Retrieval-Induced Forgetting in a Multiple-Trace Memory Model

Collin Green (cgreen@arc.nasa.gov)
Human Factors Research & Technology, NASA Ames Research Center, Moffett Field, CA 94035

Aniket Kittur (nkittur@ucla.edu)
Department of Psychology, University of California, Los Angeles, CA 90095

Abstract

Behavioral studies (e.g., Anderson, Bjork, & Bjork, 1994) suggest that competition between relevant and irrelevant memory information is resolved in part through the inhibition of competing irrelevant information, an effect dubbed Retrieval-Induced Forgetting (RIF). Green and Kittur (2004) outlined MNEM, a model of human memory that exhibits many retrieval dynamics observed in empirical studies (e.g., spacing and practice effects, forgetting over time, spontaneous recovery, serial position effects), but did not account for RIF. In this paper, we describe a modified version of MNEM that incorporates an inhibitory mechanism and simulates some basic RIF effects. Prior work that implicates a feature-specific inhibitory mechanism provides a context for the model and simulations.

Keywords: memory, modeling, similarity, retrieval, inhibition, forgetting

Introduction

This article describes a computational memory model that incorporates an inhibitory mechanism. We begin by describing retrieval-induced forgetting (RIF) and discussing one proposed form of an inhibitory mechanism (feature-based inhibition) that is suggested by behavioral data. We then describe our model, simulation results consistent with empirical findings, and our ongoing work on this project.

Retrieval-Induced Forgetting

Many details of an average day are only subtly different than details of the day before. For instance, you may park your car in the same lot every day. The location of your car within that lot probably varies from day to day. With many memories linked to your car's location in the lot, how are you able to recall the current spot? To retrieve today's parking spot, your memory system must discriminate that target memory from many related, competing memories. Research has suggested that such discrimination is facilitated by the inhibition of competing memories, and that such inhibition has lasting effects.

Anderson, Bjork, and Bjork (1994) demonstrated that retrieving a memory results in reduced subsequent recall for related memories. Their participants were required to study a set of category-exemplar pairs (e.g., *Fruit-Apple*, *Fruit-Orange*, *Tool-Hammer*, *Tool-Wrench*). The study phase provided equal encoding for all items. During a retrieval practice phase, half of the exemplars in half of the categories were practiced in a category-and-stem-cued recall task (e.g., *Fruit-Or____* cued *Orange*). Thus, there were practiced items in practiced categories (*target items*), unpracticed items in practiced categories (*competitor items*), and unpracticed items in unpracticed categories (*unpracticed items*). After a delay, there was a category-cued recall test for all items.

Results indicated that target items were recalled better than unpracticed items. Competitor items were recalled less often than unpracticed items. The latter result is surprising given that competitors were categorically related to targets; a spreading activation model would have predicted better recall for competitors than for unpracticed items. Anderson, Bjork, and Bjork (1994) explained this result in terms of an inhibitory mechanism. They postulated that during retrieval practice, competitors became active and competed with targets for retrieval. To facilitate retrieval of targets, competitors were inhibited, and inhibition had a lasting negative impact on the availability of competitors. This effect is called Retrieval-Induced Forgetting (RIF).

Similarity and the Pattern Suppression Model

Anderson and Spellman (1995) posited that RIF effects might rely on inhibition of one of two types: the associative link between a category cue and an exemplar might be weakened or the exemplar representation itself could be weakened. To choose between these hypotheses they used new, unstudied cues (called *independent probes*) during the test phase of RIF experiments to probe memory for studied items. For example, when *Fruit-Apple* was a competitor item, later recall of *Apple* could be tested with an independent probe (*Red-Ap_____*). RIF effects were observed even when independent probes were used to test memory for studied items, ruling out inhibition of the associative link as an explanation.

Anderson and Spellman (1995) concluded that inhibition acted directly on the representations of competitor items. They suggested that the inhibitory mechanism underlying RIF might act on individual features (sub-symbolic elements) of memory representations and proposed the pattern suppression model (PSM) in which the featural representation of a memory item is suppressed. The PSM postulates that features that are activated during retrieval but that are not components of the target are inhibited. Because competitor items are activated during retrieval many of their features are subject to inhibition, leading to RIF effects. The PSM also predicts that the featural overlap (similarity) of targets and competitors should affect the magnitude of RIF (see Figure 1).

