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
{knorman,enewman,gdetre}@princeton.edu
Voice: (609) 258-9694; Fax: (609) 258-1113

Submitted to *Psychological Review*
February 6, 2007

Abstract

Retrieval-induced forgetting (*RIF*) refers to the finding that retrieving a memory can impair subsequent recall of related memories. Here, we present a new model of how the brain gives rise to RIF in both semantic and episodic memory. The core of the model is a recently developed neural network learning algorithm 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 relating to RIF 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. 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

Over the past decade, several researchers (see Anderson, 2003) have argued that retrieving a memory can cause forgetting of other, competing memories. 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 three parts: In the *RIF basics* section, we describe the RIF paradigm, and we review evidence for cue-independent forgetting and competition-dependent forgetting. In the *RIF as competitor weakening* section, we briefly review Anderson's arguments regarding why RIF results are problematic for blocking and associative unlearning theories of forgetting. 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, 2006b) that leverages regular oscillations in neural feedback inhibition to strengthen weak target memories, and to weaken other (non-target) memories. The Norman et al. (2006b) paper focused on the functional properties of the oscillating algorithm (how many patterns it can store, etc.). The present manuscript focuses on the psychological implications of the oscillating algorithm.

In the *Model architecture* section, we discuss how the model is comprised of a cortical semantic memory network and a hippocampal episodic memory network, and we provide a detailed account of the structure and functioning of these networks. Crucially, the oscillating algorithm is applied to both networks, making it possible for us to simulate RIF effects in both semantic and episodic memory.

In the *RIF simulation methods* section, we describe how we constructed patterns to use in our simulations, and how we simulated each of the three phases of the typical RIF experiment (study, practice, and test).

In the *Simulations of retrieval-induced forgetting* section, we show that the oscillating algorithm can account for detailed patterns of RIF data. This section starts with a *Precis of simulations*; readers who are interested in a quick overview of our simulation results should skip ahead to the *Precis*. In *Simulation 1*, we show that the model can account for the basic RIF findings mentioned above (more RIF in high-competition vs. low-competition situations; RIF using independent cues). We also show (in subsequent simulations) that the model provides a clear account of the *boundary conditions* on these basic RIF findings. As such, the model can account for findings that are inconsistent with competition-dependence and cue-independence, as well as findings that are consistent with these principles. Throughout the *Simulations* section, simulations addressing existing findings are intermixed with simulations that generate novel, testable predictions about how different factors 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-Pe___ and must say Pear). After a delay (e.g., 20 minutes), participants' memory for all of the pairs on the study list is tested. The

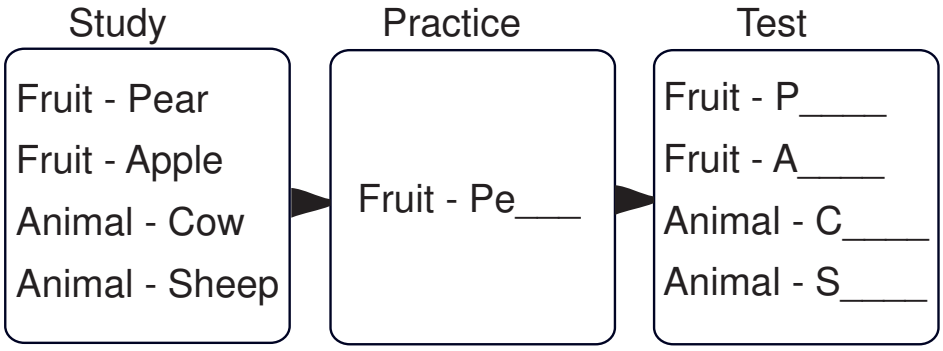

Figure 1: Flowchart diagram for Anderson's Retrieval-Induced Forgetting paradigm. See text for explanation.

paradigm is illustrated in Figure 1.

There are several notable results:

- Memory for practiced stimulus pairs (e.g., Fruit-Pear) is better than memory for control pairs that were not practiced and have no resemblance to practiced stimulus pairs (e.g., Animal-Sheep).
- Memory for non-practiced pairs that are related to practiced pairs (e.g., Fruit-Apple) is worse than memory for control pairs.
- Forgetting of pairs like Fruit-Apple is not limited to situations where Fruit is used as a retrieval cue. Forgetting also occurs when memory is tested with other cues that are related to Apple, but not to practiced stimulus pairs like Fruit-Pear. For example, forgetting is observed when Red is used to cue for Apple. Anderson calls this property *cue-independent forgetting*, although (as discussed in *Simulation 5*) some types of test cues are more effective at eliciting RIF than others.

This basic pattern (facilitated recall of the practiced pair, and cue-independent forgetting of related, non-practiced pairs) 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, 2005; Starns & Hicks, 2004). Forgetting has been observed when the "independent cue" is a related extralist word (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; Shivde & Anderson, 2001; Carter, 2004; Camp et al., 2005).

The RIF paradigm described above draws on both semantic and episodic memory (insofar as it uses pre-experimentally familiar category-exemplar pairs as stimuli). RIF has also been observed in paradigms that are more purely episodic. For example, Anderson and Bell (2001) observed cue-independent RIF for novel episodic associations between words; this finding is addressed in *Simulation 4*. Also, Ciranni and Shimamura (1999) observed RIF for novel episodic associations between colors, shapes, and locations. More recently, RIF has also been demonstrated on tests of semantic retrieval. For example, Carter (2004) demonstrated cue-independent forgetting of nonstudied semantic associates in an associate-generation paradigm. Specifically, Carter (2004) found that practicing retrieval of Clinic-Sick reduces the likelihood that participants will subsequently generate other, nonstudied associates of Clinic (e.g., Doctor), even in response to independent cues like Lawyer; this finding is addressed in *Simulation 6*. For another example of RIF in semantic memory, see Johnson and Anderson (2004).

The above examples are meant to provide a general sense of the kinds of studies that have found RIF; they are not meant to provide an exhaustive list (for other, recent examples of cue-independent forgetting, see, e.g., Shivde & Anderson, 2001; Veling & van Knippenberg, 2004; Saunders & MacLeod, 2006).

In light of the aforementioned successes, it is also worth noting a recent published failure to show

RIF using independent cues: Instead of using an independent cue that was semantically related to the competitor itself (e.g., cuing for Apple using Red), Perfect et al. (2004) paired the competitor with a semantically unrelated word (e.g., Zinc-Apple) prior to the RIF experiment, and used this "external associate" to cue memory. No RIF was observed in this condition. We discuss possible explanations for this null RIF effect in *Simulation 5*.

Evidence for competition-dependent forgetting

As stated earlier, another one of Anderson's key claims is that RIF effects are competition-dependent: Forgetting should be observed for strong competitors but not for weak competitors (Anderson, Bjork, & Bjork, 1994; Anderson, 2003). More concretely, we can define a strong competitor as an item that receives a high level of excitatory input (given a particular cue), but not enough to actually win the competition. According to this framework, practicing retrieval of Pear (using the cue Fruit-Pe___) causes forgetting of Apple because Apple receives a high level of excitatory input, but not enough to cause it to win over Pear.

The most important prediction of the competition-based account is that reducing the extent to which Apple competes with Pear (i.e., reducing the amount of excitatory input that Apple receives, relative to 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 previously studied pairs (Fruit-Pear). We will refer to this latter condition as the *extra study* condition. The intuition here is that the relative match between the cue and Pear (vs. Apple) is larger in the extra study condition than in the partial practice condition, so there should be less competition between Apple and Pear in the extra study condition. According to the competition-based view of RIF, this implies that recall of Apple should be hurt less in the extra study condition (vs. the partial practice condition). This was confirmed by Anderson and Shivde (in preparation), who found forgetting of competitors (measured using an independent cue) after partial practice but not extra study (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*.

Another way that Anderson has tested the competition-based account is by manipulating the taxonomic 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 strongly during retrieval than weak associates of Fruit (Kiwi), so strong associates should show more RIF than weak associates. This prediction was confirmed by Anderson et al. (1994) and also Bauml (1998). Both of these studies found RIF for strong associates but no RIF at all for weak associates (but see Williams & Zacks, 2001 for a failure to replicate the result). We address the effects of competitor strength on RIF in *Simulation 2*.

RIF as competitor weakening

To account for the above findings, Anderson has argued that RIF involves direct weakening of competing memory representations — that is, Apple is harder to retrieve in the paradigms described above (even with independent cues) because the *Apple representation itself* has been weakened. 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 (for examples of ratio-rule theories, see Rundus, 1973; Anderson, 1983; Raaijmakers & Shiffrin, 1981; Gillund & Shiffrin, 1984; Mensink & Raaijmakers, 1988).

- *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 representa-

tions themselves are unaffected (e.g., Melton & Irwin, 1940).

See Anderson (2003) and Anderson and Bjork (1994) for a much more detailed overview of these theories 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, Ciranni and Shimamura (1999) found a difference in competitor forgetting for partial practice vs. extra study (RIF was obtained in the former condition but not the latter) but no difference in target strengthening for partial practice vs. extra study (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, the idea that RIF involves direct weakening of competitors appears to provide a better account of extant RIF data than the blocking and associative unlearning theories described above. 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 think that blocking can also contribute to RIF in certain circumstances. We re-visit the issue of how our theory relates to competitor weakening, 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). However, this "biased competition" idea does not address the most salient aspect of RIF: Namely, that losing the competition to be retrieved has *lasting effects* on the accessibility of the losing 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; we address the time-course of RIF in more detail in the *General discussion*). To explain why losing the competition has lasting effects, our theory provides an account of how *local learning mechanisms*, operating within the networks where semantic and episodic memories are stored (cortex and hippocampus, respectively) can weaken competing memories. This approach is described in detail below.

Competitor punishment through oscillating inhibition

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. (2006b) 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. (2006b) did not address the algorithm's ability to account for this behavioral data. The goal of the present paper is

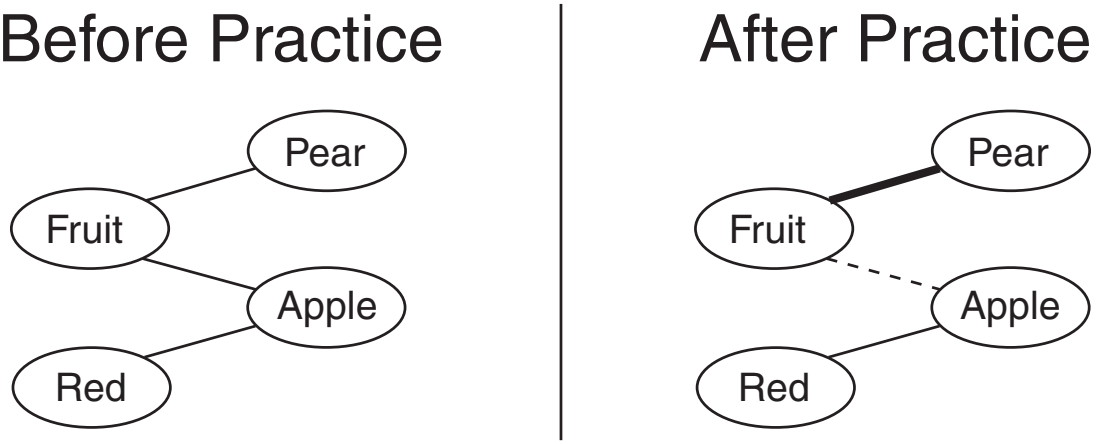

Figure 2: Illustration of associative unlearning theory (adapted from Anderson & Bjork, 1994, Figure 4). Practice of Fruit-Pe strengthens the Fruit-Pear connection and weakens the Fruit-Apple connection. This view predicts that forgetting of Apple should only be observed when using the cue Fruit (but not with other cues like Red). For evidence that contradicts this prediction, see, e.g., Anderson and Spellman (1995).

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 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.

The role of inhibition in recurrently connected networks

The network used in our simulations, like the brain itself, has recurrent connectivity: if unit X projects to unit Y, there is a path back from unit Y to unit X (although not necessarily a direct path; see, e.g., Felleman & Van Essen, 1991; Douglas, Koch, Mahowald, Martin, & Suarez, 1995).

Recurrently connected networks like this one need some way of controlling excitatory activity, so 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 (Szentágothai, 1978; Douglas et al., 1995; Douglas & Martin, 1998; 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 active, and all other units are inactive (O'Reilly & Munakata, 2000; Minai & Levy, 1994).[1]

Figure 3 provides a schematic illustration of the kWTA algorithm. First, the algorithm ranks all of the units in the layer according to the amount of excitatory input they are receiving. Next, the kWTA algorithm sets inhibition such that the inhibitory threshold (the point at which inhibition exactly balances out excitation) is located between the level of excitation received by the k^{th} unit and the level of excitation received by the $k + 1^{st}$ unit. This ensures that the top k units are above threshold and all of the other units are below threshold.

In the simulations below, we set k equal to the number of active units per layer in each studied pattern, such that (when kWTA is applied to the

[1] There are circumstances under which kWTA inhibition (as implemented in our model) can lead to slightly more or slightly fewer than k units being active; for a thorough treatment of this issue see O'Reilly and Munakata (2000). These small deviations are not important for explaining how kWTA shapes our model's behavior, so we gloss over them when discussing kWTA in the main text.

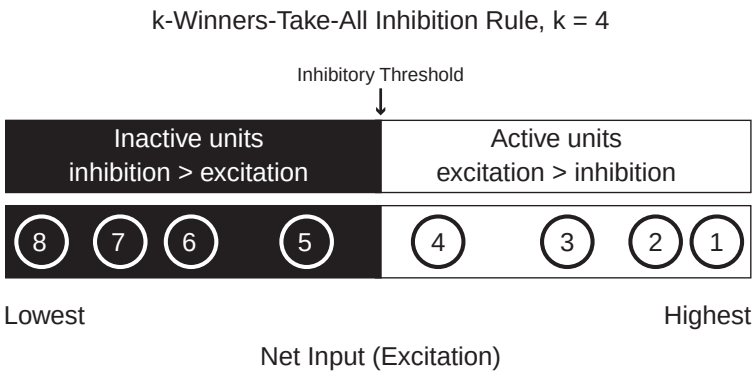

Figure 3: Illustration of key features of the k-winners-take-all (kWTA) inhibitory algorithm. The goal of the algorithm is to set inhibition such that the k units receiving the most excitatory input are active (for this example, assume that $k = 4$). To accomplish this goal, the algorithm ranks the units in a layer according to the amount of excitation that they are receiving. Next, the algorithm sets the level of inhibition such that the inhibitory threshold (the point at which inhibition exactly balances out excitation) is located between the level of excitation received by the k^{th} unit and the level of excitation received by the $k + 1^{st}$ unit. This results in a situation where the top k units (and only those units) are above threshold.

network) the best-fitting memory — and only that memory — is active. For a more detailed mathematical description of kWTA, see *Appendix A*.

Summary of the learning algorithm

The goal of the oscillating learning algorithm is to adjust synaptic weights to optimize retrieval of the target memory on subsequent trials. Because memory retrieval is a competitive process, the algorithm seeks to optimize target retrieval both by strengthening the target memory, and also by weakening competing memories. Another key learning principle 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 algorithm needs a way of identifying which parts of the target memory trace are weak. Likewise, in order to selectively punish strong competitors, the algorithm needs a way of identifying which memories are strong 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 4:

- First, **the target pattern is presented to the network**, by applying an external input to each of the units that are active 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**

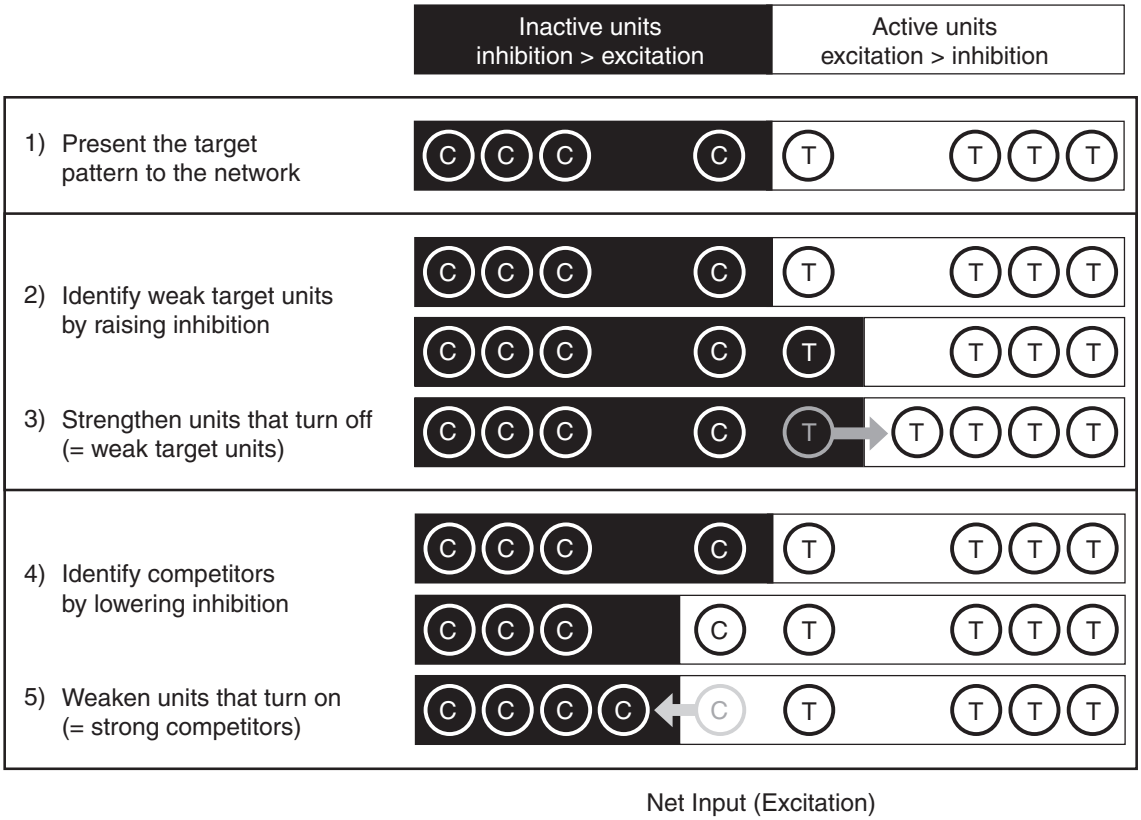

Figure 4: High-level summary of the learning algorithm. For all sub-parts of the figure, target units (labeled with a T) and competitor units (labeled with a C) are ordered according to the amount of excitatory net input they are receiving. Active units (excitation $>$ inhibition) are shown with a white background color and inactive units (inhibition $>$ excitation) are shown with a black background color. 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 equals 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.

target units), by increasing weights that connect these units to 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 activate that unit, given a partial cue) and no further strengthening will occur.

- Fourth, **the algorithm identifies competitors by lowering inhibition below the baseline level of inhibition**. Effectively, lowering inhibition reduces 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 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 that connect these units to 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. (2006b) 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).

$$dW_{ij} = \epsilon \left(X_i^+ Y_j^+ - X_i^- Y_j^- \right) \quad (1)$$

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 ϵ is the learning rate parameter.

The description of the oscillating algorithm in Figure 4 shows inhibition changing in discrete jumps (between normal, high, and low inhibition). In the actual model, we implement the learning dynamics shown in Figure 4 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 that 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. Weight changes are computed 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 more like the pattern at time $t + 1$. 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:

$$dW_{ij} = \epsilon \left(X_i(t+1) Y_j(t+1) - X_i(t) Y_j(t) \right) \quad (2)$$

If inhibition is moving away from its normal

value:

$$dW_{ij} = \epsilon \left(X_i(t) Y_j(t) - X_i(t+1) Y_j(t+1) \right) \quad (3)$$

Note that Equation 3 is the same as Equation 2, except for a change in sign. One useful to way to re-express these equations is to combine the sign change and ϵ into a single learning rate term ($lrate$):

$$dW_{ij} = lrate \left(X_i(t+1) Y_j(t+1) - X_i(t) Y_j(t) \right) \quad (4)$$

where $lrate$ takes on a positive value (ϵ) when inhibition is returning to its normal value, and $lrate$ takes on a negative value ($-\epsilon$) when inhibition is moving away from its normal value.

Figure 5 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.[2]

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 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 subsequently retrieve the competitor pattern.

[2]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 target units do not fully activate given normal inhibition, this will reduce target strengthening (see *Simulation 1.1* and *Simulation 9*). Likewise, if competitor units start to activate before inhibition is lowered, this will reduce competitor weakening (see *Simulation 2.1* and *Simulation 2.2*).

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

As discussed in Norman et al. (2006b), 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).
- 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.

Model architecture

As discussed in the *Introduction*, RIF can occur in both semantic and episodic memory. In order to encompass both types of RIF, the model used in our simulations incorporates both a semantic memory network and an episodic memory network. In keeping with prior work (e.g., McClelland, McNaughton, & O'Reilly, 1995) suggesting that cortex

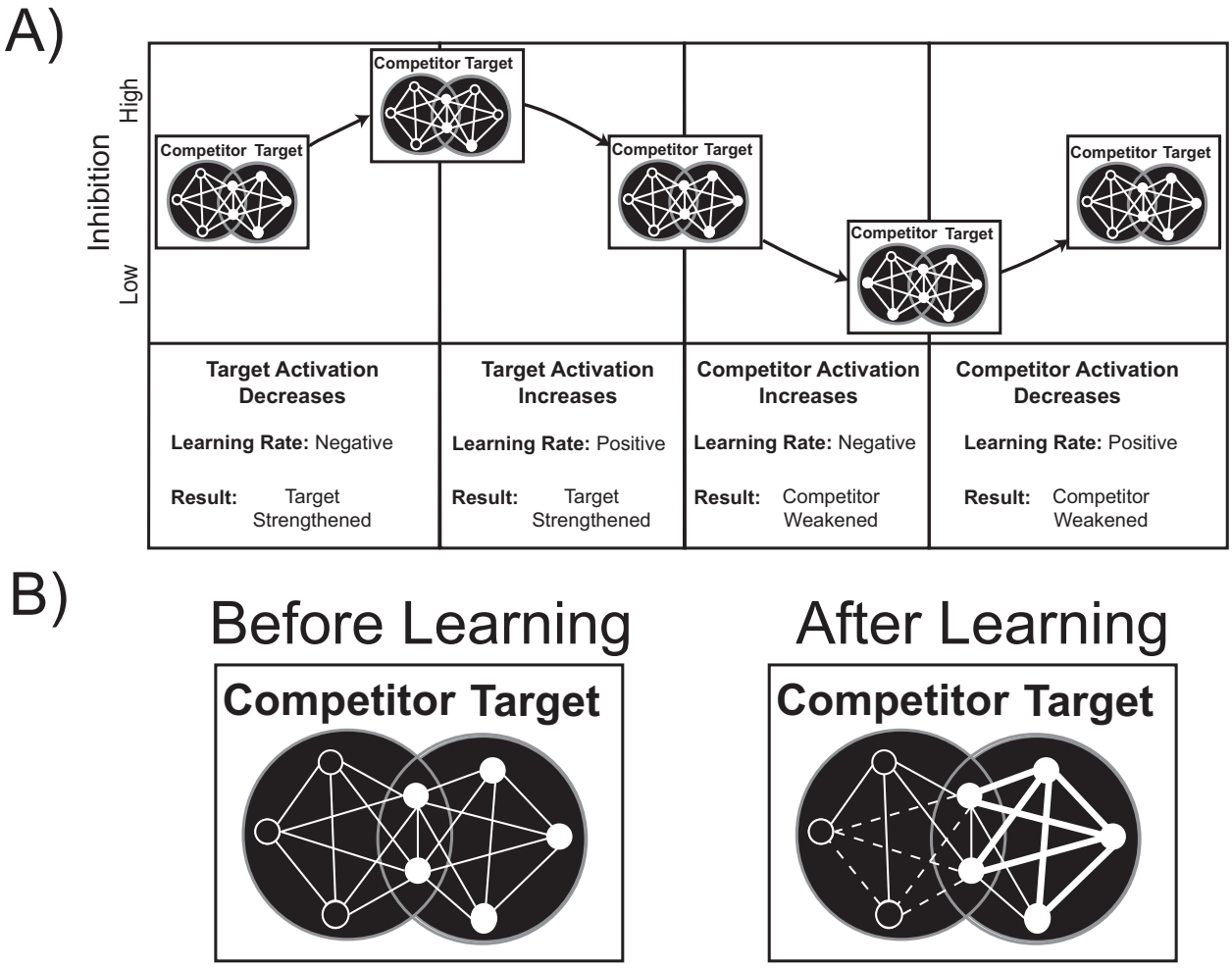

Figure 5: Summary of the oscillating learning algorithm (Norman et al., 2006b). Part A shows how target and competitor activation change during different phases of the oscillation. The target and competitor patterns are represented as interconnected sets of units (active units are represented by white circles and inactive units are presented by black circles). 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. The boxes in part A summarize how these activation changes affect network weights. To a first approximation, weight change in the model (for a particular unit) is a function of the change in that unit's activation, multiplied by the current learning rate (which is positive if inhibition is returning to its normal value, and negative if inhibition is moving away from its normal value; see Equation 4). Applying this heuristic to all four quadrants of the oscillation, the net effect of the first two quadrants is to increase weights coming into target units, and the net effect of the second two quadrants is to reduce weights coming into competitor units. Part B illustrates more specifically how the activation changes in part A affect the target and competitor representations: Target units that dropped out during the high-inhibition phase in part A become better linked to other target units; and competitor units that popped up during the low-inhibition phase in part A are cut off from the target representation (and from each other).

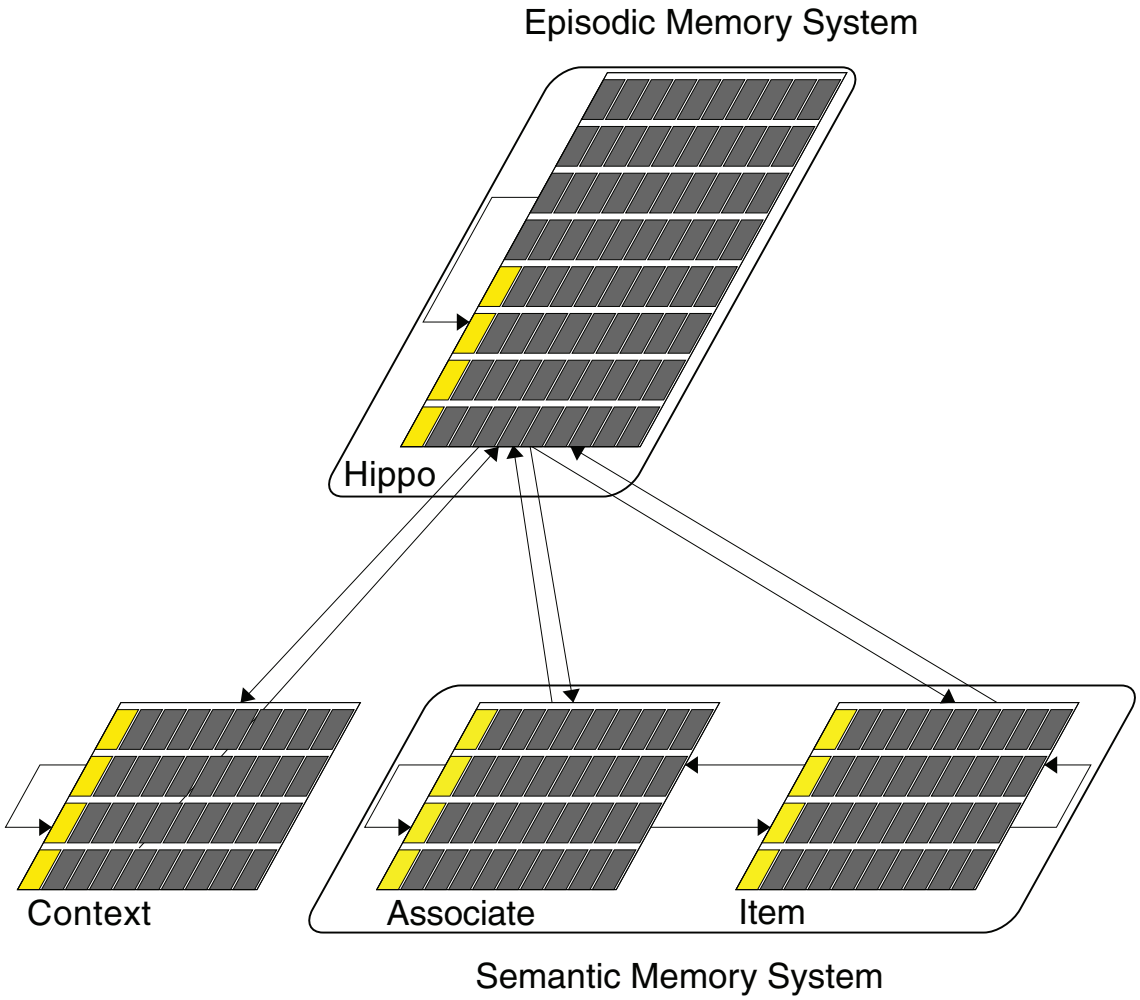

Figure 6: Diagram of the network used in our simulations. The *associate* and *item* layers constitute the network's semantic memory system: Patterns of activity in these layers directly represent the features of studied stimulus pairs (such that the first stimulus in the pair is represented in the associate layer, and the second stimulus in the pair is represented in the item layer). The item and associate layers are fully connected, such that each unit in either layer is connected to all of the units in both layers. Patterns of activity in the context layer serve as "contextual tags" (e.g., during the study phase, a fixed pattern of activity is imposed on this layer to represent the study context). Each unit in the hippocampal layer is bidirectionally connected to all of the units in the context, associate, item, and hippocampal layers (including itself). The role of the hippocampal network is to rapidly bind together co-active context, associate, and item representations in a manner that supports pattern completion (retrieval of the entire stored "episode" in response to a partial cue).

is the key structure for semantic memory and hippocampus is the key structure for episodic memory, we will refer to the semantic network as the "cortical network" and the episodic network as the "hippocampal network". However, we should emphasize that the networks used in this paper are highly simplified, relative to the more biologically-detailed cortico-hippocampal model that was used in our previous simulation work (Norman & O'Reilly, 2003; for similar models see, e.g., Hasselmo, Bodelon, & Wyble, 2002; Becker, 2005). Most of these simplifications were driven by practical necessity: The oscillating learning algorithm is highly computation-intensive because it computes weight changes at each time step (whereas most learning algorithms only factor in the final settled state of the network when changing weights; this point is discussed in the *Comparison to other models* section at the end of the paper). Thus, to keep the model from running too slowly, we tried to use the smallest and simplest possible network that allowed us to capture the relevant data. In the simulations described here, both the cortical (semantic) network and the hippocampal (episodic) network use the oscillating learning algorithm to update weights. The two systems are described in more detail below.

Cortical (semantic memory) network

The cortical semantic memory network consists of two layers, an *associate* layer and an *item* layer, consisting of 40 units apiece (see Figure 6). The semantic memory network is fully connected both within and across layers, such that each unit in the associate and item layers projects to (and receives a projection from) every unit in both layers, including itself. Our primary reason for splitting the semantic network into two layers was interpretive convenience: All of the paradigms that we simulate in this paper involve memory for stimulus pairs (e.g., Fruit-Apple), where the first stimulus is used to cue the second at test. Using a 2-layer scheme allows us to use one layer to represent the first stimulus in the pair (the *associate*: Fruit) and another layer to represent the second stimulus in the pair (the to-be-recalled *item*: Apple).

The associate-item patterns were instantiated in the model by turning on 4/40 units in each of the associate and item layers, and leaving the other units inactive (so, the 4 active units in the associate layer correspond to Fruit and the 4 active units in the item layer correspond to Apple). For more information on these patterns see the *Patterns used in the simulation* section below.

Prior to the start of the simulated RIF experiment, we pretrained a limited set of associate-item pairs into the cortical network using a simple Hebbian rule. This pretraining process was meant to capture the effects of pre-experimental experience with the stimuli that were used in the (simulated) RIF experiment. To implement pretraining, weights in the cortical network were first initialized to .5. We then ran a script that looped once through all of the patterns that we wanted to pretrain and strengthened weights between co-active units in each pattern.[3] For more details on how we implemented pretraining, see *Appendix B*.

During the simulated experiment, synchronous inhibitory oscillations were imposed on both layers (associate and item), and the oscillating learning algorithm was used to modify weights within and between layers.

Hippocampal (episodic memory) network

The hippocampal component of the model (Figure 6, top layer) is responsible for episodic memory. Specifically, the job of the hippocampal network is to rapidly memorize patterns of cortical activity in a manner that supports pattern completion (i.e., retrieval of the entire pattern, in response to a partial cue) after a single study exposure to the pattern. A key challenge for the hippocampal network is how to enact this rapid memorization without suffering from unacceptably high (*catastrophic*) levels of interference. In keeping with other hippocampal models, we posit that the hippocampus accomplishes this goal of rapid learning without catastrophic interference through its use of relatively non-overlapping, *pattern separated* representations (Marr, 1971; O'Reilly & McClelland, 1994; McClelland et al., 1995; Norman & O'Reilly, 2003; Becker, 2005).

In our previous modeling work, we used a

[3] In previous versions of the model, semantic pretraining was implemented using the oscillating learning algorithm. However, it proved to be impractical to use the oscillating learning algorithm to pretrain semantic memory for each simulated participant (it was too slow, and too difficult to precisely set memory strength values). Insofar as the focus of this paper is on simulating what happens *during* the experiment, we decided to use the simple Hebbian procedure outlined above (strengthen weights between co-active units) for pretraining. This Hebbian procedure would not work as an actual cortical learning rule (e.g., it does not have a means of decrementing weights). However, for the simplified patterns that we use in this simulation, it is a very efficient means of implanting attractors into the network.

relatively complex hippocampal model that maps closely onto the neurobiology of the hippocampus (Norman & O'Reilly, 2003). The full hippocampal model that was used by Norman and O'Reilly (2003) relies on passing activity through a "dentate gyrus" layer with a very large number of units (1600) and very sparse activity in order to enact pattern separation. Including this large dentate gyrus layer in our present model would make it run far too slowly. Thus, for this paper, we decided to radically simplify the hippocampal network, with the goal of keeping its essential properties (i.e., the ability to complete patterns after one study trial, and its use of pattern separation to reduce interference) while at the same time keeping the network as small as possible.

In this section, we first discuss the connectivity of the hippocampal network, including the role of context; next, we discuss how pattern separation is implemented in this network; and lastly we discuss learning and pattern completion in the model.

Connectivity and context

The hippocampal network used in our simulations here has 80 units. Each unit in the hippocampal layer is bidirectionally connected to all of the units in the (cortical) associate and item layers. The hippocampal layer also has full recurrent connectivity, such that each unit connects to all of the other units, including itself.

To simulate findings showing that context change between study and test can affect episodic memory (e.g., Smith, 1988), we also incorporated a separate "context" layer into the model (see Figure 6, lower left). This context layer can be viewed as representing aspects of the experimental situation other than the core semantic features of the associate and the item.[4]

The context layer contains 40 units and is bidirectionally connected to the hippocampal layer (such that each hippocampal unit receives a connection from each context unit, and sends a projection to each context unit). When simulating RIF experiments, we presented patterns with 4 active units to the context layer to represent particular contexts (e.g., we kept a particular set of 4 context units active throughout the entire study phase, to represent the fact that all of the study pairs are being presented in the "study context"). This "static context tag" mechanism was the simplest possible mechanism that we could devise that would allow us to simulate effects of context change. For reasons of simplicity, we also decided not to have the context layer oscillate, and we decided not to directly connect the context layer to the associate and item layers.[5]

All connections involving hippocampal units were initialized to zero prior to the simulation and then adjusted according to the rules outlined below.

Pattern separation: Pretraining conjunctive representations

A key property of the hippocampus is its ability to assign distinct representations to different combinations of stimuli (so it can memorize these combinations rapidly without catastrophic interference). Since the hippocampal network in this model is too small to use our standard approach to pattern separation (i.e., passing activity through a very large, very sparse "dentate gyrus" layer), we enforced pattern separation directly on the model by pretraining a unique conjunctive representation in the hippocampus for each associate-item combination.[6] These conjunctive representations were comprised of 4 active hippocampal units out of 80 total (e.g., Fruit-Apple would get its own set of 4 units; Fruit-Pear would get a different set of 4 units). For all simulations except for *Simulation 8*, the hippocampal representations corresponding to distinct associate-item pairs were completely non-overlapping.

To establish the conjunctive representation for a particular associate-item pair, we strengthened connections from active associate-layer and item-layer units to the 4 hippocampal units in the conjunctive representation. Also, to ensure robust hip-

[4] In this paper, we will remain agnostic about the neural instantiation of this context representation. In the *General discussion*, we mention that prefrontal cortex may play an especially important role in representing contextual information (Cohen & Servan-Schreiber, 1992). For additional discussion of the neural substrates of temporal context memory see Norman, Detre, and Polyn (in press).

[5] We do not want to rule out the possibility that incremental associative learning can occur between semantic features and contextual representations. We experimented with a version of the model that included direct context-associate/item connections, and decided to leave them out after finding that they greatly increase model complexity, without improving the model's ability to explain the findings discussed here.

[6] Given that it was combinatorially infeasible to pretrain a conjunctive representation for every possible associate-item combination, we focused on pretraining representations for associate-item pairs that were either semantically or episodically linked during the experiment. Specifically, we pretrained a conjunctive representation for every associate-item pair that was pretrained into semantic memory (via the cortical pretraining process described above) and/or presented during the study phase.

pocampal attractor dynamics, recurrent connections between these 4 units were strengthened. Weight values for strengthened connections were sampled from a uniform distribution with mean .95 and half-range .05 (weight-values for non-strengthened connections were kept at zero). These pretrained connections were fixed over the duration of the simulation. Importantly, while connections into the hippocampus from the associate and item layers were pretrained (giving each hippocampal unit a particular conjunctive "receptive field"), connections *out* from the hippocampus to the associate, item, and context layers were not pretrained. These connections all start out with zero strength. Without these outbound connections, activation can go into the hippocampus, but it can not feed back into cortex and support recall of cortical representations.[7]

Learning and pattern completion in the hippocampal network

During the simulated experiment, learning in the hippocampal network was focused on two sets of connections:

- Connections from the context layer to the hippocampus (which serve to bind particular associate-item pairings to the study context)
- Connections from the hippocampus back to associate and item layers (which allow the hippocampus to support pattern completion of missing pieces of associate-item pairs).

We applied the oscillating algorithm to the hippocampal layer and allowed it to modify these two sets of connections. Also, in keeping with the idea that hippocampus learns rapidly (in order to support pattern completion after a single study trial) but cortex learns more incrementally (McClelland et al., 1995; Norman & O'Reilly, 2003), we used a much higher learning rate for hippocampal connections (2.0) than for cortical connections (.05).

[7]Our use of fixed inbound connections to the hippocampus and fixed recurrents is a major simplification, relative to the Norman and O'Reilly (2003) model, which allowed both of these connections to be modified. However, we found that using fixed inbound and recurrent connections was necessary to get our simplified model to work in a robust fashion: Fixed inbound connections ensure pattern separation. Fixed recurrent connections are necessary because the oscillating algorithm requires robust attractor dynamics. In the absence of well-defined attractor states, the net input gap between the "winning" units and all of the other units in the network is small. In this situation, lowering inhibition tends to cause seizures where all of the hippocampal units activate at once.

Pattern completion in the hippocampus works in the following manner: When a partial version of a studied associate-item pair is presented, activation spreads upward in the model to the hippocampal layer, activating the hippocampal representation of that pair. If that hippocampal representation was linked back to the associate/item layers at study, then activation will flow back from the hippocampal representation to the associate/item layers and fill in the missing pieces of the cortical pattern. This process is modulated by contextual connections: If the hippocampal representation of the relevant associate-item pair was linked to the study context (during the study phase), and we cue at test with a representation of the study context, this will result in extra excitation being sent to the relevant hippocampal representation, making it more likely to activate.

Hippocampal model summary

We set out to devise the simplest possible hippocampal network that:

- instantiated the key hippocampal properties of *pattern completion* and *pattern separation*
- was compatible with the oscillating learning algorithm (in the sense that it showed robust attractor dynamics and was not too large, given the need to update every weight on every time step)

To accomplish this goal, we used a relatively small, one-layer hippocampal network and pretrained the network such that each associate-item pair that might come up in the experiment has its own "conjunctive representation" (i.e., a set of hippocampal units that are tuned to represent this particular associate-item pair). At the outset of the simulated experiment, these conjunctive representations are not contextualized (because they have not been linked to any patterns on the context layer), and they are not capable of supporting pattern completion in cortex (because they have not been linked back to the associate and item layers). During the simulated experiment, the oscillating learning algorithm can strengthen connections in order to bind hippocampal representations to the study context, and to link hippocampal representations back to the associate and item layers (so they can support pattern completion). Crucially, if a particular hippocampal representation pops up as a competitor during the practice phase, the oscillating algorithm can also weaken connections that were strengthened

at study, leading to forgetting of the episodic memory trace.

