SARKAE – Modeling the Co-Evolution of Event Memory and Knowledge

Angela B. Nelson (a2nelson@ucsd.edu)
Department of Political Science, University of California San Diego, 9500 Gilman Dr. #0521
La Jolla, CA 92093 USA

Richard M. Shiffrin (shiffrin@indiana.edu)
Department of Psychological & Brain Sciences, Indiana University, 1101 E. Tenth Street
Bloomington, IN 47405 USA

Abstract

We present an overview of a model for the co-evolution of knowledge and event memory. The model, termed SARKAE (Storing and Retrieving Knowledge and Events), describes the development of knowledge and event memories as an interactive process: knowledge is formed through the accrual of individual events, and the storage of an individual episode is dependent on prior knowledge. We reference two experiments which provide data to inform our theory: these studies involve the development of new knowledge, and then testing in transfer tasks involving episodic memory, retrieval from knowledge, and perception. The results of the transfer tasks indicate a substantial role of pure frequency or raw exposure, in opposition to the contextual diversity accounts of frequency suggested by Adelman et al (2006). An overview of the SARKAE model is presented. The model is able to account for the effects of frequency in the absence of contextual diversity.

Keywords: episodic memory; semantic memory; learning; perception; Bayesian models.

Introduction

The processes involved in the accumulation of knowledge and the formation of event memories are interdependent. Almost every study since the 1890s has shown that the way episodic (or event) memories are encoded depends on the knowledge (or semantic memory) of the individual who is encoding them. Conversely, an individual's knowledge must be formed through the episodes they encounter; this idea was the basis of the REM model's account of priming (Shiffrin & Steyvers, 1997). These interdependent processes create a feedback loop in which knowledge and episodic memory formation develop together over lifelong learning.

Studies of memory and perception in the recent past have provided strong support for the idea that memory processes are robustly influenced by prior experience with the to-be-remembered content. Priming studies, for example, have shown that prior study of a word affects how well that word is identified in a forced choice perceptual identification task (Ratcliff & McKoon, 1997). The REMI model of Schooler, Shiffrin, and Raaijmakers (2001) accounts for these effects through a process in which the lexical representation (or knowledge) of the word is changed through prior study (the "prime"); when a word is studied an event memory is formed, but in addition, novel features of the event, such as the context of the experimental setting, are added to the lexical representation of the word. When the studied word is then presented for perceptual identification, the context tends to be similar to that at study, increasing the match of the probe cues to the lexical trace, predicting a variety of measurable effects that match those observed. In other words, the knowledge that a subject has about a stimulus, and the inclusion in that knowledge of factors like the experimental context, affect the way that a stimulus is perceived.

There are many models of the storage and retrieval of event memories, and sometimes the addition to existing knowledge of information from recent events (e.g. - Raaijmakers & Shiffrin, 1981, Shiffrin & Steyvers, 1997, Howard & Kahana, 2002, Anderson, 1983). The temporal context model of Howard & Kahana for example provides an explanation for recency and contiguity effects through the storage of both item information and recent contextual information. Other models, such as the ACT-R model of Anderson (1983) also provide eloquent representations of memory storage and retrieval. A few models attempt to explain aspects of the way events produce knowledge, especially for aspects of the role played by words in language (e.g. McClelland & Rumelhart, 1981). However, most of the prior research has been aimed to explain memory and learning when knowledge has already formed (to various degrees). Previous work by Reder et al. has examined the development of knowledge on a set of pseudowords, and used the dual-process SAC model to explain their findings. Although highly relevant and useful in the development of our research, the training study and modeling by Reder et al. did not explicitly model the growth of new knowledge.