In Figure 1a, the competitor *Banana* has many features

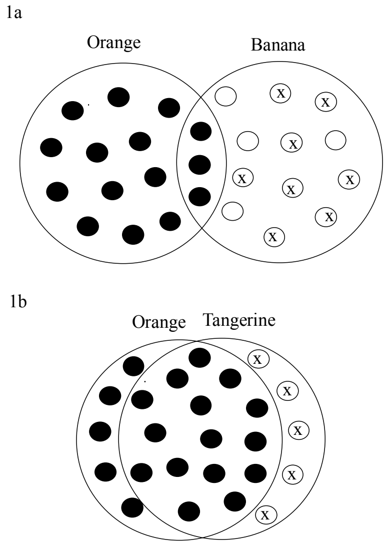


Figure 1. The Pattern Suppression Model predicts that moderate target-competitor similarity (1a) will yield more RIF than high similarity (1b). [Figure reprinted with permission from Anderson, M.C., Green, C. & McCulloch, K.C. (2000). Similarity and inhibition in long-term memory: A two-factor theory. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 26(5), 1141-1159].

which are not components of the retrieved target *Orange*. These features are inhibited during retrieval of *Orange* and later recall of *Banana* suffers. In Figure 1b, the competitor *Tangerine* overlaps *Orange* on many features, and these are strengthened rather than inhibited. As a result, subsequent RIF is diminished for *Tangerine* (see Anderson, Green, & McCulloch, 2000).

Anderson and McCulloch (1999) and Anderson, Green, and McCulloch (2000) tested and confirmed the predictions of PSM regarding similarity and RIF. Using the same RIF paradigm as Anderson, Bjork, and Bjork (1994), they showed that emphasizing common features of targets and competitors at study reduced RIF, while emphasizing unique features at study increased RIF. The work by Anderson and colleagues suggests that inhibition acts upon subsymbolic features and the total inhibition of a representation is a function of the sum of inhibition across all its features.

The central motivation for the work presented here was the formal assessment of feature-based inhibition as a basis for RIF, and thus the viability of Anderson and Spellman's theoretical ideas about feature-based inhibition in RIF. Anderson, Green, and McCulloch (2000) offer an account of how such a mechanism would work based on Anderson and Spellman's (1995) PSM. Taking the PSM as a theoretical starting point, we implemented a feature-based inhibitory mechanism in the MNEM model (Green & Kittur, 2004).

The MNEM Model

The *M*emory *N*eed *E*xpectation *M*odel (MNEM; Green & Kittur, 2004) was conceived as an implementation of Bjork and Bjork's (1992) New Theory of Disuse (NTD). NTD postulates that memory items are associated with two strengths: a storage strength, corresponding to the total level of learning of the item, and a retrieval strength, corresponding to the accessibility of the item. Storage strength only increases, decelerating as it grows. Gains in storage strength are also a function of retrieval strength such that highly accessible items benefit less from practice than items with low accessibility. Retrieval strength fluctuates up and down as a function of both current retrieval strength (highly accessible lose accessibility quickly) and storage strength (well-learned items lose accessibility slowly).

MNEM borrows some basic operations and representational conventions (including approaches to trace comparison and composition) from Hintzman's (1984, 1986, 1988) MINERVA2 model. However, MNEM and MINERVA2 are distinct in important ways, including the method by which they predict the accessibility of memory items. More detail regarding the MNEM model and its relation to MINERVA2 is presented in Green and Kittur (2004).

General Architecture

MNEM has two modules: a working memory (WM) and a long-term memory (LTM). WM consists of a single-item buffer.[1] It holds an item that is to be encoded into LTM, and also receives the information retrieved from LTM.

LTM is a collection of stored memory traces. The capacity of LTM is assumed to be unlimited (at least, very large).

Memory Representations

Memory Traces as Vectors Each memory trace in MNEM is an ordered vector of size *n*, with each element taking on a value of -1, 0, or +1. These values can be considered to indicate the absence of a feature, no information about a feature, or the presence of a feature, respectively. This format is consistent with Anderson and Spellman's (1995) description of simple feature list representations in the PSM.