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 some of the studied patterns (but not others) are presented again, either in their entirety or in partial form; and a *test* phase, which measures the network's ability to complete partial versions of studied patterns.

First, we describe how we generated the patterns that were used in the simulations. Next, we describe aspects of our procedure that were common to all three phases (study, practice, and test). Finally, we describe the different phases of the simulation in more detail.

Patterns used in the simulation

The standard RIF paradigm involves studying items from various semantic categories, where multiple items are studied per category. This was instantiated in our model using category patterns (in the associate layer) that are each linked to multiple item patterns (in the item layer). Each "category tag" in the associate layer is distinct from (i.e., has no overlap with) the category tags corresponding to other categories. Furthermore, the item-layer patterns corresponding to different studied items have zero overlap with one another (see Figure 7 for sample patterns).[8]

These semantic category-item pairs were pretrained into the network before the start of the simulated RIF experiment, via the weight pre-setting mechanism described above (in the *Cortical network* section): For semantically strong patterns, the weights between active units in the pattern were set to a high value (e.g., .90); for semantically weaker patterns, the weights between active units in the pattern were set to a lower value (e.g., .65). For specific details of the algorithm that we used to pretrain cortical weights, see *Appendix B*.

Neighbor patterns In addition to pretraining patterns that actually appear in the (simulated) experiment, we also wanted to account for the fact that other patterns exist in semantic memory that are similar to items from the experiment, but do not actually appear in the experiment. To accomplish this goal, we took each of the categorized patterns that we pretrained (for use in the experiment) and we generated another *neighbor* pattern that had 100% associate-layer overlap with that pattern (4/4 active units in common) and 75% item-layer overlap with that pattern (3/4 active units in common; see Figure 7, second row). Each of these neighbor patterns was pretrained into the cortical network prior to the simulated study phase.[9] Neighbor patterns were never presented to the network during the simulated study phase, insofar as they are meant to simulate *nonstudied*, similar patterns. Note that the retrieval cues that we use at practice and test (see Figure 7) match both the to-be-retrieved item and its neighbor equally well. This mirrors the fact that, in actual RIF experiments, retrieval cues (e.g., Fruit-A___) typically match multiple items stored in semantic memory, although they only match one studied item. For example, if you study Fruit-Apple, the cue Fruit-A___ matches the studied item Apple, but it also matches the nonstudied "neighbor" item Apricot.

Neighbor patterns contribute to the functioning of the model in two important ways: First, competition between studied items and their neighbors helps to keep recall of studied items below ceiling. Second, by influencing competitive dynamics, neighbor patterns also exert a strong influence on the *learning* that takes place on study and practice trials; this point is discussed in more detail in *Simulation 1.1*.

Finally, note that neighbor patterns were included in all of the simulations described in this paper (the only simulation that explicitly discusses their contributions is *Simulation 1.1*, but they are present in other simulations also).

General simulation procedure

This section describes our basic procedure for simulating a single trial; this procedure was the same for all three phases of the simulated experiment (study, practice, and test). We provide a substantially more detailed account of our simulation procedure (including relevant equations) in *Appendix A*.

The simulation itself was implemented using a

[8] Our use of zero overlap between item-layer patterns is a simplification; we explore the effects of higher levels of item-layer overlap in *Simulation 8*.

[9] Neighbor patterns were pretrained with semantic strength .70 (i.e., connections from shared item units to the unique neighbor item unit were set to .70, and connections from category units to the unique neighbor item unit were also set to .70).

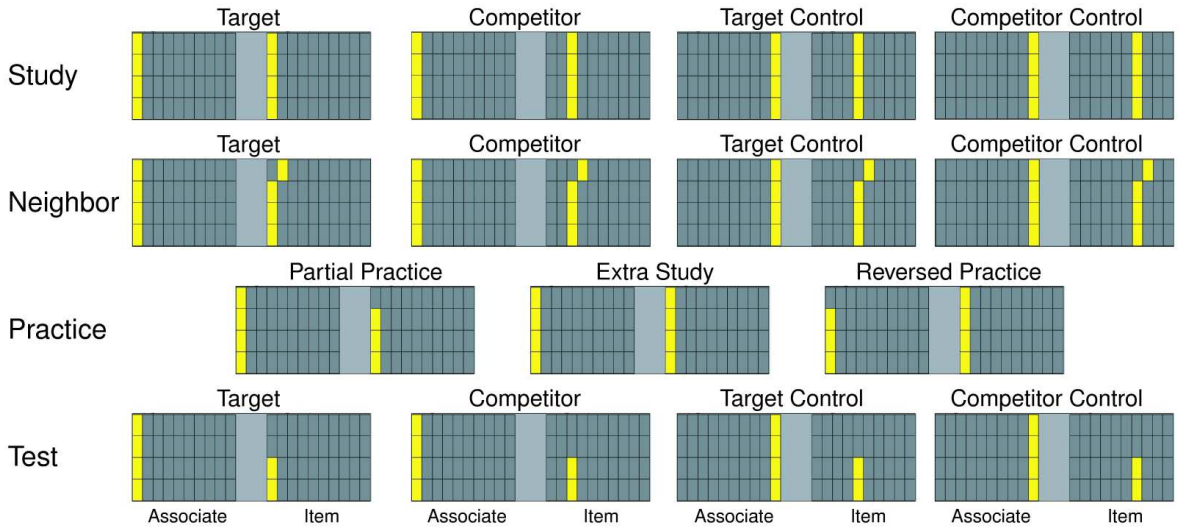

Figure 7: Figure illustrating a subset of the input patterns used during the study, practice, and test phases of *Simulation 1*; these phases are described in the *Simulation phases* section of the text. Each input pattern consists of a pattern of activity across the *associate* layer and a pattern of activity across the *item* layer. The top row shows examples of patterns that were shown at study. The *target* and *competitor* patterns come from one category and the *target control* and *competitor control* patterns come from another category. Studied patterns from the same category have 100% overlap in the associate layer but zero overlap in the item layer; studied patterns from different categories have zero overlap in both layers. The second row shows nonstudied *neighbor* patterns corresponding to each of the studied items in the top row (see the *Neighbor patterns* section for details). The third row shows examples of patterns used to probe target memory in the three different practice conditions. The fourth row shows examples of patterns used to probe memory at test.

variant of the O'Reilly and Munakata (2000) Leabra model, which includes the k-winners-take-all inhibition rule described above, as well as other useful rules governing activation propagation. The only differences between our simulations and "standard Leabra" were our addition of the inhibitory oscillation, and our use of the learning rule specified in Equation 4 (instead of the standard Leabra learning rule).

For all of our simulations, the k parameter that governs k-winners-take-all was set to $k = 4$ for each of the layers (to match the fact that associate-layer, item-layer, and context-layer patterns were all comprised of 4 active units, and that hippocampal "conjunctive representations" were also comprised of 4 active units).

On each trial, a pattern of activity (e.g., Fruit-Apple) was presented to the network by providing excitatory input to associate-layer and item-layer units that are active in that pattern. This "cue-related" input was held constant throughout the trial. The network was given 40 time steps to settle before we started to oscillate inhibition. Starting at the 40th time step, inhibition was oscillated by adding an sinusoidally varying inhibition value (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.[10]

During the trial, the model uses Equation 4 to compute a weight change value at each time step. Importantly, the model "saves up" (accumulates) these weight change values during the trial, and then applies them to the network at the end of the trial.

Simulation phases

Before the start of the simulated RIF experiment, we pretrained cortical weights (using the procedure outlined in the *Cortical network* section above) in order to implant semantic memory attractors into the network; see *Appendix B* for more details. We also pretrained hippocampal weights (using the procedure outlined in the *Hippocampal network* section above) in order to establish an appropriate set of hippocampal conjunctive representations.

[10] While the general form of the inhibitory oscillation was the same for the hippocampal network and the cortical network, the specific parameters governing the oscillation (e.g., maximum and minimum inhibition values) were slightly different in hippocampus vs. in cortex. For description of these differences see *Appendix A*.

Phase one: Study phase

During the study phase, complete patterns (i.e., 4/4 active associate units, 4/4 active item units) were presented to the network.

In most experiments, we presented two categories of patterns at study: the *practiced category* and the *control category*. As discussed above, studied items from the same category all share a common associate-layer pattern, and all have (completely) unique item-layer patterns. Items from different categories have zero overlap with one another.

The *practiced category* can be subdivided into:

- *Target* patterns. These patterns are presented at study and also during the practice phase. This condition is analogous to Fruit-Pear in Figure 1.
- *Competitor* patterns. Competitor patterns are presented at study but not at practice. This condition is analogous to Fruit-Apple in Figure 1.

The *control category* has the same number of items as the practiced category, and is structured identically to the practiced category (e.g., if the practiced category consists of items with mean semantic strength values .95, .85, .85, .85, the control category is structured this way also). This way, each item in the practiced category has a matched item in the control category. These control items are analogous to Animal-Cow and Animal-Sheep in Figure 1.

Each study trial involved presenting an associate-item pair from the study list, along with a "study context tag" (on the context layer) that was held constant throughout the entire study phase. The oscillating algorithm was applied to the network and used to update cortical and hippocampal weights.

In the simulations presented here, each item in the study list was studied once. Studied items were presented in a permuted order for each simulated participant.

Phase two: Practice phase

During the practice phase of the simulation, the target item(s) were presented to the network. As with the study phase, the oscillating algorithm was applied to the network and used to update cortical and hippocampal weights.

We explored three types of practice in the simulations reported here:

- *Partial practice* (also referred to as *retrieval practice*) involved presenting 4/4 of the active associate units, and 3/4 of the active item units.
- *Extra study* used full patterns (just like the study phase): 4/4 of the active associate units and 4/4 of the active item units were presented to the network.
- *Reversed practice* involved presenting 3/4 of the active associate (category) units and 4/4 of the active item units (e.g., after studying Fruit-Orange, reversed practice would use the cue Fr___-Orange and ask the model to recall Fruit). This reversed practice manipulation was introduced by Anderson et al. (2000a) and is discussed in more detail in *Simulation 1.1*.

In most of the simulations presented here, the target items were presented three times at practice (i.e., all of the target items were presented, then the list was presented again, then the list was presented again). Our use of three target repetitions matches the procedure typically used in RIF experiments (e.g., Anderson et al., 1994). The order of the target items was permuted with each pass through the target list.

Typically, the same "context tag" that we used at study was also presented to the network during the practice phase (but see *Simulation 5* for an exception to this rule). This allows us to capture the fact that participants are actively trying to think back to the study phase during partial practice. The influence of this context tag on retrieval was modulated by a *context scale* parameter that is described in the *Contextual cue strength* section below.

Phase three: Test phase

During the test phase, we cued recall for studied patterns using 4/4 of the active associate units and 2/4 of the active item units. Note that the test-phase partial cue (2/4 units) is slightly sparser than the practice-phase partial cue (3/4 units). This mirrors the fact that, in RIF experiments, cues at test are typically slightly sparser than cues at practice (e.g., participants might be given a 2-letter word stem at practice and a 1-letter word stem at test; e.g., Anderson et al., 1994). Using stronger cues at practice vs. test helps to ensure good recall at practice while also keeping recall at test below ceiling.

The study context tag was presented to the context layer at test (just as it typically was at practice). With a few exceptions (described below), the parameters used at test were the same as the parameters used in other phases.

Learning at test One simplification relates to the issue of learning that occurs during the test phase. Several studies have demonstrated that RIF effects can be induced by retrieval at test (see, e.g., Bauml, 1997, 1998; for further discussion of this issue see the "output interference effects" section of Anderson, 2003). However, the fact remains that learning during the test phase is not necessary to explain the vast majority of the key findings in the RIF literature.

As such, we decided to default to having learning turned off at test. This allows us to run our simulations much more quickly (since we do not have to compute weight changes at test, and we do not have to counterbalance the order in which items appear at test). Also, by removing an extra source of variance from the model, it makes it easier to draw inferences about how the practice phase is affecting stored memories. Finally, removing learning at test gives us more flexibility in how we can measure performance (e.g., as discussed below, we can test recall both before and after practice with learning turned off, and look at "pre-test - post-test" difference scores to index effects of practice).

To demonstrate that our model can account for effects of learning at test, we did run one simulation where learning at test was turned on (see *Simulation 1.1*, Figure 17 and Figure 19).

Computing recall accuracy at test As noted above, the inhibitory oscillation does not start right away on a given trial — the network is given 40 time steps to settle. We measure recall accuracy on the 39th time step (right before the onset of the oscillation).

In RIF experiments, the test phase measures recall of properties that are unique to the to-be-recalled item (e.g., the letters in the word "Apple"), rather than properties shared by practiced and non-practiced stimuli (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 activity of the one item-layer unit per pattern that is active for the to-be-recalled item but not its neighbor (see Figure 7). We call this measure *percent correct recall*.

For simulations that used our canonical two-category structure (where there was a "practiced category" and a "control category"), we measured the effects of practice-phase learning on targets and competitors by computing 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. This is the way that practice effects are typically measured in RIF experiments.

However, for some simulations (in particular, *Simulation 2.1*) it was impractical to use a two-category structure. In this case, we used a scheme where we tested recall performance *prior* to the practice phase (with learning turned off), then ran the practice phase, and then ran the test phase (with learning turned off also). In this case, we can use the difference in test performance prior to practice vs. after practice to index the effects of the practice phase on recall (with each item serving as its own control).

Contextual cue strength

In running these simulations, we discovered that we needed some way of capturing the extent to which participants were actively trying to retrieve memories from a particular context. The idea that participants can vary the extent to which they cue with contextual information has extensive precedent in the modeling literature (e.g., Gillund & Shiffrin, 1984; Shiffrin, Ratcliff, & Clark, 1990). As a simple illustration of how contextual cuing can influence behavior, participants are more likely to give a studied completion to a word-stem cue if they are specifically asked to provide completions from the study phase, vs. if they are asked to give the first completion that comes to mind (e.g., Graf, Squire, & Mandler, 1984).

In our simulations, we operationalize differences in contextual cuing by varying a parameter called *context scale*. This parameter multiplicatively modifies the strength of the projection between the context layer and the hippocampal layer (for more information on how projection scaling parameters work in the model, see *Appendix A*).

During the study phase (and during "extra study" practice trials and "reversed practice" trials) we typically set this context scale parameter to 0.0, reflecting the fact that participants are not actively trying to retrieve episodic memories during these phases (or, at least, they are not trying to do this to the same extent that they do at test). Importantly, setting the context scale parameter to zero interrupts transmission of activity from the context layer to the hippocampus, but it does not affect the network's ability to learn associations between context and hippocampal representations.

For partial practice and the test phase, we typically set context scale to 1.0, reflecting the fact that participants are more likely to try to actively "target" the study context during these phases. In *Simulation 4*, we also discuss the possibility that participants might use a higher context scale value on tests that rely purely on episodic memory, compared to tests where both semantic and episodic memory contribute.

Variability in oscillation amplitude

In our model, successful encoding depends critically on changes in activation driven by the inhibitory oscillation. To account for the fact that encoding is not always successful, the model incorporates the assumption that stimuli do not always trigger a strong inhibitory oscillation. In the simulations presented below, we use a simple "oscillatory variability" scheme where (on each trial) there is a 50% chance that the stimulus will elicit a full-sized oscillation. Otherwise, the stimulus elicits a half-sized inhibitory oscillation (i.e., the amplitude of the oscillation is multiplied by .5). The half-sized oscillation triggers smaller activation changes (on average) in hippocampus and cortex and thus triggers less learning. In particular, the half-sized oscillation is not sufficient to support formation of new hippocampal traces at study (see Figure 10 for an illustration of this point).

The idea that oscillatory amplitude varies from study trial to study trial, and that variations in oscillatory amplitude affect subsequent memory, receives strong support from the empirical literature. In particular, several studies of theta oscillations in humans have found that theta-band oscillatory power varies from trial to trial, and — crucially — that the strength of theta at encoding (for a particular stimulus) predicts subsequent retrieval success for that stimulus (Sederberg et al., 2003; Klimesch, 1999; Klimesch, Doppelmayr, Russegger, & Pachinger, 1996; Osipova, Takashima, Oostenveld, Fernandez, Maris, & Jensen, 2006).

sim #	simulation of	study phase	practice type	test cue type	key features
DEFAULTS		semantically defined categories; study both targets and competitors	partial practice	dependent cue (semantic associate + item stem)	
1.1	various	default	also extra study, reversed practice	default	manipulates practice type
1.2	various	default	also extra study, reversed practice	independent cue (semantic associate + item stem)	manipulates practice type with independent cues
2.1	Anderson, Bjork, & Bjork (1994)	default	default	default	manipulates both target and competitor semantic strength
2.2	exploratory	default	default	default	manipulates target semantic strength
2.3	exploratory	default	default	default	manipulates relative semantic strength of competitors
3	Bauml (2002)	study competitors but not targets	semantic generation of previously nonstudied targets, study of previously nonstudied targets	default	compares the effects of semantically generating vs. studying new items at practice
4	Anderson & Bell (2001)	episodically defined categories	default	independent cue (episodic associate + item stem)	practiced vs. control item sets defined by episodic associations; independent episodic cues
5	Perfect et al. (2004)	default, plus extra study phase where competitors are paired with novel associates	default	default; also, independent cue (episodic associate from another context)	compares standard test cues to external cues (episodic associations from another context)
6	Carter (2004)	study targets but not competitors	default	semantic generation with independent cue (semantic associate)	tests how retrieval practice affects subsequent semantic retrieval of a nonstudied competitor
7	Starns & Hicks (2004)	also includes semantic generation trials (to elicit nonstudied critical lures)	default	default	measures false recall of nonstudied semantic associates, and RIF for these items
8	exploratory	vary pattern overlap within category	extra study	default	measures forgetting in the extra-study condition as a function of pattern overlap
9	exploratory	default	partial practice (1, 2, or 3 units), extra study	default	tests how target semantic strength and practice cue partiality interact with target strengthening

Table 1: Overview of the simulations in the paper.

Precis of simulations

This section briefly summarizes key findings from our RIF simulations. Some simulations focus on explaining specific findings from the RIF literature, whereas other simulations (in particular, *Simulations 2.2, 2.3, 8, and 9*) explore effects of changing model parameters without trying to simulate any particular published study. Differences between the simulations are summarized in Table 1.

- In *Simulation 1.1* we address the retrieval-dependence of RIF. Specifically, we simulate the finding that forgetting of competitors occurs after partial practice but not after extra study or reversed practice (e.g., Anderson et al., 2000a). This 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 simulate the finding of *test order effects* in RIF studies: Recall is worse for category exemplars that are tested *later* in the test phase vs. earlier in the test phase (e.g., Bauml, 1998). This occurs because items tested later act as competitors during recall of items tested earlier. Finally, we simulate the finding that, even though retrieval practice hurts competitor recall more than extra study or reversed practice, these practice conditions have equivalent (beneficial) effects on target recall (e.g., Anderson et al., 2000a). We explain this finding of equivalent strengthening in terms of two opposing factors that cancel each other out: Increased competition during partial practice (vs. the other conditions) boosts target strengthening, but target misrecall during partial practice reduces target strengthening (see also *Simulation 9*).

- In *Simulation 1.2* we simulate the finding that RIF can be observed when memory is probed with independent cues (i.e., cues that did not appear at practice, and are unrelated to practiced target items; see, e.g., Anderson & Spellman, 1995; Anderson & Shivde, in preparation). Forgetting occurs for independent cues because of a combination of two factors: First, if the independent cue was paired with the competitor at study (e.g., Red-Apple), the episodic trace of that event sometimes pops up during the low inhibition phase at practice, thereby weakening the trace and harming subsequent recall. Second, pop-up of the cortical (semantic) trace of the competitor triggers incremental weakening of the competitor's cortical representation. This incremental weakening of the Apple attractor in cortex leads to subtle but measurable RIF effects in response to independent cues (see also *Simulation 6*).

- In *Simulation 2.1* we explore how the semantic strength of competitors and targets affects RIF. We replicate the pattern of results obtained by Anderson et al. (1994), whereby RIF occurs for semantically strong competitors but not semantically weak competitors, and RIF is not affected by target strength. RIF is observed for strong but not weak competitors because strong competitors pop up in semantic memory during the low inhibition phase, but weak competitors do not. Crucially, for the parameters used in this simulation, semantic pop-up is a prerequisite for episodic pop-up (so weak competitors do not pop up in episodic memory either). Because of this complete lack of pop-up, the memory traces of weak competitors are not harmed at practice, and no RIF occurs for these items.

- In *Simulation 2.2* we parametrically manipulate target strength and show that target strength actually has a nonmonotonic effect on RIF: Increasing target strength initially boosts RIF but further increases in target strength reduce RIF. This nonmonotonic pattern is observed because of two contrasting effects of target strength on competitor activation at practice. When targets are weak, competitors activate strongly, but this activation "spills over" into the high-inhibition (target strengthening) phase; this spill-over reduces RIF. The initial effect of increasing target strength is to eliminate this spill-over, thereby boosting RIF. Further increases in target strength reduce RIF by reducing the overall amount of competitor activation.

- In *Simulation 2.3* we present simulations showing effects of relative competitor strength: Increasing the strength of one competitor, relative to a second competitor, reduces RIF for the second competitor. This occurs because the baseline level of inhibition in the model is an (increasing) function of both the level of excitation of the *target* and the level of excitation of

the *strongest competitor*. As such, increasing the strength of the strongest competitor triggers an increase in baseline inhibition, which makes it less likely that other, weaker competitors will activate at practice.

- In *Simulation 3* we simulate the Bauml (2002) finding that semantic generation of nonstudied category exemplars leads to forgetting of previously studied exemplars from those categories. This occurs for the same reason that we see RIF in *Simulation 1* and *Simulation 2*: During the semantic generation phase, strong semantic competitors pop up in cortex during the low inhibition phase. This, in turn, triggers pop-up (and weakening) of the episodic representations of these competitors that were formed at study.

- In *Simulation 4* we simulate the finding from Anderson and Bell (2001) that independent-cue RIF can be observed when the "practiced" and "control" groups are defined in terms of novel episodic associations (as opposed to pre-existing semantic associations). A key finding from this simulation is that different parameter settings are required to simulate the null RIF effect for weak semantic associates observed by Anderson et al. (1994) and the presence of an RIF effect for novel episodic associates. To simulate the Anderson et al. (1994) result, we need to ensure that episodic links are *not* sufficient to trigger pop-up during the low inhibition phase at practice (otherwise weak, studied competitors will pop up at practice, leading to RIF for these items). To simulate the Anderson and Bell (2001) episodic RIF result, we need to ensure that episodic links between the retrieval cue and the competitor *are* sufficient to trigger pop-up at practice. We address this problem by positing that participants cue more strongly with context on purely episodic memory tests (vs. tests where semantic memory also contributes). In the model, we operationalize this difference by increasing the context scale parameter. This change sends extra excitation to episodic memory traces from the study context, thereby making it possible to observe pop-up of episodic associates of the practice cue (even if they do not pop up in semantic memory first).

- In *Simulation 5* we simulate the finding from Perfect et al. (2004) that not all independent cues show RIF. Specifically, RIF is not observed when the competitor is paired with a semantically unrelated "external associate" prior to the start of the RIF experiment, and the external associate is used to cue memory at test. In the model, the lack of RIF is attributable to contextual focusing during the practice phase: Cuing with the study context during the practice phase prevents episodic traces that were formed *outside* of the study context (e.g., the external associate) from activating as competitors. Because the episodic trace of the external associate does not activate during the low inhibition phase at practice, it retains its efficacy in supporting retrieval at test.

- *Simulation 6* focuses on RIF effects in semantic memory. We simulate the finding from Carter (2004) that practicing retrieval of Clinic-Sick impairs memory for nonstudied semantic associates of Clinic (such as Doctor), when memory for Doctor is tested using an independent cue ("Generate a semantic associate of Lawyer"). This effect occurs because Doctor pops up as a competitor in semantic memory when participants are practicing retrieval of the Clinic-Sick association, leading to weakening of the cortical (semantic) representation of Doctor.

- *Simulation 7* shows that, despite the model's tendency to punish semantically related competitors, it still shows robust false recall of nonstudied "critical lures" that are strongly associated with studied items (e.g., Roediger & McDermott, 1995). We also simulate the finding from Starns and Hicks (2004) that RIF is observed both for (true) recall of studied competitors and (false) recall of nonstudied critical lures. The false recall effect is largely a consequence of the model's tendency to generate critical lures at study (when given a chance to "free associate" to other items). When the model generates critical lures at study, it forms an episodic link between the critical lure and the study context. False recall can be explained in terms of the presence of this episodic trace, and RIF for the critical lure can be explained in terms of weakening of this episodic trace at practice.

- *Simulation 8* explores boundary conditions on forgetting caused by extra study. We manipulate the level of pattern overlap between same-

category items, in both the item layer and in the hippocampal layer. When overlap is low, we replicate the finding from *Simulation 1.1* that extra study does not cause forgetting. However, when overlap is sufficiently high, we start to see an effect of extra study on competitor memory (such that extra study of some category exemplars causes forgetting of other category exemplars). This occurs because increasing overlap boosts the level of net input received by the hippocampal representations of competitors, relative to the target. Eventually, the level of net input gets high enough to trigger pop-up of competitors on extra study trials, which (in turn) leads to forgetting of these items.

- *Simulation 9* explores factors that affect the amount of target strengthening that occurs at practice. We manipulate retrieval success at practice by varying the semantic strength of target items, and by varying the structure of the cue at practice (specifically, by varying the number of active item-layer units in the retrieval cue). In keeping with the idea that competition drives learning in the model, we show that optimal strengthening occurs in conditions where the target just barely wins at practice (i.e., recall accuracy at practice is high, and competition is also high).

Data fitting strategy The overall goal of this modeling work is to account for key empirical regularities in the RIF data space, and to establish 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. Unless explicitly noted, model parameters were held constant across all of the simulations presented here.

All of the simulation results that we report in the text of the paper (showing differences between conditions) are significant at $p < .001$. 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). [11]

[11] To ensure that the results reported in the paper were statistically reliable, we sometimes ran extra simulated participants to disambiguate the results of a particular simulation.

Simulation 1: Retrieval-dependence and cue-independence

This simulation addresses fundamental properties of RIF mentioned in the *Introduction*. *Simulation 1.1* explores *retrieval-dependence*: the extent to which forgetting is dependent on participants having to retrieve the target item at practice (based on partial cues). *Simulation 1.2* explores the extent to which RIF can be observed using independent cues at test.

Simulation 1.1: Basic RIF and retrieval-dependence

Background

The goal of this simulation is to explore how the structure of the cue at practice affects target strengthening and competitor weakening. Some illustrative results from Anderson et al. (2000a) are shown in Figure 8. This study used a variant of the Fruit-Apple RIF paradigm; at practice, Anderson et al. (2000a) compared partial practice (Fruit-Pe___) to reversed practice (Fr___-Pear). Reversed practice is conceptually similar to giving participants extra study of Fruit-Pear; in both cases, the item pattern (Pear) is presented outright at practice, so competition among item representations should be minimal. Thus, to the extent that RIF is competition-dependent, no RIF should be observed after reversed practice.

The left-hand panel shows that both partial practice and reversed practice improved target recall in this study to a roughly equal extent; this finding is consistent with other findings showing equal strengthening for partial practice vs. extra study (e.g., Ciranni & Shimamura, 1999). The right-hand panel shows that partial practice affected competitor recall but reversed practice did not. Below, we explore whether the model can generate this pattern of results.

Methods

The pattern structure used in this simulation is illustrated in Figure 9. As shown in the figure, two semantic categories (A and B) with 4 items apiece were pretrained into semantic memory prior to the start of the simulated RIF experiment. The semantic strength value for each of these items was sampled from a uniform distribution with mean .85 and half-range .15. The purpose of adding noise to the se-

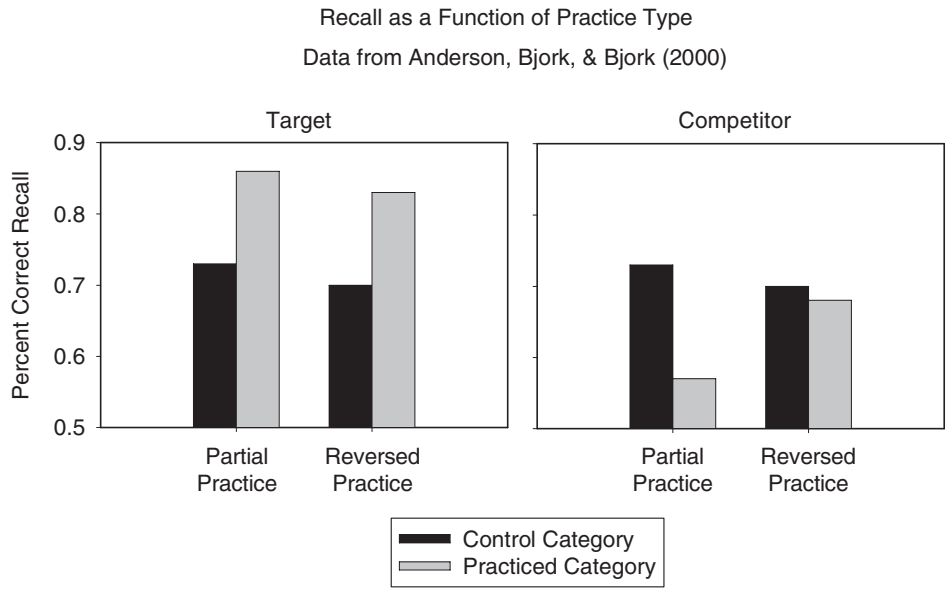

Figure 8: Data from Anderson et al. (2000a) ("tested first" condition), showing the effects of partial practice (Fruit-Pe___) and reversed practice (Fr___-Pear) on targets and competitors. This experiment used dependent cues at test (Fruit-A___). The left-hand figure shows that practice boosts target recall in both the partial practice and reversed practice conditions, to a similar degree. The right-hand figure shows that practice hurts competitor recall in the partial practice condition, but not the reversed practice condition.

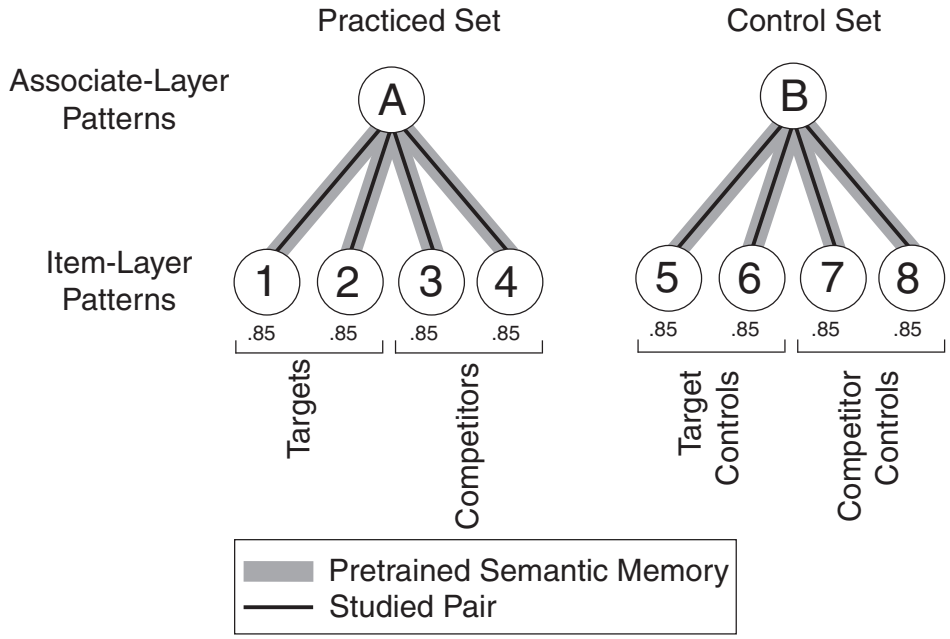

Figure 9: Illustration of the structure of the patterns used in *Simulation 1.1*. Gray bars indicate associate-item pairings that were pretrained into semantic memory prior to the start of the simulated RIF experiment. Black lines indicate associate-item pairings that were presented during the study phase of the simulated RIF experiment. Letters (A, B) are used to refer to associate-layer patterns and numbers (1-8) are used to refer to item-layer patterns. Numbers located below the "item-layer" circles indicate the mean strength of that pattern in semantic memory. The figure shows that two semantic categories with 4 items apiece were pretrained into semantic memory. All 8 patterns were presented at study.

mantic strength values was to eliminate the possibility of multiple competitors receiving the exact same level of excitatory support at practice. This situation (where no one competitor stands out above the others) is undesirable because it prevents the network from showing normal attractor dynamics — when this occurs, the network stays poised on the boundary between attractor states and none of the competitors activate strongly.

Category A served as the practiced category; this category was subdivided into 2 target items (A-1, A-2) and 2 competitor items (A-3, A-4). The other category served as the control category. All 8 category-item pairs were presented at study.

At practice, each of the 2 target items was presented 3 times. The type of practice was manipulated in a "between-simulated-subjects" fashion. We ran simulations using partial practice, extra study, and reversed practice. For partial practice trials, context scale was set to 1 (reflecting the fact that participants are deliberately thinking back to the study phase). For extra study and reversed practice trials, context scale was set to 0 (reflecting the fact that participants do not have to think back to the study phase when they are studying items; likewise, they do not have to think back to the study phase when they are retrieving category membership information).

Standard "dependent" cues (4/4 active category units and 2/4 active item units) were used at test.

Results

Activation dynamics at study

Figure 10 illustrates the activation dynamics that are present at study (averaging across trials) in the item and hippocampal layers, for both large (full-sized) oscillations and small (half-sized) oscillations. There are three important points to take away from this figure:

- *The inhibitory oscillation does not have a strong effect on item-layer activation at study.* There is a slight dip-down in activation of the target representation during the high inhibition phase, but nothing else. This result can be explained by considering the distributions of net input values associated with target units vs. other units (Figure 11). Because all of the target units are receiving strong external input (as well as strong input from each other), but none of the other item-layer units are receiving external input, the net input distribution for target units is located far above the net input distribution for other units. Given the wide separation between the distributions, the inhibitory threshold is not very close to either distribution, so raising the inhibitory threshold does not cause a strong reduction in target activation, and lowering the inhibitory threshold does not trigger activation of competitor units.[12]

- *In the hippocampus, large oscillations (but not small oscillations) cause the hippocampal representation of the target pattern to dip down.* Because the target and *target neighbor* patterns are so similar, the hippocampal representations of these items receive very similar levels of net input when the target pattern is active in cortex. The k-winners-take-all algorithm ends up placing the inhibitory threshold just below the target representation, and just above the target neighbor representation. Since the target representation's net input value is not far above threshold, its activity dips down when inhibition is raised (assuming that the oscillation is sufficiently large). This dip in target activation leads to strengthening of the context-item association, as well as strengthening of connections from the hippocampus back to the item and associate layers. Importantly, small (half-sized) oscillations are not powerful enough to displace the hippocampal representation of the target, so virtually no hippocampal learning about the target occurs on small-oscillation trials.

- *The "target neighbor" pattern pops up in the hippocampus but other items from the study list do not.* Because (as mentioned above) the hippocampal representation of the target neighbor receives strong excitatory support, this representation pops up strongly when inhibition is lowered. The hippocampal representations of other study-list items receive much less excitatory support (because they are much less similar to the target), so they do not pop up when inhibition is lowered.

In summary, the primary effect of studying a new item is strengthening of cortico-hippocampal

[12]Note that the items used in this simulation had relatively strong semantic memory traces (mean strength .85). When we use items with weaker semantic memory traces, the inhibitory oscillation has a larger effect on cortical activation at study (thereby serving to strengthen these items in semantic memory).

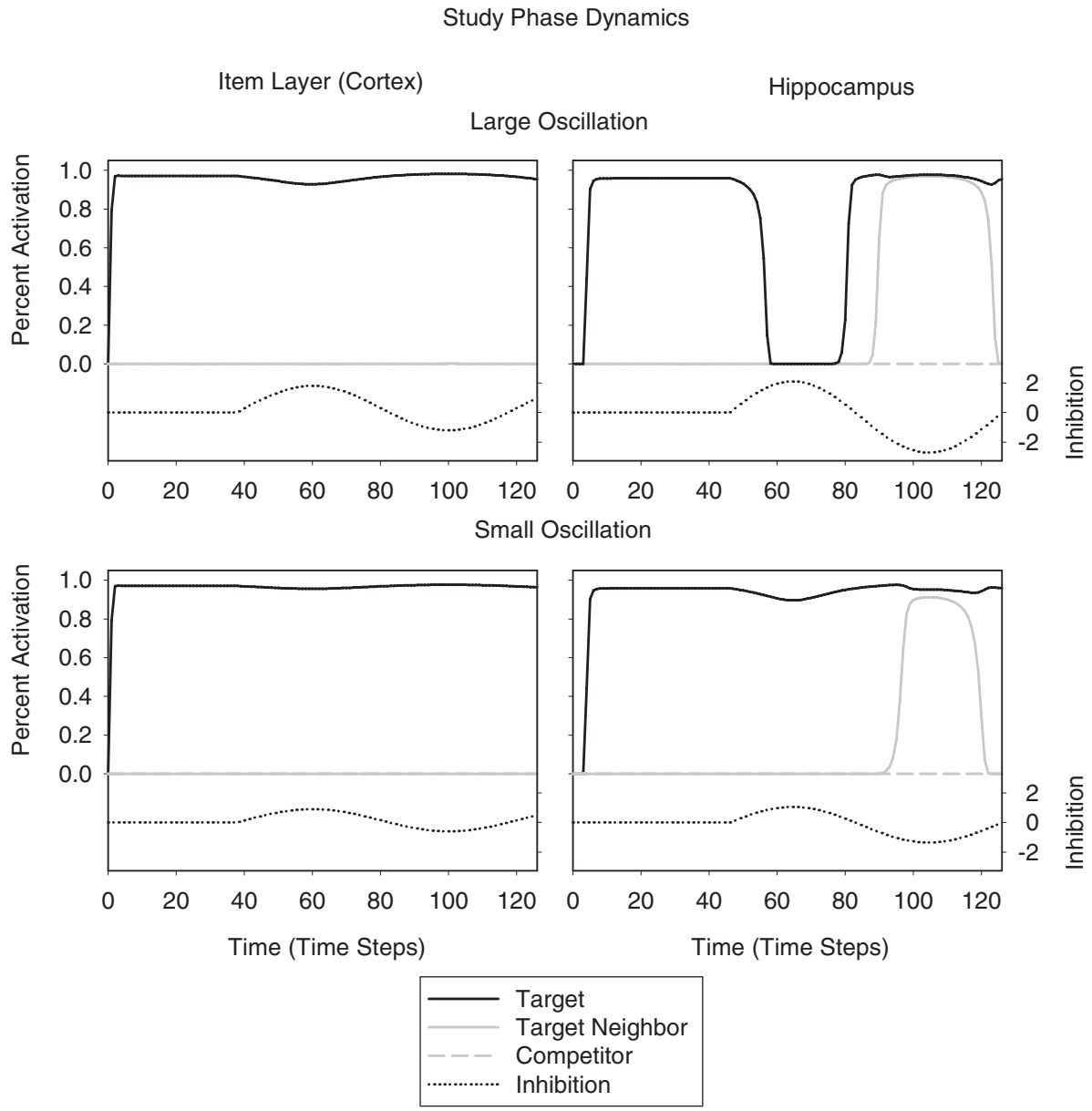

Figure 10: Plot showing average activation dynamics (over time, across the span of a trial) in the item layer and the hippocampal layer, for study trials with a large (full-size) oscillation and trials with a small (half-sized) oscillation. The solid black line plots activation of the currently studied (target) item's representation, the solid gray line plots activation of the target neighbor's representation, and the dashed gray line plots the activation of competitors (other study-list items from the practiced category). For all three lines, we only plot activation of unique features of the representation (i.e., features not shared with other items). The dotted line plots the time course of the inhibitory oscillation. The inhibitory oscillation does not have a large effect on activation in the item layer. In the hippocampus, large oscillations (but not small oscillations) result in a decrease in target activation during the high-inhibition phase. The hippocampal representation of the target neighbor pattern activates during the low inhibition phase, but the representations of other items from the target category (besides the neighbor pattern) do not activate.

