the diagram depicts Fruit and Apple as being distinct, unitized nodes in a conceptual network. Contrary to this idea, we think that it is more appropriate to represent memories as attractor states comprised of multiple, interconnected “microfeatures”. According to this view, the strength of an memory (i.e., its ability to capture activation during a competitive retrieval situation) is a function of the strength of the connections between its constituent features.

- Second, we think the idea of representing Fruit as being completely distinct from Apple is misleading. As discussed in *Simulation 3*, category cues like Fruit can be viewed as part of the apple attractor (insofar as Fruit is one of the essential features of the Apple concept; see Figure 24 for a graphical depiction of this idea).

If we accept the two premises outlined above (that the Apple memory is comprised of interconnected microfeatures, and that some of those microfeatures pertain to the Fruit concept), this implies that weakening the connection between Fruit features and other Apple features makes the Apple attractor weaker, overall. This, in turn, should lead to some amount of generalized forgetting (i.e., if the attractor is weaker, it should be generally harder to activate). The size of this generalized (“cue-independent”) impairment should be proportional to the fraction of connections that were weakened (and how much they were weakened).

Our model goes one step beyond associative unlearning theory, in the sense that weakening is not limited to connections between Fruit features and other Apple features. As discussed earlier, the learning algorithm weakens connections between features that pop up during the low inhibition phase and *all of the other features that happen to be active at the time*. Thus, in addition to weakening the connections between the features of the cue and competing features, the learning algorithm also weakens connections between the competing features themselves (see Figure 4, Figure 24, and Appendix B). The broad-based nature of this weakening helps to promote generalized RIF effects. Nonetheless, the fact that weakening only occurs between Apple features and other active features (as opposed to equally affecting all connections involving Apple features) implies that RIF effects will be larger for some cues than others. Specifically, we showed in *Simulation 3* that cues that remain inactive at practice retain their efficacy better that cues that were active at practice (see Figure 24). This property of the model may help to explain why Perfect et al. (2004) observed less RIF for novel associate cues from the pretraining phase than for more standard types of cues.

**The role of PFC and cognitive control**

PFC is not necessary for RIF Anderson’s recent writings on RIF have emphasized the role of top-down executive control (implemented by PFC) in RIF (e.g., Levy & Anderson, 2002; Anderson, 2003). We agree with the idea that PFC plays a large role in RIF. However, we disagree with the idea that PFC plays a necessary role in competitor punishment. According to our theory, RIF is a consequence of competition between memories (e.g., in the medial temporal lobes), and local learning processes that operate based on these competitive dynamics. So long as there is competition, there will be competitor punishment. PFC can influence *which memories are punished* and *to what extent* these memories are punished by biasing the retrieval competition in favor of one of the memories (as in *Simulation 4*). However, insofar as competition still occurs in the absence of PFC intervention, we expect that it should be possible to observe competitor punishment without PFC (so long as the target memory is strong enough to win without PFC intervention).
Consistent with this view, Conway and Fthenaki (2003) found intact RIF in a group of patients with frontal lobe lesions.

*How PFC contributes* In *Simulation 4*, we operationalized the contribution of PFC in terms of a simple manipulation of cue strength. According to this view, PFC can bias retrieval during the practice phase by increasing the amount of external input that is applied to the unique features of the target item. In order to simulate the competitor-punishment data in *Simulation 4*, we had to also make the assumption that PFC intervention is not perfectly efficient (i.e., biasing input from PFC is sometimes not available at the very start of a trial; thus, the system might retrieve the wrong memory at the beginning of a trial before PFC intervenes). In future work, we will explore mechanisms that might account for the delayed onset of PFC involvement (see the *Modeling the dynamics of top-down control* section below).

Importantly, in addition to ensuring that weak targets win out over strong competitors, we think that PFC plays a more pervasive role in limiting blocking effects at test. Predictions regarding how PFC damage should affect performance on RIF paradigms are described in the *Summary of predictions* section below.

*Comparison to other neural network models*

Our model is the first to address the full constellation of RIF phenomena discussed here (in particular, cue-independence, competition-dependence, and nonmonotonic retrieval practice effects). To our knowledge, the only other neural network model that has specifically tried to address RIF data (in any fashion) is a recently developed model by Oram and MacLeod (2001). Below, we provide a brief overview of the Oram and MacLeod (2001) model. We argue that, although their model can explain the basic finding that practice helps recall of the practiced item, and hurts recall of similar non-practiced items, it lacks the requisite mechanisms that would allow it to model the “competition-dependence” of RIF (as exemplified, e.g., by the finding that RIF effects are larger after partial practice vs. full practice). After discussing the Oram and MacLeod (2001) model, we discuss (in broader terms) the properties that neural network models must have in order to selectively punish strong vs. weak competitors. Finally, we discuss the possibility that the BCM learning algorithm (Bienenstock, Cooper, & Munro, 1982) might be able to account for competition-dependent learning.