Trace Similarity The similarity of any two memory traces (*A* and *B*) can be calculated as follows:

$$S(A,B) = \left(\frac{1}{N_R}\right)\sum_{j=1}^{n} A(j)B(j)\,, \tag{1}$$

where n is the number of elements in the trace and N_R indicates the number of "relevant features" in the pair of traces. Relevant features are defined as features for which at least one of the two traces contains a non-zero value. If neither trace contains any information about a feature, then that feature is not counted as relevant.

Trace Composition Two or more traces can be combined using a simple weighted average. Each feature in the composite trace is the weighted average of the corresponding features in the traces being combined. That is, to combine *m*

[1] MNEM makes no claims about the structure or function of human working memory: MNEM's WM is used solely as a buffer for input to and output from LTM.

traces (T_1 through T_m), feature j of the composite trace E, is calculated as:

$$E(j) = \frac{1}{m}\sum_{i=1}^{m} w_i T_i(j)\,, \qquad (2)$$

where w_i is a weighting factor for trace T_i. Trace composition is used to keep track of traces that have recently been targets or competitors (discussed in detail later).

Assumptions About Representation MNEM's representational assumptions are minimal. In fact, we do not make strong claims about representation here. MNEM requires only that its representations be amenable to *some* systematic similarity metric (e.g., Eq. 1), and they be systematically combinable (e.g., Eq. 2). MNEM can be implemented with any representational scheme that meets these requirements.

Encoding

To encode an item, the content of the encoded trace is copied into a newly-created LTM trace. The learning rate parameter l (where $0 < l \leq 1$) indicates the independent probability that any one feature will be copied accurately into LTM during encoding. Features that are not encoded accurately are encoded with zero value (i.e., MNEM *loses* information, but does not *distort* information).

Each instance upon which an item is studied yields a new and independently-encoded LTM trace (MNEM is a "multiple-trace" memory model). The encoding of a new trace is not affected by its similarity to items already in LTM. MNEM assigns each LTM trace a unique index which serves as a timestamp for encoding. This index is used (for now) in place of a more elaborate spatio-temporal tag for each trace. In related work, the authors are exploring how the addition of context elements to the memory trace itself may accomplish the work of tags or indices.

Retrieval

Retrieval comprises the activation of LTM traces, the marking of LTM traces as targets, competitors, or irrelevant traces, the calculation of retrieval strength, and (possibly) the construction of a composite trace from activated targets. Retrieval also includes the updating of the inhibitory mechanism in MNEM (described later).

LTM Trace Activation and Marking The activation of LTM traces involves the calculation of similarity scores for all LTM traces. The total similarity score for an LTM trace is a linear combination of its simple similarities (Eq. 1) to three other traces: the retrieval probe P; the recent targets trace (I_T); and the recent competitors trace (I_C).[2] When a retrieval probe P is introduced to the system, each item T in LTM is given a total similarity score $S_{total}(T)$, defined as:

$$S_{total}(T,P) = S(T,P) + \alpha S(T,I_T) - \beta S(T,I_C)\,, \qquad (3)$$

where α and β are parameters that weight the relative contributions of recent targets and recent competitors. That is, traces are activate more when similar to a current probe or recent targets, and less when similar to recent competitors.

LTM traces are designated as targets if their total similarity score exceeds the criterion parameter s_t. Traces with scores falling between s_t and the criterion parameter s_c are designated competitors. Traces with scores below s_c are considered irrelevant to the current retrieval. The activation of LTM results in each LTM trace being assigned to exactly one of these sets based on its total similarity score.

Calculating Retrieval Strength (RS) According to NTD, the probability of retrieving an item (its retrieval strength) is a function of how well-learned it is (its storage strength), and the time elapsed since it was last study or retrieved.

MNEM uses the average spacing and the cumulative similarity score of traces designated as targets in LTM to calculate an item's storage strength (SS). For a memory probe P, MNEM calculates the average retention interval $RI(P)$ between targets in LTM:

$$RI(P) = \frac{1}{(N_m)}\sum_{i=2}^{N_m}[index(M_i) - index(M_{i-1})]\,\cdot \qquad (4)$$

M_i is the i^{th} LTM trace marked as a target and N_m is the total number of LTM traces marked as targets.[3] (The *index()* operator simply indicates that the model is using the LTM index for a trace and not the trace itself).