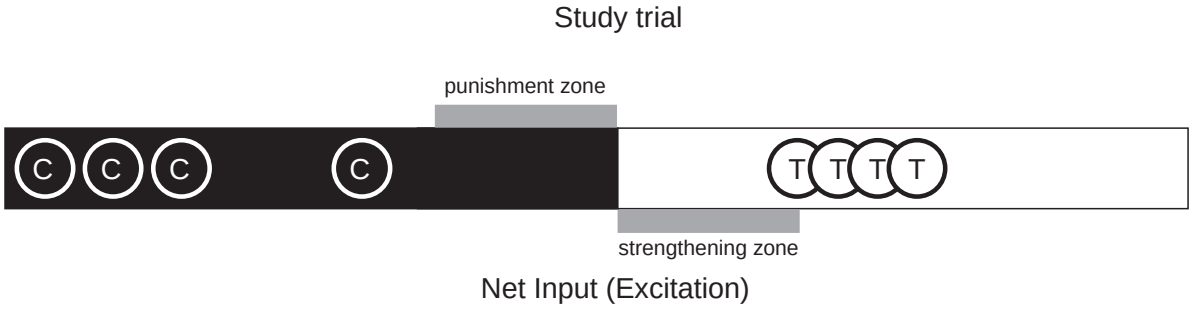

Figure 11: This figure schematically illustrates the distribution of net input scores for target units (marked with a T) and competitor units (marked with a C) in the item layer during a study trial, when inhibition is set to its normal (baseline) level. The k-winners-take-all rule places the inhibitory threshold between the k^{th} unit and the $k + 1^{st}$ unit. 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 in net input between target units and other units is large, so most target units fall outside of the strengthening zone, and competitor units fall outside of the punishment zone.

connections for that item (triggered by hippocampal dip-down during the high inhibition phase). Another important point is that studying new items does not cause forgetting of memory traces corresponding to other studied items. The key insight here is that, since the context scale parameter is set to zero at study, hippocampal competitor pop-up is determined by feature match alone (as opposed to contextual match). As such, competitor pop-up is dominated by hippocampal representations corresponding to nonstudied "neighbor" patterns (which share a very large number of features with the target), as opposed to other patterns from the study context (which have a lesser degree of feature overlap with the target).

Activation dynamics during the practice phase

We can use the same kind of activation dynamics graph to explore activation dynamics during the practice phase, as a function of practice type (partial practice, extra study, reversed practice). Figure 12 illustrates how (on average) target and competitor activation in the item layer and the hippocampal layer fluctuate over the course of partial practice trials, extra study trials, and reversed practice trials (in contrast to Figure 10, this figure and all subsequent dynamics figures collapse across large-oscillation and small-oscillation trials).

Dynamics during extra study and reversed practice Activation dynamics in the extra study and reversed practice conditions were identical (at least with regard to item-layer and hippocampal-layer activity) so they are plotted together in the bottom part of Figure 12. The overall pattern of dynamics here is the same pattern that we observed at study: In the item layer, target units do not dip down (because they are all receiving strong external input) and competitor units do not pop up. In the hippocampus, close competition between the target and the target neighbor representation causes the target representation to dip down, but hippocampal competitor representations do not receive enough support (relative to targets and target neighbors) to pop up. Thus, we expect to see episodic target strengthening, but no semantic or episodic competitor punishment in the extra study and reversed practice conditions.[13]

Dynamics during partial practice Retrieval dynamics in the partial practice condition (depicted in the top part of Figure 12) differ strongly from dynamics in the extra study and reversed practice conditions: In both the item layer and the hippocampal layer, 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

[13]The one place where reversed-practice dynamics diverge from extra-study dynamics is in the associate layer. Because the model is only given a partial cue in the associate layer during reversed practice, there is some pop-up of the "control category" pattern in the associate layer during the low inhibition phase. However, this pop-up is inconsequential to the strength of the target and competitor representations, insofar as these items were not linked to the control category in the first place.

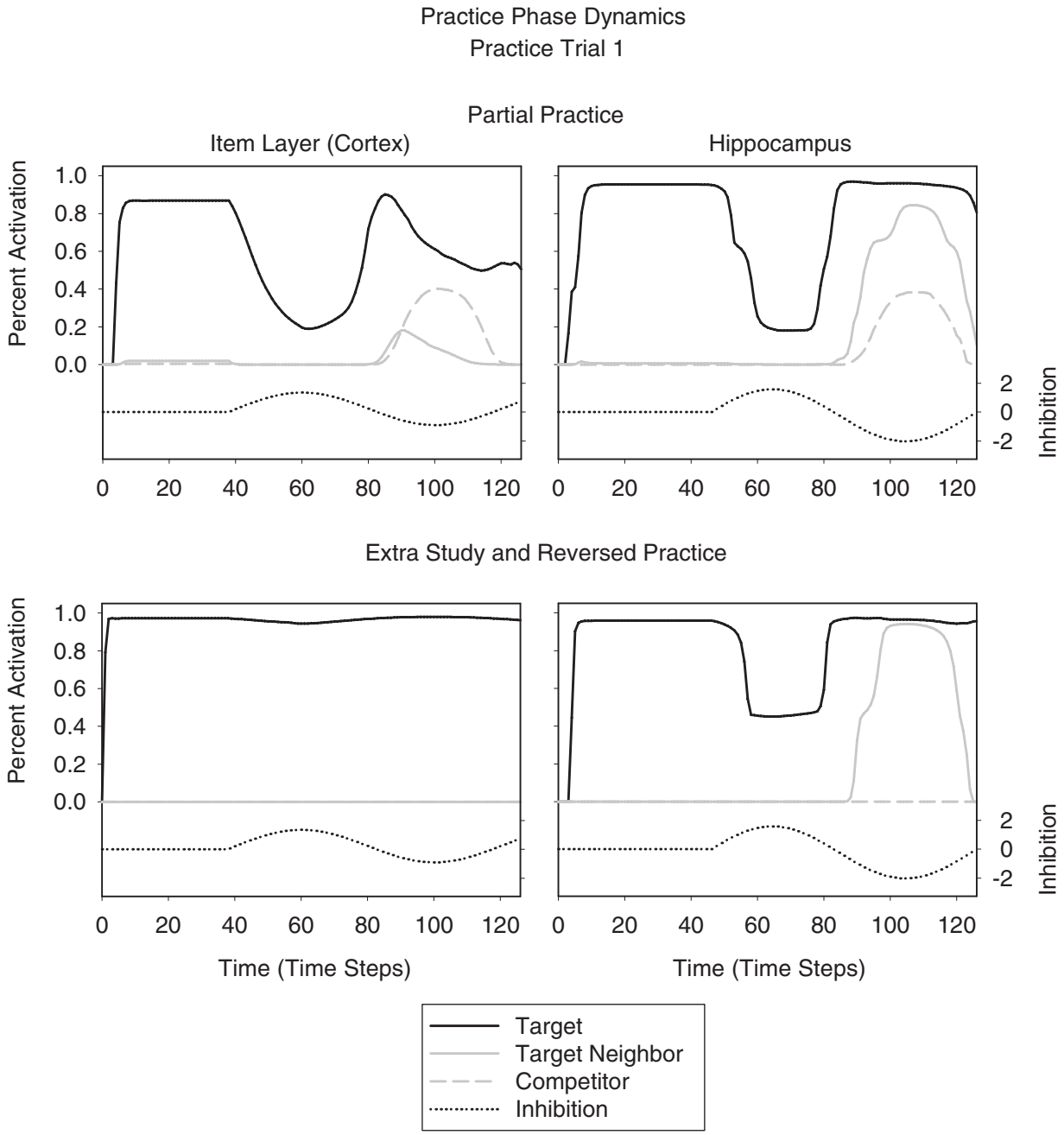

Figure 12: Plot showing average activation dynamics for partial practice, extra study, and reversed practice trials; these results are from the first practice trial (i.e., the first time this item was practiced). Extra study and reversed practice dynamics were not significantly different from one another and thus are combined in the figure. See the caption of Figure 10 for explanation of the lines in the figure; note that, here, the "competitor" line plots the activation of the *most active* of the two competitor patterns. The partial practice condition shows a large target activation dip during the high inhibition phase and a large competitor pop-up effect during the low inhibition phase, for both networks. The extra study and reversed practice conditions show a large target activation dip in the hippocampal layer, a much smaller target activation dip in the item layer, and no appreciable pop-up of studied competitor items.

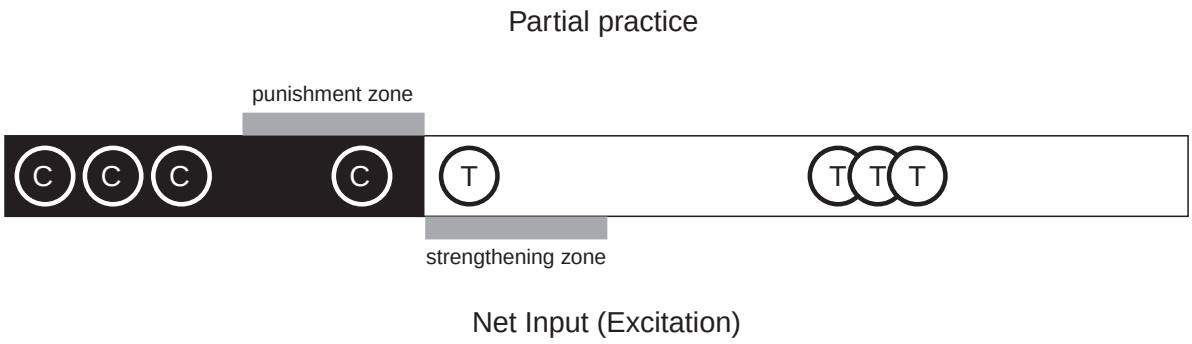

Figure 13: This figure schematically illustrates the distribution of net input scores for target units (marked with a T) and competitor units (marked with a C) in the item layer during partial practice, when inhibition is set to its normal (baseline) level. The gap between the lowest target unit and the highest other unit is smaller in the partial practice condition than in the extra study condition. As such, the weakest target unit falls into the strengthening zone, and some competitor units fall into the punishment zone.

when inhibition is lowered below its normal level.[14]

The observed dynamics in the item layer (with the target dipping down and the competitor popping up) can be explained in terms of the distribution of net input scores for target units vs. other units (shown in Figure 13). The partial practice cue provides external input to three of the four target units. On average, the remaining target unit receives only slightly more net input than other (non-target) units. 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 other unit. Because the target unit that does not receive external support is just above threshold (given normal inhibition), raising inhibition results in a strong decrease in the activation of this unit. Likewise, because strong competitor units are just below threshold, lowering inhibition results in a strong increase in the activation of these units.

The observed dynamics in the hippocampal layer are basically an echo of the cortical dynamics. When the item-layer representation of the target drops out during the high inhibition phase, the hippocampal representation of the target drops out also (because it is no longer receiving support from cortex). Furthermore, when competitor representations pop up in the item layer during the low inhibition phase, this provides strong support to competitor representations in the hippocampus, causing them to pop up also.

In terms of the oscillating learning algorithm, these dynamics have clear implications for the strength of target and competitor memories. When the cortical and hippocampal representations of the target dip down during the high-inhibition phase, this triggers target strengthening in both semantic and episodic memory. Likewise, when competitor representations pop up in cortex and hippocampus during the low-inhibition phase, this leads to competitor weakening in both semantic and episodic memory.

Effects of repeated practice on dynamics Figure 14 shows partial practice activation dynamics in the item layer and hippocampal layer, as a function of the practice trial number (i.e., whether this is the first or third time the target item has been practiced). During the first practice trial, target activation decreases sharply during the high-inhibition phase, and competitor activation increases during the low-inhibition phase. These activation changes trigger weight changes (target strengthening and competitor weakening, respectively) that reduce the size of the activation changes on subsequent practice trials. Thus, the overall effect of the learning algorithm is to "iron out the bumps" observed in the graph.

Effects of practice on target and competitor recall

Having mapped out the practice-phase dynamics, we now explore the effects of these dynamics on recall at test.

[14] On trials where the competitor has a stronger representation in semantic memory than the target, the competitor sometimes displaces the target in cortex during the low inhibition phase (see Figure 12, upper-left-hand plot). The net result of this extra "dip" in target activation is incremental strengthening of the target (since the learning rate is negative at this point in the oscillation, *increased* competitor activity *weakens* competitor weights, and *decreased* target activity *strengthens* target weights). In the context of the other weight changes that occur at practice, the effect of this extra "target dip" on target recall is negligible.

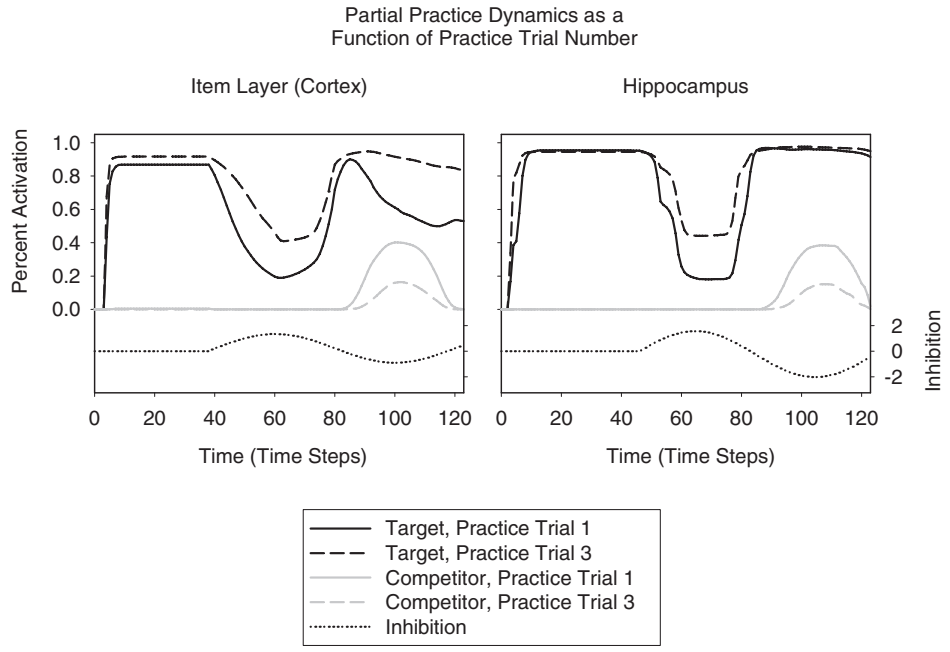

Figure 14: Plot of activation dynamics in the item layer and hippocampal layer during partial practice, as a function of practice trial (i.e., whether this is the first or third time the target has been practiced). Repeated practice reduces the extent to which the target representation dips down during the high inhibition phase, and repeated practice also reduces the extent to which the competitor representation activates during the low inhibition phase.

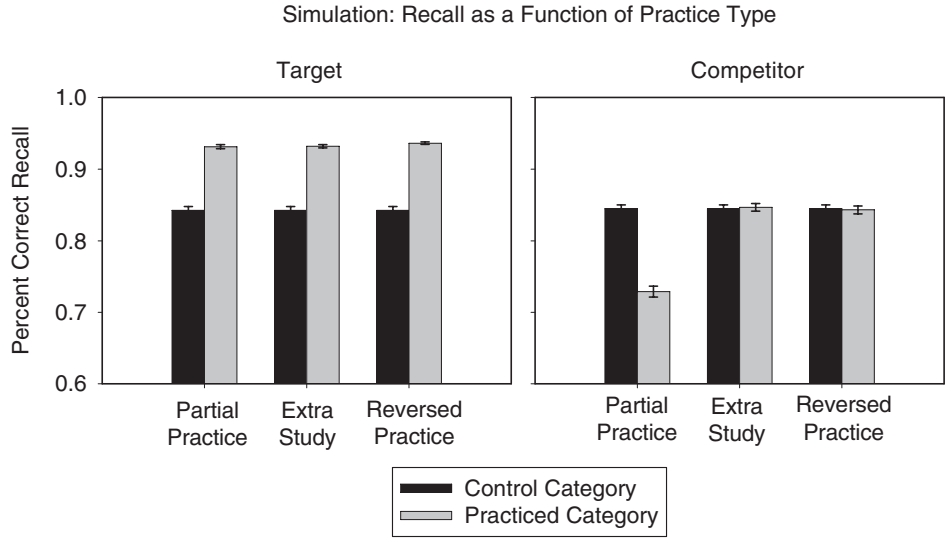

Figure 15: Graph of the effect of partial practice, extra study, and reversed practice on target and competitor recall in the model. The left side of the figure shows that all three kinds of practice boost target recall, to a roughly equal extent. The right side of the figure shows that competitor forgetting occurs in the partial practice condition but not the extra study condition or the reversed practice condition.

The left side of Figure 15 shows the effects of partial practice, extra study, and reversed practice on recall of target items in the model. Similar levels of strengthening were observed in all three conditions. This matches the widespread finding in the literature of equivalent strengthening given retrieval practice compared with either extra study or reversed practice (e.g., Ciranni & Shimamura, 1999; Anderson et al., 2000a; Anderson & Shivde, in preparation).

The right side of Figure 15 shows the effects of partial practice, extra study, and reversed practice on competitor recall in the model. Forgetting effects (relative to control items) were obtained in the partial practice condition but not the extra study condition or the reversed practice condition. This matches the findings reviewed earlier (e.g., Anderson et al., 2000a) showing that RIF is retrieval-dependent.

The competitor-recall results follow in a straightforward way from our dynamics analyses: Competitor pop-up was present for partial practice but not extra study or reversed practice, which explains why RIF was observed for the first condition (but not the other two). The relationship between the target-recall results and practice-phase dynamics is less straightforward. As shown in Figure 12, raising inhibition causes a larger "target dip" (in both the cortical and hippocampal networks) given partial practice vs. extra study or reversed practice. This is because the target representation is receiving less support from the cue in the partial practice condition vs. the other conditions. According to the oscillating learning algorithm, this larger target dip during partial practice should result in greater target strengthening in this condition.

The reason why partial practice does not yield greater target strengthening than the other conditions is because target recall accuracy (during practice) is worse in the partial practice condition than in the other conditions (mean activation of the unique part of the target representation = .87 in the partial practice condition vs. .97 in the extra study and reversed practice conditions). On trials where target recall succeeds, partial practice should yield more strengthening than extra study and reversed practice (for the reasons outlined above), but on trials where target recall fails, no target strengthening should occur.[15] For the parameters used in this simulation, these two forces (to a first approximation) cancel each other out.

Testing for blocking effects As stated in the *Summary of the learning algorithm* section, we believe that improved target recall after partial practice is attributable to target strengthening that occurs during the high-inhibition phase of the inhibitory oscillation, and that RIF is attributable to competitor weakening that occurs during the low-inhibition phase of the inhibitory oscillation. However, it is also possible that blocking effects are contributing to the observed pattern of recall data in this simulation. To the extent that items compete at recall, strengthening targets during the high-inhibition phase might indirectly hurt recall of competitors (by increasing the odds that targets will block competitor recall). Likewise, weakening competitors during the low-inhibition phase might indirectly boost recall of targets (by reducing the odds that competitors will block target recall).

To test this idea, we ran follow-up simulations where we restricted learning during partial practice to either the high-inhibition phase or the low-inhibition phase of the inhibitory oscillation (note that learning at study used both phases). The results of these simulations are shown in Figure 16: The "high-inhibition-only" simulations show a robust improvement in target recall, but no RIF, and the "low-inhibition-only" simulations show a robust RIF effect but no change in target recall.

This pattern of results (showing that it is possible to boost target recall without hurting competitor recall, and vice-versa) provides strong evidence against the idea that blocking is contributing to RIF in this simulation. Conversely, these results provide support for the idea that (in this simulation) RIF is a direct consequence of competitor-weakening that occurs during the low-inhibition phase. We re-visit the issue of blocking in *Simulation 7* and in the *General Discussion*.

Effects of context scale The above simulations show a stark difference in forgetting effects observed after partial practice (on the one hand) vs. extra study and reversed practice (on the other). There are two differences between these conditions in our simulations: Context scale is set differently (1.0 for partial practice vs. 0.0 for the other two conditions); also, the item-layer cues are structured differently (3/4 item-layer units are externally cued for partial practice, whereas all 4 item-layer units are externally cued for the other two conditions). To what extent is the difference in RIF attributable to

[15] Another factor that can reduce target strengthening in the partial practice condition is that practiced items can punish *each other*. For example, if participants practice retrieving both A-1 and A-2, A-1 might pop up as a competitor when practicing retrieval of A-2, resulting in weakening of the A-1 memory.

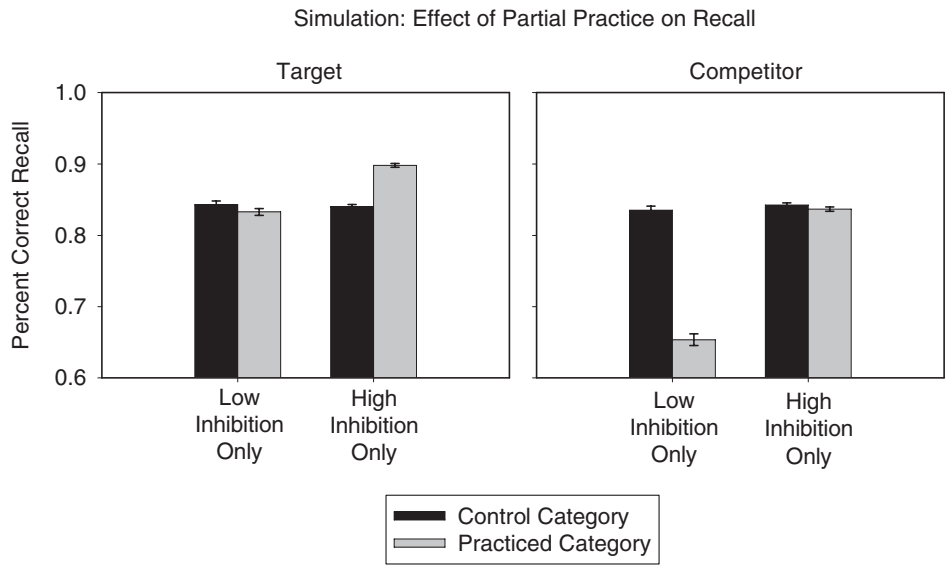

Figure 16: Graph of the effect of partial practice on target and competitor when learning is limited to the low-inhibition phase at practice and when learning is limited to the high-inhibition phase at practice. Learning during the high-inhibition phase boosts target recall without hurting competitor recall, and learning during the low-inhibition phase hurts competitor recall without boosting target recall.

the use of different context scale settings, and to what extent is the difference in RIF due to the structure of the item-layer cue? To address this question, we ran a version of the simulation where context scale was set to 1.0 throughout the entire simulation.

The results of this simulation showed the same qualitative pattern that we found in our previous RIF simulations: A robust RIF effect was observed for partial practice but not for extra study or reversed practice. This finding indicates that the partiality of the retrieval cue, on its own, is sufficient to account for the observed pattern of RIF effects.[16]

Learning at test It is also important to show that the basic pattern of RIF effects is still observed when we allow learning to occur at test. To address this question, we re-ran the above simulations with learning turned on during the test phase (as well as the study and practice phases). We tested all of the items from one category before testing any of the items from the other category; also, within the practiced category, competitors were tested before targets. For half of the simulated participants, the control category was tested before the practiced category; vice-versa for the other half.

Figure 17 shows the results of our simulations with learning at test. Overall, the results were similar to all of our previous simulations: There was equivalent strengthening of targets in the three practice conditions; there was a large RIF effect for competitors in the partial practice condition; and no RIF was observed in the other conditions.

The fact that learning was activated at test in this simulation makes it possible for us to examine test order effects. Several RIF studies have found that, when multiple items linked to the same associate appear at test (e.g., Fruit-A___, Fruit-P___), recall is better for items that are tested first vs. items that are tested last. Figure 18 illustrates this pattern, using data from Bauml (1998).[17]

To explore whether our model shows test order effects, we compared recall (at test) of the first two control items that were tested, vs. the last two control items that were tested. Results of this analysis are shown in Figure 19. The results shown are from the partial practice condition; the same pattern was observed when the practice phase involved extra study or reversed practice. As expected, recall was worse for the last two control items that were tested vs. the first two control items.

In terms of our theoretical framework, these test

[16]In this simulation, setting context scale to 1.0 at study did not have any adverse consequences. However, in *Simulation 8*, we show that using a high context scale value at study can result in massive (catastrophic) interference if there is high overlap between input patterns.

[17]Bauml (1998) also found that test order effects are larger for semantically strong items than semantically weak items; effects of semantic strength on RIF are addressed in *Simulation 2.1*.

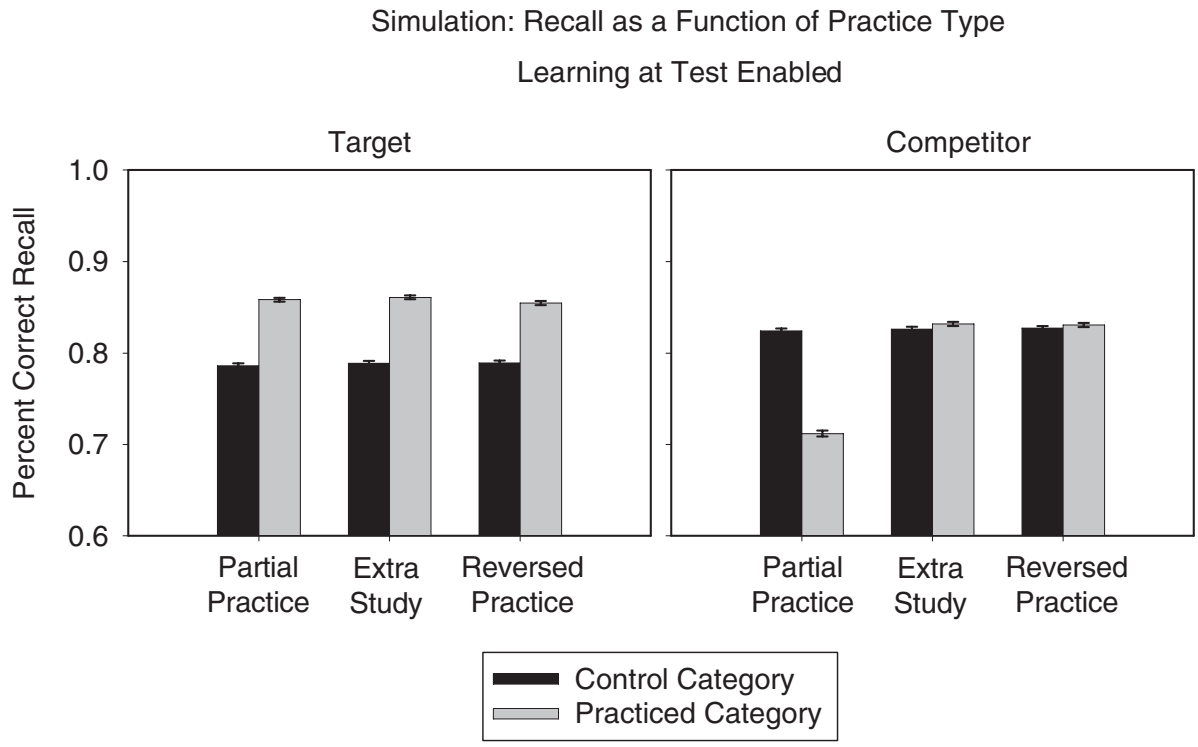

Figure 17: Graph of the effect of partial practice, extra study, and reversed practice on target recall (left-hand plot) and competitor recall (right-hand plot), when dependent cues are used at test and learning occurs at test. The results are unchanged relative to the preceding simulations: All three practice conditions lead to equivalent levels of target strengthening. For competitors, there is a large RIF effect in the partial practice condition but no forgetting effects in the extra study and reversed-practice conditions.

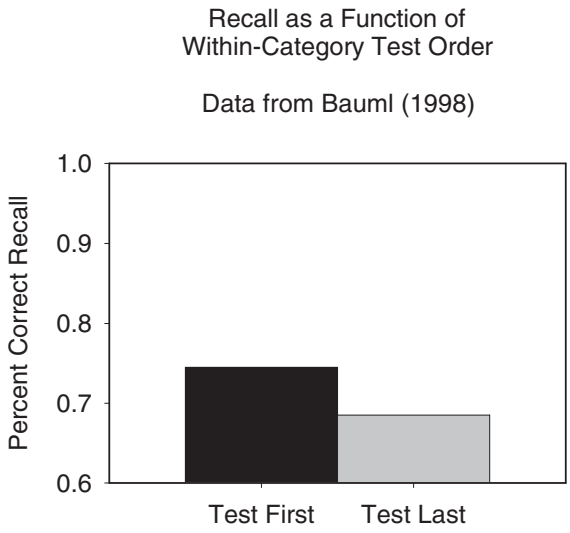

Figure 18: Results from Bauml (1998) (adapted from Figure 1, strong item condition) showing test order effects: Recall is better for the first three items that are tested from a particular category, vs. the last three items that are tested from that category.

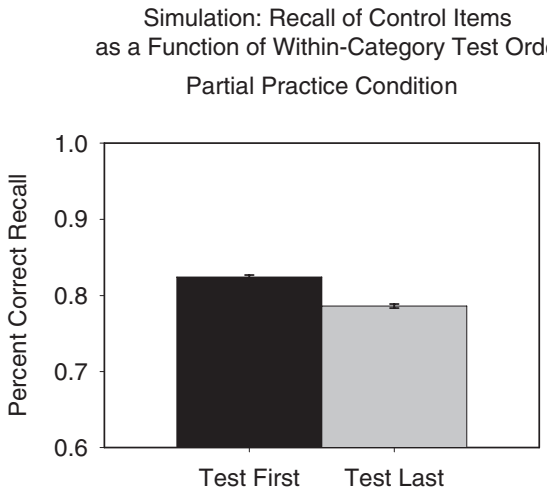

Figure 19: Graph showing test order effects. When learning is turned on at test, recall is worse for the last two control-category items that are tested, compared to the first two control-category items that are tested.

order effects can be attributed to competitor punishment occurring at test: When the first few items from a category are tested, other category exemplars pop up as competitors at retrieval and (as a result) are weakened.

Simulation 1.2: Cue-independent forgetting

Background

The previous simulations explored RIF with dependent cues (i.e., where the same cue was used at practice and test). In this simulation, we explore the critical issue of whether the model shows RIF when it is probed at test with an independent cue (in this case, a semantic associate of the to-be-recalled item that was not itself presented at practice). As discussed in the *Introduction*, several studies have observed RIF with independent cues (Figure 20 shows representative results from Anderson & Shivde, in preparation) and the presence of this "cue-independent" effect is a critical constraint on theories of RIF.

Methods

The small size of the network being used here (and our constraint that studied item-layer patterns should not overlap) places limits on the number of patterns that we can accommodate in our simulations. In order to accommodate the use of independent cues, we had to use smaller categories in this simulation (2 items per category) than in the preceding simulations.

Figure 21 illustrates the structure of the patterns used in this simulation. The key difference between this simulation and *Simulation 1.1* is that, in addition to the A and B categories, we pretrained two additional categories (C and D) that overlap with A and B, respectively. Crucially, the competitor item (2) is semantically linked to both category A and category C. Likewise, the competitor control item (5) is semantically linked to both category B and category D. When pretraining patterns, each pattern's semantic strength value was set to .85.[18]

All 8 pretrained pairings (A-1, A-2, C-2, C-3, B-4, B-5, D-5, D-6) were presented at study. The target item (A-1) was presented three times at practice. As in the preceding simulations, we also manipulated practice type in a "between-simulated-subjects" fashion (partial practice vs. extra study vs. reversed practice).

At test, we probed for the competitor item (2) using category C plus two item units. This constitutes an independent cue insofar as category C did not appear at practice. We also probed recall of the competitor control item (5) using category D plus two item units.

[18] In simulations (like this one) where there is just one competitor item, it is not necessary to add noise to semantic strength values at pretraining. The main purpose of adding noise to semantic strength values in *Simulation 1.1* was to "break ties" between competitors, and there is no possibility of a tie if there is only one competitor. Nonetheless, to match *Simulation 1.1*, we also ran a variant of this simulation where semantic strength values were sampled from a uniform distribution with mean .85 and half-range .15; the results of this simulation were qualitatively identical to the results reported here.

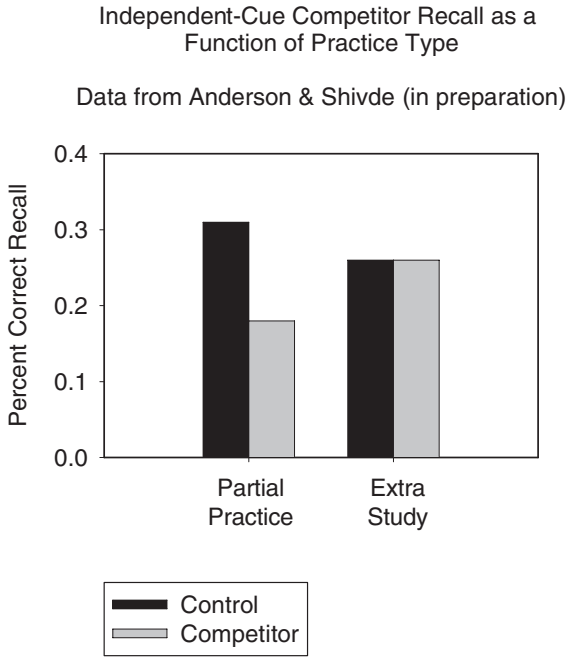

Figure 20: Data from Anderson and Shivde (in preparation), showing the effects of partial practice and extra study on competitor recall, when memory is tested using independent cues (semantic associates of the competitor that were not presented at practice). Partial practice impairs competitor recall using independent cues, but extra study does not.

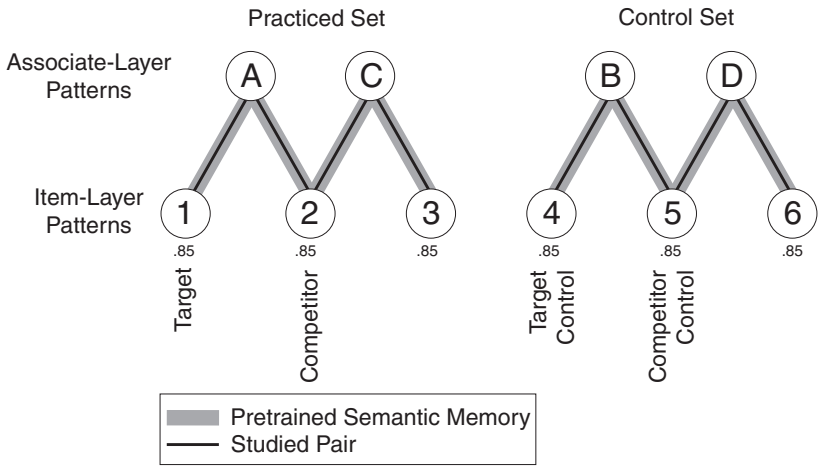

Figure 21: Illustration of the structure of patterns used in *Simulation 1.2*. As in Figure 9, gray bars indicate pairings that were pretrained into semantic memory, black lines indicate pairings that were presented at study, and numbers below the "item-layer" circles indicate mean semantic strength values for those items. The key feature of this design is the inclusion of additional ("independent") category cues C and D that can be used to access the competitor and the competitor control, respectively.

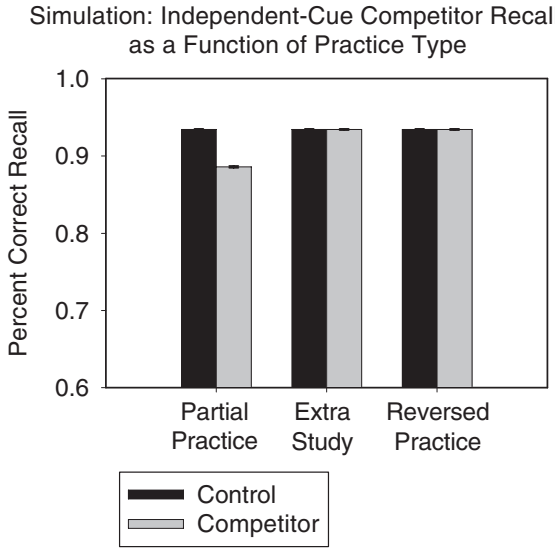

Figure 22: Graph of the effect of partial practice, extra study, and reversed practice on competitor recall in the model, when independent cues (semantic associates of the target word) were used to probe memory at test. RIF was observed in the partial practice condition but not the extra study condition or the reversed practice condition.

Results

The results of the independent-cue RIF simulation are shown in Figure 22. In keeping with the findings of Anderson and Shivde (in preparation; Anderson & Spellman, 1995; Anderson et al., 2000b) and many others, we observed a robust RIF effect given independent cues (semantic associates of the target word). As in our dependent-cue simulations above, the RIF effect was observed given partial practice but not given extra study or reversed practice.

The independent-cue RIF effect can be explained in the following manner: When the A-1 (partial) cue is presented at practice, the competitor pattern (2) activates in the item layer during the low-inhibition phase. This triggers hippocampal pop-up of the A-2 hippocampal representation. It also (to a lesser degree) triggers hippocampal pop-up of the C-2 hippocampal representation (if C-2 was encoded in the hippocampus at study). To quantify competitor pop-up in the hippocampal layer, we measured the activation of hippocampal representations at the trough of the inhibitory oscillation (i.e., when inhibition was lowest, and competitor activation was at its peak) during the first practice trial. Peak activation of the A-2 hippocampal representation was .58 (SEM .01) and peak activation of the C-2 hippocampal representation was .17 (SEM .01). Thus, we end up seeing hippocampal pop-up (and punishment) of *both* traces that could possibly support recall of the 2 pattern at test. This, in turn, results in diminished recall of the 2 pattern using both the A and C cues.

In addition to the hippocampal weakening described above, pop-up of the competitor's cortical representation should weaken recall of this representation, which (in turn) should incrementally reduce recall of the competitor, regardless of the cue. To get a rough estimate of how much hippocampal weakening vs. cortical weakening were contributing to the observed independent-cue RIF effect, we ran one variant of the simulation where cortical learning was turned off at practice, and another variant where hippocampal learning was turned off at practice. With both hippocampal learning and cortical learning at practice, the size of the RIF effect was .048 (SEM .001). With hippocampal learning (but not cortical learning) at practice, the size of the RIF effect was .034 (SEM .001). With cortical learning (but not hippocampal learning) at practice, the size of the RIF effect was .008 (SEM .001). Taken together, these results show that both cortical and hippocampal learning reliably contribute to RIF, but the effects of hippocampal learning are proportionally much larger. This result is a straightforward consequence of the fact that the hippocampal learning rate is larger than the cortical learning rate in these simulations (2.0 vs. .05).

Simulation 1: Discussion

In *Simulation 1*, we showed that the model captures several key aspects of the RIF data space:

- All three practice conditions (partial practice, extra study, and reversed practice) boost retrieval of the target item, as evidenced by better recall of this item vs. control items.
- Partial practice leads to RIF (as evidenced by worse recall of the competitor than control items) but extra study and reversed practice do not cause forgetting of the competitor.
- Given that we used an independent cue to probe for the competitor in *Simulation 1.2*, our results confirm that competitor-punishment can be obtained in the model 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___). This independent-cue RIF effect arises because of two factors: Pop-up (and weakening) of the hippocampal trace corresponding to the independent cue-competitor pairing; and also pop-up (and weakening) of the competitor's representation in cortex.

Boundary conditions

This simulation provides a mechanistic account of why competitor punishment is larger after partial practice ("retrieval practice") vs. extra study. In the model, there is nothing special about partial practice *per se*. Rather, the key determinant of learning in the model is the *gap in excitatory net input* between target units and competitor units. Competitor punishment was smaller in the extra study (vs. partial practice) condition because the gap in net input between targets and competitors was larger in this condition (compare Figure 11 to Figure 13). This view implies that it should be possible to get competitor punishment effects after extra study if we could reduce the gap in net input between the target pattern and competitors. One way to reduce this gap is to increase feature overlap between targets and competitors. To the extent that targets and competitors share features, anything that excites the target representation will tend to excite the competitor representation as well. We explore how feature overlap interacts with the effects of extra study in *Simulation 8*.

Another boundary condition relates to target strengthening effects. In *Simulation 1.1*, we found equivalent target strengthening after partial practice vs. extra study. This finding is consistent with extant data (e.g., Ciranni & Shimamura, 1999) but inconsistent (at least on the surface) with the idea that more learning should occur in conditions where there is high competition (e.g., partial practice) vs. low competition (extra study). We argued that the higher competition in the partial-practice condition (which should boost strengthening) was offset by target misrecall in the partial-practice condition (which should reduce strengthening). This leads to the prediction that, if we could increase the odds of targets being recalled successfully during partial practice (e.g., by strengthening their representations in semantic memory) we would see greater strengthening during partial practice vs. in the other conditions. Conversely, if we reduced the odds of targets being recalled successfully, we would see less strengthening during partial practice vs. other conditions. These predictions are addressed in *Simulation 9*.

Simulation 2: Effects of competitor strength and target strength on RIF

In this simulation, we explore how competitor strength and target strength interact with RIF. In *Simulation 2.1* we simulate results from a study by Anderson et al. (1994) that orthogonally manipulated competitor and target strength. In *Simulation 2.2*, we parametrically explore effects of target strength on RIF, and in *Simulation 2.3* we explore how adjusting the strength of competitors *relative to each other* affects RIF.