*The Oram & MacLeod (2001) model of RIF* This model consists of a two-layer network, where input nodes (each corresponding to a specific item) are connected in a feedforward, diffuse fashion to a set of “memory nodes” that serve as an internal representation of the inputs. Connections in the model are modified according to simple Hebbian learning principles, whereby connections between active input units and active memory nodes are strengthened, and connections between inactive input units and active memory nodes are weakened (for additional background on this kind of learning rule, see O’Reilly & Munakata, 2000 and Grossberg, 1976). In the Oram and MacLeod (2001) model, items that are grouped together at study end up getting linked to a shared set of memory nodes. Subsequently, when one item from the group is practiced, this has two effects:

- Connections between the practiced item’s (active) input node and the shared memory nodes are strengthened.
- Connections between the non-practiced items’ (inactive) input nodes and the shared memory nodes are weakened.

This fact allows Oram and MacLeod (2001) to explain facilitated recall of the practiced item, and impaired recall of non-practiced items from the same group. Oram and MacLeod (2001) do not try to address any of the more complex RIF phenomena described in this paper (e.g., data indicating cue-independence and competition-dependence). A possible problem with the Oram and MacLeod (2001) model (with regard to simulating these more complex phenomena) is that it lacks an explicit mechanism for punishing strong vs. weak competitors; the Hebbian learning rule used in the model weakens (in a non-selective fashion) connections from all inactive input units, instead of specifically targeting strong competitors. In light of this fact, it seems very unlikely to us that the Oram and MacLeod (2001) model will be able to provide a principled and comprehensive account of the learning phenomena that (in our model) we explain in terms of competition-dependent learning.

*How to get competition-dependent learning* Rather than limit ourselves to models that specifically have mentioned RIF, it is worth considering more broadly whether there are other neural network learning principles that could account for the “competition-dependence” of RIF.

Most learning algorithms for rate-coded neural networks (e.g., the Hebbian rule used by Oram & MacLeod, 2001, and O’Reilly’s Leabra algorithm; O’Reilly & Munakata, 2000) learn based on the final settled state of the network, without factoring in the patterns of activation that are present (possibly transiently) during the settling process. The problem with this approach is that the final settled state of the network can be very similar for high-competition and low-competition trials (making it difficult to enact differential learning in these situations). For example, consider the finding that RIF is larger after full practice (Fruit-Pear) than after partial practice (Fruit-Pe____) (e.g., Anderson & Shivde, in preparation). If we...
assume that the final state of the network is the same in both cases (“Fruit-Pear”) then there is no way for algorithms like Leabra to enact more punishment in the partial-cue condition than the full-cue condition.

There are two possible responses to this problem: One solution is to allow competitors to be active in the final settled state of the network. For example, in the full vs. partial practice example, one could argue that competitors (Apple) remain weakly active in the Fruit-Pear condition, but not in the Fruit-Pear condition. Another solution (which we chose) is to use a learning algorithm that is sensitive to states of network activation that occur prior to the final settled state. Algorithms that have this property can learn based on competitor activation, even if this activation is transient.

Once a learning algorithm has solved this problem of how to “detect” (possibly transient) differences between high-competition and low-competition trials, it has to solve the problem of how to punish the competitor without also hurting the target item. The mere fact that the competitor is active can not be sufficient to trigger punishment of that item (or else all active representations will be punished, not just the competitor). The oscillating algorithm solves this problem by changing the sign of the learning rule based on the phase of the inhibitory oscillation; thus, increased activation that occurs during the start of low-inhibition phase (when the network is “peeking” below threshold) has a different consequence than increased activation that occurs during the end of the high-inhibition phase (when the target representation is coming back on).

The BCM algorithm and competitor punishment
Another algorithm besides ours that could, in principle, solve the problem of competitor punishment is the BCM algorithm (Bienenstock et al., 1982). Like the simple Hebbian learning algorithm used by Oram and MacLeod (2001), BCM strengthens connections between active sending units and strongly active receiving units. The critical property of BCM, with respect to competitor punishment, is that it reduces synaptic weights from active sending units that are activated above zero but below its average level of activation. Put simply: When an input pattern elicits weak activation in a receiving unit, the connections between the input pattern and the (weakly activated) receiving unit are weakened. So, in the “partial practice” example, if we posit that Fruit-Pear elicits strong activation of Pear, weak activation of Apple, and no activation of Shoe, the BCM algorithm will strengthen connections to Pear, weaken connections to Apple, and it will not affect connections to Shoe. This property suggests that it is worth exploring whether BCM can account for the full range of RIF findings discussed in this paper.\footnote{While (to our knowledge) no one has used BCM to address RIF data, some studies have used BCM to address competitive learning phenomena in other domains. For example, Gotts and Plaut (2005) show that BCM can account for data from a perceptual negative priming paradigm, where participants are asked to attend to a visual stimulus and ignore another (simultaneously presented) visual stimulus. Negative priming refers to the effect of ignoring a stimulus on participants’ ability to (subsequently) respond to that stimulus; see Fox, 1995 for a review.}