MNEM also calculates the sum of all targets' total similarity scores as an indicator of the overall similarity of the probe item to items in LTM.[4] That is, the base rate *BR(P)* for an item P is calculated as:

$$BR(P) = \sum_{i=1}^{m} S_{total}(M_i, P)\,, \qquad (5)$$

where $S_{total}(M_i,\ P)$ is the total similarity score of the ith marked item in LTM (Eq. 3).

The product of *RI(P)* and *BR(P)* corresponds to SS (which increases with the frequency or spacing of study events). Retrieval strength is defined as a ratio of storage strength to the interval elapsed since the last study event occurred. The interval elapsed since the last study event, or current interval *CI(P)*, is defined:

$$CI(P) = index(P) - index(M_{\max})\,, \qquad (6)$$

where $index(M_{max})$ indicates the index of the most recently encoded target trace. Also, *index(P)*, the index for the probe item, is simply set to the current time step (which is equal to the number of traces in LTM plus one: $N_{ltm} + 1$).

So, for a probe P, retrieval strength *RS(P)* is defined:

[2] I_T and I_C are described in detail in the section "Tracking Inhibition". See Eqs. 9a, 9b, 10a, and 10b.

[3] When only a single trace in LTM is marked, the average retention interval is defaulted to a value of one.

[4] In this respect, MNEM's new base rate calculation is similar to MINERVA2's intensity calculation (Hintzman, 1984).

$$RS(P) = \frac{RI(P) \cdot BR(P)}{CI(P)} . \qquad (7a)$$

To accurately compare retrieval strengths across items, the raw retrieval strength *RS(P)* from Eq. 7a is converted to normalized retrieval strength $RS_N(P)$:

$$RS_N(P) = \frac{\log(RS(P)+1)}{\log(RS_{\max}(P)+1)} , \qquad (7b)$$

where $RS_{max}(P)$ is the maximum retrieval strength that *P* could attain (retrieval strength at immediate recall).

Retrieved Information Retrieval strength is a measure of the probability of retrieving LTM information given a probe. However, memory also involves the reconstruction of the remembered content. MNEM reconstructs content by creating a composite of the target traces in LTM. Eq. 2 describes a general weighted-average method for combining traces, and MNEM implements this method using the target traces' total similarity scores as weights:

$$E(j) = \frac{1}{m} \sum_{i=1}^{m} S_{total}(T_i, P) * T_i(j) . \qquad (8)$$

We do not address reconstructed memory content in the simulations reported here.

Tracking Inhibition

To implement a PSM-like inhibitory mechanism in MNEM, we modified the similarity calculation used to mark LTM traces (resulting in Eq. 3). This requires tracking the content of recent targets and competitors. Each retrieval event is associated with a set of targets and a set of competitors. The traces in each of these sets are combined into single vectors. Targets from the most recent retrieval are represented by a trace R_T, where each feature of R_T is a weighted average of the corresponding features in each of the *m* LTM traces marked as targets (a version of Eq. 2):

$$R_T(j) = \sum_{i=1}^{m} S_{total}(T_i, P) \cdot T_i(j) , \qquad (9a)$$

Competitors from the most recent retrieval are represented by a trace R_C, where each feature of R_C is a weighted average of the corresponding features in each of the *k* LTM traces marked as competitors:

$$R_C(j) = \sum_{i=1}^{k} S_{total}(C_i, P) \cdot C_i(j) , \qquad (9b)$$

R_T and R_C describe the features of targets and competitors from the most recent retrieval event. The "running count" of such features from several recent retrieval events is tracked in traces I_T and I_C. Each time a new retrieval event occurs, the entries of I_T and I_C are updated with the contributions of R_T and R_C:

$$I_T(j) = \frac{pI_T(j) + R_T(j)}{p+1} , \qquad (10a)$$

where *p* is a parameter controlling how quickly the vector representing recent targets changes. And likewise:

$$I_C(j) = \frac{qI_C(j) + R_C(j)}{q+1} , \qquad (10b)$$

where *q* is a parameter controlling how quickly the vector representing recent competitors changes.[5]

In summary, each retrieval event is associated with a set of competitor traces, and these are combined into a trace R_C via weighted averaging. The model tracks a trace I_C, which is a simple composite (straight average) of the traces R_C over the most recent q+1 retrievals. In subsequent retrieval events, I_C is used to mark LTM, and then modified by the that marking. The same procedure holds for recent targets (tracked with traces R_T and I_T).