Simulation 2.1: Simulation of Anderson, Bjork, and Bjork (1994)

Background

The first RIF experiment to explore effects of target strength and competitor strength in detail was Anderson et al. (1994). As mentioned earlier, Anderson et al. (1994) found that partial practice of items like Fruit-Pear led to RIF for semantically strong competitors (e.g., Fruit-Apple) but not semantically weak competitors (e.g., Fruit-Kiwi; 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 led to forgetting of subsequently-tested strong items but not subsequently-tested weak items. With regard to target strength: In the same study where they

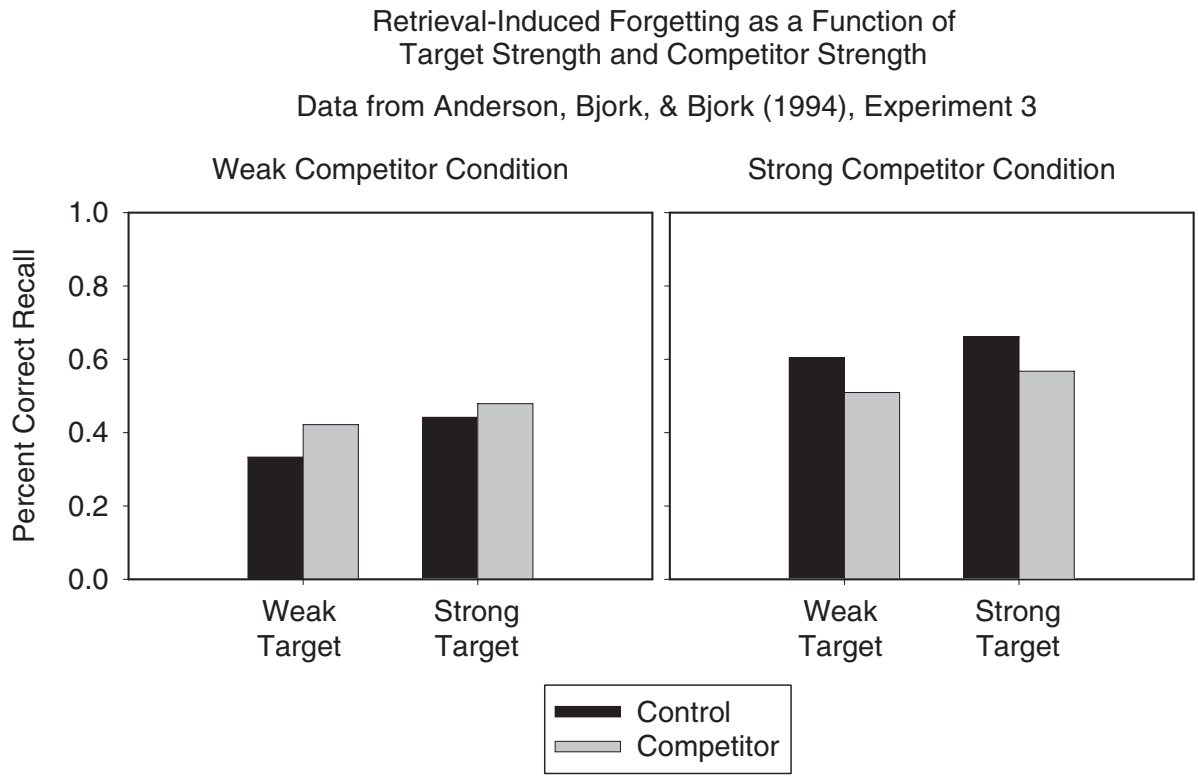

Figure 23: Graph of results from Anderson et al. (1994), Experiment 3, showing how competitor strength and target strength interact with RIF. There is RIF for strong competitors but not weak competitors (in both the Weak Target and Strong Target conditions). RIF effects are of similar size in the Strong Target condition and the Weak Target condition.

manipulated the semantic strength of competitors, Anderson et al. (1994) also manipulated the semantic strength of target items, and found no effect of target strength on RIF. The data from Anderson et al. (1994), Experiment 3 (showing the pattern described above) are shown in Figure 23.[19]

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 *Summary of the learning algorithm* section). Figure 24 schematically illustrates the amount of net input received by target units, units belonging to strong competitors, and units belonging to weak competitors. Units belonging to strong competitors receive more input from the retrieval cue than units belonging to weak competitors. 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 between the k^{th} unit (typically, the weakest target unit) and the $k + 1^{st}$ unit. 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 farther below the inhibitory threshold in the strong-target condition, they are less likely to activate when inhibition is lowered, so they are less likely to be punished.

Figure 25 illustrates hypothetical net input distributions given a strong target vs. a weak target.

In summary, based on Figure 24 and Figure 25, we would expect more RIF for strong vs. weak competitors, and less RIF 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: As per Figure 25, we show that competitor pop-up is larger given weak vs. strong targets. However, when targets are weak, we also show that competitor activation starts to "spill over" into the high-inhibition (target strengthening) phase of the oscillation, reducing RIF. In this simulation, the spill-over effect cancels out the effects of greater (overall) competitor activation in the weak target condition, thereby making it possible for us to simulate the null effect of target strength on RIF observed by Anderson et al. (1994).

Methods

In Anderson et al. (1994), Experiment 3, target strength and competitor strength were manipulated in a between-subjects fashion. The same semantic categories were used in all conditions. The four conditions of their experiment were defined by orthogonally crossing the following two factors:

- whether strong or weak items from these categories served as targets, and
- whether strong or weak items from these categories served as competitors

We set out to mirror this design in our simulations. To do this, we needed semantic categories that had more than one weak item (so we could simultaneously have a weak target and a weak competitor) and more than one strong item (so we could simultaneously have a strong target and a strong competitor). We settled on using 8 items per category, with 4 strong items and 4 weak items. Having 4 strong items helps to spread out the competitor weakening that occurs at practice, such that no single item suffers a disproportionate amount of semantic weakening.

With 8-item categories, there is no room to fit patterns for two categories ($8 * 2 = 16$ total items,

[19] Figure 23 shows a numerical trend towards a reversed RIF effect for weak competitors, but this effect was not consistent across experiments in Anderson et al. (1994).

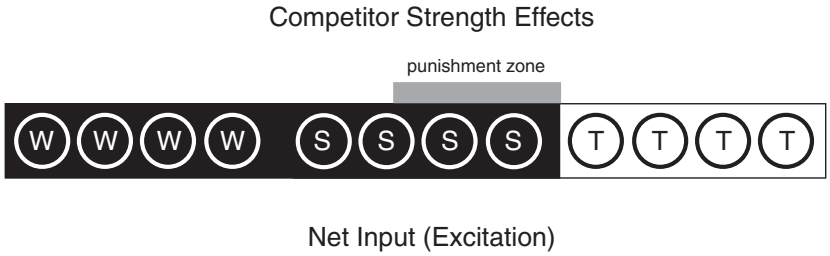

Figure 24: Schematic illustration of the distribution of net input scores for target units (marked with a T), units belonging to strong competitors (marked with an S), and units belonging to weak competitors (marked with a W). Units belonging to strong competitors are closer to the inhibitory threshold, which in turn should lead to greater punishment for strong vs. weak competitors.

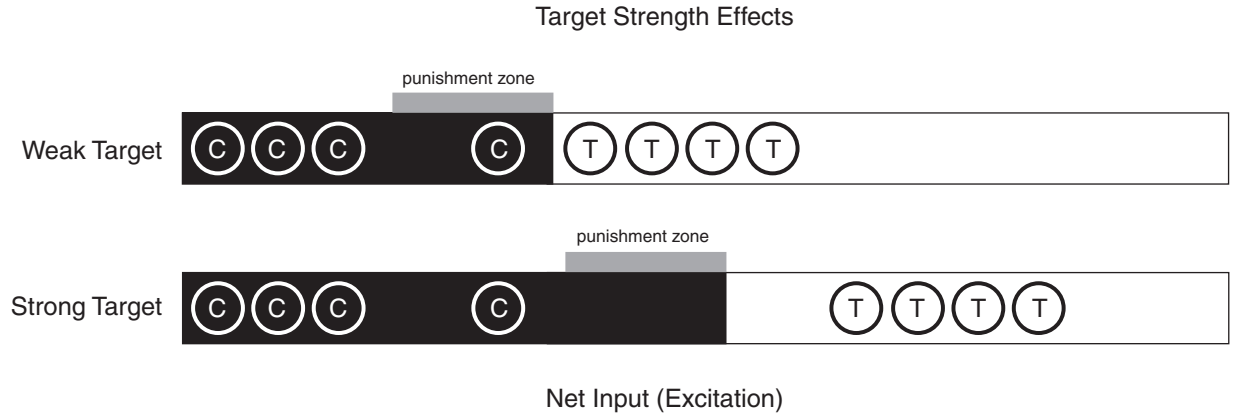

Figure 25: The figure schematically illustrates the distribution of net input scores for target units (marked with a T) and competitor units (marked with a C) for weak targets (upper bar) and strong targets (lower bar). Competitors are closer to the inhibitory threshold in the weak target condition than the strong target condition, so they are more likely to be activate and be punished in the weak target condition.

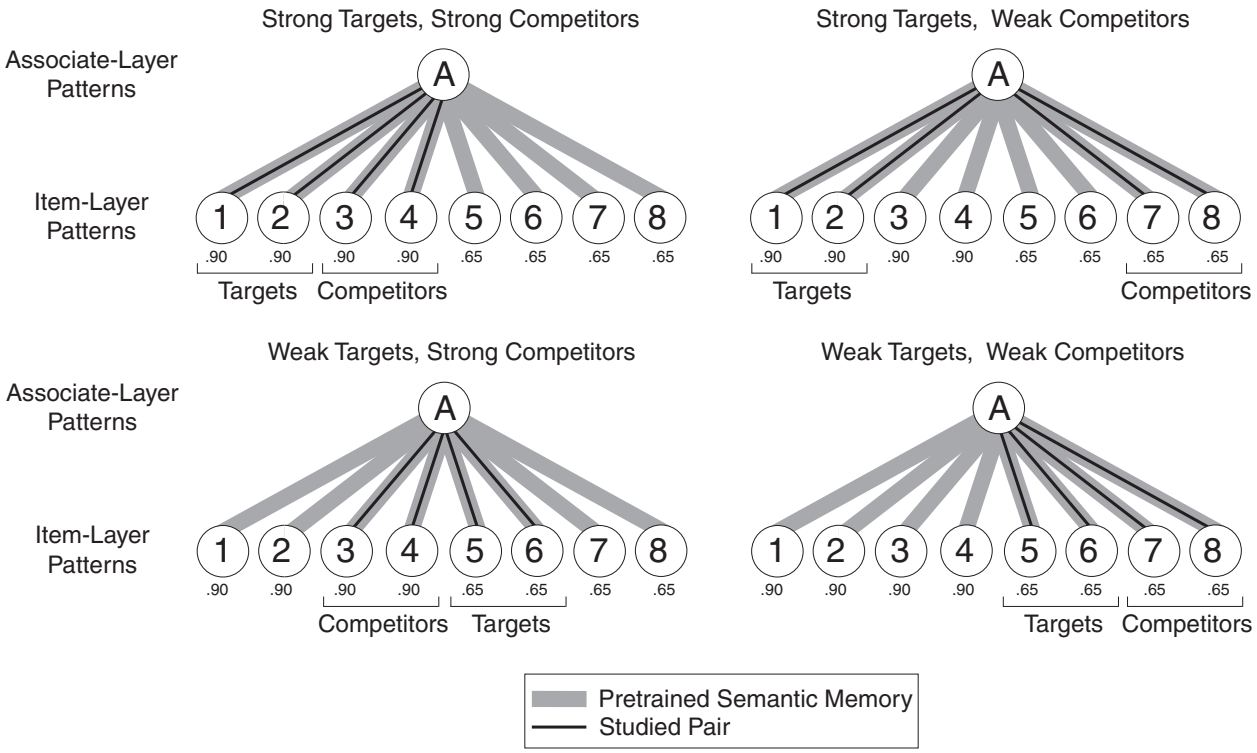

Figure 26: Illustration of the structure of the patterns used in *Simulation 2.1*. Gray bars indicate semantically pretrained pairings, black lines indicate pairings that were presented at study, and numbers below the "item-layer" circles indicate the semantic strength of each item. There were four conditions, defined by orthogonally crossing target strength (weak/strong) and competitor strength (weak/strong). Semantic pretraining was the same in all four conditions: There was one category, paired with 4 strong items (strength .90) and 4 weak items (strength .65). The only difference between the conditions is which two items were used as targets and which two items were used as competitors.

plus 16 neighbors) into the item layer without allowing overlap between item patterns. Rather than use overlapping item-layer patterns, we decided that it would be simpler to forego our standard "two category" procedure and use a single category.[20] Since we did not have a control category in this simulation, we measured RIF by testing recall performance (with learning turned off) before practice and after practice, and then computing the pretest - posttest difference score.

Figure 26 illustrates the structure of the patterns used in the simulation. During pretraining, we sampled semantic strength values for the 4 *weak* category exemplars from a uniform distribution with mean .65, half-range .10, and we sampled semantic strength values for the 4 *strong* category exemplars from a uniform distribution with mean .90, half-range .10.

In all of the conditions, 4 items were presented at study (2 targets and 2 competitors). The only difference between the conditions was whether strong or weak items were used as targets, and whether strong or weak items were used as competitors. We used partial practice during the practice phase. Finally, as per Anderson et al. (1994), we used dependent cues (our standard cues: 4/4 associate units, and 2/4 item units) at test.

Results

Results from these conditions are shown in Figure 27. 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 affect RIF. As per Anderson et al. (1994), there was no RIF whatsoever for weak competitors.

Effects of competitor strength

The finding of greater RIF for strong vs. weak competitors (in the model) can be explained in terms of the principles expressed in Figure 24. Semantically strong competitors are closer to the inhibitory threshold in cortex, so they show a larger increase in cortical activation when inhibition is lowered. This cortical pop-up for strong competitors triggers hippocampal pop-up for these competitors also. Collapsing across the Strong Target and Weak Target conditions, peak competitor activation in cortex (at the trough of the inhibitory oscillation) during the first practice epoch was .21 on average for strong competitors (SEM .00) and .00 for weak competitors (SEM .00). Hippocampal pop-up results were very similar: .21 for strong competitors (SEM .00) and .00 for weak competitors (SEM .00).

Another key to explaining the null RIF effect for weak competitors is that, for the parameters used here, hippocampal pop-up only occurs if cortical pop-up occurs first. More concretely: The hippocampal representation of the competitor needs support from the item-layer representation of the competitor in order to have enough excitatory support (in aggregate) to trigger pop-up. Thus, the fact that weak competitors do not pop up in the item layer ensures that these competitors will not pop up in the hippocampus either.

Effects of using a higher context scale value One way to underscore the importance of this dynamic (whereby cortical pop-up is a permissive condition for hippocampal pop-up) is to change the model's parameters such that hippocampal pop-up of the competitor can occur on its own. Specifically, we ran simulations where we increased the context scale parameter at practice and test from 1.0 to 1.25. This change selectively boosts the excitation of episodic traces from the study phase, making it more likely that these traces will pop up when inhibition is lowered. Whereas weak competitors did not show any pop-up (in cortex or hippocampus) for context scale 1.0, they show a significant pop-up effect in both networks for context scale 1.25; pop-up starts in the hippocampus and spreads back to cortex. Collapsing across the Strong Target and Weak Target conditions, peak competitor activation in the hippocampus was .24 on average for strong competitors and .10 for weak competitors (cortical pop-up results were very similar: .22 for strong competitors and .05 for weak competitors). This pop-up of weak competitors results in a substantial RIF effect for weak competitors, illustrated in Figure 28.[21] We re-visit the issue of how context scale interacts with episodic RIF in *Simulation 4*.

Effects of target strength

With regard to target strength effects: Earlier, we had argued that increasing target strength should reduce competitor pop-up and competitor punishment

[20] We address the issue of item-layer overlap in *Simulation 8* and in the *General discussion*.

[21] In this simulation, the RIF effect is even larger for weak competitors than strong competitors. This is a consequence of the fact that strong competitors can sometimes be retrieved correctly via semantic memory if their episodic trace is damaged, but weak competitors can not — if their episodic trace is damaged they are almost always forgotten.

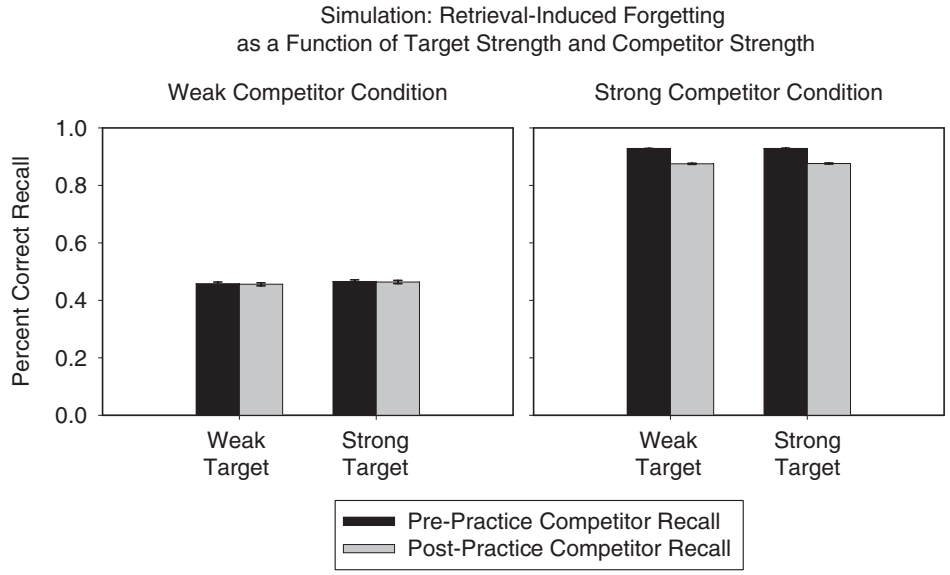

Figure 27: Graph of how competitor strength and target strength affect RIF in the model. In this simulation, RIF is affected by competitor strength (there is a robust RIF effect for strong competitors but no RIF effect for weak competitors), but target strength has no effect on RIF.

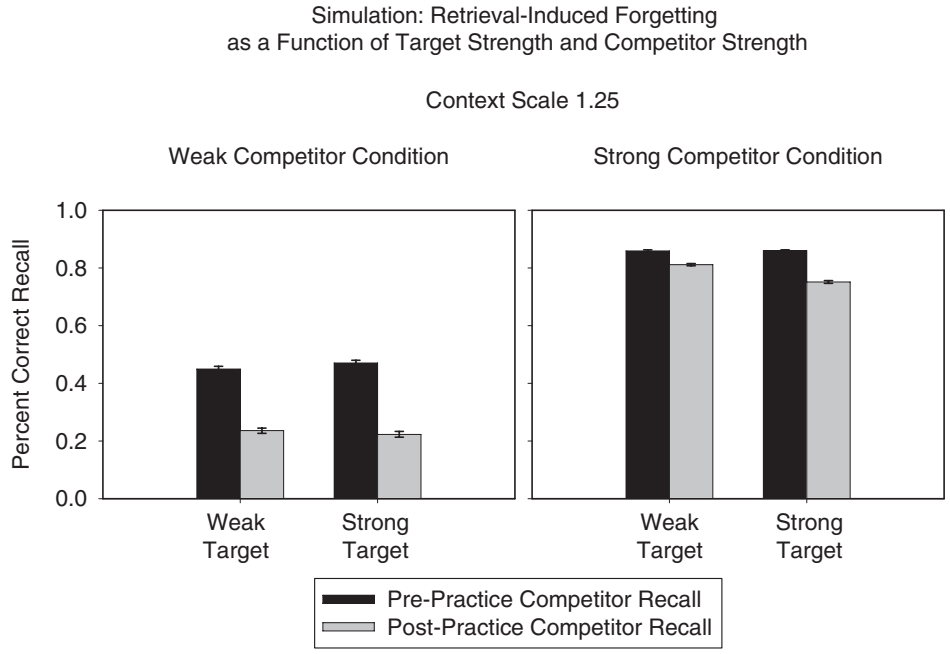

Figure 28: Graph of how competitor strength and target strength affect RIF in the model, given context scale 1.25. Unlike the context scale 1.0 simulations shown earlier (which showed a null RIF effect for weak competitors), the context scale 1.25 simulations show a very large RIF effect for weak competitors (even larger than the RIF effect for strong competitors).

(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 did not affect the size of the RIF effect.

To explore why we did not observe an effect of target strength on RIF, we plotted dynamics graphs showing competitor activation in cortex (over the course of the first partial practice trial) for the "weak target, strong competitor" condition, and the "strong target, strong competitor" condition (Figure 29). As in our previous dynamics graphs, the "competitor activation" line shows the activation of the most active of the two (strong) competitors on a given trial.

The first point to make about the graph is that, for the "weak target, strong competitor" condition, the competitor starts to activate before the onset of the low-inhibition phase. The fact that some competitor activation takes place during the end of the high-inhibition (strengthening) phase, instead of during the low-inhibition (weakening) phase, should reduce RIF. Increasing the strength of the target has two effects on competitor activation:

- First, it pushes back competitor activation so it occurs later in the trial. This has the effect of boosting competitor punishment (by ensuring that all of the competitor activation occurs during the low-inhibition phase).
- Second, as discussed above, increasing target strength reduces the overall magnitude of competitor activation during the low inhibition phase, which should reduce competitor punishment

For the parameters used in this simulation, these two effects cancel each other out, resulting in a null overall effect of target strength on RIF.

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 affect RIF) we now explore boundary conditions on these findings. First, in *Simulation 2.2*, we show that increasing target strength does reduce RIF if we use a more powerful target strength manipulation. Next, in *Simulation 2.3* 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.

Simulation 2.2: Boundary conditions on the null target strength effect

Methods

To parametrically map out the effects of target strength on RIF, we used a simpler paradigm in which the model was pretrained on two categories, each comprised of two items (the practiced category was comprised of one target and one competitor item; the control category was comprised of one target control and one competitor control). The paradigm is illustrated in Figure 30.

The competitor item and its control in the other category were pretrained with mean semantic strength .85. The semantic strength of the target item (and its control in the other category) was varied in a "between-simulated-subjects" fashion from .65 to .95 in steps of .05.[22]

The target item was practiced once, using our usual partial practice procedure. Our decision to use one practice trial (instead of three) stems from our desire to precisely control target strength — insofar as each practice trial changes both target strength and competitor strength, item strength values that are present on the second practice trial (and subsequent practice trials) might deviate considerably from the original item strength settings.

Our decision to use one target item (instead of two) was also driven by our desire to keep the target strength manipulation as pure as possible. Consider a situation where there are multiple target items (say, A-1 and A-2). In this situation, strengthening the two target items affects retrieval dynamics during "A-1" practice trials in two, qualitatively distinct ways: The strengthening manipulation boosts the strength of the currently-practiced item (A-1), but it also boosts the extent to which A-2 competes with A-1. Put another way: Increasing the strength of multiple targets also has the side-effect of changing the "competitive landscape" that is present when any one of those targets is practiced. Limiting ourselves to a single target item gets rid of this side effect and allows us to observe (without any confounds) the effect of changing target strength on RIF.

[22] To smooth out the curve relating target strength to RIF, we added noise sampled from a uniform distribution with mean 0, half-range .05 to the semantic strength values of targets, competitors, and their controls. The same qualitative pattern is present if we do not add noise.

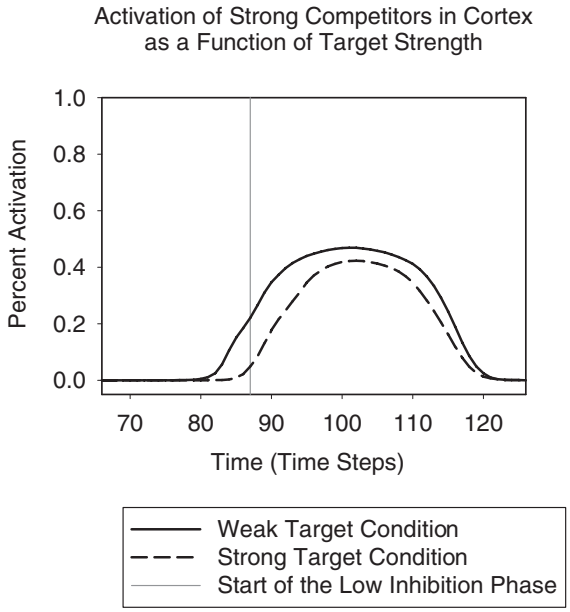

Figure 29: This plot shows competitor activation dynamics in cortex (during the first partial practice trial) for the "weak target, strong competitor" condition and the "strong target, strong competitor" condition. In the weak target condition, the competitor starts to activate before the onset of the low inhibition phase. Increasing target strength makes competitor activation occur later in the trial, and it also reduces the overall amount of competitor activation.

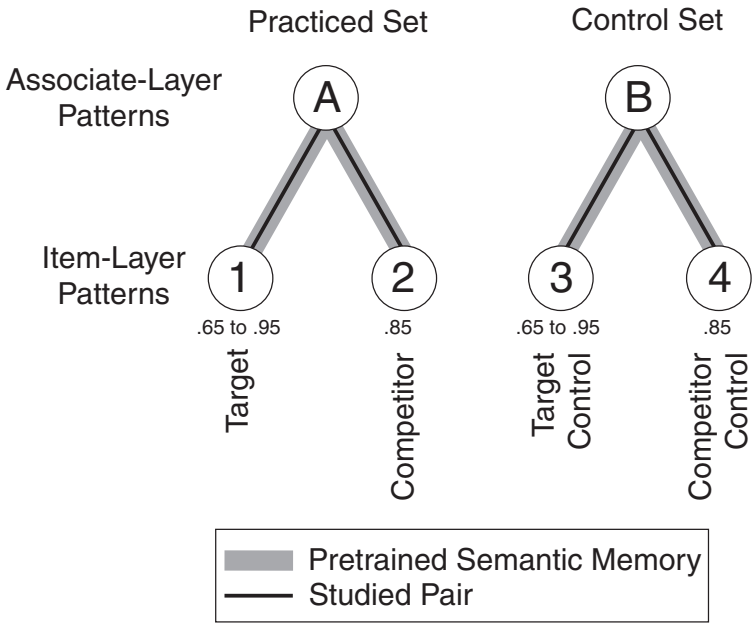

Figure 30: Illustration of the structure of patterns used in *Simulation 2.2*. Gray bars indicate pairings that were pretrained into semantic memory, black lines indicate pairings that were presented at study, and numbers below the item-layer circles indicate the mean strength of that pattern in semantic memory. Target strength was varied from .65 to .95 and competitor strength was held constant at .85.

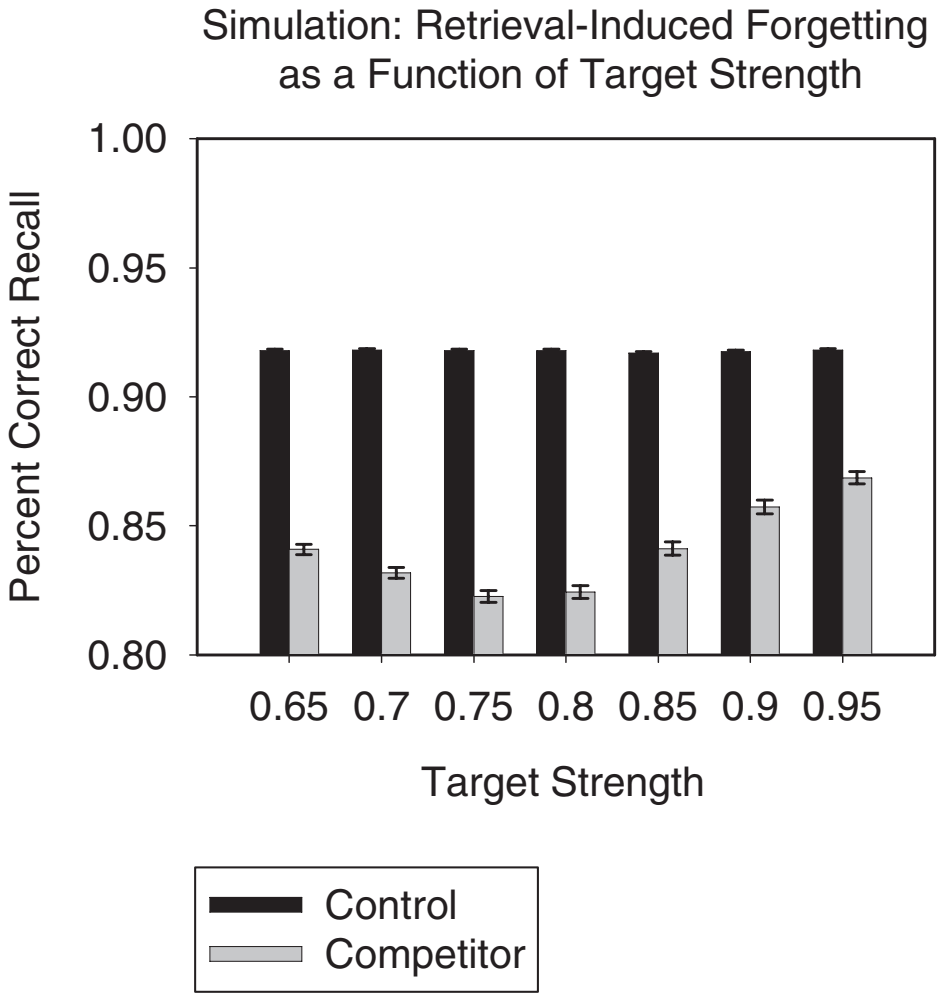

Figure 31: Graph of how target strength affects RIF. The gray bars indicate competitor recall, and black bars indicate recall of the corresponding control items. The effect of target strength is nonmonotonic: Going from target strength .65 to target strength .75, RIF increases. However, further increases in target strength beyond this point start to reduce RIF.

Results

Figure 31 plots the effect of target strength on competitor recall. Crucially, the figure shows that increasing target strength has a nonmonotonic effect on RIF. Increasing target strength from .65 to .75 boosts RIF, but additional increases in target strength reduce competitor punishment. This graph reconciles the observed null effect of target strength on RIF in our simulation of Anderson et al. (1994) with our prediction that, asymptotically, increasing target strength should reduce RIF.

The nonmonotonic pattern observed here can be explained in terms of the two effects of target strength mentioned earlier: Increasing target strength causes competitor activation to occur later (ensuring that it falls entirely within the low-inhibition phase) and it also reduces the overall amount of competitor activation.

These two competing influences are shown in Figure 32, which plots competitor activation in cortex at the onset of the low-inhibition phase ("premature activation") and competitor activation at the peak of the low-inhibition phase. Competitor punishment in the model is a function of how much competitor activation *changes* during the low inhibition phase. Thus, the difference between initial competitor activation and peak competitor activation should be a good predictor of RIF. In keeping with this view, the difference between initial and peak activation shows the same nonmonotonic pattern that was present in the RIF results (Figure 31). At first, increasing target strength boosts the "peak - initial" difference, by reducing the amount of competitor activation that is present at the start of the low-inhibition phase. Subsequent increases in target strength reduce the "peak - initial" difference by reducing the peak level of competitor activation.

Simulation 2.3: Effects of relative competitor strength

Our explanation of competitor strength effects (e.g., in Figure 24) 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*. This occurs because the k-winners-take-all inhibition rule factors in the level of excitatory support for both target and competitor units when computing inhibition. Specifically, as discussed in the *Role of inhibition* section and shown in Figure 3, k-winners-take-all places the inhibitory threshold between the k^{th} most excited unit (typically, this is the weakest target unit) and the $k + 1^{st}$ most excited unit (typically, this is the strongest competitor unit). As such, any manipulation that increases the amount of excitation received by the strongest competitor will have the effect of boosting the inhibitory threshold computed by kWTA, thereby making it less likely that other (less well-supported) competitors will pop up at practice.

We can demonstrate this point about relative competitor strength by holding the strength of some competitors constant and manipulating the strength of other competitors.

Methods

The design of *Simulation 2.3* is illustrated in Figure 33. Like *Simulation 1.1*, this simulation used two categories with 4 items apiece (two targets, two competitors). For the practiced category, the two targets had a mean strength of .85; one competitor (the *fixed-strength* competitor) had a fixed mean strength of .85; for the other competitor (the *variable-strength* competitor), mean strength was varied from .65 to .95 in steps of .10. Strength values for the control category were matched to strength values for the practiced category. For items in both the practiced and control categories, uniform noise with mean 0 and half-range .10 was added to items' semantic strength values during pretraining.

Results

Figure 34 shows the results of the simulation: As discussed above, raising the strength of the variable-strength competitors reduces RIF for the fixed-strength competitors.

Figure 35 provides further insight into the results of the relative-competitor-strength simulations. The figure plots the peak activation (during the low inhibition phase, in cortex) of the variable-strength competitor and the fixed-strength competitor, as a function of the strength of the variable-strength competitor: As the variable-strength competitor is strengthened, pop-up of this item increases, and pop-up of the fixed-strength competitor decreases. This decrease in pop-up for the fixed-strength competitor explains the decrease in RIF shown in Figure 34.

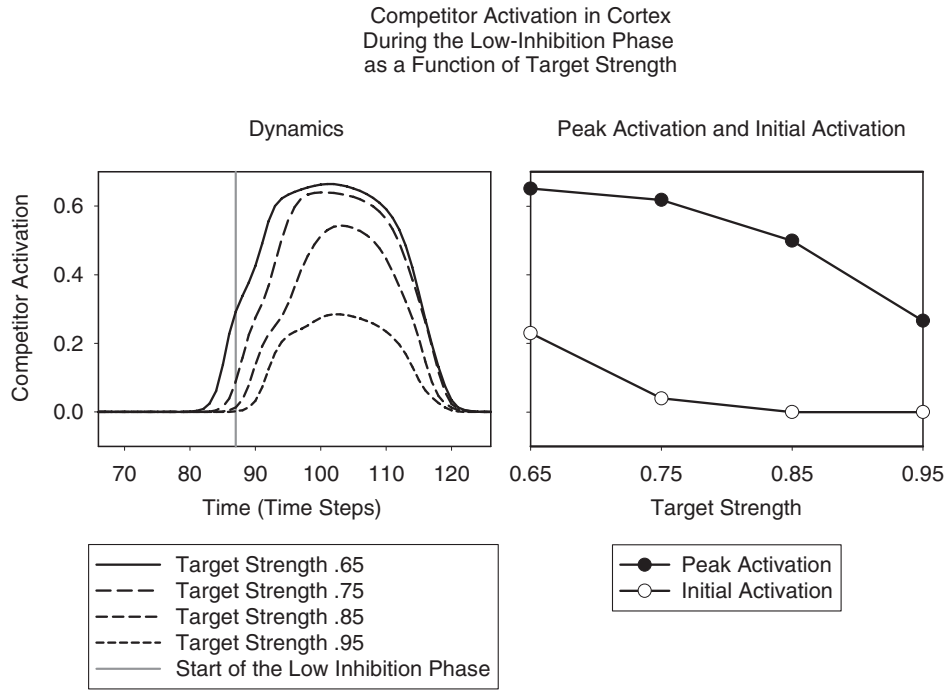

Figure 32: Plot of competitor activation in cortex during the low-inhibition phase, as a function of target strength. The left-hand figure shows competitor activation as a function of time, for target strength values .65, .75, .85, and .95. The right-hand figure re-plots this data, showing the activation of the competitor at the *onset* of the low inhibition phase, and the *peak* activation of the competitor (at the middle of the low inhibition phase), as a function of target strength. For weak target strength values, the competitor activates strongly (its peak activation is high) but it also starts to activate early, before the onset of the low inhibition phase. The primary effect of raising target strength from .65 to .75 is to make competitor activation occur later (without much change in peak competitor activation). Further increases in target strength reduce peak competitor activation.

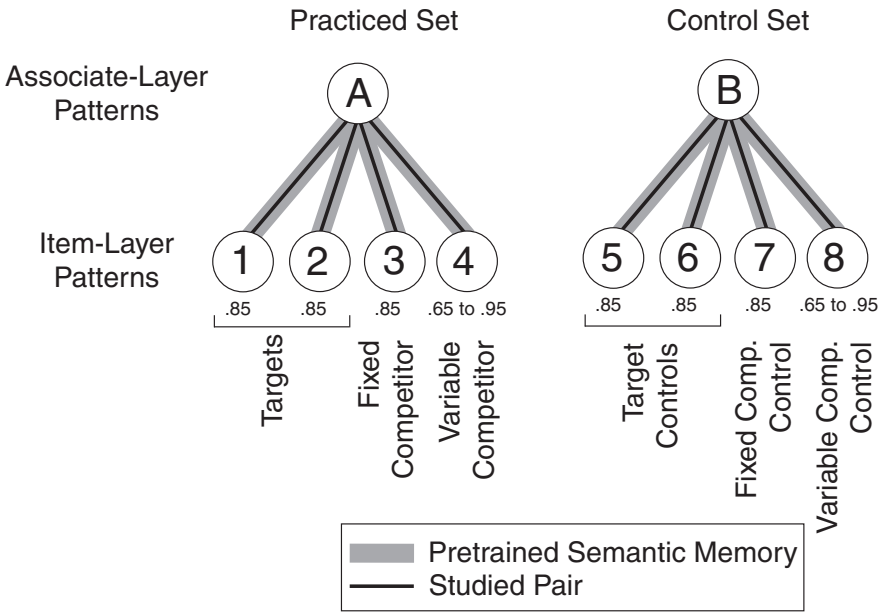

Figure 33: Illustration of the structure of the patterns used in *Simulation 2.3*. Gray bars indicate pairings that were pretrained into semantic memory, black lines indicate pairings that were presented at study, and numbers below the item-layer circles indicate the mean strength of that pattern in semantic memory. The design is the same as the design used in *Simulation 1.1*, except we varied the semantic strength of one of the competitors from .65 to .95 (the mean semantic strength of the other competitor was fixed at .85).

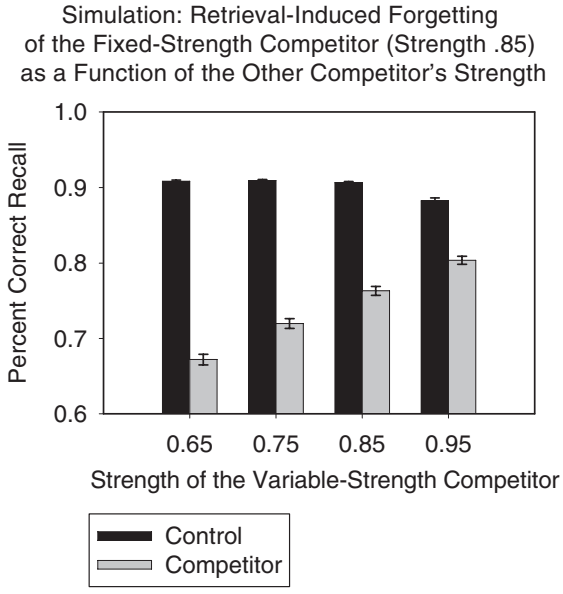

Figure 34: Plot of how RIF for the fixed-strength competitor (strength .85) varies as a function of the strength of the other (variable-strength) competitors. The target had strength .85; the variable-strength competitor's strength ranged from .65 to .95 (step .10). As the variable-strength competitor is strengthened, RIF for fixed-strength competitor decreases. This illustrates how RIF is affected by the strength of the competitor relative to other competitors (in addition to the strength of the competitor relative to the target).

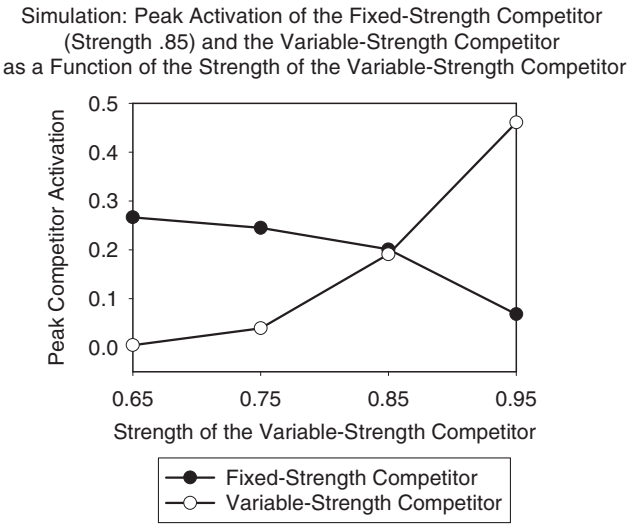

Figure 35: Plot of the peak activation ("pop-up") of the fixed-strength and variable-strength competitor in cortex during the low-inhibition phase at practice, as a function of the strength of the variable-strength competitor. As the variable-strength competitor is strengthened, pop-up for this competitor increases, and pop-up for the fixed-strength competitor decreases (thereby explaining the decrease in RIF shown in Figure 34).

Effects of relative competitor strength in our simulation of Anderson et al. (1994)

These ideas about relative competitor strength might also help to explain the lack of RIF for weak competitors in *Simulation 2.1*. Specifically, the idea that strong competitors can occlude weaker competitors suggests that, if we lowered the strength of the "strong" competitors in *Simulation 2.1*, we might start to see some cortical pop-up of weak competitors.