Comparison to abstract computational models of memory
Abstract memory models like SAM (Search of Associative Memory; Raaijmakers & Shiffrin, 1981) and REM (Retrieving Efficiently from Memory; Shiffrin & Steyvers, 1997) have proved to be very useful in understanding interference effects in memory (for a recent review, see Raaijmakers, 2005; see also Reder, Nhouyvanisvong, Schunn, Ayers, Angstadt, & Hiraki, 2000 for description of another relevant model). These models posit that memory traces are placed in a long-term store at study, without any sort of structural interference between memory traces. At test, cues activate stored traces to varying degrees, and these activated traces compete to be the one that gets retrieved. Although no published papers have specifically addressed the RIF phenomena described here using models like SAM and REM, we can discuss (in a general sense) the relationship between the kinds of explanations that are offered by these models, and the explanations that are provided in this paper.

The hallmark of the abstract-modeling approach, as applied to forgetting data, has been to show that phenomena that (previously) were thought to be attributable to unlearning (e.g., retroactive interference in AB-AC interference paradigms; Barnes & Underwood, 1959) can actually be explained by ratio-rule models (Mensink & Raaijmakers, 1988). This work is very important — in addition to giving the field a more robust appreciation for the power of ratio-rule models, it has also led researchers to think more carefully about the role of retrieval cues...
in determining forgetting effects (in particular, the role of contextual cues, and how temporal drift in contextual cues can cause forgetting; Mensink & Raaijmakers, 1988; Howard & Kahana, 2002).

The major substantive difference between our modeling approach and the abstract-modeling approach is that our model incorporates “unlearning” at the synaptic level. While we appreciate the analytic utility of trying to explain as much data as possible using blocking alone (Mensink & Raaijmakers, 1988), without positing any kind of unlearning, there is extensive evidence for activity-dependent synaptic weakening in the brain (e.g., Malenka & Bear, 2004). Furthermore, at a very basic level, network models (like ours) need bidirectional plasticity (both strengthening and weakening) in order to function properly (O’Reilly & Munakata, 2000). 18

While our inclusion of unlearning mechanisms may appear to put us into direct conflict with models like Mensink and Raaijmakers (1988), we do not think this is the case — the models operate at different levels. As discussed many times throughout this paper, we strongly agree with the claim that blocking can cause forgetting; the competitive retrieval dynamics that give rise to blocking are a deep-seated property of our model (and of recurrently-connected neural network models, more generally). As such, abstract “ratio-rule” models can be viewed as providing a high-level account of competitive retrieval dynamics in network models like ours. The goal of the work presented here is to flesh out this high-level account with a more mechanismically detailed theory of how that competition transpires in the brain, and how these dynamics interact with low-level synaptic weakening (and strengthening) processes.

Summary of predictions

This section provides a brief overview of the novel model predictions discussed in the main part of the paper. Each prediction is linked back to the section of the paper where it was first discussed.

Effects of full vs. partial practice

- It should be possible to observe RIF in the full practice (“extra study”) condition if the target is extremely weak and the competitor is extremely strong (discussed in Simulation 1).

- The generation effect (more strengthening for partial vs. full cues) should interact in a nonmonotonic fashion with the strength of the target (relative to competitors). According to the model, the generation effect should be maximal when the target is just strong enough win cleanly over the competitor during partial practice. If the target is too weak, leading to imperfect target retrieval dynamics during partial practice, the model predicts a reverse generation effect: more strengthening after full vs. partial practice (see Simulation 2, Figure 17). Likewise, if the target is too strong, this should reduce the generation effect, by reducing the overall amount of competition-based learning that occurs at practice (in the limit, if target strength is already at ceiling prior to practice, no further strengthening will occur in the partial practice condition or the full practice condition).

Competitor strength effects

- Manipulations that boost blocking (by reducing the amount of unique information in the test cue, and increasing the amount of shared information) should reduce the competitor strength effect, by selectively boosting RIF for the weak competitor (see Simulation 2, Figure 19 and Figure 21).

- In the model, competitor punishment is a function of the strength of the competitor relative to the target, and also the strength of the competitor relative to other competitors. As such, selectively increasing the strength of other competitors should reduce RIF for non-strengthened competitors (see Simulation 2, Figure 22).

Target strength effects

- Target strength should have a nonmonotonic effect on RIF: When target recall is imperfect, increasing target strength can boost RIF; as soon as the target is strong enough to cleanly win over competing memories, further increases in target strength should reduce RIF (see Simulation 2, Figure 18).