An aspect of our implementation that bears additional mention is the special status of the inhibitory mechanism: it is not emergent from the basic dynamics of the model, but instead involves special structures and processes on top of the MNEM architecture. At first glance, this may seem inelegant. However, there is behavioral evidence that the inhibitory mechanism underlying RIF can be willfully invoked, and may therefore be related to special executive processes not solely related to memory (Anderson & Green, 2001).

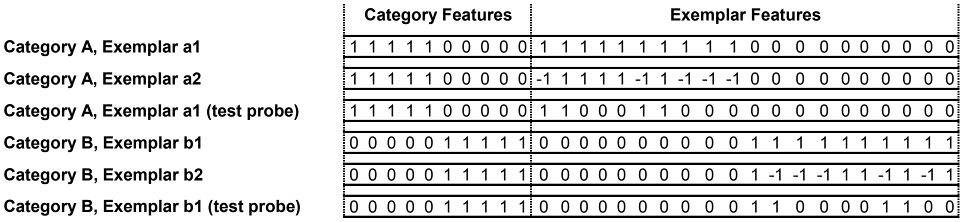
Simulation of RIF with Similarity Effects

Simulations based on the Anderson, Green, and McCulloch (2000) experiments were run. Each simulation involved four items (two members of each of two categories). One item was a target (practiced item from a practiced category), one a competitor (unpracticed item from a practiced category), and two were unpracticed items (from an unpracticed category). During the study phase, each item was studied three times with equal spacing between study events. In the retrieval practice phase, the target item was tested and restudied three times (we assumed successful retrieval practice and the equivalent of another study exposure resulting from that practice). RS_N for all four items was tracked from the beginning of study until well after the end of retrieval practice (including time when testing would occur).

The items were 30-element traces. Elements one through ten corresponded to a category label and elements 11 through 30 corresponded to exemplar descriptions. Items from different categories were orthogonal. Items within a category were not orthogonal: they had identical category label features but differed (to varying degrees) in the features of the specific exemplars. During the retrieval practice and to assess retrieval strength[6], we used test probes with

[5] I_T and I_C equate to averages of the R_T and R_C vectors calculated over the most recent p+1 and q+1 retrievals, respectively.

[6] We have included in MNEM a function that evaluates RS_N for a test item without affecting the state of the model.



	Category Features	Exemplar Features
Category A, Exemplar a1	1 1 1 1 1 0 0 0 0 0	1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0
Category A, Exemplar a2	1 1 1 1 1 0 0 0 0 0	-1 1 1 1 1 -1 1 -1 -1 -1 0 0 0 0 0 0 0 0 0 0
Category A, Exemplar a1 (test probe)	1 1 1 1 1 0 0 0 0 0	1 1 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0
Category B, Exemplar b1	0 0 0 0 0 1 1 1 1 1	0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1
Category B, Exemplar b2	0 0 0 0 0 1 1 1 1 1	0 0 0 0 0 0 0 0 0 0 1 -1 -1 -1 1 1 -1 1 -1 1
Category B, Exemplar b1 (test probe)	0 0 0 0 0 1 1 1 1 1	0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 1 1 0 0

Figure 2. Examples of items used in simulations. The first ten elements of each trace describe features of the item's category, and the remaining twenty elements describe features of the exemplar itself. Items from different categories are orthogonal. At test, most of the exemplar features were removed to mimic a category-and-stem-cued recall method.