To test this idea, we took the "weak target, weak competitor" condition from *Simulation 2.1* (where the 4 weak category exemplars were presented at study, and the 4 strong category exemplars were nonstudied), and varied the strength of the 4 nonstudied category exemplars. The average strength of these nonstudied items was varied from .90 (the value used in *Simulation 2.1*) all the way down to .60, in increments of .10. Based on the results shown in Figure 34, we expected that reducing the strength of these nonstudied, strong competitors should boost pop-up (and RIF) for studied, weak competitors.

Figure 36 shows the results of our simulation. As expected, we found that reducing the strength of the four nonstudied items boosts RIF for the studied, weak competitors. When nonstudied-item strength was set to .90 (the value we used for strong items in *Simulation 2.1*), there was no RIF for the weak (strength .65) competitors. When nonstudied-item strength was reduced, a strong RIF effect emerged for the weak competitors (driven by pop-up of these items during the low-inhibition phase). This finding underscores that, when trying to predict RIF effects, the "weakness" of a particular competitor should always be computed *relative* to other competitors: *Simulation 2.1* showed that "weak" competitors are not strong enough to trigger pop-up and RIF in the presence of other, much stronger category exemplars; however — as shown in this simulation — the very same competitors *are* strong enough to trigger pop-up and RIF, when other (nonstudied) category exemplars are relatively weak.

Summary and discussion of Simulation 2

Competitor strength These simulations 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. One clear prediction from Figure 34 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. The results shown in Figure 36 also suggest that it should be possible to observe RIF for semantically weak competitors in situations where these items are not occluded by stronger competitors.

Target strength The target strength simulation results presented here are consistent with the

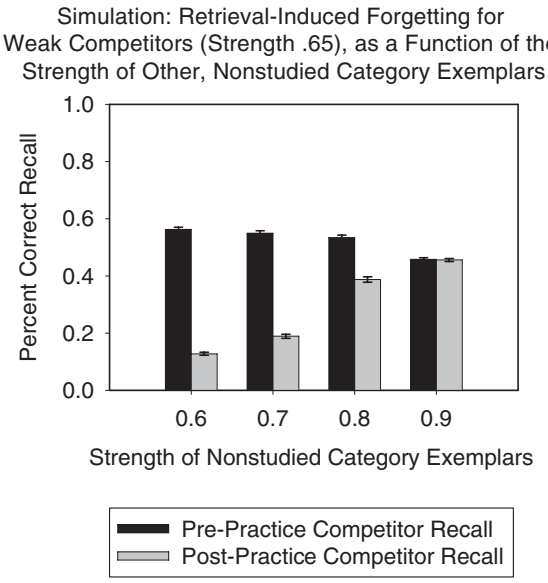

Figure 36: Plot of how RIF in the "weak target, weak competitor" condition of *Simulation 2.1* varies, as a function of the strength of the four nonstudied category exemplars. When these nonstudied items have much stronger semantic representations than the four studied items (studied strength .65, nonstudied strength .90), the nonstudied, strong items occlude the studied, weak items, preventing them from popping up at practice, and thus preventing RIF for these items. Weakening the four nonstudied items increases the odds that studied competitors will pop up at practice, thereby boosting RIF for these items.

idea, expressed earlier, that RIF should asymptotically decrease as targets are strengthened (see Figure 25). Also, our simulation results add an important boundary condition on this effect: In situations where the target is particularly weak, the competitor may start to pop up prematurely (before the start of the low-inhibition phase), thereby reducing RIF. When this happens, increasing target strength can actually boost RIF, by causing competitor activation to occur later (so it is fully confined to the low-inhibition phase). Our analytic simulations suggest that the true shape of the curve relating target strength to RIF is nonmonotonic: Going from very weak to slightly stronger targets reduces "premature pop-up" of the competitor, boosting RIF. Further increases in target strength reduce RIF by reducing the overall amount of competitor pop-up. Thus, the null effect of target strength on RIF observed by Anderson et al. (1994) (and replicated in *Simulation 2.1*) may be a consequence of the particular points on the target strength continuum that were sampled in that experiment, rather than being a parameter-independent property of RIF.

This account leads to the following prediction: By selecting appropriate target strength values for "weak" and "strong" targets, such that weak targets are close to the peak of the curve shown in Figure 31, and strong targets are located on the right side of the curve (i.e., extremely strong), it should be possible to demonstrate a robust reduction in RIF with increasing target strength.

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 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 the simulation finding (shown in Figure 31) that, as target strength increases, competitor punishment asymptotically starts to decrease. 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; see Figure 16). 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: Semantic generation can cause episodic RIF

Background

The previous simulations focused on the effects of episodic retrieval practice (i.e., actively trying to find a studied completion for a partial cue) on subsequent recall. Bauml (2002) asked a different, related question: How does semantic generation (i.e., generating a completion in semantic memory for a partial cue) affect memory for related studied items? The design that Bauml (2002) used was very similar to the "standard RIF" paradigm used in *Simulation 1.1*: First, participants studied category-exemplar pairs. Then, during the "practice" phase, participants were given partial cues that could be completed using previously nonstudied exemplars from studied categories, and they were asked to semantically generate those items. The practice phase was framed as a separate task from the study phase. Participants were not asked to think back to the study phase at all (nor would it help if they did think back, since none of the to-be-generated items were presented at study). At test, participants were asked to retrieve pairs from the initial study phase using dependent cues. Thus, the study phase and test phase were identical to the standard RIF paradigm illustrated in Figure 1. The only difference was the practice procedure. Bauml (2002) also included a control condition where participants simply studied new exemplars from studied categories at practice (instead of semantically generating these exemplars).

Figure 37 shows the results from the Bauml (2002) experiment. Semantic generation of new exemplars from studied categories led to RIF for previously studied items, but mere presentation of those exemplars did not cause forgetting.[23] The goal of this simulation is to explore whether the model can accommodate this finding.

[23] A recent study by Racsmany and Conway (2006) (Experiment 6) also looked at effects of semantic generation on recall of previously studied category exemplars, and failed to find an RIF effect. The studies used different materials, and there were also several differences in the paradigms that were used. For example, Racsmany and Conway (2006) asked participants to respond as quickly as possible during the generation test, whereas Bauml (2002) gave participants 7 seconds to reply. Further research is needed to address which of these differences was responsible for the observed difference in RIF.

Methods

Figure 38 illustrates the structure of the patterns used in *Simulation 3*. The procedure that we used for this simulation was very similar to the procedure that we used in the *Simulation 1.1*: As in *Simulation 1.1*, we pretrained two categories with 4 exemplars apiece; the semantic strength value for each of these items was sampled from a uniform distribution with mean .85 and half-range .15. However, unlike *Simulation 1.1* (where all 4 items from each category were presented at study), here we only presented 2/4 items from each category at study.

There were two practice conditions that were manipulated in a "between-simulated-subjects" fashion:

- In one condition (the *semantic generation* condition), the model was given partial cues matching the nonstudied items from one category.
- In the other condition (the *extra study* condition), the model was given full cues matching the nonstudied items from one category.

In both practice conditions, the model was given three presentations of each of the two practice cues (as per our usual procedure). The context scale parameter was set to zero for both practice conditions (reflecting the fact that, in both conditions, participants were not actively thinking back to the study phase).

Also, we used a different context tag at practice from the context tag that was present at study. This mirrors the fact that (in the experiment) the practice phase was framed as a completely separate task from the study phase.[24]

At test, we activated the "study context" tag in the context layer, and we used our standard dependent cues (4/4 associate units, 2/4 item units) to probe for the studied items.

Results and discussion

Figure 39 shows the results of our simulation, which match the Bauml (2002) results: RIF is present after semantic generation but not after extra study.

[24] Because context scale is set to zero at practice, changing the context tag between study and practice does not affect the results of this simulation; the same pattern of results is observed when identical context tags are used at study and practice.

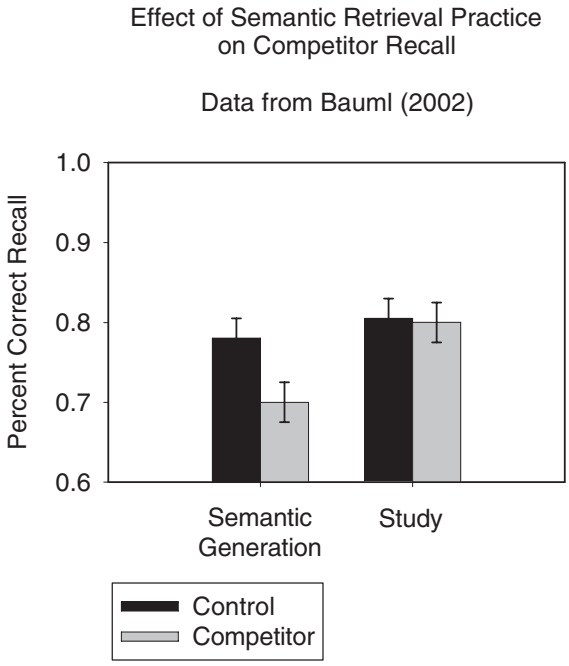

Figure 37: Results from Bauml (2002) (adapted from Figure 1 of that paper). Semantically generating non-studied exemplars from studied categories leads to RIF for studied category exemplars, but simply studying these new exemplars (instead of semantically generating them) does not lead to RIF.

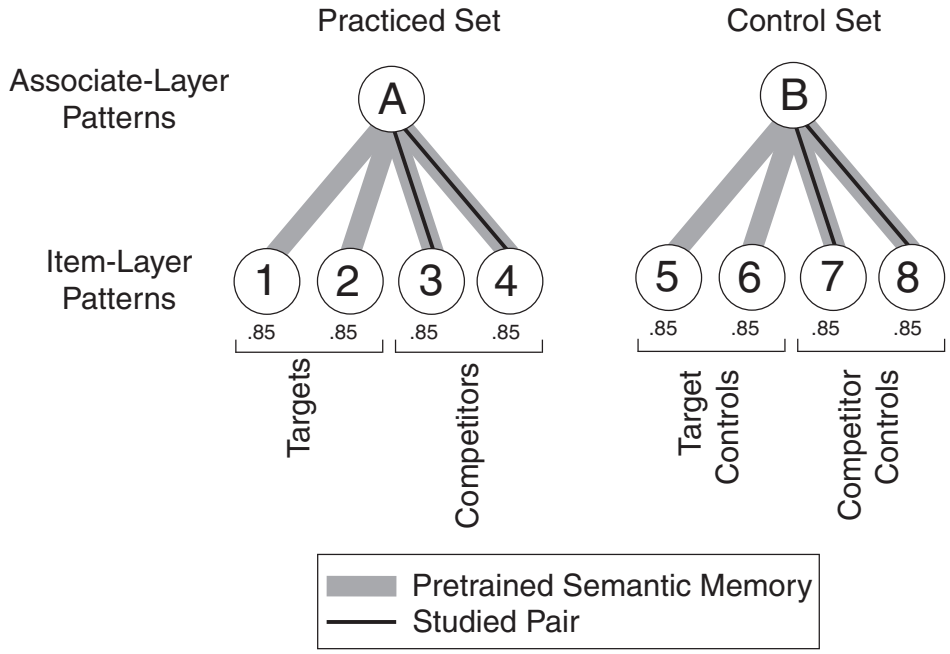

Figure 38: Illustration of the structure of the patterns used in *Simulation 3*. Gray bars indicate pairings that were pretrained into semantic memory, black lines indicate pairings that were presented at study, and numbers below the item-layer circles indicate the mean strength of that pattern in semantic memory. The pattern structure was the same as *Simulation 1.1*, except only the competitors were studied, not the targets.

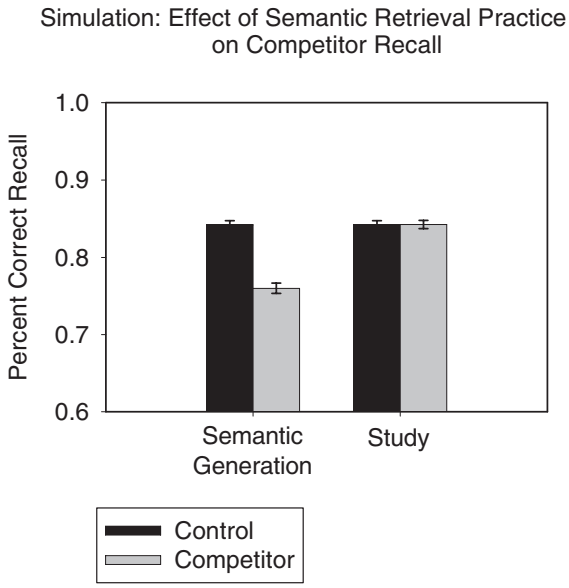

Figure 39: Simulation of Bauml (2002), exploring how semantic generation of items from a category affects recall of previously studied items. As in the Bauml (2002) experiment, semantic generation of new category exemplars causes RIF for previously studied category exemplars, but simply studying those new exemplars does not cause forgetting.

The reason why RIF occurs after semantic generation is very similar to the reason why RIF occurs after partial practice in *Simulation 1* and *Simulation 2*: When inhibition is lowered, items that are semantically associated with the category cue start to become active in cortex. If one of these semantic associates happens to be an item that was studied, this triggers activation of the hippocampal trace of that item (from the study phase). This pop-up of the hippocampal trace during the low-inhibition phase leads to RIF for the hippocampal trace.

Likewise, the reason why RIF does not occur after extra study in this simulation is identical to the reason why RIF does not occur after extra study in *Simulation 1*: When all 4 item units are externally cued (and the item's representation is strong in semantic memory), the practiced item's representation in cortex is far enough above threshold (and the competing items' representations are far enough below threshold) that no competitor pop-up occurs during the low inhibition phase (see Figure 13).

Boundary conditions

Overall, the dynamics in this simulation are quite similar to the dynamics observed in previous simulations. As such, the points made above (in *Simulation 2*) regarding effects of target and competitor strength also apply here. For example, in situations where a category includes both strong and weak exemplars, semantic generation (of either strong or weak exemplars) does not cause RIF for weak category exemplars in the model.[25]

Simulation 4: RIF for novel episodic associations

Background

Simulations 1, 2, and 3 used a paradigm where participants were asked to remember pre-experimentally associated pairs (e.g., Fruit-Apple). However, as mentioned in the *Introduction*, RIF effects can also be observed when novel pairings are used at study (forcing participants to rely entirely on episodic memory).

For example, Anderson and Bell (2001) had participants study sentences like "The teacher lifted the violin". The pairings of sentence frames ("teacher lifted") and objects ("violin") were deliberately selected to minimize obvious semantic relationships, so participants could not rely on semantic memory in this experiment. Later, participants were asked to retrieve "violin" using cues like "The teacher lifted the v___".

[25]To validate this point, we ran a variant of *Simulation 2.1* where targets were not studied, context scale was set to zero at practice (to simulate semantic generation), and different context patterns were used at study and practice. As in *Simulation 2.1*, no RIF was observed for studied weak competitors.

The Anderson and Bell (2001) study used a standard study-practice-test RIF design. The key difference between the Anderson and Bell (2001) study, on the one hand, and the studies simulated in *Simulations 1, 2, and 3*, on the other, relates to how the "practiced" and "control" sets were defined at study. In *Simulations 1, 2, and 3*, the "practiced" and "control" sets were defined by virtue of common *semantic* associations (i.e., items from the practiced set came from one semantic category, and items from the control set came from another semantic category). In contrast, in the Anderson and Bell (2001) study, the "practiced" and "control" sets were defined by virtue of common *episodic* associations. For example, in Anderson and Bell (2001), some words were paired at study with the sentence frame, "The actor is looking at", and other words were paired at study with the sentence frame "The teacher is lifting". During the practice phase, participants might practice retrieving some of the "teacher is lifting" words but none of the the "actor is looking at" words.

The basic question being addressed by Anderson and Bell (2001) is the same as in previous simulations: How does practicing retrieving some items from the practiced set affect retrieval of other items from the practiced set?

Figure 40 shows results from Anderson and Bell (2001), Experiment 4b. This experiment is especially informative because it used independent cues at test (e.g., study "actor looking at tulip", "actor looking at violin", "teacher lifting violin"; practice "actor looking at tu___"; test with "teaching lifting v___") and found a significant RIF effect.

Below, we demonstrate that we can replicate this finding of independent-cue RIF for novel associations in the model. We also describe important boundary conditions on this effect, relating to settings of the context scale parameter.

Effects of context scale

In *Simulation 2.1*, we discussed how the context scale parameter affects competitive dynamics during practice: When context scale is set to 1.0, hippocampal pop-up only occurs if the item pops up first in semantic memory. However, when context scale is set to 1.25, hippocampal traces can pop up on their own, without pop-up occurring first in the item layer. Put another way: With context scale 1.0, *only strong semantic associates are punished*, but with context scale 1.25, strong semantic links are not necessary to trigger pop-up and punishment.

Taken together, these results have strong implications for our simulations of the Anderson and Bell (2001) paradigm. Insofar as competitors are episodically (but not semantically) related to the retrieval cue in this paradigm, our previous explorations suggest that competitor pop-up (and RIF) should be observed given context scale 1.25, but not given context scale 1.0.

To test this idea, we ran two sets of simulations: one set where we used context scale 1.0 during practice and test and another set using context scale 1.25 during practice and test.

Methods

Figure 41 illustrates the structure of the patterns used in this simulation. During semantic pretraining, 8 different associate-layer patterns were linked in a 1-to-1 fashion to 8 different item-layer patterns. At study, the model was given novel pairings of these pretrained associates and items: The target (1) and competitor (2) were paired with associate A, and the target control (3) and competitor control (4) were paired with associate B. The competitor and the competitor control were also paired with other associates (C and D, respectively) that could be used as independent probes at test.

Note that, with this procedure, four of the associate-layer patterns used during pretraining (E, F, G, H) do not appear at study, and four of the item-layer patterns used during pretraining (5, 6, 7, 8) do not appear at study either. The purpose of pretraining semantic links between studied items and nonstudied associate patterns (E-1, F-2, G-3, and H-4) was to mirror the procedure used by Anderson and Bell (2001), Experiment 4b, whereby items used at study all came from different semantic categories; these four pairs were all pretrained using our "standard" semantic strength value of .85.[26] The purpose of pretraining semantic links between the "episodic cues" used at study and nonstudied item patterns (A-5, B-6, C-7, and D-8) was to capture the fact that episodic cues used in experiments like Anderson and Bell (2001) have strong semantic links to other, nonstudied items. These four pairs were pretrained using a semantic strength value of .95.[27]

[26] We also ran a version of the simulation where semantic strength values for these pairs were sampled from a uniform distribution with mean .85 and half-range .15. The results of that simulation were qualitatively identical to the results presented here.

[27] Semantic associates of episodic cues play an important role in network dynamics. In the model, if the "episodic cue" used at practice (cue A) is *not* strongly linked to any items in semantic memory, all of the units in the item layer tend to pop up at

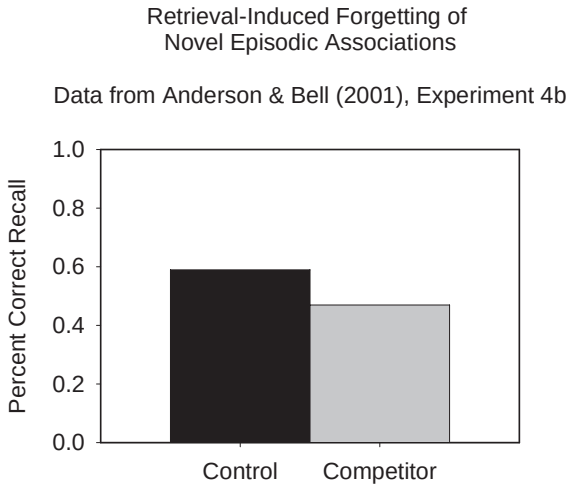

Figure 40: Results from Anderson and Bell (2001) (Experiment 4b), showing RIF effects driven by novel episodic associations. This study used verbal materials (sentences like "the actor is looking at the tulip") and independent cues at test.

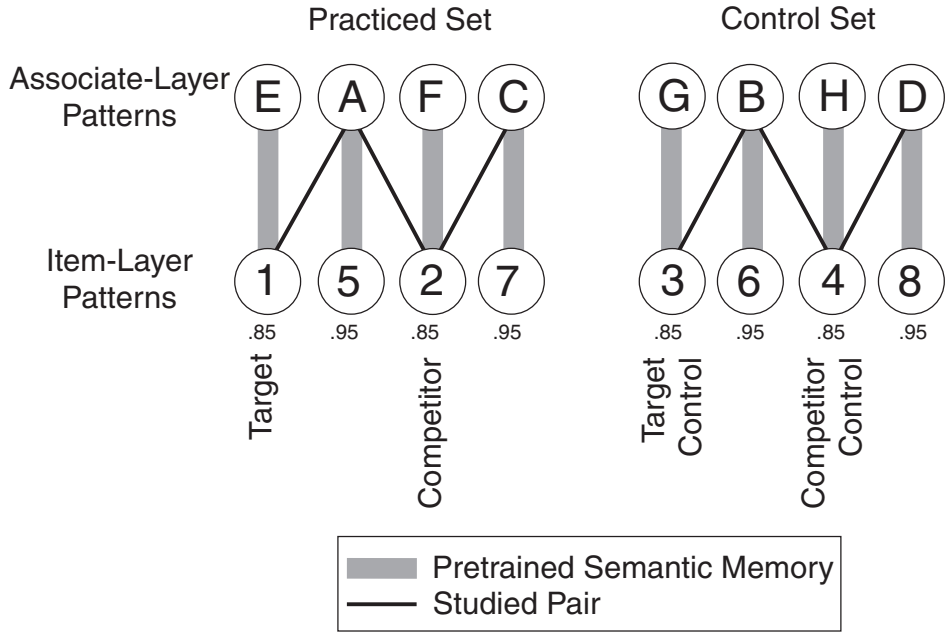

Figure 41: Illustration of the structure of the patterns used in *Simulation 4*. Gray bars indicate pairings that were pretrained into semantic memory, black lines indicate pairings that were presented at study, and numbers below the item-layer circles indicate the mean strength of that pattern in semantic memory. During semantic pretraining, 8 different associate-layer patterns were linked in a 1-to-1 fashion with 8 different item-layer patterns. At study, the model was given novel pairings of previously pretrained associates and items.

During the practice phase, we probed for the target three times using our standard partial practice cue (associate A plus three item units). For comparison purposes, we also included an extra study practice condition. During the test phase, we used our standard "associate-plus-2-item-unit" cues to probe recall for studied patterns.

Results

Figure 42 shows the effects of partial practice on independent-cue competitor recall, as a function of context scale. In the context scale 1.0 condition, the model did not show any RIF for independent cues, but the model showed a robust RIF effect in the context scale 1.25 condition. The results for dependent cues (not shown here) were the same as the results for independent cues: RIF for context scale 1.25 but not context scale 1.0. Finally, the results of the extra study simulations (not shown here) were consistent with all of our previous extra study simulations — no forgetting effect was observed for extra study, regardless of context scale. Overall, these results are consistent with our expectation that higher context scale values are needed in order to trigger episodically-mediated RIF.

The finding that RIF occurs for both dependent and independent cues in the model (for context scale 1.25) is, in large part, a consequence of the fact that *both* the "dependent cue" hippocampal representation (A-2) and the "independent cue" hippocampal representation (C-2) tend to pop up during the low-inhibition phase at practice.

Dynamics The dynamics of competitor pop-up at practice (given context scale 1.25) are illustrated in Figure 43.[28] In our previous simulations (with semantically related competitors) cortical competitor pop-up was responsible for triggering hippocampal competitor pop-up. This simulation shows the opposite pattern: During partial practice of A-1, the hippocampal representation of A-2 (the "dependent cue competitor") pops up first; this triggers activation of the cortical representation of the competitor (2). Once the cortical representation of item 2 pops up, this activates the hippocampal representation of C-2 (the "independent cue competitor"). This process, whereby activation travels from cortex to hippocampus to cortex, and then back to the hippocampus, allows the model to "find" and then weaken the hippocampal trace of the independent cue, even though the independent cue (C-2) has zero cortical overlap with the target (A-1).

Roles of hippocampal vs. cortical weakening To explore how much of the independent-cue RIF effect is attributable to weakening of hippocampal vs. cortical traces, we ran the same analysis that we ran in *Simulation 1.2*, where we measured RIF with hippocampal vs. cortical learning turned off at practice. The results of this analysis indicated that, in this simulation, RIF was entirely attributable to hippocampal weakening: The RIF effect for hippocampal-learning-only (.11) was virtually identical to the RIF effect with both hippocampal and cortical learning enabled, and the RIF effect for cortical-learning-only was not significantly different from zero. The fact that cortical weakening made a small but reliable contribution to RIF in *Simulation 1.2* but not here can be explained in terms of the idea that semantic associations were contributing to recall in *Simulation 1.2* but not here. The key feature of the current simulation paradigm is that episodic traces are both *necessary* and *sufficient* for recall: If there is not an intact episodic trace, the competitor will not be recalled properly, regardless of the strength of the cortical representation. Likewise, if the model has an intact episodic trace for the competitor, recall will be successful, regardless of whether the competitor's cortical representation has been weakened.[29]

Higher context scale values As a final note, we also ran simulations with context scale at prac-

once during the low inhibition phase, because there is no input from the associate layer to tip the balance in favor of one attractor or the other. Pretraining a semantic link between cue A and item 5 helps to break the tie between item-layer units (such that the initial wave of activation during the low-inhibition phase consists of item 5 becoming active, instead of all of the item-layer units becoming active). Note that this pop-up of item 5 causes weakening of the A-5 memory. Using a higher-than-usual semantic strength value (.95) for associations like A-5 helps to ensure that A-5 association *stays* strong enough to influence model dynamics on later practice trials, even if this association undergoes some weakening on earlier practice trials.

[28]Note that other items besides the competitor pop up at practice. In particular, given the cue A-1, item 5 (which was semantically linked to A at pretraining) tends to pop up during the low inhibition phase. Since pop-up of item 5 is not directly relevant to explaining cue-independent forgetting of the competitor, we do not discuss it further.

[29]This latter claim depends on our use of a small cortical learning rate. With our standard cortical learning rate (.05), cortical learning at practice can incrementally weaken the cortical representation of the competitor, but these changes are too small to damage the overall viability of the representation (i.e., even after weakening, the competitor still exists as an attractor state in the cortical network). If we use a much larger cortical learning rate (.20), cortical pop-up at practice can catastrophically damage the cortical representation of the competitor, such that recall is impaired even the presence of an intact episodic trace.

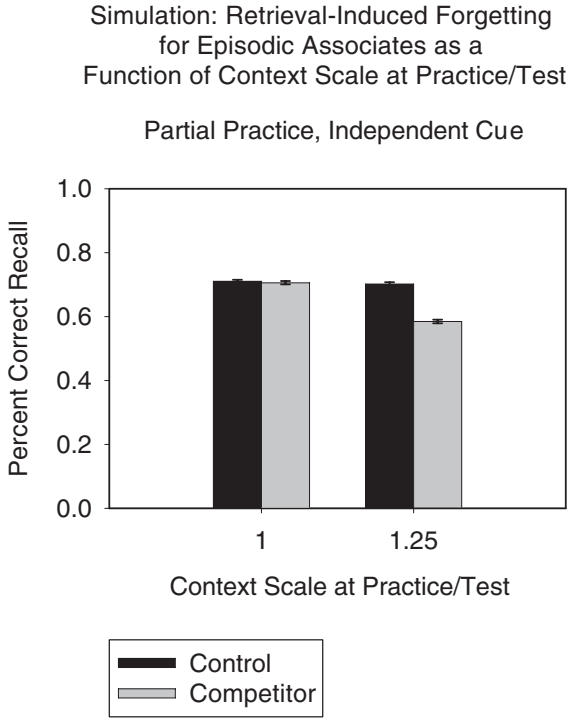

Figure 42: Simulation of independent-cue RIF effects after partial practice, when the practiced and control categories are defined by episodic associations. The left-hand plot shows RIF when context scale (during partial practice and test) is set to its default value (1.0) and the right-hand plot shows RIF when context scale is set to a higher value (1.25). RIF is observed with context scale set to 1.25 but not with context scale set to 1.0.

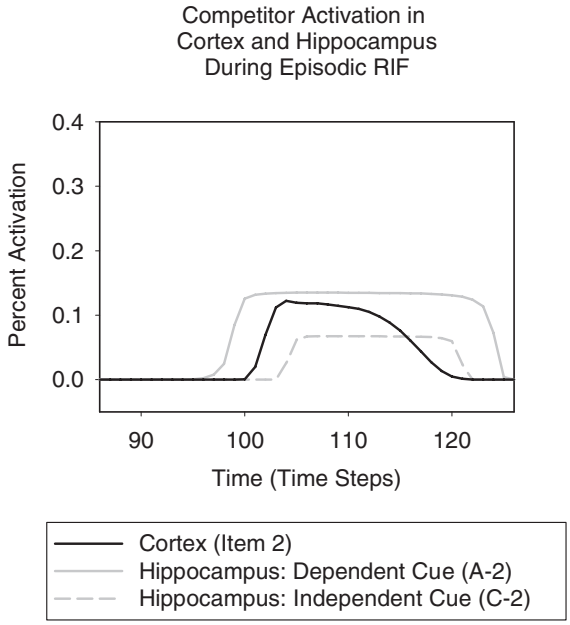

Figure 43: Plot of competitor pop-up in our episodic-memory RIF simulations, during the low inhibition phase of the partial practice condition (on the first practice epoch). The black line plots pop-up of the cortical representation of the competitor (item 2). The solid gray line plots pop-up of the episodic representation of "dependent cue-competitor" pair (A-2), and the dashed gray line plots pop-up of the episodic representation of the "independent cue-competitor" pair (C-2). Unlike previous simulations, where competitor pop-up occurred in cortex first (followed by hippocampus), pop-up in this simulation occurs first in the hippocampus. See text for discussion.

tice/test set to even higher values (1.5 and 1.75). Results for context scale 1.5 were similar to results for context scale 1.25 (pop-up of competitors, but no pop-up of control items at practice). When context scale was raised to 1.75, control items started to pop up at practice, in addition to competitors. Another way of framing this point is that, if context scale is set high enough, *merely having been linked to the study context* becomes sufficient to trigger pop-up, even if the item in question has no association whatsoever with the associate-layer and item-layer features being used as a practice cue. Pop-up of control items in this condition leads to forgetting of these items. This result may help to explain why forgetting of control items has sometimes been observed in the RIF literature (e.g., Tsukimoto & Kawaguchi, 2001).

Discussion

When comparing the results of this simulation to the results of *Simulation 2.1*, we see an interesting pattern:

- To simulate the finding of null RIF for semantically weak competitors (e.g., Anderson et al., 1994), context scale must be set to 1.0 (not 1.25) at practice. This parameter setting ensures that episodic links are *not* sufficient to trigger competitor pop-up.
- To simulate the finding of RIF for novel associates of the practice cue (e.g., Anderson & Bell, 2001), context scale must be set to 1.25 (not 1.0). This parameter setting ensures that episodic links between the practice cue and the competitor *are* sufficient to trigger competitor pop-up.

Given that different context scale settings are needed to simulate these findings, this raises the question of why participants would cue more strongly with context in Anderson and Bell (2001) compared to Anderson et al. (1994). One possible explanation is that participants modulate their (episodic) context scale value based on the contribution of semantic memory: Intuitively, episodic cuing is less important on tests where participants can "fall back" on semantic memory vs. on tests where participants are forced to rely entirely on episodic memory. According to this view, participants may have used a lower context scale value in the Anderson et al. (1994) Fruit-Apple paradigm than in the Anderson and Bell (2001) novel sentences paradigm because they could draw upon semantic memory in the former case but not the latter. We describe a way of testing these ideas about context scale and RIF in the next section.

Boundary conditions

The results of our context scale manipulations in *Simulation 2.1* and *Simulation 4* suggest that RIF for weak semantic associates and novel episodic associates should be very sensitive to how strongly participants cue with context at practice. Specifically:

- Increasing contextual cuing in RIF paradigms that use semantically-related category-exemplar pairs (e.g., Anderson et al., 1994) should result in RIF occurring for both strong category exemplars and weak category exemplars (see Figure 28).
- Reducing contextual cuing in RIF paradigms that use novel episodic associates (e.g., Anderson & Bell, 2001) should eliminate RIF for these items (see Figure 42).

One way to address these questions would be to use a hybrid episodic-semantic paradigm where a given cue (Fruit) is paired at study with some semantically related items (Apple, Pear, Kiwi) as well as some unrelated items (Shark, Helicopter, Eraser). To manipulate the extent to which participants cue with context at practice, one could manipulate (at practice) whether the practiced items are all semantically related to the cue (e.g., Fruit-Pear), or whether they are all semantically unrelated to the cue (e.g., Fruit-Shark). If all of the practiced items are semantically related to the retrieval cue, we expect that participants will use a relatively low context scale value at practice (akin to context scale 1.0 in our simulations). In this condition, as per the results of *Simulation 2.1* and *Simulation 4* (context scale 1.0 condition), we would expect to find RIF for strong semantic competitors (Fruit-Apple) but not for weak semantic competitors (Fruit-Kiwi) or semantically unrelated competitors (Fruit-Shark). Conversely, if all of the practiced items are semantically *unrelated* to the retrieval cue (thereby forcing participants to rely entirely on episodic memory), we expect that participants will use a relatively high context scale value (akin to context scale 1.25 in our simulations). In this condition, as per the results of *Simulation 2.1* and *Simulation 4* (context scale 1.25 condition) we would expect to observe RIF for all three types of studied competitors: strong semantic

competitors, weak semantic competitors, and also semantically unrelated competitors.

Simulation 5: Effects of context change on independent-cue RIF

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 that are unrelated to practiced items (see *Simulation 1.2* and *Simulation 4*). 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 Fruit-Apple RIF procedure by including a *novel associate study phase*, where each category exemplar was paired with a unique, semantically unrelated word cue (e.g., Zinc-Apple). Following this phase, participants were given a *standard study phase* where they studied category-exemplar pairs (Fruit-Pear, Fruit-Apple). Next, participants were given partial practice using category-plus-fragment cues (e.g., cue for Fruit-Pear using Fruit-_e_r). Finally, at test, Perfect et al. (2004) compared recall using two different types of cues:

- category-plus-fragment cues (e.g., test for Apple using Fruit-__p__); we will call this the *standard* cue condition
- cues from the novel associate study phase (e.g., test for Apple using Zinc-___); we will call this the *external* cue condition

Note that the first type of cue is a dependent cue. The second type of cue is an independent cue because Zinc is unrelated to practiced stimulus pairs (e.g., Fruit-Pear).

Perfect et al. (2004) found RIF using standard category-plus-fragment cues but failed to find any RIF when they tested using external cues from the novel associate study phase (Zinc). Figure 44 shows the results from Perfect et al. (2004), Experiment 3.

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 novel associate study 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, since other studies have found RIF using novel associates as cues (see *Simulation 4* above) 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 association was learned during the main study phase, whereas in Perfect et al. (2004) (Experiment 3) the novel association was learned outside of the main study phase. As such, one of the main goals of this simulation was to address the role of contextual information in modulating RIF.

Below, we show that — in keeping with the Perfect et al. (2004) data — the model shows RIF for standard cues but no RIF for external cues. At a high level, our explanation for the null external-cue RIF effect is as follows. Consider the competitor word Apple:

- During the novel associate study phase, participants form an episodic trace linking Zinc, Apple, and a "novel associate context" tag.
- During the standard study phase, participants form an episodic trace linking Fruit, Apple, and a "standard study context" tag.
- At practice, participants are given a cue like Fruit-_e_r (if Pear is a target). Also, they are explicitly asked to think back to the standard study phase, which should lead to reinstatement of the "standard study context" tag. When inhibition is lowered at practice, Apple pops up in cortex as a semantic competitor. The combination of Fruit, Apple and "standard study context" being active is an excellent match to the "Fruit + Apple + standard study context" episodic trace, and a relatively poor match to the "Zinc + Apple + novel associate context" episodic trace. As such, the Fruit-Apple episodic trace tends to pop up strongly in the hippocampus, but the Zinc-Apple trace does not. Because the Zinc-Apple episodic trace does not pop up as a competitor, it is not punished.

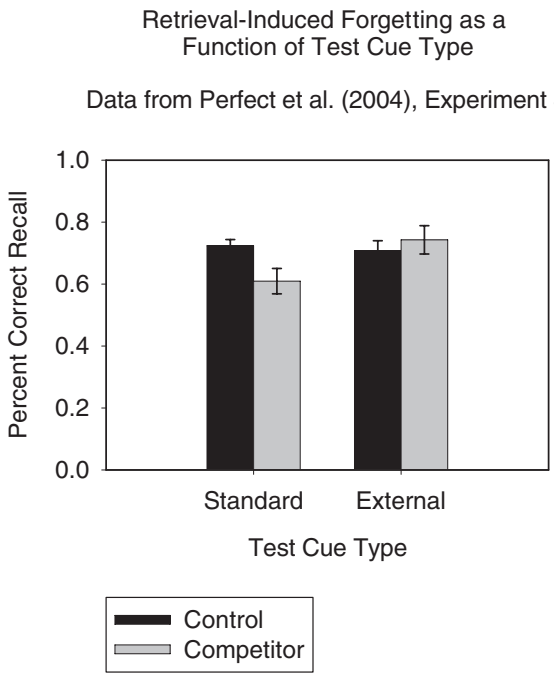

Figure 44: Graph of data from Perfect et al. (2004), Experiment 3, showing RIF when memory for the competitor (Apple) is tested with a category-plus-fragment cue (the Standard condition) vs. when memory is tested with a semantically unrelated word (e.g., Zinc) that was associated with Apple during the novel associate study 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.

- At test, when participants are cued with Zinc and asked to think back to the novel associate study phase (i.e., to reinstate the "novel associate context" tag), they can use their fully intact Zinc-Apple episodic trace to retrieve the missing associate (Apple).

In summary: This paradigm resembles *Simulation 1.2* insofar as semantically categorized items are used at study, and it resembles *Simulation 4* insofar as the independent cue is a novel episodic associate. The key difference is that, here, the independent cue is studied outside of the standard study phase. At practice, when participants cue with the "standard study context" tag, the independent-cue hippocampal trace is at a competitive disadvantage, relative to traces of items that were presented during the standard study phase. As such, the independent-cue hippocampal trace does not pop up (and is not punished).

Methods

Figure 45 illustrates the structure of the patterns that we used in *Simulation 5*. In this simulation, we semantically pretrained two categories (A and B) with two items apiece (using semantic strength .85).[30] In addition to pretraining these two categories, we also semantically pretrained two additional associate-layer patterns (C and D). These associate-layer patterns were used as "external associates" (analogous to Zinc) during the novel associate study phase, described below.[31]

For this simulation, the study phase was broken into two parts:

- First, the model was given a "novel associate study phase" in which it was given novel pairings of semantically unrelated items (analogous to Zinc-Apple): Associate C was paired

[30] We also ran a variant of this simulation where semantic strength values were sampled from a uniform distribution with mean .85 and half-range .15. The results of this simulation were qualitatively identical to the results reported here.

[31] As per the procedure used in *Simulation 4*, the two "external associate" patterns (C and D) were each paired during semantic pretraining with items (5 and 6, respectively) that did not appear elsewhere in the simulation. We included items 5 and 6 at pretraining to simulate the fact that external associates like Zinc have strong semantic links to other, nonstudied items (e.g., Tungsten). The C-5 and D-6 pairings both used semantic strength .95 (but note that strength .85 yields qualitatively identical results).

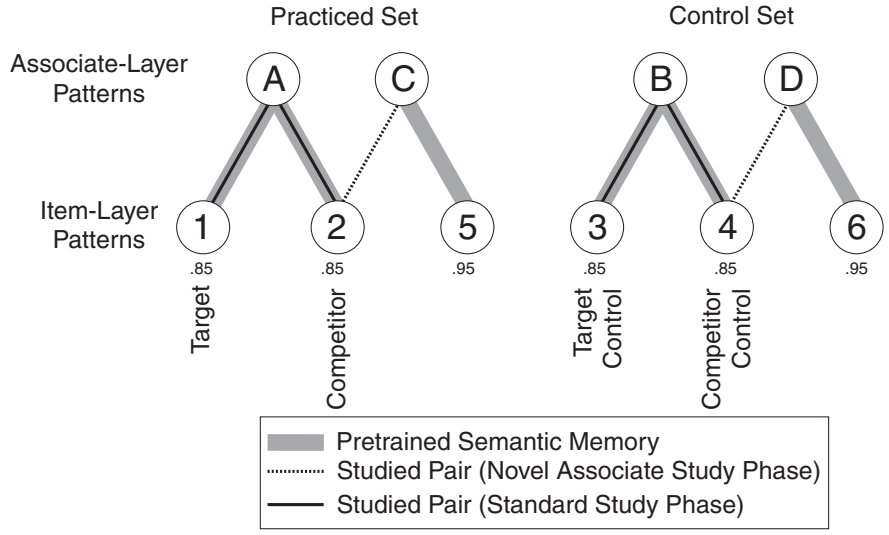

Figure 45: Illustration of the structure of the patterns used in *Simulation 5*. Gray bars indicate semantically pretrained pairings, dotted lines indicate pairings presented during the novel associate study phase, and black lines indicate pairings presented during the standard study phase. Numbers below the item-layer circles indicate the mean strength of that pattern in semantic memory. During the novel associate study phase, the model studied pairings of semantically unrelated items and associates: Associate C was paired with the competitor item (2), and associate D was paired with the competitor control item (4). During the standard study phase, the model studied semantically related category-item pairs from category A and category B.

with the competitor item (2) and associate D was paired with the competitor control item (4). A fixed "novel associate context" pattern was active in the context layer during this phase.