RIF using external cues

- According to the model, the amount of RIF elicited by a particular “independent” cue is — in part — a function of whether that cue is retrieved during the low-inhibition phase at practice. In Simulation 3, we explained Perfect et al.’s (2004) finding of less RIF for external cues (e.g., “zinc-apple”) than standard cues in terms of the idea that, at practice, participants could use contextual information to focus recall on the study phase (and block retrieval of external associations learned during the pretraining phase); see Figure 24 and Figure 25. This implies that manipulations that make it more difficult

18Most neural networks are set up to optimize some kind of mapping between inputs and outputs — to accomplish this goal, the network needs to be able to weaken connections when target outputs are too strong, in addition to strengthening connections when target outputs are too weak. If there is no weakening mechanism to counteract strengthening, the synapses in the network will eventually saturate, which (in turn) will lead to catastrophic recall failure.
to use “contextual blocking” (e.g., temporally intermixing the learning of external associations with the main study phase) should boost the amount of RIF elicited by external cues.

**RIF effects when the competitor is strong enough to displace the target**

These predictions apply to situations where (in the absence of PFC intervention) the competitor is strong enough to displace the target memory.

- The asymptotic effect of target practice on competitor recall should be highly dependent on the similarity of the competitor and the target. For low similarity values, the asymptotic effect should be facilitation of the competitor; for high similarity values, the asymptotic effect should be significant weakening of the competitor, far below its pre-practice memory strength value (see **Simulation 4**, Figure 29).

- The asymptotic effect of target practice on competitor recall should also depend on how strongly the unique features of the target are cued at practice. Reducing the amount of external input that is applied to these features should tilt the practice curve upward, so net facilitation of the competitor is observed instead of net punishment (see **Simulation 4**, Figure 29).

### Effects of PFC damage/dysfunction

As mentioned above, we think that PFC plays a major role in imposing retrieval constraints on the system. As such, any of the above predictions that relate to reduced cue specificity or cue strength should be applicable to populations with PFC damage/dysfunction (e.g., elderly adults). For example, Figure 21 shows how reducing cue specificity leads to an overall increase in RIF, but the increase is highest for the weak (vs. strong) competitor in the dependent-cue condition (vs. the independent-cue condition). This implies that populations with PFC damage/dysfunction (who, by hypothesis, show reduced cue specificity) should show a differential increase in RIF for weak competitors in the dependent-cue condition. Also, Figure 29 shows that reducing the amount of input that is applied to unique target features at practice results in net facilitation (as opposed to punishment) of the competitor in the Johnson and Anderson (2004) homograph paradigm. As such, populations with PFC dysfunction should show a monotonic increase in recall of the dominant (noun) homograph meaning, as a function of the number of verb practice trials.

### Neurophysiological predictions

If the link between the oscillating algorithm and theta oscillations (as described in the *Theta oscillations* section above) is valid, the model can be used to make predictions regarding the fine-grained activation dynamics of target and competitor representations. According to the model, the activation of competitor representations should increase at a fixed phase of theta (corresponding to the “low inhibition” phase), and the activation of the target representation should dip at a fixed phase of theta (corresponding to the “high inhibition” phase) that is 180 degrees out of phase with the “competitor bump.” The idea that activation dynamics (with respect to theta) should vary for items receiving high levels of net input (targets) vs. items receiving less net input (competitors) receives some support from the rat navigation electrophysiology literature: Several studies have found that a place cell will fire during a specific theta phase when the rat is in the preferred place of the cell, and that the firing will shift phases as the rat moves from this preferred location (see, e.g., O’Keefe & Recce, 1993; Yamaguchi, Aota, McNaughton, & Lipa, 2002; see also Mehta, Lee, & Wilson, 2002).

The model predicts that the theta-locked “competitor bump” and “target dip” for a given stimulus should both decrease in size as a function of experience with that stimulus (see Figure 7). Importantly, the model also predicts that the size of the competitor bump can be used to predict RIF — a large “competitor bump” should result in extensive punishment of that competitor, and a smaller bump should lead to less punishment.

Testing the above predictions will require methodological advances in neural recording: Specifically, we will need a means of reading out the instantaneous activation of the target and competitor representations, and relating these activation dynamics to theta. One way to accomplish this goal is to use pattern classification algorithms, applied to thin time slices of electrophysiology data (on the order of milliseconds) to isolate the “neural signatures” of the target and competitor representations. Once the pattern classifier is trained, it can be used to track the activity of these representations over time (and across phases of theta). Pattern-classification studies meeting these desiderata are underway now in our laboratory.

### Challenges for the model

In this section, we discuss several important challenges for the model, and ways that the model could be modified to address these challenges.

#### Multiple levels of representation

In this paper, we have discussed examples of RIF occurring in lexical-semantic memory (e.g., Johnson & Anderson, 2004) and episodic memory (e.g., Anderson & Bell, 2001). Anderson (2003) discusses how RIF may
also extend to lower-level perceptual representations, although evidence is presently lacking. It also seems possible that competitor-punishment can occur for high-level task representations in prefrontal cortex (e.g., Mayr & Keele, 2000) found that switching away from a task makes it hard to return to that task later; at the moment of the switch, the new task can be viewed as the “target” and the old task can be viewed as a “competitor” that needs to be punished.