Our aim is the development of a model that begins to explain the interacting growth of event memory and knowledge, as they influence both memory storage and retrieval. This co-evolution of the two systems was the focus of the REM-II model, created by Mueller and Shiffrin (2006). In this model, knowledge (or semantic memory) is represented as an accumulation of the co-occurrence of features: Features that are present in an episodic event are coded as occurring together in a matrix representation of semantic memory. REM-II is a quite powerful model, but a simplified version is sufficient to explain the basic concepts

by which event memory and knowledge co-develop, and is sufficient to model the empirical results presented in this paper. However, even a simplified model when applied to five different tasks spanning the range of learning, memory, and perception can grow to appear quite complex. The simplified model uses a representation in which each (separate) trace, whether an event trace or a knowledge trace, is a vector of feature values. Rather than term the model some other variant of REM, we use the terminology "Storing and Retrieving Knowledge and Events", abbreviated SARKAE.

Role of Experience and Frequency in Cognition

If one hopes to develop a theory in which events accumulate to form knowledge, then it is critical to understand the role of event frequency. Such effects are omnipresent in memory and perception tasks, but the processes responsible for such effects remain in debate.

Researchers have explored the effects of experience in various ways, typically by analyzing existing knowledge, identifying stimuli with different histories of experience, and using the stimuli with different frequencies in memory and perception tasks. The great majority of such investigations use words as stimuli: Words are categorized based on their frequency. Frequency is defined as normative occurrence in the environment, and these frequencies are estimated from various databases of typically textual materials. Words differing in frequency are then tested and exhibit a variety of consistent differences. These are termed the 'Word Frequency Effect', especially when found in recognition memory (Glanzer & Adams, 1985). In episodic recognition memory tasks, words that occur rarely in the environment are recognized *better* than words that occur frequently in the environment. Word frequency has also been shown to have effects on recall performance (high frequency words are recalled better), and perceptual tasks such as lexical decision and perceptual identification (forced choice, etc.).

However, given that word frequency is correlated with so many other variables (e.g. meaning, regularity of spelling, length of the word, and virtually every other characteristic one can measure for words), it is hard to know whether experience per se is responsible for the observed effects. In fact, a current debate concerns whether frequency per se or context effects are the primary cause of the observed findings. Adelman, Brown, and Quesada (2006) for example suggest that the diversity of contexts in which a word has been seen is a more accurate predictor of word frequency effects than the actual frequency of the word. By analyzing a large corpora of texts separated both by word frequency and contextual diversity (the number of documents in which a word was present), they concluded that it was the contextual diversity of an item, not the word frequency, that affected response times in word naming and lexical decision for three separate data sets. The difficulty of assessing the cause of frequency effects for words is one reason we chose to vary frequency of training of novel characters in the present studies. By training novel stimuli we can control with far greater precision the factors correlated with frequency and thereby properly constrain the theory. The studies referenced in this article create experience differences over a fairly lengthy period of training in two quite different tasks, one based on visual search, and the other based on perceptual matching.

In order to control for the confounds produced by word stimuli, our studies use stimuli that are far less related to existing language and numeric knowledge, and far less likely to bring with them existing frequency correlations: Chinese characters. (We select participants for whom such stimuli are unfamiliar). The first study used a visual search task in training. This task was based loosely on that of Shiffrin and Lightfoot (1997). Different Chinese characters appeared with widely differing frequencies during training. Following training, the subjects completed various recognition memory and perception tasks different from the training task, using both the trained characters and new characters as stimuli.

In the interest of space, this first experiment using the visual search training will not be discussed in detail. It is sufficient to mention that the crucial finding of this study was that substantial frequency effects occurred for all transfer tasks. What is more relevant to the discussion of the no-context experiment described below (as well as the SARKAE model) is that the visual search task used for training varied character frequency, but the randomization of trials and foils ensured that higher frequency characters most often occurred in the spatial and temporal vicinity of other higher frequency characters. Thus frequency per se was correlated with what could be termed character context, temporal context, or character diversity. As mentioned previously, Adelman et al. (2006) proposed that only the diversity of contexts in which an item occurs is responsible for most frequency effects. The confounding of frequency and character context made inference about causal mechanisms uncertain, and hence led to the design of the No-Context Experiment described below.