- Next, the model was given a "standard study phase" (corresponding to the study phase of the simulated RIF experiment), in which the model was given semantically related category-item pairs: A-1, A-2, B-3, and B-4. A "standard study context" pattern (completely distinct from the "novel associate context" pattern) was active in the context layer during this phase.

The practice phase followed our standard partial practice procedure (with semantic-category-plus-three-unit cues). The "standard study context" pattern was presented to the context layer during this phase (since participants were asked in the experiment to think back to the study phase). As in *Simulation 1.2* and *Simulation 4*, the model was given 3 trials of partial practice with the target (A-1). Context scale was set to 1.0 at practice (since recall on this test can be supported by both semantic and episodic memory).

Finally, the model was given two tests:

- First, we tested recall for the A-1, A-2, B-3, and B-4 pairings, using our standard test cues (4/4 associate-layer units, 2/4 item-layer units). Context scale was set to 1.0 (because both episodic and semantic memory can contribute to recall on this test), and the "standard study context" pattern was presented to the context layer.

- Second, we tested recall of the competitor and the competitor control using external associates. For this test, "the novel associate context" pattern was presented to the context layer (since participants were instructed to think back to the novel associate study phase). In keeping with the procedure used by Perfect et al. (2004), we cued with the associate on its own (Zinc-___). Also, in keeping with the principles for context-scale-setting outlined in *Simulation 4*, we set context scale to 1.25 for this test (insofar as this is a pure test of episodic memory — semantic memory can not be used to support performance).[32]

[32]The same pattern of results was obtained when we used context scale 1.0.

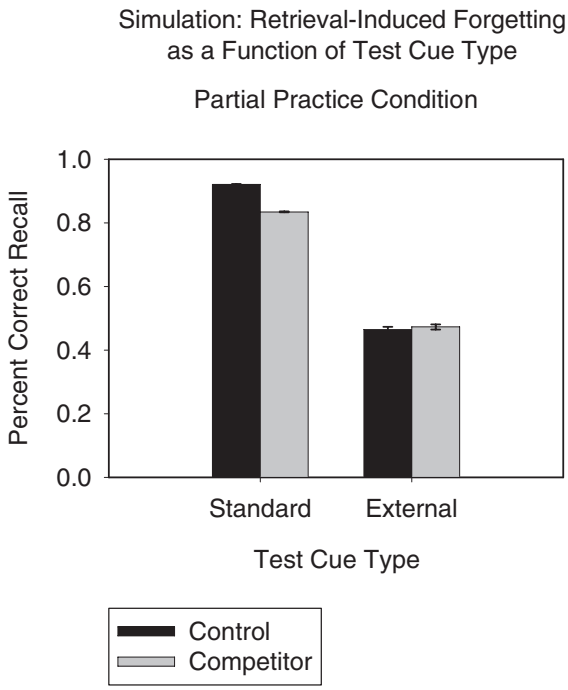

Figure 46: Simulation of Perfect et al. (2004) Experiment 3. Memory was tested using a standard dependent cue, or with an external cue (i.e., a semantically unrelated item that was paired with the target during the novel associate study phase, and was not presented during the standard study phase). RIF is observed in the model (after partial practice) in the standard-cue condition but not the external-cue condition.

Results and discussion

The results of the simulation are shown in Figure 46: In keeping with the results of Perfect et al. (2004), robust RIF was observed for the standard cue but not the external cue.[33]

These results are consistent with 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 novel associate study phase do not yield RIF, even in situations where more standard types of cues yield robust RIF effects. The model's explanation for this finding is that hippocampal traces corresponding to external associations do not activate at practice, because they do not match the contextual cue that is active at practice. Since these hippocampal traces do not activate, they are not punished, so they retain their efficacy in supporting recall at test. We ran additional analyses of network dynamics during the first practice trial to confirm this explanation of the model's behavior. As expected, the cortical representation of the competitor showed robust pop-up during the low-inhibition phase (peak activation = .59 on average, SEM = .01). Crucially, while the hippocampal representation of the standard cue-competitor pair (A-2) also showed robust pop-up (peak activation = .60, SEM = .01), the hippocampal representation of the external cue-competitor pair (C-2) did not pop up at all (peak activation = .00, SEM = .00).

These results match our finding from *Simulation 4* that cortical pop-up (on its own) is not sufficient to cause forgetting on tests of memory for novel associations — success or failure on these tests is entirely a function of whether the episodic memory trace is intact. A useful way of summarizing the results of *Simulation 1.2*, *Simulation 4*, and *Simulation 5* is that the effect of cortical weakening on recall is an (increasing) function of how much the model is relying on semantic (vs. episodic) memory at test: When semantic memory and episodic memory are both contributing (as in *Simulation 1.2*),

[33] Overall levels of recall were higher for standard cues than external cues because the model can fall back on semantic recall for standard cues, but not for external cues. Recall in the external-cue condition closely tracks the probability of successful episodic encoding (which defaults to 50% in our model). To better match recall in the standard vs. external cue conditions, we ran additional simulations where we increased the encoding success rate for external associations from 50% all the way up to 100%. This manipulation boosted the overall level of recall for external cues (so it was similar to the level of recall for standard cues) but the overall pattern of RIF effects was unchanged — the RIF effect for external cues was close to zero in all of these simulations.

the effect of cortical weakening will be small (but nonzero). When semantic memory is making *no* contribution, the effect of cortical weakening will be null. This view suggests that the most sensitive way to measure cortical weakening effects would be to set up a paradigm where participants do not episodically encode the to-be-retrieved item at all (so there is no episodic trace to get in the way, and participants are forced to rely entirely on semantic memory). This point is addressed in more detail in *Simulation 6*.

We should also point out that other factors might contribute to the null external-cue RIF effect, besides the contextual factors outlined above. For example, it is possible that participants encode Apple using different semantic features in the presence of Zinc vs. the presence of Fruit (M. C. Anderson, personal communication; see also the discussion of "transfer-inappropriate testing" effects in Anderson, 2003). At a high level, this idea has a lot in common with the explanation that we provided above. In our account and Anderson's account, the pattern of neural activity is different when participants study Zinc-Apple vs. when Apple pops up as a semantic competitor at practice. Our account posits that different "contextual tags" are active whereas the Anderson account posits that different Apple features are active. In both cases, the difference (be it contextual or semantic) creates a mismatch between features that are active at practice and features that were encoded during the novel associate study phase, and this difference prevents the Zinc-Apple episodic trace from being damaged at practice. These two accounts are not mutually exclusive, although it should be possible to tease them apart experimentally (see discussion below).

Boundary conditions

We have argued that the key factor driving the null RIF effect in Perfect et al. (2004) is that the "Zinc + Apple + novel associate context" episodic trace is a poor match for the retrieval cues that were present at practice. As such, the Zinc-Apple trace does not pop up as a competitor at practice and (consequently) it is not punished.

One prediction that comes out of this view is that, if the external associate is studied in the *same context* as the standard associate (i.e., Zinc-Apple and Fruit-Apple are studied as part of the same study list), this will remove the "contextual mismatch" factor that was blocking retrieval of Zinc-Apple at practice — when participants cue with the study-phase context, it will now be *pulling in* the Zinc-Apple trace, instead of blocking it out. As a result, Zinc-Apple pop-up should increase, leading to external-cue RIF. [34]

This prediction differentiates our context-centered view from the view that different Apple features are active for Zinc-Apple vs. Fruit-Apple. According to the latter view, the null RIF effect should persist even when Zinc-Apple and Fruit-Apple are studied in the same context (insofar as there will still be semantic feature mismatch between the Apple representation that pops up in response to Fruit-_e_r at practice, and the Apple representation that was active when studying Zinc-Apple; this mismatch should prevent pop-up of the Zinc-Apple trace and thus prevent RIF).

To test the viability of our prediction that removing contextual mismatch will boost Zinc-Apple RIF, we ran a simulation that was identical to our previous simulation of Perfect et al. (2004), except the same context tag was used throughout the simulation.

The results of this simulation are shown in Figure 47. In keeping with our expectations, there was a large RIF effect for external associates (as well as standard associates) in this simulation. This RIF effect is driven by the fact that Zinc-Apple now shows robust pop-up during the low inhibition phase (peak activation = .19, SEM = .01).

Simulation 6: RIF in semantic memory

Background

In most RIF studies, participants are explicitly asked to retrieve studied items on the final test; all of the paradigms that we have simulated up to this point fall into this category. In this simulation, we address the finding that RIF can also be observed on semantic generation tests (Carter, 2004; Johnson & Anderson, 2004).

Experiment 2 from Carter (2004) provides a clear illustration of semantic RIF. The paradigm used in this study was briefly described in the *Introduction*, and is summarized in Figure 48. Carter

[34] It is 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 the "face" episodic traces would have activated at practice, thus their efficacy as retrieval cues should be relatively preserved.

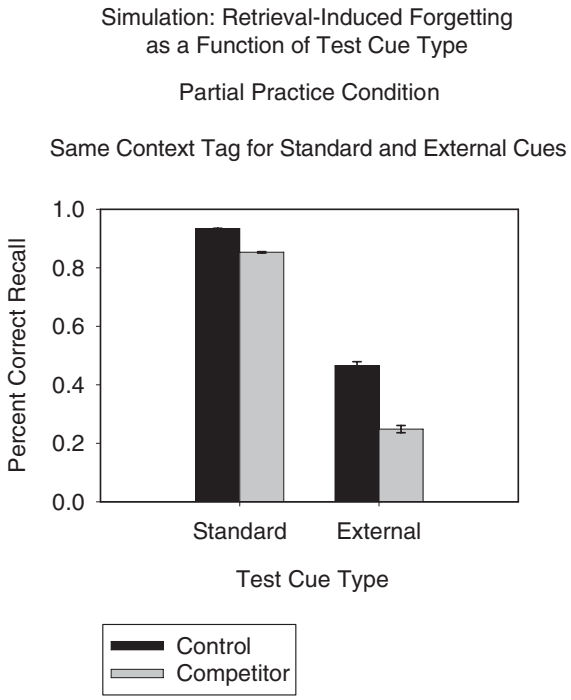

Figure 47: Results from a variant of our Perfect et al. (2004) simulation, in which the same context tag was used throughout the simulation. In this situation (where Zinc-Apple and Fruit-Apple are studied in the same context), we observe a robust RIF effect for both standard and external cues.

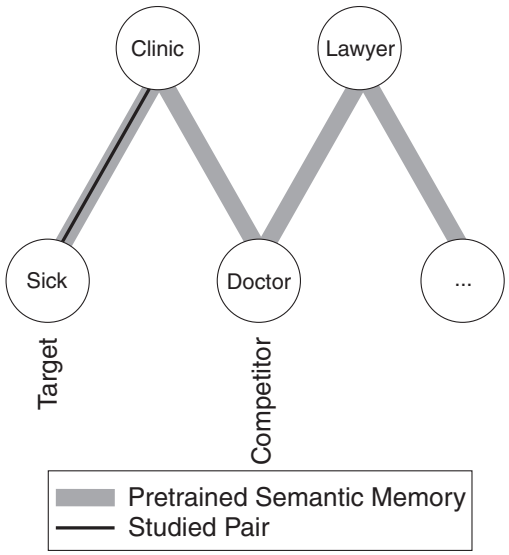

Figure 48: Illustration of the stimuli used by Carter (2004) (Experiment 2). Gray bars indicate pre-existing semantic relationships and black lines indicate pairings that appear at study. The key question addressed by the Carter (2004) study is whether practicing retrieval of studied pairs like Clinic-Sick will impair recall of nonstudied associates of Clinic (e.g., Doctor), when recall is tested using an independent cue that is also nonstudied (Lawyer).

(2004) used words like Clinic that have multiple strong associates (e.g., Sick and Doctor). Participants studied one of these associate pairs (Clinic-Sick) but not the other (Clinic-Doctor). At practice, participants were asked to retrieve Sick, using Clinic-Si___ as a cue. During this retrieval attempt, nonstudied associates of Clinic (Doctor) compete with recall of the studied associate. At test, memory for Doctor was probed by giving participants the independent cue Lawyer (which, like Clinic, is semantically linked to Doctor) and asking them to generate a semantic associate.

Figure 49 shows the data from Carter (2004), Experiment 2. The results show a clear RIF effect: Practicing retrieval of one semantic associate to Clinic (Sick) leads to forgetting of other, nonstudied semantic associates of Clinic (e.g., Doctor). We set out to simulate this finding of robust semantic RIF here.

Methods

Figure 50 illustrates the structure of the patterns used in this simulation. We used the same semantic pretraining structure that we used in *Simulation 1.2* (our previous simulation using semantically-related independent cues). The key property of this structure is that the competitor item (2) is semantically linked with two separate associates (A and C). This mirrors the property of the Carter (2004) experiment whereby Doctor (the competitor) is an associate of both Clinic and Lawyer. All items were semantically pretrained with mean strength .85.

The target (A-1) and target control (B-4) were presented at study; in keeping with Carter (2004), the model was never given a chance to study the competitor. During the practice phase, the model was given 3 trials of partial practice for the target pattern (A-1).

At test, we probed for recall of the competitor and the competitor control using associate-only cues (i.e., no item-layer units were cued). Associate C was used to probe for the competitor and associate D was used to probe for the competitor control. These are independent cues insofar as C and D are unrelated to stimuli that were presented at practice. Our use of associate-only cues at test mirrors Carter (2004)'s use of single-word test cues (like Lawyer). Context scale was set to zero at test to reflect the fact that participants were doing semantic generation (not episodic retrieval).

In the absence of any practice, the model is roughly equally likely to recall the two items (2 and 3) that were paired with associate C during semantic pretraining. The same is true for the control items (the model is equally likely to recall the two items, 5 and 6, that were paired with associate D during semantic pretraining). The key question is whether cortical pop-up of the competitor (2) during practice will weaken its semantic representation enough to "tip the balance" away from the competitor, toward the other item (3) associated with cue C.

One parameter that is important in this simulation is the variability (across items) of semantic strength values that are assigned at pretraining. If item strength variance is set to 0 (i.e., all items have weights set to .85 exactly), this constitutes a best-case scenario for detecting subtle changes to cortical weights. In this situation, the model is poised on a "knife edge" where items 2 and 3 are precisely balanced in association strength (given cue C) at the outset of the experiment, and any weakening of item 2's weights will cause the model to favor item 3 at test. A more realistic scenario is to use a small but nonzero item strength variance value (such as .05). This captures the idea that the competitor is similar in strength to other associates of cue C, but the strength values of these items are not identical. In this situation, RIF should be smaller than the zero-variance condition: On some trials, the competitor might start out weaker than the other associate (in which case it will not be recalled before or after practice); on other trials, the competitor might start out substantially stronger than the other associate, such that (even after weakening) it is still stronger and thus is not forgotten. To explore the robustness of the RIF effect in this simulation, we decided to run some simulations with item strength variance 0 and some simulations with item strength variance .05.

Results and discussion

Figure 51 shows the results of our simulation. In keeping with the results of Carter (2004), robust RIF is observed after partial practice. This RIF effect is observed because the competitor pops up in semantic memory at practice. This incrementally weakens the cortical representation of the competitor and makes it less likely that the competitor will be generated in response to an independent semantic cue at test.[35]

[35] In keeping with the idea that RIF is driven by cortical weakening in this simulation, follow-up simulations showed that turning off cortical learning at practice completely eliminates RIF, whereas turning off hippocampal learning at practice has no effect on RIF.

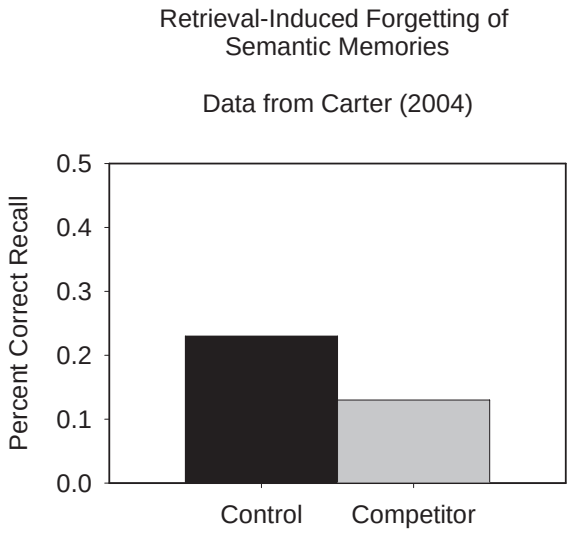

Figure 49: Results of Experiment 2 from Carter (2004), showing how practicing retrieval of studied pairs like Clinic-Sick affects the semantic representations of other (nonstudied) semantic associates of Clinic such as Doctor. Semantic memory for Doctor was tested by taking another item associated to Doctor (e.g., Lawyer) and then asking participants to semantically generate an associate to this cue. The results show a robust RIF effect for semantic memory: Practicing retrieval of Clinic-Sick leads to decreased semantic generation of Doctor.

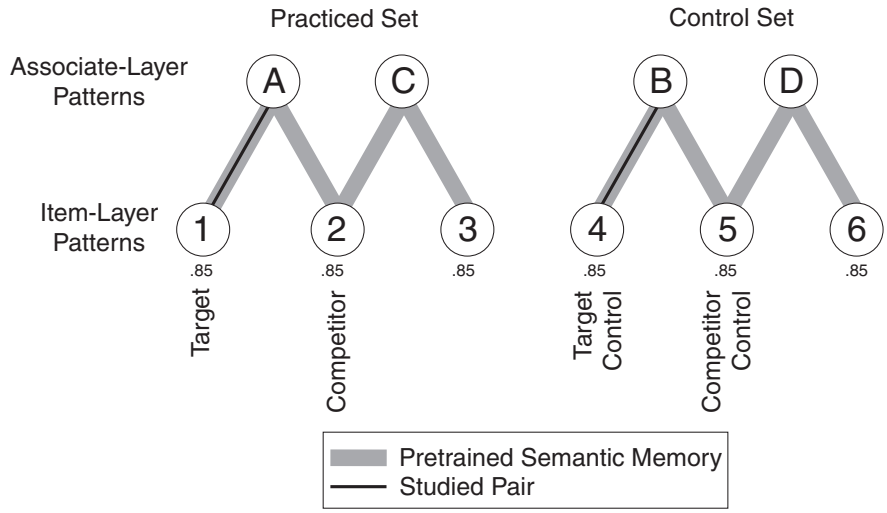

Figure 50: Illustration of the structure of the patterns used in *Simulation 6*. Gray bars indicate pairings that were pretrained into semantic memory, black lines indicate pairings that were presented at study, and numbers below the item-layer circles indicate the mean strength of that pattern in semantic memory. The patterns used during semantic pretraining in this simulation were identical the patterns used in *Simulation 1.2* (our previous simulation using semantically-related independent cues). A key difference between this simulation and *Simulation 1.2* is that — in this simulation — only the target (and target control) were presented at study; the model was never given a chance to study the competitor.

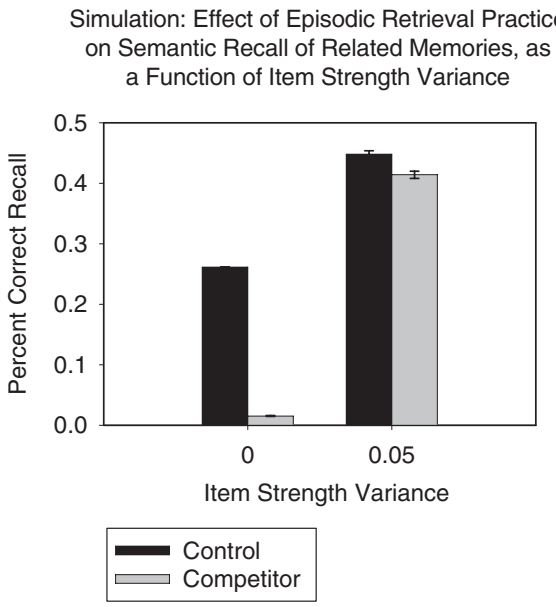

Figure 51: Simulation of Carter (2004), showing cue-independent RIF effects in semantic memory. Partial practice of the target leads to cue-independent RIF of a nonstudied competitor. The size of the RIF effect is modulated by the amount of variance that is present in the strength of the items. With zero item strength variance (such that all items start out equivalently strong), the two associates of the independent cue (the competitor and the other associate) are precisely balanced in strength and any weakening of the competitor's weights tips the balance away from the competitor. With item strength variance .05, the RIF effect is smaller but still highly reliable.

As expected, the size of the partial-practice RIF effect is modulated by the amount of item strength variance that is built into the model. When all items start out matched in strength (i.e., no item strength variance), tweaking the competitor reliably "tips the balance" of recall away from the competitor and causes a massive RIF effect. Adding .05 noise to the item strength values reduces RIF. However, even with .05 noise, the RIF effect is still highly reliable.

The main contribution of this simulation is to illustrate how relatively subtle cortical weakening effects can have a large effect on behavioral recall performance. Taken together with the results of *Simulation 1.2*, the results of this simulation also show how the effects of cortical weakening on recall are modulated by the structure of the final recall test. In *Simulation 1.2*, we showed that cortical weakening has a relatively minor effect on recall performance when the model can rely on both episodic and semantic memory at test. The results of the present simulation show that, when we force the model to rely *entirely* on semantic memory at test (by setting context scale to zero, and by testing recall of nonstudied competitors), the same level of cortical weakening has a much larger effect on recall performance.

Simulation 7: False recall and RIF

Background

There is an enormous psychological literature showing that studying items can cause false recall of other, semantically related items. Much of this evidence was obtained using variants of the Deese-Roediger-McDermott (DRM) paradigm (Roediger & McDermott, 1995), in which participants study lists of items that are all associated with a nonpresented "critical lure" word (we will refer to these lists as "DRM lists"). The finding that studying a list of items can *boost* recall of semantic associates poses a challenge for theories (like ours) that posit that semantic competitors are punished. The goal of this simulation is to explore false recall effects in the context of our model. Specifically, we want to show that the model can generate these false recall effects, and we also want to explore how false recall effects interact with the competitor-punishment mechanisms described in this paper.

In recent years, several behavioral studies have started to explore this intersection between false recall and memory weakening (e.g., Kimball & Bjork, 2002). The most relevant of these studies is Ex-

periment 2 from Starns and Hicks (2004). In this study, items from DRM lists were studied in the form of paired associates, where the list item that most strongly cues the critical lure was used as the first item in the pair, and other list items were paired with this item. For example, in the DRM list corresponding to Shirt, the strongest associate is Blouse, followed (later in the list) by Sleeves and Buttons; for this list, participants might study Blouse-Sleeves and Blouse-Buttons. After studying these pairs, participants were given a retrieval practice phase where they had to retrieve items from some of the studied lists using partial fragment cues (e.g., Blouse-S_e_v_s). Finally, participants were given a category-cued recall test ("name all of the studied words that were paired with Blouse"). Thus, apart from the fact that category cues were used at test (instead of category-plus-partial-item cues), the paradigm used by Starns and Hicks (2004) was the same as the "standard RIF" paradigm that we used in *Simulation 1.1*. Based on their use of well-established DRM lists, Starns and Hicks (2004) expected to see robust false recall of the critical non-studied item (Shirt) on the final test. Furthermore, based on their use of a standard RIF design, they expected to see RIF effects for studied competitors (i.e., studied, non-practiced items from practiced categories). The key question is whether the retrieval practice manipulation would result in RIF of critical lures (just like it results in RIF of studied competitors).

Figure 52 shows the results of Starns and Hicks (2004), Experiment 2: There is a robust false recall effect for critical lures, and (more importantly) RIF is present for both studied competitors and nonstudied critical lures.

Below, we present our simulation of Starns and Hicks (2004). To anticipate our results, we can explain the key findings from Starns and Hicks (2004) (false recall of critical lures, and RIF for these items) in the following manner:

- The most important assumption built into our simulation is that, on some trials during the study phase, participants "free associate" to the cue word (Blouse). We simulate this by cuing with Blouse only (i.e., no accompanying item information). On some proportion of these trials, the critical lure wins the competition to be retrieved (i.e., it behaves like a target on a normal study trial) — it appears at the start of the trial, disappears during the high inhibition phase, and reappears at the end of the high inhibition phase. These dynamics lead to the critical lure getting linked to the study context in the hippocampus (just as if it had been actually studied). This contextual link boosts the odds that the critical lure will be recalled at test.[36]

- These "free association" trials are intermixed with trials where the model studies actual list pairs (e.g., Blouse-Sleeves). These study trials behave just like study trials in our other simulations: Because the item representation is receiving strong support from the external cue, no competitor pop-up occurs. This lack of competitor pop-up ensures that newly-formed associations between critical lures and the study context are not damaged.

- During partial practice (unlike study trials) semantic competitors activate when inhibition is lowered. Since the critical lure is the strongest of these competitors, it has a high likelihood of activating in cortex during the low inhibition phase, which (in turn) causes the critical lure's hippocampal representation to activate as well. This cortical and hippocampal pop-up of the critical lure during partial practice results in RIF for the critical lure. In particular, hippocampal pop-up — if it occurs — damages the critical lure-study context association formed during the study phase, which reduces the odds that the critical lure will activate on the category-cued recall test.

Methods

Figure 53 illustrates the structure of the patterns used in this simulation. As in *Simulation 1.1*, we pretrained two 4-item categories. In each category, three items were pretrained with mean strength .85. Also, one item per category (the *critical lure*) was pretrained with mean strength .95. For items in both the practiced and control categories, uniform noise with mean 0 and half-range .15 was added to items' semantic strength values during pretraining.[37]

In keeping with Figure 53, we use A-1, A-2, A-3, B-5, B-6, and B-7 to refer to the normal-strength

[36] The idea that generation of associates at study contributes to false recall can be traced back to the *Implicit Associative Response* theory developed by Underwood (1965). For additional discussion of this theory and related theories, see Roediger, McDermott, and Robinson (1998).

[37] We also ran simulations using lower levels of noise, and the results were qualitatively identical to the results presented here.

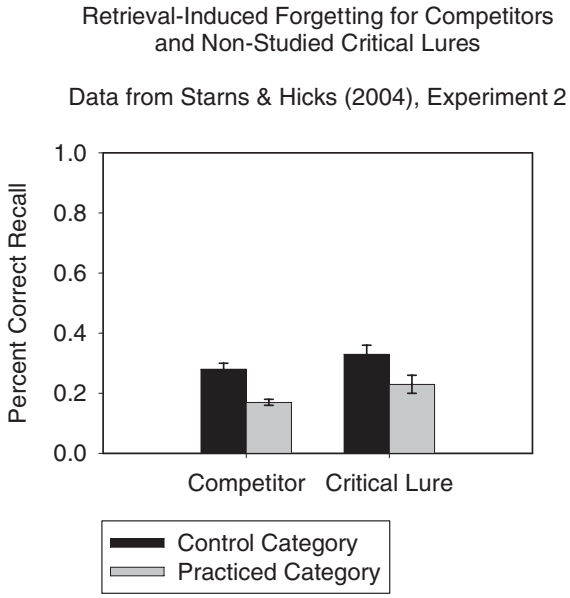

Figure 52: Results from Starns and Hicks (2004), Experiment 2, which explored the effects of retrieval practice on recall of studied competitors (i.e., studied, non-practiced items from practiced categories) and also nonstudied "critical lures" that are semantically related to studied items. The study found that partial retrieval practice results in RIF for both studied competitors and nonstudied critical lures.

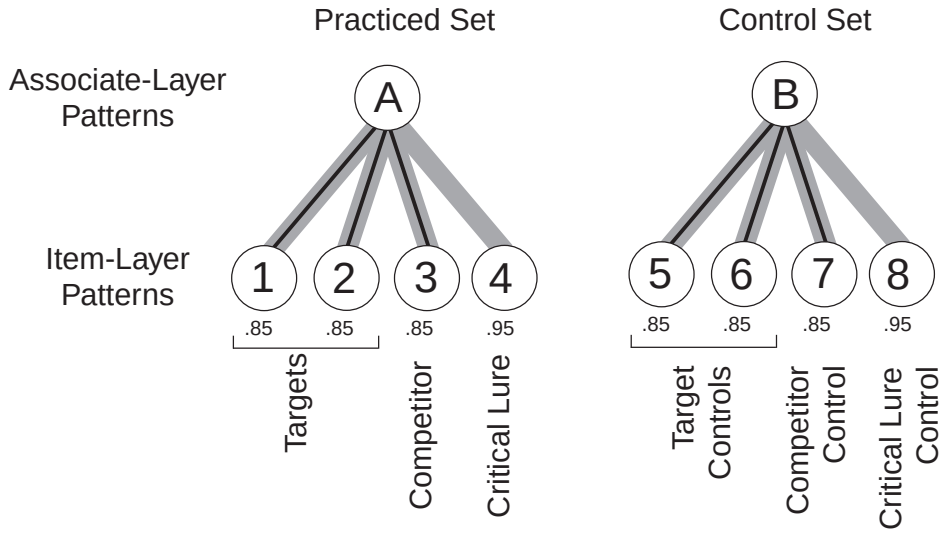

Figure 53: Illustration of the structure of the patterns used in *Simulation 7*. Gray bars indicate pairings that were pretrained into semantic memory, black lines indicate pairings that were presented at study, and numbers below the item-layer circles indicate the mean strength of that pattern in semantic memory. As in *Simulation 1.1*, the model was pretrained with two 4-item semantic categories. In this simulation, one item in each category (the *critical lure*) was semantically stronger than the others (strength .95 vs. strength .85). At study, the model was presented with the three "normal strength" items from each category, but not the critical lure or its control.

(.85) items, and we use A-4 and B-8 to refer to the critical lures.

During the study phase, we included one "generation trial" per category where we presented the category cue by itself, without any item-layer features. As discussed above, the model tends to recall the critical lure from the cued category on these generation trials (causing it to become associated with the study context).

For each category, the generation trial was followed by study of the 3 weaker items from each category, so the ordering of items at study was A- (generation trial), A-1, A-2, A-3, B- (generation trial), B-1, B-2, B-3.[38]

During the practice phase, the model was given our standard partial practice procedure for the two target items from category A (A-1 and A-2).

At test, to mirror the category-cued-recall procedure used by Starns and Hicks (2004), we cued recall with the category pattern (4/4 units) but no item units.[39]

Results

Figure 54 shows the results of our simulation of Starns and Hicks (2004). As per the results of Starns and Hicks (2004), Experiment 2, we found a robust false recall effect for critical lures, and we observed RIF for both nonstudied critical lures and studied competitors.[40]

To test the idea that "generation trials" at study play a key role in engendering false recall, we also ran another condition that was identical to the above simulation, except it did not include generation trials at study. In the absence of generation trials, the false recall rate for critical lures (after study, but without practice) was .13, compared to .28 when generation trials were included. As discussed above, generation trials foster false recall by providing an opportunity for critical lures to get linked to the study context.

Blocking effects Our use of category-only cues in this simulation provides an opportunity to revisit the issue of blocking effects (i.e., whether strengthening targets, in and of itself, causes forgetting of competitors). Anderson et al. (1994) argue that blocking effects should be larger with category cues than with category-plus-item-feature cues. The gist of their argument is that category cues match strengthened and non-strengthened items equally well; as such, there is nothing to prevent strengthened targets from coming to mind at test and blocking recall of non-strengthened items.[41] To explore whether blocking was contributing to RIF in this simulation, we ran follow-up simulations where we limited learning to the high-inhibition phase or the low-inhibition phase at practice. In keeping with the idea that blocking occurs on category-cued recall tests, we observed a significant RIF effect when we limited learning to the high-inhibition (target-strengthening) phase at practice. However, in support of the idea that competitor weakening *also* contributes to RIF, we observed a significant RIF effect when we limited learning to the low-inhibition (competitor-weakening) phase at practice. The contributions of blocking and competitor weakening were roughly equal in this simulation (e.g., for studied competitors, high-inhibition-phase practice reduced recall by .13, SEM = .01, and low-inhibition-phase practice reduced recall by .12, SEM = .01).

Discussion

Importantly, this simulation is not meant to be a definitive account of false recall in the DRM

[38] The results of the simulation were qualitatively identical regardless of the location of the generation trial in the study list.

[39] Note that fully simulating category-cued recall data would entail specifying a mechanism for generating multiple items, based on a single category cue (which would, in turn, entail simulating strategic recall-organization processes that are not the main focus of this paper). As things stand, our model generates a single response to the category cue, and we use the probability of recalling a given item (on that one recall attempt) as our index of recall performance. Thus, our category-cued-recall procedure should not be viewed as an exact simulation of category cued recall as it occurs in real life, but rather as a simple index of the likelihood that an item will come to mind, when participants cue with the study context plus the category cue.

[40] In comparing the Starns and Hicks (2004) results (Figure 52) to our simulation results (Figure 54), it is apparent that the RIF effect observed in this simulation is larger than the RIF effect observed in Starns and Hicks (2004). This difference is a consequence of the way that we simulated category-cued recall — specifically, the fact that we only used a single retrieval attempt per category cue at test. Learning at practice might result in a situation where a particular competitor is no longer receiving the *most* net input at practice (and hence is not retrieved on the first attempt), but the competitor might still be receiving *enough* support to be recalled on subsequent retrieval attempts. The real memory system's ability to cast about for additional items at test (which is lacking in the current model) should act to soften RIF effects by scooping up extra items (i.e., items receiving the second-most and third-most net input, and so on) in addition to the item receiving the most net input.

[41] Anderson et al. (1994) use this idea to explain why they observed RIF for weak competitors in Experiment 1 (which used category cues) but not Experiments 2 and 3 (which used category-plus-letter-stem cues).

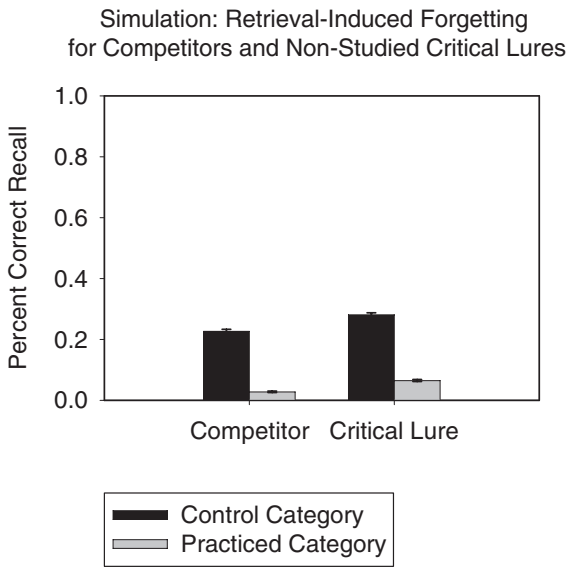

Figure 54: Plot showing how partial practice affects retrieval of studied competitors and nonstudied critical lures in the model. A large RIF effect is observed for both types of items.

paradigm. There are numerous important factors that are not yet incorporated in the model (for example, the model does not have the ability to integrate across items at study and "notice" common themes among those items). Rather, the goal of this simulation is to make two simple points:

- First, the model shows false recall in a DRM-like paradigm despite its tendency to punish related items. This occurs because generation ("free-association") trials at study cause the critical lure to become episodically associated with the study context (Underwood, 1965). Study of actual list items does not damage these context-to-critical-lure connections because (as discussed in *Simulation 1.1*) competitor pop-up does not occur on study trials.

- Second, the model shows RIF effects for non-presented critical lures, for the same reason that it shows RIF effects for studied competitors: During the practice phase, activation spreads to semantically associated items in cortex, including the critical lure. This triggers pop-up of the critical lure's hippocampal representation, which weakens the context-to-critical-lure association formed at study and (through this) causes forgetting of the critical lure at test.

At this point, it is useful to consider similarities and differences between *Simulation 6* and *Simulation 7*. Both simulations looked at RIF for non-studied semantic associates of the practice cue. The most important difference between the simulations is that the final recall test was an episodic retrieval test (context scale = 1) in *Simulation 7*, whereas it was a semantic generation test (context scale = 0) in *Simulation 6*. The fact that context scale was set to 1 on the final test in *Simulation 7* means that episodic memory traces will exert a strong influence on recall performance on the final test; these episodic influences will work to mask the more subtle kinds of semantic weakening that we focused on in our discussion of *Simulation 6*. Conversely, the fact that context scale was set to 0 on the final test in *Simulation 6* means that any kind of contextual/episodic learning that occurs at study is largely irrelevant to performance on the final test.[42]

[42]Importantly, although we framed *Simulation 6* as a simulation of Carter (2004), and we framed *Simulation 7* as a simulation of Starns and Hicks (2004), we are open to the possibility that the mechanisms described in *Simulation 6* might contribute to RIF in Starns and Hicks (2004), if participants adopt a semantic generation strategy at test. Likewise, we are open to the possibility that the mechanisms described in *Simulation 7* might contribute to RIF in Carter (2004), if participants generate semantic associates (e.g., Doctor) at study and adopt an episodic retrieval strategy at test.

Simulation 8: Extra study can cause forgetting given high pattern overlap

Background

As discussed above, several experiments have found that extra study (during the practice phase) does not cause forgetting of competitors on cued-recall tests (e.g., Blaxton & Neely, 1983; Bauml, 1996, 1997; Ciranni & Shimamura, 1999; Anderson et al., 2000a; Shivde & Anderson, 2001; Bauml, 2002). However, contrary to these findings, some experiments have found that of extra study of some list items *does* impair cued recall of other list items. For example, Ratcliff, Clark, and Shiffrin (1990), Experiment 6 found that extra study of some pairs of unrelated words impairs cued recall of other pairs of unrelated words; for a similar result, see Kahana, Rizzuto, and Schneider (2005).

Other relevant evidence comes from Norman (2002), who found that extra study of some items impairs recognition sensitivity for other items on a plurality recognition test; this test requires participants to remember whether they studied words in singular or plural form (Hintzman, Curran, & Oppy, 1992). Also, Verde and Rotello (2004) found that extra study of some items impairs recognition sensitivity for other items on an associative recognition test. Both plurality recognition and associative recognition load very heavily on retrieval of specific details (e.g., Hintzman & Curran, 1994; Curran, 2000; Yonelinas, 1997; Hockley, 1999). As such, the fact that extra study led to forgetting on plurality and associative recognition tests suggests that extra study can impair cued recall.

Finally, Anderson and Bell (2001), Experiment 5 used the sentence stimuli described in *Simulation 4* ("the actor is looking at the tulip") and found that extra study caused forgetting of competitors (see also Shivde & Anderson, 2001).

It is possible that some of these findings might be attributable to experimental confounds or other, strategic factors. For example, Bauml (1997) argued that the Ratcliff et al. (1990) cued-recall forgetting effect might be attributable to output-order confounds. Also, Kahana et al. (2005) point out that their experiment did not control for study-test lag. Finally, Anderson and Bell (2001) argue that the extra study forgetting effect that they observed might be attributable to participants covertly enacting retrieval practice during the extra study phase. When participants' results were binned according to their self-reported use of a covert retrieval strategy, participants who reported using covert retrieval during the "extra study" phase showed a significant forgetting effect, and participants who did not report using covert retrieval showed a smaller, non-significant forgetting effect.

All of the above points indicate that it is appropriate to be skeptical of findings of forgetting after extra study. Nonetheless, some of the studies reviewed above (in particular, the Norman, 2002 study and the Verde & Rotello, 2004 study) are free of obvious confounds, and both studies used demanding encoding tasks that should minimize participants' ability to covertly rehearse during extra study trials.

As such, it seems to be worth exploring (using the model) whether there are boundary conditions on the null extra-study interference effect for cued recall. In particular, we decided to focus on the issue of *pattern overlap*: How many features (on average) do participants' representations of studied items have in common with one another? One of the most salient features of the Norman (2002) and Verde and Rotello (2004) studies mentioned above is that both studies intentionally used stimulus/encoding task combinations that were designed to create highly overlapping traces: Norman (2002) asked participants to try to picture whether each object could fit inside a small box (so participants ended up picturing the box on almost every trial). Verde and Rotello (2004) gave participants unrelated word pairs and asked participants to form integrative images; crucially, individual words appeared in more than one pair, so (for example) if two studied pairs were Ostrich-Umbrella and Ostrich-Computer, participants would end up picturing an ostrich on both trials. The Anderson and Bell (2001) study also asked participants to form images and rate them for vividness. Overall, these results suggest that having participants form representations that overlap strongly across stimuli might be important for triggering forgetting.