Given the fact that RIF can occur for multiple different kinds of representations, identifying the level(s) of representation where competition is occurring in a particular paradigm is critical for making accurate predictions about RIF. For example, if competition is primarily occurring at a conceptual level, we would expect RIF to be observed when memory is probed using conceptual implicit memory tests (see, e.g., Perfect, Moulin, Conway, & Perry, 2002). However, if competition is primarily occurring at the episodic level (such that episodic memories are punished, but conceptual memories are largely intact), we would expect minimal levels of RIF on conceptual implicit memory tests (after controlling for explicit contamination; see, e.g., Camp et al., in press).

In its present (one-layer) form, the model does not provide a natural way of exploring how competition might differ in episodic vs. semantic memory systems, or how these systems might interact. The present version of the model is more “semantic” than “episodic” — it can simulate the gradual acquisition of semantic knowledge (via multiple, interleaved presentations of the to-be-learned patterns), but it lacks the requisite machinery for rapid, one-trial memorization of specific episodes. Also, since the model lacks a hierarchical structure, it lacks a principled way of predicting where in the cortical representation hierarchy (from low-level perceptual representations to high-level representations of concepts and tasks) competition is most likely to occur.

The only way to deal with these issues is to gradually flesh out the model so that the structure of the network more accurately reflects the structure of processing in the brain. Specifically, we could use the Complementary Learning Systems (CLS) architecture that we have used in prior simulation work (e.g., Norman & O’Reilly, 2003; O’Reilly & Rudy, 2001). The key features of this architecture are a hierarchical, multilayer cortical system that is responsible for gradually extracting statistical regularities from the environment, and (at the top of the cortical hierarchy) a hippocampal network that is responsible for rapidly memorizing patterns of cortical activity, in a manner that supports subsequent recall of those patterns based on partial cues. According to the CLS framework, cortex assigns similar representations to stimuli that share features (like our one-layer network does now) whereas hippocampus is biased to assign distinct representations to stimuli, regardless of their similarity; the latter property makes it possible for hippocampus to rapidly memorize episodes without suffering from catastrophic interference. We have already made progress toward implementing the oscillating algorithm in the CLS architecture: Norman et al. (in press) implemented the oscillating algorithm in a two-layer hierarchical cortical network, consisting of an input-output layer (where patterns were presented to the network) that is bidirectionally connected to a hidden layer. We are also in the process of implementing the algorithm in a hippocampal architecture (see Norman & O’Reilly, 2003 for a description of the hippocampal network that we are using; see also Norman, Newman, & Perotte, 2005 for ideas regarding how to implement the oscillating-inhibition algorithm in this network). Once the oscillating algorithm has been implemented in both the cortical and hippocampal networks, this combined architecture should be capable of simulating RIF effects at multiple levels of representation (for more discussion of interactions between RIF in semantic and episodic memory, see Bauml, 2002).

Learning hidden representations in hierarchical networks One important difference between multi-layer, hierarchical networks and our single-layer network is that — in hierarchical networks — only some of the layers (at the bottom of the hierarchy) receive external input. The other layers (e.g., the hidden layer in our 2-layer cortical network) are free to develop their own representations of input patterns. Norman et al. (in press) describe how the oscillating-inhibition learning rule works in this kind of multi-layer network. Specifically, they describe how — in addition to strengthening and weakening representations — the learning algorithm also changes the structure of hidden representations elicited by input patterns, in order to facilitate subsequent recall of these input patterns. For example, consider the case of two similar input patterns (A and B) that are repeatedly presented in an interleaved fashion. Initially, A will pop up as a competitor when B is studied, and B will pop up as a competitor when A is studied. When A activates as a competitor (on B trials), the competitor-punishment mechanism will dissociate the unique features of A from the hidden representation elicited by B (likewise, the competitor-punishment mechanism will dissociate the unique features of B from the hidden representation elicited by A). The net result of these changes is differentiation (Shiffrin, Ratcliff, & Clark, 1990; McClelland & Chappell, 1998; Norman & O’Reilly, 2003): As training progresses, the hidden representations of A and B will move farther and farther apart, until they are sufficiently distant that A no longer pops up as a competitor on B trials, and vice-versa. This differentiation...
process should have testable consequences (e.g., stimulus A should be less effective in priming stimulus B).

**Time-course of RIF**

Another challenge for the model is simulating data on the time-course of RIF. In the model, target strengthening and competitor punishment are both enacted through the same mechanism: modification of synaptic weights. This implies that, in principle, it should be possible to observe competitor-punishment effects that are as long-lasting as target-strengthening effects.

This view is challenged by a study conducted by MacLeod and Macrae (2001). In that study, MacLeod and Macrae (2001) manipulated the length of the interval between the end of the practice phase and the beginning of the test phase: In the “short delay” condition, this interval lasted 5 minutes; in the “long delay” condition, this interval lasted 24 hours. MacLeod and Macrae (2001) found robust competitor punishment and target strengthening after a 5-minute delay; after the 24 hour delay, target strengthening was largely intact but the RIF effect was gone. As things stand, this is the only study (that we know of) that has used delays lasting longer than a few hours to examine RIF.