No-Context Experiment

The no-context experiment used a training paradigm not involving visual search. Participants were trained using a same vs. different judgment task: A character was presented briefly twice in succession, and half the time the two presentations varied slightly in size, rotation, or contrast. The participant judged whether the two presentations were exactly the same or varied slightly in one of these three dimensions. Thus a character was its 'own' context. Further, to remove the possibility that the test character on the previous trial might provide context for the present trial, one fixed 'control' character, different from any of the experimental characters, was tested using the same judgment task between every two experimental character judgments. This extremely high frequency character was not subsequently used in the post-training transfer tasks. If

context is carried forward from the previous trial during training, the context that is carried forward for the experimental characters of different frequency will be equated, because the previous character is always the same one. The no-context experiment used the same frequency distribution (given below) as the visual search training experiment. By removing characters that provide context on any given trial, and by holding constant the character context on the preceding trial, it is plausible to assume that the confound between context and frequency is mostly if not totally eliminated.

Training Methods

Participants. Seven participants, recruited with an email advertisement, participated in the experiment for monetary compensation. All participants reported no prior experience with Chinese characters.

Design and Stimuli. The occurrence of the characters in the same/different task was manipulated to produce four frequency conditions which varied in a ratio of 1::3::9::27. For each subject, a set of 32 characters was selected randomly from a pool of approximately 200 characters. From these 32 characters, 8 were assigned to each frequency condition. In order to keep the complexity of the characters reasonable, all the characters in the pool were composed of 7 strokes or less. In order to fully eliminate context from the training, one "super-high frequency" item was also randomly chosen, making the entire training set 33 characters. This character appeared as a "buffer" item every other trial, and was not used as a stimulus in the post-training tasks.

Procedure. Each trial consisted of two brief (500 ms) presentations of a single Chinese character, which subtended a visual angle of approximately 4.3 x 4.3 degrees. The two presentations of the character were either identical or varied slightly in size, rotation, or contrast of the character. Only one of these three dimensions varied at a time. There were three levels of each variable (size: small, medium, large; rotation: left, straight, right; contrast: dark, normal, light), and the change between each of these levels varied based on a staircase algorithm. The staircase rules were as follows: when the subject answered two rotation-difference trials correctly, the rotation factor (i.e. – the difference in angle between the three levels) decreased by a given amount. If they got a rotation-different trial wrong, the rotation factor increased by a given amount. This staircase was done separately for each of the three variables. In this way, subjects were kept at approximately 75% accuracy. Subjects completed 12 training sessions, approximately 3 per week. There were a total of 1060 trials for sessions 1-11, and 1140 trials for session 12.

Training Results

Since the training paradigm used a staircase algorithm to keep subjects at approximately 75% accuracy, the results of

training were analyzed by examining the change factors for size, rotation, and contrast. If the subjects are showing improvement at the same/different discrimination, then the change in variable (size, rotation, or contrast) needed to keep them at 75% should decrease over session. Figure 1 shows the mean rotation, contrast, and size changes required (averaged over all subjects) as a function of training session. The results indicate that subjects were becoming more efficient at the task as training progressed, as indicated by the decrease in variable change over session.
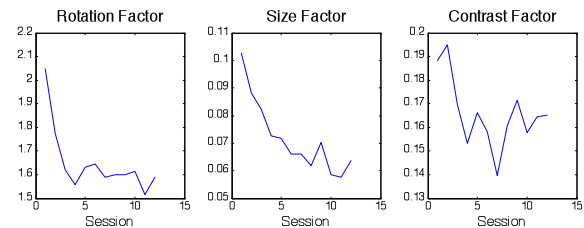

Figure 1: Mean change in rotation (panel A) size (panel B) and contrast (panel C) needed to obtain 75% accuracy as a function of training session. Rotation factor is measured in degrees, size factor in percentage size difference, and contrast factor in percentage contrast difference.