In the simulations below, we vary overlap by varying the number of cortical (item-layer) units shared by stimuli in the experiment. Also, Norman and O'Reilly (2003) discussed how the hippocampus' ability to assign distinct conjunctive codes to overlapping stimuli can break down under conditions of high cortical overlap. Thus, in addition to manipulating cortical pattern overlap, we also manipulate the degree of overlap between hippocampal traces.

Methods

The methods for this simulation were the same as the methods that we used in *Simulation 1.1* (see Figure 9), except for the changes noted below.

The major difference between this simulation and *Simulation 1.1* is that we manipulated the level of cortical and hippocampal overlap within a given stimulus category. Specifically, the level of overlap within a category was manipulated in the cortical "item" layer and the hippocampal layer. As in previous simulations, the level of overlap between same-category items in the associate layer was 100%. We included the following overlap conditions:

- 0% item-layer overlap, 0% hippocampal overlap (this matches our previous simulations)
- 25% item-layer overlap (1/4 units), 0% hippocampal overlap
- 50% cortical overlap (2/4 units), 0% hippocampal overlap
- 50% cortical overlap, 25% hippocampal overlap (1/4 units)
- 50% cortical overlap, 50% hippocampal overlap (2/4 units)

Another difference between these simulations and *Simulation 1.1* is that we used 3/4 item units to cue recall at test (instead of 2/4 units). The third unit ensured that each pattern in the 50% overlap condition would be cued with at least one unit that was unique to that pattern.

For these simulations, we only looked at the effects of extra study at practice (i.e., we did not run partial practice or reversed practice simulations). Also, we took the opportunity to add another condition (crossed with the overlap manipulation) where we used context scale 1.0 during the study phase and during extra study practice trials (instead of our usual study context scale value of 0.0). Previously (in *Simulation 1.1*) we showed that increasing context scale at study did not have a large effect on performance given low overlap. Here, we show that increasing context scale has a very large effect given higher levels of pattern overlap.

Results and discussion

Figure 55 shows the effects of extra study on competitor recall, as a function of cortical and hippocampal pattern overlap. The left-hand side of the figure shows results when context scale at study — and during extra study practice trials — is set to our default value of 0.0. The right-hand side of the figure shows results when context scale at study — and during extra study practice trials — is set to 1.0 (the same value that we normally use for partial practice and test trials).

The simulation results show that, when context scale is set to zero, the null extra-study forgetting effect is reasonably robust to cortical overlap. Forgetting effects were either null (for 25% cortical overlap) or modest (for 50% cortical overlap, and 0% or 25% hippocampal overlap) until we reached 50% cortical overlap and 50% hippocampal overlap, at which point we observed catastrophic forgetting. When context scale is set to 1.0, the results are very different: There is a small but significant forgetting effect with 25% overlap, and increasing overlap beyond this point leads to catastrophic forgetting.

The extra-study forgetting effects observed in this simulation are driven by hippocampal pop-up of competitors. In the 0% overlap condition, there is an enormous gap in the level of excitatory input received by target vs. competitor representations on extra study trials; given the large size of this gap in excitatory input, there is no competitor pop-up (and no RIF) in this condition. Increasing target-competitor overlap boosts the level of excitatory input that the competitor receives when the target is active. Once the level of support for the hippocampal competitor representation is sufficiently high, this representation starts to pop up when inhibition is lowered, which (in turn) leads to forgetting of the competitor. Using context scale 1.0 on extra study trials boosts competitor pop-up even further, by providing additional excitatory input to the hippocampal representations of previously studied items (including competitor items).

There are several important conclusions to be gleaned from this simulation:

- For our default parameters (i.e., context scale 0 at study), the null extra-study forgetting effect is robust to the presence of some cortical overlap between patterns. This is important insofar as, in real experiments, it is likely that there will be overlap between patterns of cortical activity elicited by different items.
- If overlap is high enough, and especially if there is a high level of overlap in the hippocampus (indicating that the level of cortical overlap is overwhelming the hippocampus' ability

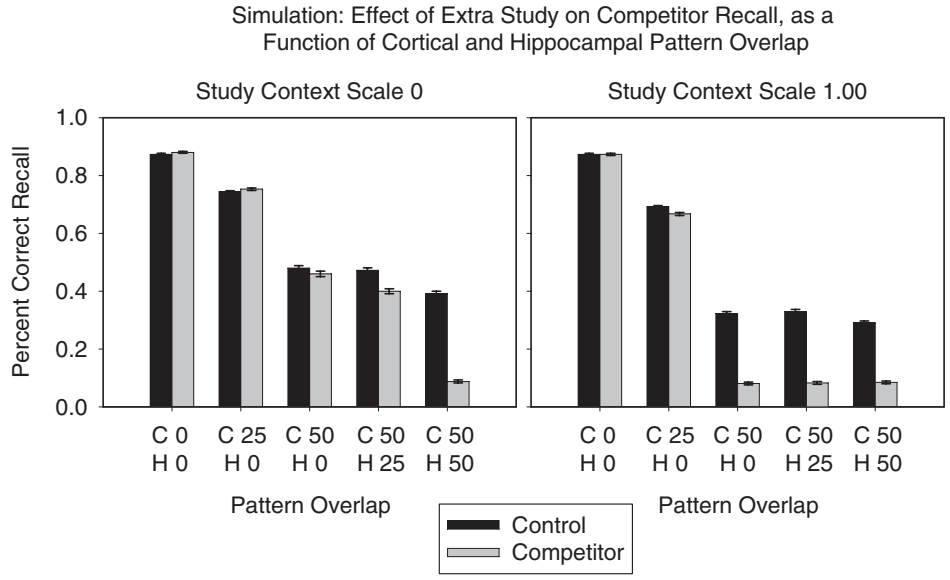

Figure 55: Plot showing the effects of extra study on competitor recall, as a function of representational overlap between items within the same category. The left-hand plot shows results for when context scale is set to its default value (0) during study; the right-hand plot shows results for when context scale at study is set to a higher value (1.0). X-axis labels indicate the degree of cortical (item-layer) overlap and hippocampal overlap (e.g., "C25 H0" = 25% item-layer overlap, 0% hippocampal overlap). When context scale is set to 0, no forgetting is observed for 0% and 25% item-layer overlap; a very small forgetting effect is observed for 50% item-layer overlap; a somewhat larger forgetting effect is observed for 50% item-layer overlap and 25% hippocampal overlap, and a massive forgetting effect is observed with 50% item-layer overlap and 50% hippocampal overlap. When context scale is set to 1.0, a small, significant forgetting effect is observed with 25% item-layer overlap. Higher levels of overlap yield massive forgetting.

to keep patterns separate), the model predicts that forgetting effects will start to emerge in the extra-study condition. This is consistent with findings, e.g., from Norman (2002), indicating that extra study can cause forgetting in situations where participants are encouraged to encode stimuli in a rich, highly overlapping fashion.

- The results from the context scale 1.0 condition illustrate the benefits of using a context scale value lower than 1.0 on study trials (instead of keeping it at 1.0 throughout all of the phases of the simulation). When context scale is set to 1.0 at study, we observe unrealistically high levels of interference: A significant forgetting effect is observed even for relatively modest levels of overlap (25% in cortex), and higher levels of overlap lead to catastrophic forgetting.

We should emphasize that our explanation of forgetting after extra study (i.e., that it is driven by high representational overlap) is not mutually exclusive with the "covert retrieval practice" explanation set forth by Anderson and Bell (2001). The main contribution of our simulation is to specify conditions where extra study might lead to forgetting, even if subjects do not deliberately try to rehearse items from the study phase. One way to get at the "image overlap" idea in a more controlled fashion would be to run a variant of Anderson and Bell (2001) where we present pictures to go along with the sentences (e.g., we could show a picture of a teacher lifting a violin) and then vary the similarity of the pictures.

Finally, we should note that some studies using a standard RIF paradigm have manipulated target-competitor similarity; all of these studies have found that increasing target-competitor similarity *reduces* RIF (e.g., Anderson et al., 2000b; Bauml & Hartinger, 2002). Importantly, these studies all used a partial practice procedure, whereas our pattern-overlap simulations (described above) used an extra study procedure. In the *General discussion*, we revisit this issue and discuss how increasing similarity can have different effects depending on whether the practice phase uses partial practice or extra study.

Simulation 9: Competition-dependent target strengthening

Background

In *Simulation 1.1*, we argued that the equivalent strengthening observed for partial practice vs. extra study was due to two countervailing forces:

- When the target is recalled successfully during partial practice, strengthening effects should be larger in the partial practice condition than in the extra study condition (because there is more competition in the former condition than the latter).
- However, recall is not always successful during partial practice; when the target is not recalled successfully, it is not strengthened (and can even be punished, if it pops up during the low-inhibition phase).

This view implies that equivalent strengthening for partial practice vs. extra study is not a parameter-independent regularity of memory, and that we should be able to unmask a competition-dependent target strengthening advantage for partial practice vs. extra study by boosting recall success during partial practice. To address this prediction, we manipulated recall success at practice in two ways:

- The first way that we manipulated recall success at practice was to vary the semantic strength of target items. Strengthening the target's semantic trace increases the odds that the model will be able to "fill in" based on semantic memory, in situations where the target's episodic trace is weak.
- The second way that we manipulated recall success was to vary the "partiality" of the retrieval cue at practice — holding target strength constant, the model was cued with all 4 associate-layer units and either 1, 2, 3, or 4 item-units. Using a sparser retrieval cue should lead to worse target recall.

For both of these manipulations, we expected that conditions associated with relatively poor target recall would show greater strengthening after extra study than partial practice, and conditions associated with relatively good target recall would show greater strengthening after partial practice than extra study.

Methods

In this simulation, we used the exact same paradigm that we used to parametrically assess how target strength interacts with RIF in *Simulation 2.2* (see Figure 30). The only difference is that, in addition to looking at partial practice effects, we also included an extra study condition. Target strength (set during pretraining) was varied from .65 to .80 in steps of .05.

During partial practice, the default was to use cues comprised of all 4 associate-layer units and 3/4 item-layer units. We also ran additional simulations (given target strength .75) where we manipulated the number of item-layer units that were used to cue recall at practice (from 1/4 units all the way up to 4/4 units).

Results

Effects of target strength Figure 56 shows the results of our target strength manipulation. These results confirm our assertion that (in the model) the relative amount of strengthening for partial practice vs. extra study depends on target strength. For weak targets (where misrecall at practice is more prevalent), more strengthening occurs for extra study vs. partial practice. For stronger targets (which are more likely to be recalled accurately at practice), more strengthening occurs for partial practice vs. extra study.[43]

Effects of cue partiality Figure 57 shows the results of simulations where we held target strength constant at .75 and manipulated the number of item units that were included in the practice cue (from 1 unit all the way up to 4 units). Context scale was held constant at 1.0 across all of the practice conditions. The data show an interesting nonmonotonic pattern whereby moving from a 4-unit (full) practice cue to a 3-unit partial practice cue boosts target strengthening, but moving from 3-unit cues to 2-unit cues and 1-unit cues leads to a decrease in target strengthening. These results can be explained as follows:

- 3-unit partial practice results in the highest amount of strengthening because the 3-unit cue is just barely strong enough to support accurate target recall. In this situation, the target comes on at the start of the trial but dips down extensively (in both cortex and hippocampus) when inhibition is raised, resulting in robust strengthening (see Figure 12, top).
- Using a full (4-unit) cue reduces strengthening because the target is *too* well-specified (so it does not dip down as much during the high inhibition phase; see Figure 12, middle).
- Using a sparser partial practice cue (with 1 or 2 item units) reduces strengthening by reducing the odds that the target will be recalled correctly in the first place.[44]

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 cue-independent forgetting, to effects of competitor and target strength, to effects of partial practice vs. extra study, to RIF for novel episodic associations (see the *Precis of Simulations* section in the *Introduction* for a more complete list of results). 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 blocking vs. weakening contribute to forgetting in our model; how "associative unlearning" theories of RIF can be reconciled with theories that posit weakening of the competitor itself; the contributions of episodic vs. semantic learning to RIF in our model; the context-dependence of RIF; the role of prefrontal cortex and top-down executive control

[43] To give a rough idea of how target strength affects recall accuracy at practice, moving from target strength .65 to target strength .75 boosts percent correct recall at practice from .52 (SEM .01) to .80 (SEM .01).

[44] To give a rough idea of how cue partiality affects recall accuracy at practice, moving from a cue with 3 item units to a cue with 1 item unit reduces % correct recall at practice from .80 (SEM .01) to .56 (SEM .01).

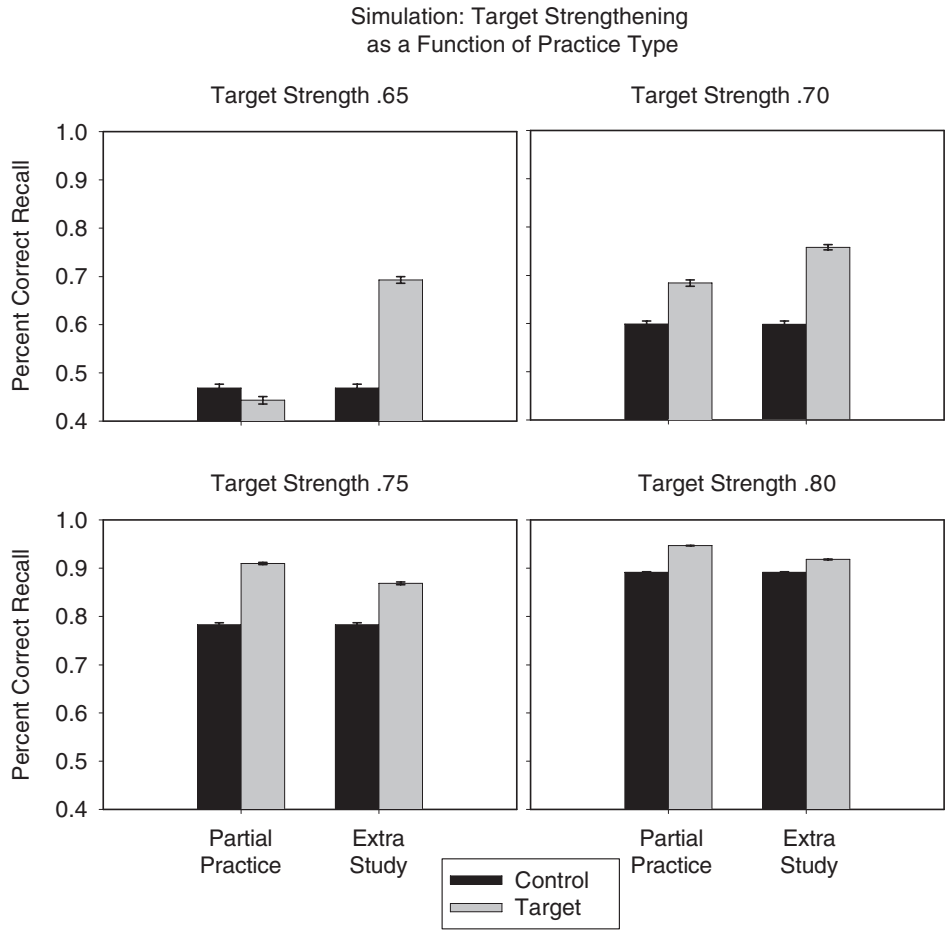

Figure 56: Simulation results showing the effects of partial practice vs. extra study on target recall, as a function of the strength of the target representation in semantic memory. For weak target strength values (.65 and .70) extra study leads to more strengthening than partial practice. For higher target strength values (.75 and .80) partial practice leads to more strengthening than extra study.

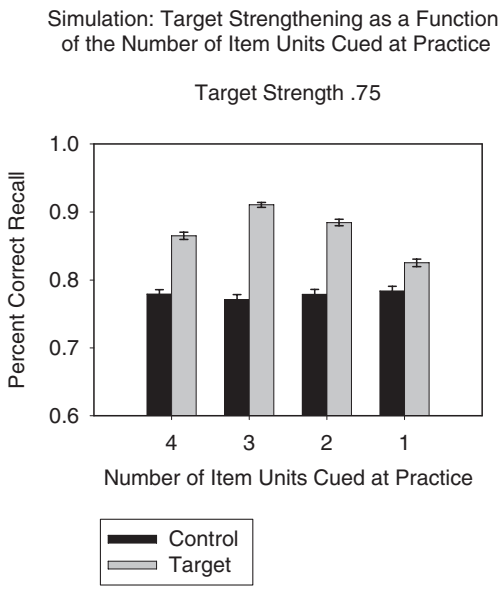

Figure 57: Simulation results showing how "cue partiality" during practice (i.e., the number of item-layer units included in the practice cue) interacts with target strengthening. Moving from a full cue (4 out of 4 units) to a partial cue (3 out of 4 units) boosts strengthening, but further reductions in the number of cued units reduce strengthening.

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.
- Third, we discuss some challenges for the model: how to account for the effects of target-competitor similarity and integration on RIF; and how to account for data on the (possibly transient) time-course of RIF. We also discuss various ways in which the model could be improved (e.g., by adding on a "prefrontal cortex" layer and exploring how it interacts with other structures during memory retrieval).
- Fourth, we discuss other applications of the model (besides modeling RIF data). Specifically, we discuss our attempts to characterize the functional properties of the oscillating learning algorithm (e.g., how many patterns it can store, compared to other algorithms; Norman et al., 2006b). We also discuss other psychological domains 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 practice vs. reversed practice or extra study (e.g., Anderson et al., 2000a; Ciranni & Shimamura, 1999; see *Simulation 1.1*, Figure 13 and Figure 15).
- The finding that strong competitors are punished more than weak competitors (e.g., Anderson et al., 1994; see *Simulation 2.1*, Figure 24 and Figure 27).

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 RIF.

- The finding that extra study and partial practice can lead to equivalent levels of target strengthening, even though there is more competition in the partial practice condition (for an example, see Ciranni & Shimamura, 1999).

In both cases, we were able to reconcile these findings with the competitive-learning framework by carefully analyzing network dynamics during the practice phase.

- With regard to the effect of target strength on RIF: We observed that, when the target's semantic memory representation is very weak, the competitor starts to pop up early (before the start of the low inhibition phase); this "premature pop-up" reduces RIF. In this situation, increasing target strength boosts RIF by making competitor pop-up occur later. Once the target is strong enough to eliminate premature pop-up of the competitor, the model behaves in full accordance with the competitive learning framework: Further increases in target strength reduce RIF, by reducing the overall amount of competitor pop-up (see *Simulation 2.2*, Figure 31 and Figure 32).

- With regard to effects of partial practice on target recall: We observed that increased competition in the partial practice phase (which boosts target strengthening) was being offset by inaccurate recall of target items (which reduces target strengthening). We manipulated recall accuracy during partial practice (by adjusting target strength and cue specificity) and showed that — when recall accuracy at practice is high — the model behaves in accordance with the competitive learning framework: Partial practice leads to more strengthening than extra study (see *Simulation 9*, Figure 56 and Figure 57). This simulation result might help to explain empirical findings (from outside of the RIF literature) showing better learning when participants generate the to-be-learned stimuli based on partial cues, as opposed to merely viewing the stimuli (the *generation effect*; for discussion of this effect see Slamecka & Graf, 1978).

Perhaps the most important contribution of this competitive-learning framework is that it provides a straightforward way of characterizing boundary conditions on RIF. These predictions are reviewed in the *Summary of predictions* section below.

Forgetting via weakening of attractor states

Blocking vs. weakening As discussed in the *Introduction*, theories 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. In simulations using "category-plus-item-feature" cues at test, RIF in the model appears to be driven *entirely* by weakening of stored traces, and not at all by blocking. To illustrate this point, we ran simulations showing that RIF is present when we limit learning at practice to the "low inhibition" (competitor weakening) phase, but not when we limit learning at practice to the "high inhibition" (target strengthening) phase (see *Simulation 1.1*, Figure 16). We also showed in *Simulation 7* that, when we use category-only cues at test, both blocking and trace weakening contribute to RIF in the model.

The total lack of observed blocking in *Simulation 1.1* merits further explanation: Insofar as recall is a competitive process in our model, how is it possible to strengthen target items without impairing recall of other (non-strengthened) items? The fact that target strengthening is not sufficient to cause forgetting in our model (given category-plus-item-feature cues) can be explained in terms of the following ideas:

- If we rank memories according to the amount of excitatory support (net input) they receive, recall success is a function of whether the net input received by the sought-after memory exceeds the *maximum* of all of the other net input values. Blocking occurs when learning at practice boosts the maximum net input value associated with other items, to the point where it "leapfrogs" over the net input value for the sought-after item.

- Because of the very high learning rate that we are using in the hippocampal model, episodic memory strength can come close to its maximal value after a single study presentation.

- If we assume that some members of the practiced category were encoded into episodic

memory at study, then additional learning at practice might result in practiced target items *matching* or *slightly exceeding* these already-encoded items in strength. However, because of ceiling effects on episodic memory strength, it is unlikely that practiced targets will substantially exceed these other items in memory strength.[45]

- Since the practice phase does not substantially affect the *maximum* strength of other items from the practiced category, blocking effects should be small or nonexistent.

Given these points, why did we observe blocking on the category-cued recall test in *Simulation 7*? The key point here is that category cues (taken by themselves) match strengthened targets and non-strengthened competitors equally well. In this situation, both targets and competitors receive very similar levels of net input from the cue — the system is effectively balanced on a knife edge between multiple memory states. When the system is in this unstable state, very small changes in target strength (at practice) can tip the balance in favor of recalling the strengthened target at test.

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.

As discussed in the *Introduction*, the simple associative unlearning hypothesis depicted in Figure 2 is falsified by the presence of RIF for cues other than Fruit. However, we think that it is possible to reconcile the idea of associative unlearning with Anderson's "inhibitory" theory by moving away from unitary concept nodes, toward a distributed-pattern approach to representing concepts. Specifically, in our model, memories are represented as attractor states comprised of multiple, interconnected microfeatures. At practice, the learning algorithm acts to weaken associations coming in to competitor features that "pop up" during the low-inhibition phase (see Figure 5). The net effect of this associative weakening is to make the competitor a weaker attractor overall, leading to generalized forgetting. Thus, at a functional level, the competitor acts as if it has been inhibited (it is generally less prone to become active), but the *mechanism* of this inhibition is associative weakening, operating at the level of microfeatures.[46]

Contributions of episodic vs. semantic memory to RIF

One of the central claims of our model is that both hippocampal (episodic) and cortical (semantic) learning can contribute to independent-cue RIF. In the model, the precise contributions of these two types of learning depend on the details of the paradigm being simulated. In paradigms that tap only episodic memory (e.g., *Simulation 4*), independent-cue RIF is driven entirely by weakening of hippocampal traces. In paradigms that tap only semantic memory (e.g., *Simulation 6*), independent-cue RIF is driven entirely by weakening of cortical traces. In paradigms where both episodic and semantic memory contribute (e.g., *Simulation 1.2*), independent-cue RIF is driven by a combination of hippocampal and cortical weakening, but (proportionally) hippocampal weakening contributes more to RIF than cortical weakening. This is a consequence of the fact that the learning rate is larger in the hippocampal network than the cortical network.

Another important point to take away from these simulations is that relatively subtle changes in the structure of the retrieval cue can have a large effect on whether episodic associates of the cue are punished (for additional discussion of this point, see Anderson, 2003). In particular, we showed that small changes to the context scale parameter at practice can change the observed pattern of RIF results: With context scale set to 1.0, episodic competitor pop-up occurs *only if* the competitor pops up first in semantic memory. This dynamic limits competi-

[45] It is important to note that episodic memory traces do not completely saturate after one learning trial, and semantic memory strength can increase at practice also. However, these effects are relatively subtle compared to the basic effect of whether or not an item was encoded into episodic memory.

[46] The idea of relating RIF to "distributed feature" models is not new: Many of Anderson's papers use distributed feature diagrams to illustrate RIF effects (e.g., Anderson et al., 2000b; Anderson & Spellman, 1995). The main difference between our "attractor weakening" idea and the Anderson distributed-feature theory is that our model focuses on strengthening and weakening of *connections* between features (and the effect of these changes on attractor dynamics), whereas the Anderson's distributed-feature diagrams focus on inhibition and strengthening of the features themselves.

tor punishment to strong semantic associates of the cue, thereby helping to explain why Anderson et al. (1994) and Bauml (1998) found a null RIF effect for weak semantic associates of the cue (see *Simulation 2.1*). However, with context scale set to 1.25, episodic associates of the cue can pop up on their own. This dynamic is important for explaining how RIF can occur in purely episodic paradigms (see *Simulation 4*).

Context-dependence of RIF

Several recent discussions of RIF have argued that RIF is "context-dependent" (e.g., Perfect et al., 2004; Racsmany & Conway, 2006). Different authors use this term in slightly different ways. The key unifying claim is that RIF involves weakening or inhibition of context-sensitive episodic memories from the study phase. As such, changing context between the initial learning phase and subsequent phases of the experiment should reduce RIF.

Our model shows context-dependent RIF effects because the oscillating algorithm weakens context-dependent hippocampal memories. *Simulation 5* provides a useful illustration of the context-dependent nature of RIF in our model: Changing the context representation between the novel associate study phase and the practice phase effectively prevents episodic traces from the novel associate study phase from popping up at practice, thereby protecting them from punishment.

However, it is also important to emphasize that RIF is not completely context-dependent in the model. As discussed throughout the paper, the oscillating algorithm weakens traces that pop up in the hippocampal network and also in the cortical network (i.e., the associate and item layers). Insofar as the cortical network is not directly connected to the context layer, the model predicts that cortically-mediated RIF effects (like the semantic RIF effect that we showed in *Simulation 6*) should still be observed when context is changed between study and test. Another point is that, while recall in the hippocampal component of our model is *modulated* by context, contextual match is not a *strict prerequisite* for hippocampal recall. To the extent that it is possible to access hippocampal traces outside of the original context, weakening the hippocampal trace should result in some degree of generalized (i.e., context-independent) impairment.

How prefrontal cortex contributes to RIF

Anderson's recent writings on RIF have emphasized the role of top-down executive control (implemented by prefrontal cortex) in RIF (e.g., Levy & Anderson, 2002; Anderson, 2003). According to this view, prefrontal cortex (*PFC*) acts to suppress competing memory traces during retrieval; these suppression effects linger after the trial is over, resulting in RIF. We agree with the idea that prefrontal cortex plays a large role in RIF. However, we do not think that PFC plays a necessary role in competitor weakening. 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 weakening. PFC can influence *which memories are weakened* and *to what extent* these memories are weakened by biasing the retrieval competition in favor of one of the memories. There is a large neuropsychological literature (e.g., Schacter, 1987) and neuroimaging literature (e.g., Fletcher & Henson, 2001) showing that PFC helps to "target" memories from particular temporal contexts. As such, we would expect PFC to play a key role in focusing retrieval on study-phase memories in RIF experiments (a process that is captured in our model in a very crude way with the "context scale" parameter). More generally, we think that PFC plays a critical role in minimizing blocking at test, by helping to focus attention on features of the retrieval cue that are especially diagnostic (i.e., features that match the sought-after item but not other, competing items). We discuss PFC contributions to RIF in more detail in the *Model improvements* section below.

Comparison to other neural network models

Our model is the first to address the full constellation of RIF phenomena discussed here. To our knowledge, the only other published neural network model that has specifically tried to address RIF data 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 related 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. extra study). 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 the more complex RIF phenomena described in this paper (e.g., data indicating competition-dependence). A possible problem with the Oram and MacLeod (2001) model in this regard is that 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 unlikely to us that the Oram and MacLeod (2001) model will be able to provide a comprehensive account of the learning phenomena that 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 (i.e., networks that represent unit activity in terms of a scalar output value, rather than simulating actual spiking neurons) 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. This is true of the Hebbian rule used by Oram and MacLeod (2001), and O'Reilly's Leabra algorithm (O'Reilly & Munakata, 2000). 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 partial practice (Fruit-Pe___) than after extra study (Fruit-Pear) (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 extra study vs. partial practice example, one could tune the network to allow competitors (Apple) to remain weakly active in the Fruit-Pe___ condition, but not in the Fruit-Pear condition; this adjustment could be coupled with a learning rule that selectively weakens representations that are weakly (vs. strongly) active; this is the tactic taken by the BCM algorithm (described below). 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.

Another key challenge for models of RIF is how to weaken the competitor without also weakening 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 (see Equation 4). Because of these phase-dependent changes in the sign of the learning rule, increased activation that occurs during the start of low-inhibition phase (when the network is "peeking" below threshold) has a different effect on synaptic weights 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 when the receiving unit's activation is *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-Pe___ 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.[47] One potential issue is that previous applications of BCM have focused on feedforward self-organizing networks (e.g., models of the development of receptive fields in visual cortex; Bienenstock et al., 1982) and it is unclear whether BCM is up to the basic task of memorizing large numbers of overlapping patterns (so they can be completed based on partial cues) in a recurrently connected network.[48] It is also worth noting that BCM's form of competitor punishment and the oscillating algorithm's form of competitor punishment are not mutually exclusive: It is possible that combining the algorithms would result in better performance than either algorithm taken in isolation. We will explore ways of integrating BCM with the oscillating learning algorithm in future research.

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 were previously attributed to unlearning (e.g., retroactive interference in the AB-AC interference paradigm; 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 particular, the role of contextual cues) in determining forgetting effects (e.g., Mensink & Raaijmakers, 1988; Howard & Kahana, 2002).

Our model deviates sharply from the approach taken by abstract models, insofar as our model incorporates a synaptic-level unlearning process, and it posits that synaptic weakening is a major cause of forgetting (although blocking can also contribute, in situations where retrieval cues are relatively ambiguous; see *Simulation 7*). While we appreciate the analytic utility of trying to explain as much data as possible without positing any kind of trace weakening, there is abundant evidence for activity-dependent synaptic weakening in the brain (e.g., Malenka & Bear, 2004), and it stands to reason that this synaptic weakening has functional conse-

[47] 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.

[48] By way of comparison, we have demonstrated (in work published elsewhere: Norman et al., 2006b) that the oscillating algorithm is capable of memorizing large numbers of overlapping patterns in a multi-layer cortical network; this work is discussed briefly in the *Functional properties of the learning algorithm* section, below.

quences. Our work can be construed as an attempt to better understand when memory-weakening occurs, and how it affects performance on semantic and episodic memory tests. In future work, it will be valuable to assess whether ratio-rule models can account for the findings described in this paper without positing any kind of competition-dependent synaptic weakening mechanism.

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.

Target strength effects

- Target strength should have a nonmonotonic effect on RIF: When targets are very weak, increasing target strength should boost RIF (by delaying the onset of competitor activation so it lines up better with the low-inhibition phase of the oscillation). Further increases in target strength should reduce RIF, by reducing the overall amount of competitor activation (see *Simulation 2.2*, Figure 31). As discussed in *Simulation 2.2* one explanation for the null target strength effect observed by Anderson et al. (1994) is that their "weak target" and "strong target" conditions happened to fall on the rising and falling sides (respectively) of this nonmonotonic curve.

Competitor strength effects

- 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, strengthening some competitors can reduce RIF for non-strengthened competitors (see *Simulation 2.3*, Figure 34). When testing this prediction, it is important to recognize that the competitive space encompasses both studied and nonstudied semantic associates of the cue. For example, in Figure 36 we showed that increasing the semantic strength of nonstudied competitors can reduce RIF for studied competitors.

RIF using external cues

- In *Simulation 5*, we explained Perfect et al. (2004) finding of null RIF given novel-associate cues (e.g., null RIF when cuing for Apple using Zinc) in terms of *contextual focusing* at practice. Specifically, we argued that participants use contextual information at practice to focus retrieval on the standard study phase. Insofar as the Zinc-Apple trace was formed *outside* of the standard study phase, focusing retrieval on the standard study phase effectively blocks pop-up (and weakening) of the Zinc-Apple trace. This view implies that manipulations that make it more difficult to contextually "block out" the Zinc-Apple trace at practice (e.g., having participants study Zinc-Apple and Fruit-Apple as part of the same list) should boost the amount of RIF elicited by Zinc (see Figure 46 and Figure 47).

Forgetting after extra study

- In *Simulation 8*, we showed that extra study can lead to forgetting of other studied items if the level of pattern overlap between targets and competitors (in cortex and in the hippocampus) is high (see Figure 55). One way to test this would be to present pictures along with sentences in the Anderson and Bell (2001) paradigm (e.g., a picture of the teacher lifting the violin) and then vary the similarity of the pictures.

Effects of partial practice vs. extra study on target recall

- In *Simulation 9*, we showed that it should be possible to observe more target strengthening after partial practice vs. extra study, if we engineer a situation where the target is *just barely strong enough* to be retrieved correctly during partial practice. We showed how it is possible to manipulate the semantic strength of the target and the specificity of the retrieval cue in order to generate optimal dynamics for strengthening. If the target is too weak (and/or the cue is too vague) to support accurate recall at practice, this diminishes strengthening. Conversely, if the target is retrieved too easily (as in the extra study condition) this also diminishes strengthening, by reducing the overall amount of competitor pop-up that occurs at practice (see Figure 56 and Figure 57).

Effects of context cue strength on episodic RIF and semantic RIF

- To reconcile the finding of RIF for novel episodic associations in Anderson and Bell

(2001) with the null RIF effect for weak semantic associates in Anderson et al. (1994), we had to posit that participants cue more strongly with context when trying to recall novel episodic associations, vs. when trying to recall studied items that are semantically related to the cue. In the *Boundary conditions* section of *Simulation 4*, we highlighted two novel implications of this view: Participants who try to retrieve *novel episodic* associates of a cue will also show RIF for studied weak semantic associates of the cue. Also, participants who try to retrieve *semantic* associates of a cue will not show RIF for episodic associates of the cue.

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 14). 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 (for preliminary results, see Newman & Norman, 2006).

Challenges for the model

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

Effects of target-competitor integration and similarity

As reviewed by Anderson (2003), several extant studies have explored how *target-competitor integration* (i.e., how strongly the target's features are linked to the competitor's features) and *target-competitor similarity* (i.e., how many features target and competitor have in common) interact with RIF. These studies have generally found that increasing target-competitor integration or similarity reduces RIF. For example, Anderson et al. (2000b) had participants study two exemplars from a category at the same time (e.g., Red-Tomato and Red-Brick), where one category exemplar (e.g., Tomato) was a target and the other was a competitor (e.g., Brick); participants were asked to either find *similarities* or *differences* between the two items. RIF (after partial practice) was observed in the "find differences" condition but not in the "find similarities" condition. More recently, Goodmon (2005) took the materials from a study that had failed to obtain RIF (Butler, Williams, & Zacks, 2001), and showed that RIF effects emerge after partial practice when the materials are re-arranged to minimize target-competitor association strength (for other examples of how target-competitor similarity/integration can reduce RIF, see, e.g., Anderson & McCulloch, 1999; Anderson & Bell, 2001; Bauml & Hartinger, 2002).

Simulating these findings is an important challenge for our model. In particular, we need to reconcile the above findings (showing that increasing similarity/integration *reduces* the amount of forgetting caused by partial practice) with the results of *Simulation 8*, which showed that — in the model —

boosting pattern similarity *increases* the amount of forgetting caused by extra study. Below, we discuss how (in terms of our modeling framework) boosting target-competitor similarity could have opposite effects in extra-study and partial-practice paradigms, boosting forgetting in the former case, and reducing forgetting in the latter case. Then, we discuss how issues with k-winners-take-all inhibition make it difficult to simulate these results in our model (as it currently stands), and we discuss ways of remedying this problem

A competitive-learning account of integration and similarity effects As with other manipulations discussed in this paper, seemingly contradictory results can be sorted out when one carefully considers how similarity/integration manipulations affect the level of excitatory support received by competitors (relative to targets) in the model. Increasing target-competitor integration (association strength) and target-competitor similarity (feature overlap) should both increase the amount of excitatory input received by competitor units when the target is active, thereby "narrowing the gap" in excitatory support between the target and the competitor. The key difference between the extra study condition and the partial practice condition is the size of the target-competitor gap, *prior* to increasing similarity/integration.

Extra study condition As discussed in *Simulation 8*, competitors do not receive enough support to pop up on extra study trials when target-competitor overlap is low. Increasing target-competitor overlap boosts excitatory support for competitors, to the point where competitors start to pop up (and show RIF).

Partial practice condition The situation is very different for partial practice. On partial practice trials, the "net input gap" between targets and competitors in the model is small enough to trigger competitor pop-up (see Figure 13), even if there is absolutely no feature overlap or integration between targets and competitors in the item layer. In this situation, boosting target-competitor similarity or integration will narrow the net input gap between target and competitor representations even more. If the competitor receives a sufficiently high level of support (relative to the target) we should observe a situation like we observed in the "weak target, strong competitor" condition of *Simulation 2.1*, where the competitor starts to pop up *before* the onset of the low inhibition phase. As discussed in *Simulation 2.1* and *Simulation 2.2*, this premature pop-up should reduce RIF. In the limiting case, if the competitor and target are receiving nearly equal levels of support (e.g., due to extremely strong target-competitor integration), one might imagine that the competitor and the target would act as a single "functional unit" — coming on together at the start of the trial, dipping down together during the high inhibition phase, and then staying on together during the low inhibition phase. In this case (where the competitor's activation dynamics match the target's activation dynamics), we might expect the competitor to show *strengthening*, not weakening, in the high-integration condition. This pattern was observed by Anderson et al. (2000b).

In summary, our learning framework predicts that increasing similarity/integration when excitatory support for the competitor is *relatively low* can *boost* forgetting by triggering pop-up of the competitor (this is what happened in *Simulation 8*). However, increasing similarity/integration when excitatory support for the competitor is *already high* can *reduce* forgetting by increasing the odds that the competitor will activate before the start of the low inhibition phase. This latter fact may help explain why Anderson et al. (2000b) and others have found less RIF with increasing target-competitor integration.

Problems with k-winners-take-all inhibition Importantly, while our learning framework can (in principle) account for reduced RIF with increased target-competitor similarity/integration, there are ways in which the behavior of the actually-implemented model diverges from the idealized account described above. We mentioned above that — with sufficiently high levels of target-competitor integration — the competitor and target should act as a single functional unit. However, it is not possible to simulate this dynamic using the k-winners-take-all (kWTA) inhibitory algorithm. As discussed earlier, the kWTA algorithm enforces a rigid limit on the number of units that can be strongly active at once, when inhibition is set to its normal ("baseline") value. In our simulations, kWTA is parameterized to allow 4 units (i.e., a single item) to be active given normal inhibition, and there is no way to adaptively expand this limit to allow the target and competitor to be active at the same time (regardless of how much mutual support there is between the target and the competitor).

The most straightforward way to remedy this problem is to replace the kWTA inhibitory algorithm with explicitly simulated inhibitory interneu-

rons. 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 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 allow the target and competitor to act as a single "functional unit" if the target and competitor representations are strongly interconnected.

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 (for a similar result, see Saunders & MacLeod, 2002). As things stand, these two studies are the only ones (that we know of) that have used delays lasting longer than a few hours to examine RIF, so it is unclear whether the "no RIF after 24 hours" reflects a general principle that applies across all RIF paradigms (not just the paradigms used in the studies cited above).

One way to account for decreased RIF after a delay is to appeal to the context-dependence of RIF: To the extent that RIF is context-dependent, and elapsed time is correlated with change in the participant's "mental context", this implies that elapsed time should reduce RIF. As discussed above, our model predicts that it should be possible to observe *some* RIF after a context change, but these effects might be small and thus hard to detect.

Another possible explanation of null RIF after a 24-hour delay relates to the effects of sleep on memory representations. Recently, Norman, Newman, and Perotte (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, if an attractor has been weakened (but still exists in the network), this process is capable of activating the damaged attractor and then fixing it (by oscillating inhibition to locate weak parts of the memory, and then strengthening these weak parts). Furthermore, Norman et al. (2005) argued that this autonomous attractor-repair process occurs during REM sleep.[49] If this theory is correct, it is possible that participants in the 24-hour delay condition of the MacLeod and Macrae (2001) and Saunders and MacLeod (2002) studies fail to show RIF because REM sleep (during the 24-hour retention interval) repaired the attractor damage that occurred during the (pre-sleep) practice session. This view implies that if we un-confound the effects of time and REM sleep, we should find that REM sleep sharply reduces RIF, but time *per se* does not differentially interact with competitor-punishment vs. target-strengthening effects.

Model improvements

Above, we described how kWTA inhibition impedes the model's ability to fully account for target-competitor similarity and integration effects, and how kWTA could be replaced by more realistic forms of inhibition. In this section, we evaluate other simplifications built into the model and discuss ways in which we can move beyond these simplifications.