If the MacLeod and Macrae (2001) finding (showing that RIF effects do not last across a single-day retention interval) reflects a general principle, and is not just a consequence of the specific circumstances of that study, this would be very problematic for our model. However, at this point in time, there are other possibilities that need to be pursued before positing (as a general principle) that RIF effects are transient. One interesting possibility relates to the effects of sleep on memory representations. Recently, Norman et al. (2005) presented simulations showing how the oscillating learning algorithm can be used to autonomously repair damaged attractor states: If noise is injected into a trained network (with no other external input), that noise will coalesce into stored attractor states. Norman et al. (2005) showed that, when inhibition is lowered, the competitor will not be active given normal inhibition, and thus does not “pop up” when inhibition is lowered, the competitor will not be punished.

Unfortunately, it is not possible to simulate this dynamic using kWTA, because of the rigidly enforced limit on the number of units that can be active. The most straightforward way to remedy this problem is to replace the kWTA inhibitory algorithm with explicitly simulated inhibitory interneurons. While this will increase the complexity of the model (and the complexity of the activation dynamics generated by the model), neural network researchers have made great strides in recent years toward understanding how to generate stable activation dynamics using a mixture of excitatory and inhibitory neurons (e.g., Wang, 2002). In networks with explicitly simulated inhibitory interneurons, the amount of activation elicited by a given input is an emergent property of interactions between excitatory and inhibitory interneurons (instead of being directly legislated by the inhibitory

---

19See Anderson (2003) for additional discussion of how target-competitor associations can mitigate RIF effects.
algorithm, as is the case with kWTA). As such, we expect that this architecture will have sufficient flexibility (in terms of the number of neurons that are allowed to be active) to account for target-competitor integration effects.

**Modeling the dynamics of top-down control**

Our current method of simulating PFC involvement in recall (i.e., varying cue strength), allows us to vary the degree of PFC involvement at a particular point in time, but it does not allow us simulate the fine-grained temporal dynamics of PFC involvement. To address this problem, we will explicitly implement a simple network architecture for conflict detection and cognitive control, as proposed by Botvinick et al. (2001). In that paper, Botvinick et al. (2001) propose that the function of anterior cingulate cortex (ACC) is to detect conflict between representations (where conflict is operationalized as “co-activity of incompatible representations”). When ACC detects conflict, this causes PFC to activate, which (in turn) serves to resolve the conflict. For example, in the Johnson and Anderson (2004) homograph paradigm, ACC would be set up to detect co-activity of the noun and verb representations. When co-activity is detected, this would trigger PFC activity, which would selectively boost activation of the verb representation (resolving the conflict). We expect that this model will allow us to generate much richer predictions about the dynamics of PFC intervention in memory retrieval, and how these dynamics influence learning. Having an explicitly simulated PFC will also allow us to directly simulate the effects of PFC damage/dysfunction on RIF.

**Other applications of the model**

The work presented here constitutes a first step toward understanding the neural basis of competitor-punishment, and we are currently working to further our understanding of the learning algorithm (and its relation to neural and behavioral data) in several different ways. One approach has been to assess the functional properties of the algorithm: Do the same features of the algorithm that help us explain RIF (in particular, its ability to punish competitors) also help the algorithm do a better job of memorizing patterns? Another approach has been to apply the model to psychological domains other than RIF. These two approaches are briefly reviewed below.

**Functional properties of the learning algorithm**

Norman et al. (in press) showed that, apart from its useful psychological properties, the oscillating algorithm also has desirable functional properties: Using a two-layer version of the model (i.e., with a hidden layer that is bidirectionally connected to the input layer; see the *Multiple levels of representation* section above), Norman et al. (in press) found that the oscillating algorithm outperforms several other algorithms (e.g., Leabra; O’Reilly & Munakata, 2000) at storing and retrieving correlated input patterns. As discussed by Norman et al. (in press), the oscillating algorithm’s good performance on these pattern memorization tasks is directly attributable to its ability to punish competing memories. Whenever the hidden-layer representations of different patterns blend together, they start to compete with one another at retrieval, and the competitor-punishment mechanism pushes them apart. In this manner, the oscillating algorithm manages to keep representations from completely merging into one another in the hidden layer, even when inputs overlap strongly. This extra degree of pattern separation helps to ensure that memories can be accurately stored and accessed even in difficult situations (e.g., when there are many similar memories stored in the system, and the cue only slightly favors one memory over the other). Importantly, Norman et al. (in press) also show that the oscillating algorithm’s ability to keep patterns separate does not compromise its ability to generalize to input cues that resemble (but do not exactly match) stored patterns. Unlike other network architectures (e.g., the hippocampal model from Norman & O’Reilly, 2003) that automatically assign distinct hidden representations to similar inputs, the oscillating algorithm (as applied to a cortical network architecture) is only concerned that memories observe a “minimum separation” from one another. So long as this constraint is met, memories are free to overlap according to their similarity (thereby allowing the network to enact similarity-based generalization).