Post-training Tasks

Following the training, the subjects completed three post-training tasks: pseudo-lexical decision, episodic recognition, and forced-choice perceptual identification. Testing was carried out again six weeks after training. A programming error, discovered after the immediate transfer tasks, caused the forced choice data to be very noisy and essentially uninformative. These results are therefore neither reported nor analyzed. Also, because forced choice results were not available for immediate test, forced choice testing was omitted for the delayed testing at six weeks.

Pseudo-lexical Decision

Design and Procedure. Subjects viewed one list, which contained all 32 trained characters (excluding the buffer item), as well as 32 new characters. Each of these characters occurred 3 times throughout the list, making the total length of the list 192 characters. Subjects were presented with a single character on the screen, and were asked to decide (by keypress) as quickly as possible whether they had ever seen that character during any of the previous training sessions.

Results. Response time and accuracy were measured for each frequency condition, as well as new items. The results for the trained items when tested shortly after training was completed (2-3 days) are shown in Figure 2. A contrast analysis showed that there was a significant negative relationship between frequency and response time ($t(6)=-2.97, p=.03$), and a significant positive relationship between frequency and accuracy ($t(6)=2.90, p=.03$).

Response time and accuracy were measured again (for 6 of the 7 subjects) approximately 6 weeks after the previous test session. The results followed the same qualitative pattern as they did 6 weeks prior: there was a significant negative relationship between response time and frequency ($t(5)=-2.45$, $p=.058$), and a significant positive relationship between accuracy and frequency ($t(5)=2.44$, $p=.059$, see Figure 2). A contrast analysis showed that there was no significant difference in the magnitude of the effects that occurred in the shortly after training and those that occurred after the 6 week delay for either accuracy ($t(5)=1.14$, $p=.31$) or response time ($t(5)=.51$, $p=.63$).
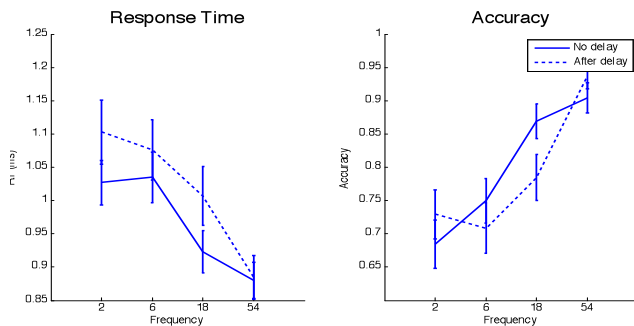

Figure 2: Mean response time (panel A) and accuracy (panel B) for all subjects in the lexical decision task as a function of frequency. The solid line shows the results when the test was administered after a very short delay (2-3 days), the dashed line corresponds to the data following a 6 week delay.

Discussion. The results of the lexical decision task showed that the absence of character-context during training did not eliminate the effects of frequency on speed and accuracy of decision. Therefore, it follows that there must be some mechanism other than the context present during training that is causing improved recognition that high frequency characters are present in knowledge. In addition, this frequency effect showed little signs of reduction over six weeks.

Episodic Recognition

Design and Procedure. The task consisted of eight pairs of study and test lists. Each study list contained eight trained characters (two from each frequency category) and eight untrained characters. Each test list contained all the items from the study list as well as 16 unstudied items, which included eight trained characters (two from each frequency category) and eight untrained characters. The first four items on the test list were always untrained characters, providing a buffer for the items of interest (trained characters). Subjects viewed each item on the study list for 1000 milliseconds, presented one at a time on the screen. Following the study list, the subjects were presented with the items on the test list one by one, and for each item had to

respond whether the character had been present on the list they had just studied. Subjects were instructed to 'reset' their memory in between each list, and answer 'old' to an item on the test list only if it had been present on the most recent study list.

Results. The data from the episodic recognition task were analyzed by examining the hit rates (correctly identifying a studied item as old) and false alarm rates (incorrectly identifying an unstudied item as old). The hit and false alarm rates (averaged over all subjects) are plotted as a function of frequency in figure 3. When tested shortly after the completion of training, false alarms significantly increased as frequency increased (panel A, $t(6) =3.19$, $p=.02$). There was also a marginally significant decrease in d' due to frequency ($t(6)=-1.86$, $p=.11$). The hit rate analysis however showed no significant effect of frequency.