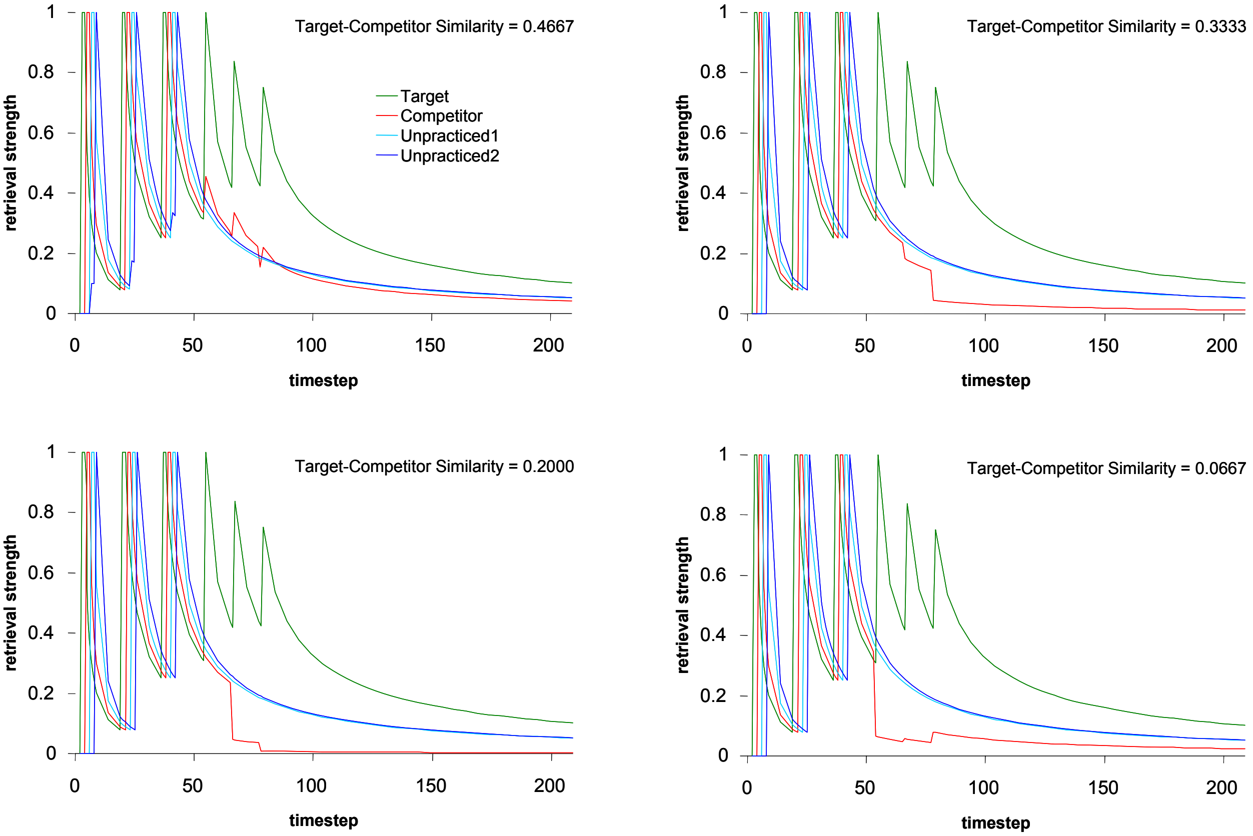


Figure 3. RS_N for target, competitor, and unpracticed items at four levels of target-competitor similarity. The magnitude of RIF was affected by the similarity of target items to competitor items. High similarity (upper-left panel) protected competitor items from RIF relative to moderate similarity. RIF increased as similarity decreased (upper-right and lower-left panels), until competitors became very dissimilar at which point RIF began to decrease again (lower-right panel).

intact category features, but without most of their exemplar features (see Figure 2).

The PSM predicts that the similarity of target and competitor items will influence the magnitude of RIF. Competitors that are very similar to target items will be protected from RIF. As target-competitor similarity decreases RIF should also decrease (but remain present as long as similarity is greater than zero). Simulations were run in which target-competitor similarity was 0.4667, 0.3333, 0.2000, and 0.0667. Other parameter settings were: $l = 1.0$, $s_t = 0.50$, $s_c = 0.01$, $\alpha = 0.5$, $\beta = 0.75$, $p = 3$, $q = 3$.

The data are shown in Figure 3 and qualitatively match the results of Anderson, Green, and McCulloch (2000).[7] The

[7] MNEM predicts lower recall rates than observed in humans. This may be because humans have pre-experimental associations between categories and exemplars (non-zero guessing rates) and because human mental representations are richer than those used here.

magnitudes of simulated RIF effects changed with target-competitor similarity. In the 0.2000 similarity simulation, competitor RS_N is approximately 0.125 below the unpracticed items' RS_N at time step 100 (20 time steps after retrieval practice has ended). This difference corresponds to RIF. As similarity is increased, RIF is reduced (virtually no RIF when similarity is set at 0.4667). Decreasing similarity also diminishes RIF. Figure 4 illustrates how the magnitude of RIF changes with the similarity of target and competitor traces. In addition, RIF effects diminished over time in our simulations. The three different curves show the magnitude of RIF effects at 100, 150, and 200 time steps.