**Other psychological data**

In this paper, we focused on a particular set of RIF results because we thought they were especially constraining, and also illustrative of the model’s unique properties. However, the RIF findings discussed here constitute only a small fraction of the space of findings from memory paradigms (and other types of paradigms) that could — in principle — be addressed by the model.

In one line of work, we have started to simulate familiarity-based recognition using a 2-layer version of the model, operationalizing familiarity in terms of the size of the decrease in activation during the high inhibition phase at training. As stimuli become more familiar, the size of the decrease in activation during the high inhibition phase gets smaller (see Figure 7 for an illustration of this phenomenon). Preliminary simulations show that the model’s capacity for supporting familiarity-based discrimination (operationalized in terms of the number of familiar and unfamiliar patterns that can be discriminated) is much higher than the capacity of a comparably
sized network using a simple Hebbian architecture (Norman et al., 2005). Future work will explore whether the oscillating-algorithm familiarity model can account for the full range of list-learning interference results (e.g., the null recognition list strength effect observed by Ratcliff, Clark, & Shiffrin, 1990) that were previously addressed, using a simple Hebbian familiarity model, by Norman and O’Reilly (2003).

In another line of work, we have started to use the model to account for data on how sleep affects memory for motor sequences (Walker, Brakefield, Hobson, & Stickgold, 2003), and — more generally — on how learning during sleep can help to mitigate memory interference (Norman et al., 2005). Future work will also address data relating to cognitive dissonance reduction (e.g., Freedman, 1965), negative priming effects in object perception (e.g., DeSchepper & Treisman, 1996), and backward inhibition effects in task switching (e.g., Mayr & Keele, 2000).

Conclusions

In the simulations presented in this paper, we showed that the oscillating-inhibition model can account for key qualitative regularities in the RIF data space (e.g., more RIF for strong vs. weak competitors). The model also provides a principled account of boundary conditions on these regularities. To our knowledge, this is the first computational model to address the set of RIF phenomena discussed here. However, we also realize that the model has a long way to go before it provides a comprehensive account of how the brain gives rise to RIF. As discussed in the Challenges for the model section above, we need to incorporate significantly more neurobiological detail in the model (e.g., we need a hippocampal network to simulate RIF in episodic memory; we need to explicitly simulate inhibitory interneurons to account for target-competitor integration effects). Also, in addition to testing behavioral predictions of the model, we need to start testing neural predictions (both coarse-grained predictions about, e.g., PFC damage, and fine-grained predictions regarding how target and competitor activation should be linked to theta phase). Overall, we believe that a convergent approach using behavioral constraints, neural constraints, and functional constraints (showing that our model learns efficiently, relative to other algorithms) will result in the most progress toward solving the puzzle of retrieval-induced forgetting.

Acknowledgments

This work was supported by NIH grant RO1 MH069456, awarded to KAN. We thank Tim Curran and Michael Kahana for their very insightful comments on an earlier draft of this manuscript.

Appendix A: Details of the oscillating learning algorithm

This appendix provides details of how the oscillating learning algorithm was instantiated in the simulations reported here. For more information on the oscillating algorithm and its functional properties, see Norman et al. (in press).

Our oscillating-algorithm simulations were implemented using a modified version of O’Reilly’s Leabra algorithm (O’Reilly & Munakata, 2000). Apart from a small number of changes listed below (most importantly, relating to the weight update algorithm, and how we added an oscillating component to inhibition) all other aspects of the algorithm used here were identical to Leabra. For a more detailed description of the Leabra algorithm, see O’Reilly and Munakata (2000).

As per the Leabra algorithm, we only explicitly simulated excitatory units and excitatory connections between these units; we did not explicitly simulate inhibitory interneurons. As mentioned in the main text, excitatory activity was controlled by means of a k-winners-take-all (kWTA) inhibitory mechanism (O’Reilly & Munakata, 2000; Minai & Levy, 1994). The kWTA rule sets inhibition such that at most $k$ units show activation values $> .25$. The $k$ parameter was set to match the number of units in the input patterns (for most simulations, $k = 8$).

To implement the inhibitory oscillation required for the learning algorithm, we used the following procedure: First, at each time step, we used the kWTA algorithm to compute a baseline (normal) level of inhibition. Then, we added an oscillating component to the baseline inhibition value. The oscillating component of inhibition was varied in a sinusoidal fashion from $\min = -1.21$ to $\max = 1.96$. As per Equation 2 and Equation 3, the sign of the learning rate parameter was shifted from positive to negative depending on whether inhibition was moving toward its normal (midpoint) value, or away from its normal value.

At the start of each training trial, the target pattern was presented to the network (by presenting an external input to each of the units in the target pattern; this external input was applied in a constant fashion throughout the entire trial). The network was given 20 time steps to settle into a stable state before the onset of the inhibitory oscillation. Over the course of a given trial, inhibition was oscillated once from its normal value up to the high inhibition value, then down to the low inhibition value, then back to normal. The period of the inhibitory oscil-
lation was set to 80 time steps. Figure 30 shows how inhibition was oscillated on each trial, and how the sign of the learning rate parameter was changed as a function of the phase of the inhibitory oscillation.