Cortical network In our current model, each item has a single, unified cortical representation. However, in the actual brain, cortex represents items in a *hierarchical* fashion, with low-level perceptual features represented at the bottom of the hierarchy, and more abstract concepts represented at the top of the hierarchy; each layer of the hierarchy works

[49] For discussion of how this REM-sleep attractor-repair process can help to protect stored knowledge (so it is not catastrophically "swept away" by new learning) see Norman et al. (2005).

to extract statistical regularities in the layer(s) below it. In light of this fact, we have started to explore how the oscillating algorithm works in hierarchical networks. One advantage of this approach is that it allows us to make more principled predictions about *where in the hierarchy* competition (and RIF) should occur. For example, if competition is taking place between conceptual representations (but not perceptual representations), we might expect RIF to be observed when memory is probed using conceptual implicit memory tests but not perceptual implicit memory tests (see, e.g., Perfect, Moulin, Conway, & Perry, 2002; for more discussion of this point see Anderson, 2003).

Another important difference between hierarchical models of cortex and our current cortical network is that — in hierarchical networks — only some of the layers (at the bottom of the hierarchy) receive external input. The other layers are free to develop their own representations of input patterns. Norman et al. (2006b) describe how the oscillating-inhibition learning algorithm works in a multi-layer network (consisting of an input/output layer that is bidirectionally connected to a hidden layer). 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 et al., 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).

Hippocampal network The hippocampal model used in this paper is also highly simplified, relative to other published hippocampal models: It only consists of one layer (instead of multiple layers, corresponding to different hippocampal subregions), it restricts learning to a relatively small number of projections, and it externally enforces pattern separation (rather than having pattern separation be an emergent property of the model). These simplifications were necessary in order to keep the speed and complexity of the model within acceptable bounds. However, with the advent of faster computers, and given our improved understanding of how the model works, we can start to consider ways of bridging the gap between our simplified hippocampal model and more complex, biologically realistic models (e.g., Hasselmo et al., 2002; Norman & O'Reilly, 2003; Becker, 2005). Using a hippocampal model that maps more closely onto the actual neurobiology of the hippocampus would have several benefits: It would make it easier to use the model to address the vast empirical literature on hippocampal theta oscillations and learning (e.g., Hyman et al., 2003). It would also make it easier to relate our model to other theoretical accounts of hippocampal theta (e.g., Hasselmo's idea that theta oscillations optimize hippocampal dynamics for encoding vs. retrieval; see Norman et al., 2005 for discussion of how our theory relates to the Hasselmo et al., 2002 model).

Modeling the dynamics of top-down control

At present, the model does not include a means of simulating top-down control (via PFC). As discussed above, we believe that PFC plays a major role in shaping competitive dynamics and (through this) shaping which memories are punished and which memories are strengthened. PFC should be especially important in situations where the target is much weaker than the competitor. In these situations, PFC can ensure that the (weaker) target wins by sending extra activation to the target representation (Miller & Cohen, 2001).

The simplest way to simulate PFC involvement at retrieval is to include an additional input projection that provides support to features of the target memory; see Norman, Newman, and Detre (2006a) for some preliminary simulations of PFC contributions to RIF using this method. This method allows us to vary the degree of PFC involvement on a particular trial. However, it does not allow us simulate the fine-grained temporal dynamics of PFC involvement. To address this problem, we plan to implement a simple network architecture for conflict detection and cognitive control, as proposed by Botvinick, Braver, Barch, Carter, and Cohen (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").[50] When ACC detects conflict, this causes PFC to activate, which (in turn) serves to resolve the conflict. For example, in a recent study by Johnson and Anderson (2004), participants were given homographs like Prune with dominant noun meanings (the fruit "prune") and subordinate verb meanings ("trim"), and were asked to complete word fragments that matched the subordinate verb meaning. In this situation, 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 detailed predictions about the dynamics of PFC intervention in memory retrieval, and how these dynamics influence learning.

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. (2006b) showed that, apart from its useful psychological properties, the oscillating algorithm also has desirable functional properties: Using the hierarchical cortical network described above (i.e., with a hidden layer that is bidirectionally connected to the input/output layer) Norman et al. (2006b) found that the oscillating algorithm outperforms several other algorithms (e.g., backpropagation and Leabra) at storing and retrieving correlated input patterns. For example, when given 200 patterns to memorize (with average between-pattern feature overlap of 57%, and noisy retrieval cues), a version of the oscillating algorithm with 40 hidden units can correctly recall approximately 100 of these patterns (based on partial cues), whereas a comparably-sized Leabra network recalls fewer than 10 patterns. As discussed by Norman et al. (2006b), 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. The key point to be gleaned from this discussion is that the exact same attribute (selective weakening of close competitors) that helps the model account for RIF data also helps the model do a better job according to purely functional criteria.

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 hierarchical version of the cortical network, operationalizing familiarity in terms of the size of the "dip" in target activation during the high inhibition phase. As stimuli are presented repeatedly (making them more familiar), the dip in target activation during the high inhibition phase gets smaller. Norman et al. (2005) presented simulations showing 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 the Norman and O'Reilly (2003) cortical familiarity model (which does not oscillate inhibition, and uses a simple Hebbian learning rule to adjust weights). Future work will explore whether the oscillating-algorithm familiarity model can account for the full range of list-learning interference results that were previously addressed using the Norman and O'Reilly (2003) familiarity model (e.g., the null recognition list strength effect observed by Ratcliff et al., 1990).

[50] For a model of how ACC learns to detect conflict, see Brown and Braver (2005).

Another important future direction for the model is to simulate results from the classical paired-associate-learning literature. As mentioned above, abstract mathematical models have successfully simulated data from the AB-AC paired-associate learning paradigm (e.g., Barnes & Underwood, 1959) without positing any kind of trace weakening process (Mensink & Raaijmakers, 1988). Furthermore, there are certain facets of this data space that appear to directly contradict the predictions of unlearning models. For example, associative unlearning theory (Melton & Irwin, 1940) predicts that, in AB-AC learning paradigms, learning a new association (e.g., soldier-army) should directly cause forgetting of previously learned associations involving that cue (e.g., soldier-gun). However, several analyses have found that — across stimuli — learning of the second association is statistically independent from forgetting of the first association (for discussion of this point, see Martin, 1971; Greeno, James, DaPolito, & Polson, 1978; Mensink & Raaijmakers, 1988; Chappell & Humphreys, 1994; Kahana, 2000). It will be very informative to see how well our model can account for this "AB-AC independence" finding, and others like it.[51]

Finally, we also plan to use the model to address other psychological phenomena (outside of the domain of declarative memory) that may involve competitor-weakening, including 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 full 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 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 (e.g., 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.

[51] Prior simulation results from Mensink and Raaijmakers (1988) and others suggest that *gradual drift* in contextual representations is a major cause of forgetting in classical paired-associate-learning paradigms. As such, properly simulating results from these paradigms may require us to replace our "static tag" contextual representations with contextual representations that evolve over time. For discussion of mechanisms of contextual drift see Howard and Kahana (2002), and for discussion of how these mechanisms could be implemented in neural network models see Norman et al. (in press).

Acknowledgments

This work was supported by NIH grant RO1 MH069456, awarded to KAN. We thank Michael Anderson, Tim Curran, Michael Kahana, Lynn Nadel, Joel Quamme, Per Sederberg, and two anonymous reviewers for their very insightful comments on an earlier draft of this manuscript. KAN would also like to thank Randy O'Reilly for his mentorship in all things relating to neural networks.

References

Ackley, D. H., Hinton, G. E., & Sejnowski, T. J. (1985). A learning algorithm for Boltzmann machines. *Cognitive Science*, *9*, 147–169.

Anderson, J. R. (1983). *The architecture of cognition.* Cambridge, MA: Harvard University Press.

Anderson, M. C. (2003). Rethinking interference theory: Executive control and the mechanisms of forgetting. *Journal of Memory and Language*, *49*, 415–445.

Anderson, M. C., & Bell, T. (2001). Forgetting our facts: the role of inhibitory processes in the loss of propositional knowledge. *Journal of Experimental Psychology: General*, *130(3)*, 544–570.

Anderson, M. C., Bjork, E. L., & Bjork, R. A. (2000a). Retrieval-induced forgetting: Evidence for a recall-specific mechanism. *Memory & Cognition*, *28*, 522.

Anderson, M. C., & Bjork, R. A. (1994). Mechanisms of inhibition in long-term memory: A new taxonomy. In D. Dagenbach, & T. H. Carr (Eds.), *Inhibitory processes in attention, memory, and language* (pp. 265–325). San Diego: Academic Press.

Anderson, M. C., Bjork, R. A., & Bjork, E. L. (1994). Remembering can cause forgetting: Retrieval dynamics in long-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *5*, 1063–1087.

Anderson, M. C., Green, C., & McCulloch, K. C. (2000b). Similarity and inhibition in long-term memory: Evidence for a two-factor theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 1141–1159.

Anderson, M. C., & McCulloch, K. C. (1999). Integration as a general boundary condition on retrieval-induced forgetting. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 608–629.

Anderson, M. C., & Shivde, G. S. (in preparation). Strength is not enough: Evidence against a blocking theory of retrieval-induced forgetting.

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

Barnes, J. M., & Underwood, B. J. (1959). Fate of first-list associations in transfer theory. *Journal of Experimental Psychology*, *58*, 97–105.

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

Bauml, K. (1997). The list-strength effect: Strength-dependent competition or suppression? *Psychonomic Bulletin and Review*, *4*, 260–264.

Bauml, K. (1998). Strong items get suppressed, weak items do not: The role of item strength in output interference. *Psychonomic Bulletin and Review*, *5(3)*, 459–463.

Bauml, K. H. (2002). Semantic generation can cause episodic forgetting. *Psychological Science*, *13*(4), 356–60.

Bauml, K.-H., & Hartinger, A. (2002). On the role of item similarity in retrieval-induced forgetting. *Memory*, *10*(3), 215–224.

Becker, S. (2005). A computational principle for hippocampal learning and neurogenesis. *Hippocampus*, *15*(6), 722–38.

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

Blaxton, T. A., & Neely, J. H. (1983). Inhibition from semantically related primes: Evidence of a category-specific retrieval inhibition. *Memory and Cognition*, *11*, 500–510.

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review*, *108*, 624–652.

Brown, J. W., & Braver, T. S. (2005). Learned predictions of error likelihood in the anterior cingulate cortex. *Science*, *307*(5712), 1118–21.

Butler, K. M., Williams, C. C., & Zacks, R. T. (2001). A limit on retrieval-induced forgetting. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 1314–1319.

Buzsaki, G. (2002). Theta oscillations in the hippocampus. *Neuron*, *33*, 325–340.

Camp, G., Pecher, D., & Schmidt, H. G. (2005). Retrieval-induced forgetting in implicit memory tests: the role of test awareness. *Psychonomic Bulletin and Review*, *12*(3), 490–494.

Carter, K. L. (2004). *Investigating semantic inhibition using a modified independent cue task.* PhD thesis, University of Kansas, Lawrence, KS.

Chappell, M., & Humphreys, M. S. (1994). An auto-associative neural network for sparse representations: Analysis and application to models of recognition and cued recall. *Psychological Review*, *101*, 103–128.

Ciranni, M. A., & Shimamura, A. P. (1999). Retrieval-induced forgetting in episodic memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 1403.

Cohen, J. D., & Servan-Schreiber, D. (1992). Context, cortex, and dopamine: A connectionist approach to behavior and biology in schizophrenia. *Psychological Review*, *99*, 45–77.

Curran, T. (2000). Brain potentials of recollection and familiarity. *Memory and Cognition*, *28*, 923.

DeSchepper, B., & Treisman, A. (1996). Visual memory for novel shapes: implicit coding without attention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 27–47.

Douglas, R. J., Koch, C., Mahowald, M., Martin, K. A. C., & Suarez, H. H. (1995). Recurrent excitation in neocortical circuits. *Science*, *269*, 981.

Douglas, R. J., & Martin, K. A. C. (1998). Neocortex. In G. M. Shepherd (Ed.), *The synaptic organization of the brain* (Chap. 12, pp. 459–509). Oxford: Oxford University Press.

Ekstrom, A. D., Caplan, J. B., Ho, E., Shattuck, K., Fried, I., & Kahana, M. J. (2005). Human hippocampal theta activity during virtual navigation. *Hippocampus*, *15*, 881–889.

Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, *1*, 1–47.

Fletcher, P. C., & Henson, R. N. (2001). Frontal lobes and human memory: insights from functional neuroimaging. *Brain*, *124*(Pt 5), 849–81.

Fox, E. (1995). Negative priming from ignored distractors in visual selection. *Psychonomic Bulletin and Review*, *2*, 145–173.

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review*, *91*, 1–67.

Goodmon, L. (2005). *The influence of pre-existing memories on retrieval-induced forgetting*. PhD thesis, University of South Florida, Tampa, FL.

Gotts, S. J., & Plaut, D. C. (2005). Neural mechanisms underlying positive and negative repetition priming. *Poster presented at the Annual Meeting of the Cognitive Neuroscience Society*.

Graf, P., Squire, L. R., & Mandler, G. (1984). The information that amnesic patients do not forget. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*, 164–178.

Greeno, J. G., James, C. T., DaPolito, F. J., & Polson, P. G. (1978). *Associative learning: A cognitive analysis*. Englewood Cliffs, NJ: Prentice-Hall.

Grossberg, S. (1976). Adaptive pattern classification and universal recoding I: Parallel development and coding of neural feature detectors. *Biological Cybernetics*, *23*, 121–134.

Hasselmo, M. E., Bodelon, C., & Wyble, B. P. (2002). A proposed function for hippocampal theta rhythm: Separate phases of encoding and retrieval enhance reversal of prior learning. *Neural Computation*, *14*, 793–818.

Hinton, G. E. (1989). Deterministic Boltzmann learning performs steepest descent in weight-space. *Neural Computation*, *1*, 143–150.

Hinton, G. E., & Sejnowski, T. J. (1986). Learning and relearning in Boltzmann machines. In D. E. Rumelhart, J. L. McClelland, & PDP Research Group (Eds.), *Parallel distributed processing. Volume 1: Foundations* (Chap. 7, pp. 282–317). Cambridge, MA: MIT Press.

Hintzman, D. L., & Curran, T. (1994). Retrieval dynamics of recognition and frequency judgments: Evidence for separate processes of familiarity and recall. *Journal of Memory and Language*, *33*, 1–18.

Hintzman, D. L., Curran, T., & Oppy, B. (1992). Effects of similiarity and repetition on memory: Registration without learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*, 667–680.

Hockley, W. E. (1999). Familiarity and recollection in item and associative recognition. *Memory and Cognition*, *27*, 657.

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*, 6470–6477.

Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology*, *46*, 269–299.

Huerta, P. T., & Lisman, J. E. (1996). Synaptic plasticity during the cholinergic theta-frequency oscillation in vitro. *Hippocampus*, *49*, 58–61.

Hyman, J. M., Wyble, B. P., Goyal, V., Rossi, C. A., & Hasselmo, M. E. (2003). Stimulation in hippocampal region CA1 in behaving rats yields long-term potentiation when delivered to the peak of theta and long-term depression when delivered to the trough. *Journal of Neuroscience*, *23*, 11725–11731.

Johnson, S. K., & Anderson, M. C. (2004). The role of inhibitory control in forgetting semantic knowledge. *Psychological Science*, *15*, 448–453.

Kahana, M. J. (2000). Contingency analyses of memory. In E. Tulving, & F. Craik (Eds.), *The oxford handbook of memory* (pp. 59–72). New York: Oxford University Press.

Kahana, M. J., Rizzuto, D. S., & Schneider, A. R. (2005). Theoretical correlations and measured correlations: relating recognition and recall in four distributed memory models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(5), 933–953.

Kahana, M. J., Seelig, D., & Madsen, J. R. (2001). Theta returns. *Current Opinion in Neurobiology*, *11*, 739–44.

Kimball, D. R., & Bjork, R. A. (2002). Influences of intentional and unintentional forgetting on false memories. *Journal of Experimental Psychology: General*, *131*(1), 116–130.

Klimesch, W. (1999). EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis. *Brain Res Brain Res Rev*, *29*(2-3), 169–195.

Klimesch, W., Doppelmayr, M., Russegger, H., & Pachinger, T. (1996). Theta band power in the human scalp EEG and the encoding of new information. *Neuroreport*, *7*(7), 1235–40.

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

MacLeod, C., & Macrae, N. (2001). Gone but not forgotten: The transient nature of retrieval-induced forgetting. *Psychological Science*, *12*, 148–152.

Malenka, R. C., & Bear, M. F. (2004). LTP and LTD: an embarrassment of riches. *Neuron*, *44*, 5–21.

Marr, D. (1971). Simple memory: A theory for archicortex. *Philosophical Transactions of the Royal Society (London) B*, *262*, 23–81.

Martin, E. (1971). Verbal learning theory and independent retrieval phenomena. *Psychological Review*, *78*, 314–332.

Mayr, U., & Keele, S. (2000). Changing internal constraints on action: the role of backward inhibition. *Journal of Experimental Psychology: General*, *1*, 4–26.

McClelland, J. L., & Chappell, M. (1998). Familiarity breeds differentiation: a subjective-likelihood approach to the effects of experience in recognition memory. *Psychological Review*, *105*, 724.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, *102*, 419–457.

McGeoch, J. A. (1936). Studies in retroactive inhibition: VII: Retroactive inhibition as a function of the length and frequency of presentation of the interpolated lists. *Journal of Experimental Psychology*, *19*, 674–693.

Mehta, M. R., Lee, A. K., & Wilson, M. A. (2002). Role of experience and oscillations in transforming a rate code into a temporal code. *Nature*, *416*, 741–745.

Melton, A. W., & Irwin, J. M. (1940). The influence of degree of interpolated learning on retroactive inhibition and the overt transfer of specific responses. *American Journal of Psychology*, *3*, 173–203.

Mensink, G., & Raaijmakers, J. G. (1988). A model for interference and forgetting. *Psychological Review*, *95*, 434–455.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, *24*, 167–202.

Minai, A. A., & Levy, W. B. (1994). Setting the activity level in sparse random networks. *Neural Computation*, *6*, 85–99.

Movellan, J. R. (1990). Contrastive Hebbian learning in the continuous Hopfield model. In D. S. Tourtezky, G. E. Hinton, & T. J. Sejnowski (Eds.), *Proceedings of the 1989 Connectionist Models Summer School* (pp. 10–17).

Newman, E. L., & Norman, K. A. (2006). Tracking the sub-trial dynamics of cognitive competition. *Society for Neuroscience Abstracts*.

Norman, K. A. (2002). Differential effects of list strength on recollection and familiarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28(6)*, 1083–1094.

Norman, K. A., Detre, G. J., & Polyn, S. M. (in press). Computational models of episodic memory. In R. Sun (Ed.), *The Cambridge handbook of computational cognitive modeling*. Cambridge University Press.

Norman, K. A., Newman, E. L., & Detre, G. J. (2006a). *A neural network model of retrieval-induced forgetting* (Technical Report 06-1). Princeton, NJ: Princeton University, Center for the Study of Brain, Mind, and Behavior.

Norman, K. A., Newman, E. L., Detre, G. J., & Polyn, S. M. (2006b). How inhibitory oscillations can train neural networks and punish competitors. *Neural Computation*, *18*(7), 1577–1610.

Norman, K. A., Newman, E. L., & Perotte, A. J. (2005). Methods for reducing interference in the complementary learning systems model: Oscillating inhibition and autonomous memory rehearsal. *Neural Networks*, *18*, 1212–1228.

Norman, K. A., & O'Reilly, R. C. (2003). Modeling hippocampal and neocortical contributions to recognition memory: A complementary-learning-systems approach. *Psychological Review*, *104*, 611–646.

O'Keefe, J., & Recce, M. L. (1993). Phase relationship between hippocampal place units and the EEG theta rhythm. *Hippocampus*, *3*, 317–30.

Oram, M. W., & MacLeod, M. D. (2001, August). Remembering to forget: Modeling inhibitory and competitive mechanisms in human memory. *Cognitive Science Society Annual Meeting*.

O'Reilly, R. C., & McClelland, J. L. (1994). Hippocampal conjunctive encoding, storage, and recall: Avoiding a tradeoff. *Hippocampus*, *4*(6), 661–682.

O'Reilly, R. C., & Munakata, Y. (2000). *Computational explorations in cognitive neuroscience: Understanding the mind by simulating the brain*. Cambridge, MA: MIT Press.

Osipova, D., Takashima, A., Oostenveld, R., Fernandez, G., Maris, E., & Jensen, O. (2006). Theta and gamma oscillations predict encoding and retrieval of declarative memory. *Journal of Neuroscience*, *26*(28), 7523–7531.

Perfect, T., Stark, L., Tree, J., Moulin, C., Ahmed, L., & Hutter, R. (2004). Transfer appropriate forgetting : The cue-dependent nature of retrieval-induced forgetting. *Journal of Memory and Language*, *51*, 399–417.

Perfect, T. J., Moulin, C. J. A., Conway, M. A., & Perry, E. (2002). Assessing the inhibitory account of retrieval induced forgetting with implicit memory tests. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 1111–1119.

Raaijmakers, J. G. W. (2005). Modeling implicit and explicit memory. In C. Izawa, & N. Ohta (Eds.), *Human learning and memory: Advances in theory and application* (pp. 85–105). Mahwah, NJ: Erlbaum.

Raaijmakers, J. G. W., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review*, *88*, 93–134.

Racsmany, M., & Conway, M. A. (2006). Episodic inhibition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(1), 44–57.

Ratcliff, R., Clark, S., & Shiffrin, R. M. (1990). The list strength effect: I. Data and discussion. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*, 163–178.

Reder, L. M., Nhouyvanisvong, A., Schunn, C. D., Ayers, M. S., Angstadt, P., & Hiraki, K. A. (2000). A mechanistic account of the mirror effect for word frequency: A computational model of remember-know judgments in a continuous recognition paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 294–320.

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 803–814.

Roediger, H. L., McDermott, K. B., & Robinson, K. J. (1998). The role of associative processes in creating false memories. In M. A. Conway, S. E. Gathercole, & C. Cornoldi (Eds.), *Theories of memory II* (pp. 187–245). Hove, Sussex: Psychology Press.

Rundus, D. (1973). Negative effects of using list items as retrieval cues. *Journal of Verbal Learning and Verbal Behavior*, *12*, 43–50.

Saunders, J., & MacLeod, M. D. (2002). New evidence on the suggestibility of memory: The role of retrieval-induced forgetting in eyewitness misinformation effects. *Journal of Experimental Psychology: Applied*, *8*, 127–142.

Saunders, J., & MacLeod, M. D. (2006). Can inhibition resolve retrieval competition through the control of spreading activation? *Memory and Cognition*, *34*(2), 307–322.

Schacter, D. L. (1987). Memory, amnesia, and frontal lobe dysfunction. *Psychobiology*, *15*, 21–36.

Seager, M. A., Johnson, L. D., Chabot, E. S., Asaka, Y., & Berry, S. D. (2002). Oscillatory brain states and learning: Impact of hippocampal theta-contingent training. *Proceedings of the National Academy of Sciences*, *99*, 1616–1620.

Sederberg, P., Kahana, M. J., Howard, M. W., Donner, E. J., & Madsen, J. R. (2003). Theta and gamma oscillations during encoding predict subsequent recall. *Journal of Neuroscience*, *23*, 10809–10814.

Shiffrin, R. M., Ratcliff, R., & Clark, S. (1990). The list strength effect: II. Theoretical mechanisms. *Journal of Experimental Psychology: Learning Memory and Cognition*, *16*, 179–195.

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

Shivde, G., & Anderson, M. C. (2001). The role of inhibition in meaning selection: Insights from retrieval-induced forgetting. In D. S. Gorfein (Ed.), *On the consequences of meaning selection: Perspectives on resolving lexical ambiguity* (pp. 175–190). Washington, D. C.: American Psychological Association.

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory*, *6*, 592–604.

Smith, S. M. (1988). Environmental context-dependent memory. In G. M. Davies, & D. M. Thomson (Eds.), *Memory in context: Context in memory.* (pp. 13–34). Oxford, England: John Wiley & Sons.

Starns, J. J., & Hicks, J. L. (2004). Episodic generation can cause semantic forgetting: retrieval-induced forgetting of false memories. *Memory and Cognition*, *32*(4), 602–609.

Szentágothai, J. (1978). The neuron network of the cerebral cortex: A functional interpretation. *Proceedings of the Royal Society (London) B*, *201*, 219–248.

Toth, K., Freund, T. F., & Miles, R. (1997). Disinhibition of rat hippocampal pyramidal cells by GABAergic afferents from the septum. *Journal of Physiology*, *500*, 463–474.

Tsukimoto, T., & Kawaguchi, J. (2001). Retrieval-induced forgetting: Is the baseline in the retrieval-practice paradigm true? *Poster presented at the International Conference on Memory, Valencia, Spain*.

Underwood, B. J. (1965). False recognition produced by implicit verbal responses. *Journal of Experimental Psychology*, *70*, 122–129.

Veling, H., & van Knippenberg, A. (2004). Remembering can cause inhibition: Retrieval-induced inhibition as a cue independent process. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*, 315–318.

Verde, M. F., & Rotello, C. M. (2004). Strong memories obscure weak memories in associative recognition. *Psychonomic Bulletin and Review*, *11*(6), 1062–1066.

Wang, X. J. (2002). Probabilistic decision making by slow reverberation in cortical circuits. *Neuron*, *36*(5), 955–68.

Williams, C., & Zacks, R. (2001). Is retrieval-induced forgetting an inhibitory process? *Journal of Psychology*, *114*, 329–354.

Yamaguchi, Y., Aota, Y., McNaughton, B. L., & Lipa, P. (2002). Bimodality of theta phase precession in hippocampal place cells in freely running rats. *Journal of Neurophysiology*, *87*, 2629–2642.

Yonelinas, A. P. (1997). Recognition memory ROCs for item and associative information: The contribution of recollection and familiarity. *Memory and Cognition*, *25*, 747–763.

Appendix A: Algorithm details

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. (2006b).

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). Parts of this Appendix were adapted from Appendix A of Norman and O'Reilly (2003).

Pseudocode

The pseudocode for the algorithm that we used is given here, showing how the pieces of the algorithm (described in more detail in subsequent sections) fit together. Parts of the learning algorithm that differ from the standard Leabra procedure are marked in boldface.

Outer loop: Iterate over events (trials) within an epoch. For each event, settle over time steps of updating:

1. At start of settling, for all units:
   (a) Initialize all state variables (activation, V_m, etc).
   (b) Apply external patterns.
2. During each time step of settling:
   (a) Compute excitatory net input (g_e, eq 7).
   (b) Compute kWTA inhibition g_i^{kWTA} for each layer, based on g_i^{Θ} (eq 10):
      i. Sort the n units into two groups based on g_i^{Θ}: top k and remaining $k+1$ to n.
      ii. Set inhibitory conductance g_i^{kWTA} between g_k^{Θ} and g_{k+1}^{Θ} (eq 9).
   (c) **Compute overall inhibition by combining kWTA inhibition with an oscillating component (eq 11 and eq 13).**
   (d) Compute point-neuron activation combining excitatory input and inhibition (eq 5).
3. Update the weights (based on linear current weight values), for all connections:
   (a) **Compute weight changes according to the oscillating algorithm (eq 4).**
   (b) Increment the weights and apply contrast-enhancement (eq 13).

Point neuron activation function

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 described in the main text (and detailed below) inhibition was controlled by means of a k-winners-take-all (*kWTA*) inhibitory mechanism (O'Reilly & Munakata, 2000; Minai & Levy, 1994), which was modified by an oscillating-inhibition component.

To simulate excitatory neurons, Leabra uses a *point neuron* activation function that models the electrophysiological properties of real neurons, while simplifying their geometry to a single point.

The membrane potential V_m is updated as a function of ionic conductances g with reversal (driving) potentials E as follows:

$$\frac{dV_m(t)}{dt} = \tau \sum_c g_c(t)\overline{g_c}(E_c - V_m(t)) \qquad (5)$$

with 3 channels (c) corresponding to: e excitatory input; l leak current; and i inhibitory input. Following electrophysiological convention, the overall conductance is decomposed into a time-varying component $g_c(t)$ computed as a function of the dynamic state of the network, and a constant $\overline{g_c}$ that controls the relative influence of the different conductances. The equilibrium potential can be written in a simplified form by setting the excitatory driving potential (E_e) to 1 and the leak and inhibitory driving potentials (E_l and E_i) of 0:

$$V_m^{\infty} = \frac{g_e\overline{g_e}}{g_e\overline{g_e} + g_l\overline{g_l} + g_i\overline{g_i}} \qquad (6)$$

which shows that the neuron is computing a balance between excitation and the opposing forces of leak and inhibition. This equilibrium form of the equation can be understood in terms of a Bayesian decision making framework (O'Reilly & Munakata, 2000).

The excitatory net input/conductance $g_e(t)$ is computed as a function of sending activations times the weight values. This value is computed separately for each projection k coming into a unit (where a "projection" is the set of connections coming from a particular layer):

$$g_{e_k}(t) = \frac{1}{\alpha_k} \frac{r_k}{\sum_p r_p} \langle x_i w_{ij} \rangle_k \quad (7)$$

In the above equation, $\frac{1}{\alpha_k}$ is a normalizing term based on the expected activity level of the sending projection, and r_k is a projection scaling factor that determines the influence of this particular projection, relative to all of the other projections. We discuss these projection scaling factors and their significance in the *Projection scaling parameters* section below. The overall excitatory net input value for a unit $g_e(t)$ is computed by summing together all of the projection-specific $g_{e_k}(t)$ terms.

Cue-related inputs (i.e., inputs from the stimulus pattern that are directly applied to the network) are factored into the computation of $g_e(t)$ just like any other projection. These inputs are applied starting on the first time step of the trial and stay on (at a constant value) throughout the trial.

The inhibitory conductance is computed by combining the level of inhibition computed by kWTA with an oscillating component, as described in the next two sections. Leak is a constant.

Activation communicated to other cells (y_j) is a thresholded (Θ) sigmoidal function of the membrane potential with gain parameter γ:

$$y_j(t) = \frac{1}{\left(1 + \frac{1}{\gamma [V_m(t) - \Theta]_+}\right)} \quad (8)$$

where $[x]_+$ is a threshold function that returns 0 if $x < 0$ and x if $x > 0$. This sharply-thresholded function is convolved with a Gaussian noise kernel ($\sigma = .005$), which reflects the intrinsic processing noise of biological neurons.

k-Winners-Take-All inhibition

Leabra uses a kWTA function to achieve sparse distributed representations (c.f., Minai & Levy, 1994). kWTA is applied separate to each layer. A uniform level of inhibitory current for all units in the layer is computed as follows:

$$g_i^{kWTA}(t) = g_{k+1}^{\Theta} + q(g_k^{\Theta} - g_{k+1}^{\Theta}) \quad (9)$$

where $0 < q < 1$ is a parameter for setting the inhibition between the upper bound of g_k^{Θ} and the lower bound of g_{k+1}^{Θ}. These boundary inhibition values are computed as a function of the level of inhibition necessary to keep a unit right at threshold:

$$g_i^{\Theta} = \frac{g_e^* \bar{g}_e (E_e - \Theta) + g_l \bar{g}_l (E_l - \Theta)}{\Theta - E_i} \quad (10)$$

where g_e^* is the excitatory net input.

In the basic version of the kWTA function used here, g_k^{Θ} and g_{k+1}^{Θ} are set to the threshold inhibition value for the k^{th} and $k+1^{st}$ most excited units, respectively. Thus, the inhibition is placed exactly to allow k units to be above threshold, and the remainder below threshold. In our simulations, the q parameter is set to .325, allowing the k^{th} unit to be sufficiently above the inhibitory threshold. We should emphasize that, when membrane potential is at threshold, unit activation in the model = .25. As such, the kWTA algorithm places a firm upper bound on the number of units showing activation > .25, but it does not set an upper bound on the number of weakly active units (i.e., units showing activation between 0 and .25).

The k parameter in cortex was set to match the number of active units per layer in the input patterns (k = 4) and the k parameter in hippocampus was set to match the number of active units in the pretrained "conjunctive representations" (k = 4 also).

Inhibitory oscillation

The overall inhibitory current $g_i(t)$ is computed by combining the level of inhibition computed by kWTA $g_i^{kWTA}(t)$ with an oscillating inhibitory component $g_i^O(t)$:

$$g_i(t) = g_i^{kWTA}(t) + g_i^O(t) \quad (11)$$

The oscillating inhibitory current, $g_i^O(t)$, is set to zero for the initial part of the trial, in order to give the network time to settle. A parameter O_{onset} determines the number of time steps to wait before starting the inhibitory oscillation, such that if $t <= O_{onset}$, then $g_i^O(t) = 0$, and if $t > O_{onset}$ then $g_i^O(t)$ is set according to the following equation:

$$g_i^O(t) = \frac{O_{max} - O_{min}}{2} \sin\left(\frac{2\pi}{O_T} t + \frac{2\pi}{360} O_\theta\right) + \frac{O_{max} + O_{min}}{2} \quad (12)$$

In the above equation, O_T, O_θ, O_{max}, and O_{min}, are the period (in time steps), phase offset (in de-

Layer	O_{max}	O_{min}	O_θ	O_T	O_{onset}
Hippocampus	2.1	-2.7	-200	80	47
Associate/Item	1.8	-1.2	-180	80	39

Table 2: Parameters defining the hippocampal and cortical inhibitory oscillations.

grees), maximum magnitude, and minimum magnitude of the oscillating inhibitory current respectively.

We used different parameters for the hippocampal inhibitory oscillation and the cortical (i.e., associate-layer and item-layer) inhibitory oscillation. O_{max}, O_{min}, and O_θ for hippocampus and cortex were iteratively adjusted (by hand) to maximize qualitative fit to existing RIF data. These parameters are listed in Table 2, and the oscillations are plotted in Figure 58. Note that the hippocampal and cortical oscillations have the same period but the cortical oscillation is slightly offset in phase relative to the hippocampal oscillation (it starts earlier and peaks earlier).

The total length of each trial was 127 time steps. Factoring in the delay in the start of the oscillation, and the 80-time-step period of the oscillation, 127 item steps is enough time for inhibition to be oscillated once from its normal value up to the high inhibition value, then down to the low inhibition value, then back to normal.

Weight adjustment

At each time step (starting at the onset of the hippocampal inhibitory oscillation) weight updates were calculated using Equation 4:

$$dW_{ij} = lrate\,(X_i(t+1)Y_j(t+1) - X_i(t)Y_j(t))$$

where $lrate$ takes on a positive value (ϵ) when the inhibitory oscillation is moving toward its midpoint value, and $lrate$ takes on a negative value ($-\epsilon$) when the inhibitory oscillation is moving away from its midpoint value. Figure 58 illustrates these $lrate$ changes.[52] The ϵ learning rate parameter was set to .05 for connections within the cortical network (i.e., item-item, item-associate, associate-item, and associate-associate); ϵ was set to 2.0 for connections between the cortical network and the hippocampal network.

[52] Note that $lrate$ changes are aligned with the peak and trough of the hippocampal inhibitory oscillation instead of the cortical inhibitory oscillation. We experimented with several different ways of aligning $lrate$ changes, and this was the configuration that worked best.

From	To	Scale
Item	Hippo	2.00
Assoc	Hippo	0.75
Hippo	Hippo	1.50
Context	Hippo	variable
Hippo	Item	0.50
Assoc	Item	0.66
Item	Item	1.25
Hippo	Assoc	0.50
Item	Assoc	0.66
Assoc	Assoc	1.25
Hippo	Context	1.00

Table 3: Projection scaling parameters for the model. These scaling factors determine the relative influence of the different projections coming into a layer.

Note that, while weight updates were calculated at each time step during the trial, these weight updates were not applied until the end of the trial.[53]

Weight contrast enhancement

Leabra includes a weight contrast enhancement function that magnifies the stronger weights and shrinks the smaller ones in a parametric, continuous fashion. This contrast enhancement is achieved by passing the linear weight values computed by the learning algorithm through a sigmoidal nonlinearity of the following form:

$$\hat{w}_{ij} = \frac{1}{1 + \left(\theta \frac{w_{ij}}{1 - w_{ij}}\right)^{-\gamma}} \tag{13}$$

where $\hat{w}_{ij}$ is the contrast-enhanced weight value, and the sigmoidal function is parameterized by an offset θ and a gain γ (standard defaults of 1.25 and 6, respectively, used here).

Projection scaling parameters

Table 3 lists the scaling parameters that determine the relative influence of different projections within the model (see Equation 7 for a precise description of how these projection scaling parameters influence excitatory net input values). Although the complexity of the model makes it impossible to exhaustively search the space of scaling-parameter settings, we did manage to search through a very wide range of scaling parameter configurations before settling on this particular set of parameters. The most important aspects of this particular parameter

[53] Another difference between our algorithm and the standard implementation of Leabra is that our algorithm does not include adjustable bias weights, whereas the standard version of Leabra does include these weights.

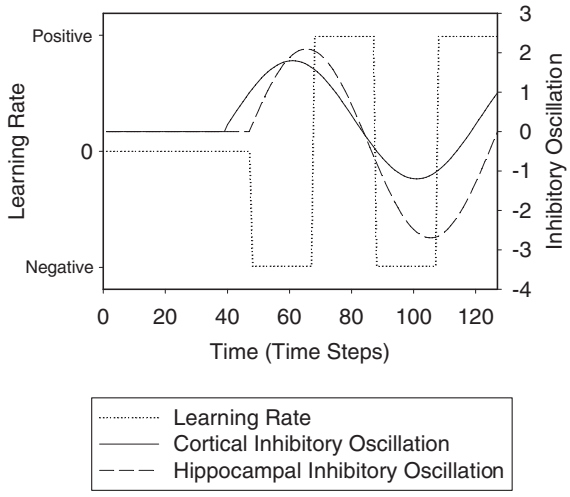

Figure 58: Illustration of how inhibition was oscillated on each trial. At each time step, the hippocampal "inhibitory oscillation" value was added to the inhibition value computed by the kWTA algorithm for the hippocampal layer. Likewise, the cortical "inhibitory oscillation" value was added to the inhibition values computed by the kWTA algorithm for the associate and item layers. The graph also shows how the sign of the learning rate was varied over the course of the inhibitory oscillation.

set, with regard to generating the dynamics outlined in the main part of the paper, were our use of high projection scaling values for recurrent projections (in both hippocampus and cortex), and our use of a high scaling value for the item-to-hippocampus projection.

With regard to recurrent projections: Using a high projection scaling value for recurrents helps to ensure well-delineated pop-up of competitors during the low-inhibition phase — a limited number of units pop up strongly, and most units do not pop up at all. When a lower projection scaling value is used for recurrents, competitor pop-up is much more diffuse (i.e., we tend to observe weak pop-up of a large number of units). In the limiting case, if the recurrents are too weak, lowering inhibition causes all of the units in the layer to start to activate; this diffuse wave of activation can trigger a seizure in the network.

With regard to item-to-hippocampus projection: Using a large scaling value on this projection (relative to the associate-to-hippocampus projection) is important for getting robust pop-up of hippocampal traces corresponding to independent cues. For example, consider what occurs in *Simulation 1.2*: In this simulation, the competitor item (2) is paired with two associates (A-2 and C-2) at study. When the model is cued with a partial version of the target (A-1) at practice, item 2 pops up as a semantic competitor. Using a strong item-to-hippocampus projection scaling factor ensures that semantic pop-up of item 2 will trigger pop-up of *all* of the hippocampal traces from the study phase that contain item 2 (i.e., both A-2 and C-2). Without this strong item-to-hippocampus scaling factor, the hippocampal representation of A-2 pops up (because it receives support from both the associate layer and the item layer at practice) but the hippocampal representation of C-2 does not.

Other parameters

All of the parameters (governing underlying model dynamics) 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), q (the parameter in Equation 9 that determines whether kWTA places the inhibitory threshold relatively close to the target units, or relatively close to competing units), and τ (the time constant parameter in Equation 5 that governs updating of the membrane potential). stm_gain was set to 0.6, q was set to 0.325, and τ was set to .15.

Appendix B: Details of semantic pretraining

This appendix contains pseudocode describing how weights in the cortical and network were pretrained (for each simulated participant) prior to the start of the simulated RIF experiment. The goal of this process was to implant a set of associate-item pairings into the cortical network (to simulate pre-experimental experience with the stimuli used in the RIF experiment).

1. Pretraining representations in the associate layer
    (a) Initialize all associate-layer recurrent connections by setting them to .5.
    (b) For each associate-layer pattern that is used in the simulation, set weights between co-active units in the associate layer to .95.
2. Pretraining representations in the item layer
    (a) Initialize all item-layer recurrent connections by setting them to .5.
    (b) For each item-layer pattern that is used in the simulation (e.g., Apple):
        i. Sample a semantic strength value for that item from a uniform distribution with mean μ and half-range σ. These μ and σ parameters can vary across simulations, and μ can also vary across conditions within a simulation (e.g., in *Simulation 2*).
        ii. Set weights between co-active units in the item layer to that item's semantic strength value.
3. Pretraining associate-item and item-associate connections
    (a) Initialize all item-associate and associate-item connections by setting them to .5.
    (b) For each associate-item pairing in the pretraining set, set weights between co-active pairs of item-layer and associate-layer units (i.e., pairs comprised of one active item unit and one active associate-layer unit) to the item's semantic strength value.