Another notable feature of our data is the decrease in peak target RS_N as retrieval practice proceeds through several repetitions. The decrease is counterintuitive in that the target is becoming less accessible with more retrieval practice. It is important to keep in mind that the data do not account for any contribution of WM to retrieval. Thus, one possibility is that reductions in the accessibility of a target LTM trace *do* occur as retrieval practice proceeds, but that these decreases are offset by substantial contributions of WM. The (longer-term) accessibility of target traces *does* increase with repeated retrieval practice (hence the difference between target and unpracticed items).

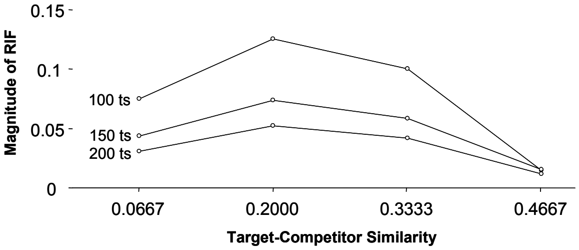


Figure 4. The magnitude of RIF changed with target-competitor similarity. High target-competitor similarity protected competitors from RIF. The magnitude of RIF diminished as the delay from retrieval practice increased, but RIF peaked at moderate similarity regardless of delay.

Conclusions

Our simulation results match those from behavioral experiments, and demonstrate a computational instantiation of the PSM's feature inhibition explanation of RIF. Admittedly, the simulations reported here do not thoroughly explore MNEM's parameter space and different settings may yield results that less accurately resemble the human data. Simulations with a wider range of parameters must be evaluated both to understand how sensitive these results are and to explore the predictions MNEM makes about the influence that noise, distortion, and the relative strengths of excitation and inhibition exert on RIF. One line of ongoing work with MNEM focuses on developing a better understanding of the way that parameter settings influence RIF and other memory dynamics in the model.

We are also working on a more sophisticated approach to inhibition in which traces are less subject to rigid criteria in the search for target, competitor, and irrelevant LTM items. Another line of research is making use of the model's ability to reconstruct memory items from degraded probes. In particular, we are exploring the way that the inhibitory mechanism described here influences the content of retrieved (reconstructed) traces.

Acknowledgments

Part of this research was conducted while the first author held a National Research Council Research Associateship Award at NASA Ames Research Center. We thank Li Li, Guy Pyrzak, Alonso Vera, and Ben Storm for their comments on this paper.

References

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, 20*, 1063-1087.

Anderson, M.C., & Green, C. (2001). Suppressing unwanted memories by executive control. *Nature*, 410(6826), 366-369.

Anderson, M.C., Green, C., & McCulloch, K.C. (2000). Similarity and inhibition in long term memory: Evidence for a two-factor theory. *Journal of Experimental Psychology: Learning, Memory and Cognition, 26(5)*, 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., & Spellman, B.A. (1995). On the status of inhibitory mechanisms in cognition: Memory retrieval as a model case. *Psychological Review, 102*, 68-100.

Bjork, R.A. & Bjork, E.L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In A. Healy, S. Kosslyn, & R. Schiffrin (Eds.), *From Learning Processes to Cognitive Processes: Essays in Honor of William K. Estes* (Vol. 2, pp.35-67). Hillsdale, NJ: Erlbaum.

Green, C. & Kittur, A. (2004). A multiple-trace memory model exhibiting realistic retrieval dynamics. In K. Forbus, D. Gentner, and T. Reiger (Eds.): *Proceedings of the 26th Annual Meeting of the Cognitive Science Society*. Mahwah, NJ: Erlbaum.

Hintzman, D.L. (1984). MINERVA2: A simulation model of human memory. *Behavior Research Methods, Instruments, & Computers, 16*, 96-101.

Hintzman, D.L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review, 93(4)*, 411-428.

Hintzman, D.L. (1988). Judgements of frequency and recognition memory in a multiple-trace memory model. *Psychological Review, 95(4)*, 528-551.