At each time step (starting at the beginning of the inhibitory oscillation) weight updates were calculated using Equation 2 and Equation 3. However, these weight updates were not applied until the end of the trial.

**Basic Network Parameters**

At the beginning of each simulation, all of the weights were set to random values from the uniform distribution centered on .5 with range = .4. The initial weight values were symmetric, such that the initial weight from unit $i$ to unit $j$ was equivalent to the initial weight from unit $j$ to unit $i$. This symmetry was maintained through learning because the weight update equations are symmetric.

All of the other parameters shared by the oscillating algorithm and Leabra were set to their Leabra default values, except for $stm\_gain$ (which determines the overall influence of external inputs that are applied to the network, relative to the influence of collateral connections between units) and $i\_kwta\_pt$ (the parameter that determines whether kWTA places the inhibitory threshold relatively close to the target units, or relatively close to competing units; this parameter is discussed in more detail in the main text): $stm\_gain$ was set to 0.4, and $i\_kwta\_pt$ was set to 0.325.

**Initializing the Network**

For each simulated participant, after weights were set to their initial (random) values, but before the start of the actual RIF simulation, the network was given an initialization phase consisting of 10 epochs of training with 36 randomly generated 8-unit patterns (using a learning rate of .02). These patterns had no systematic overlap with the patterns used in the simulated RIF experiment (although there was some degree of random overlap), and a new set of patterns was generated for each simulated participant.

The purpose of this initialization phase was to introduce some variance across non-target units in how strongly they compete at test. If we omit this initialization procedure, then (at the start of the RIF simulation) the net input scores for non-target units tend to be very similar to one another (i.e., no one unit competes much more than the others). Thus, when inhibition is decreased low enough to activate one of the non-target units, they all tend to activate at once, leading to a seizure in the network. Because of random overlap between the RIF study patterns and the 36 initialization patterns, the initialization procedure ensures that some non-target units will compete more than others, right from the start of the RIF simulation. This makes it possible to decrease inhibition low enough to activate some additional units, without causing all of the units to activate at once. The exact number of training patterns and epochs used during the initialization phase does not matter, so long as the initialization phase meets the goal of introducing variance in the net input distribution for non-target units.

**Appendix B: Effects of partial practice on network weights**

To validate our claims about how the oscillating algorithm affects target and competitor representations (as shown, e.g., in Figure 4), we directly measured how practice affects network weights in Simulation 1. Figure 31 shows the average weight value, computed immediately before and immediately after partial practice, for all possible connection types involving target and competitor units (e.g., weights connecting unique competitor units to each other; weights connecting unique target units to units shared by the competitor and the target pattern; etc.). These analyses show that the learning algorithm weakens connections between unique competitor units and all of the other units that are active during the low-inhibition phase (shared units; other unique competitor units; and unique target units). As discussed in the Competitor punishment through oscillating inhibition section, this decrease in the interconnectivity of the competitor representation makes it harder to activate the competitor pattern, regardless of the retrieval cue. Note that connections between unique competitor units and units that are not part of any study pattern (and thus stay inactive during the low inhibition phase) are unaffected by practice. With regard to the target representation: On average, the pairwise connections between units in the target pattern are strengthened as a result of partial practice. This provides a straightforward explanation of why practice improves target recall.
Figure 30: Illustration of how inhibition was oscillated on each trial. At each time step, the “inhibitory oscillation” component depicted on this graph was added to the value of inhibition computed by the kWTA algorithm. The graph also shows how the sign of the learning rate was set to a positive value when the inhibitory oscillation was moving toward its midpoint, and it was set to a negative value when the inhibitory oscillation was moving away from its midpoint.
Figure 31: Average network weight values for different types of connections, measured immediately before the practice phase and immediately after the practice phase in Simulation 1. We are interested in connections between four kinds of units: Units that are unique to the competitor (C units), units that are unique to the target (T units), units that are shared by the target and competitor (S units), and units that are inactive for all study patterns (I units). The bar graphs show the average weight value for various combinations of unit types (for example, C-S = weights connecting unique competitor units to shared units; C-C = weights connecting unique competitor units to each other). The learning algorithm has two primary effects: First, it weakens connections between competitor units and other units that are active during the low-inhibition phase (shared, competitor, and target units), as shown in the C-S, C-C, and C-T bars. Second, it strengthens connections between target units and other units in the target pattern (shared units and target units), as shown in the T-S and T-T bars. Connections involving inactive units (C-I and T-I) are unaffected.
References


Holscher, C., Anwyl, R., & Rowan, M. J. (1997). Stimulation on the positive phase of hippocampal theta rhythm induces long-term potentiation that can be depotentiated by stimulation on the negative phase in area CA1 in vivo. Journal of Neuroscience, 17, 6470.