Six of the seven subjects were tested again following a six-week delay. The results of the delayed test are shown in panel B of figure 3. Statistical analyses showed no significant effect of frequency on hit rates, false alarm rates, or d'. Furthermore, a contrast analysis showed that there was a significant difference in the magnitude of the false alarm rate effect found immediately after training compared to the effect found after a 6 week delay: the increase in false alarms due to increased frequency was (marginally) significantly larger immediately after training ($t(5)=2.11$, $p=.09$).
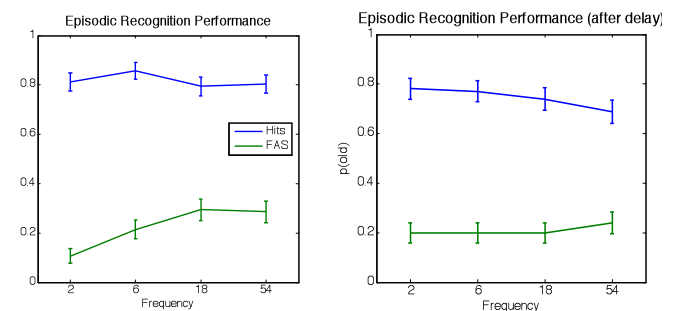

Figure 3: Episodic Recognition Results soon after training (Panel A) and after a 6-week delay (Panel B). Hit rates are shown in blue, false alarm rates in green.

Discussion. When tested shortly after the completion of training, the results in the episodic recognition task are similar to results found in our previous visual search training experiment and in normative word frequency studies: as frequency increases, d' decreases. In the current study, this is due more to an increase in false alarm rates than a decrease in hit rates with higher frequency items. Unlike some previous studies, the no-context training experiment did not show a significant effect of frequency on hit rates.

Unlike the lexical decision task which showed a large persistence of frequency effects after a six week delay, the d' effect and false alarm rate effect found in episodic recognition were largely reduced and possibly absent when

subjects were re-tested after delay. Both the existence of frequency effects in recognition, and the reduction with delay call into serious question the modeling processes used to account for recognition in the one factor model applied to our visual search training experiment. That model assumed poorer performance for high frequency test items was due to increased confusions with traces of list items, because those traces were more similar to the high frequency test probes. The present design should have eliminated such similarity differences. In addition, within list confusions should not have decreased if a recognition task was carried out at a six week remove from training, because the relevant episodic traces should have been those stored in the just seen study list. Thus the elaborated SARKAE model provides an explicit role for frequency per se (especially to explain pseudo-lexical decision findings) and an elaborated model for recognition. Due to spatial limitations, in this paper we present only an overview of the theory that is the foundation of the SARKAE model, with examples of how the theory is implemented to explain our experimental results.

SARKAE – Theoretical Overview

A fundamental storage assumption in SARKAE allows both event memories and knowledge to develop in concert: Each storage episode produces both: 1) an event trace; 2) additional information added to traces in memory that are brought to mind due to similarity to the present event. Such a prior trace can include a previous event trace (the basis for the start of knowledge accumulation), or a developing or mature knowledge trace. There is no fundamental distinction between event traces and knowledge traces in this view. Instead there is a continuum: traces are stored initially for each single event; some of these are retrieved (when a similar new event occurs), gain additional information, and are re-stored. As this process continues over successive occurrences of similar events, a rich knowledge trace results.

In SARKAE, accumulation of knowledge about an item or concept (e.g. for words, its lexical entry) includes features of the surrounding context that is present at the time of learning. Specifically, knowledge traces develop during learning by storing features that come both from the physical properties of the item or concept being learned, and also from the context surrounding the item during learning, both types of storage being modified and governed by attentional focus. These context features arise from other (attended) events nearby in time and the environment, and from the various components of internal and external context that numerous investigators have discussed for many years. For example, during training, when a character is presented, physical features of that item as well as surrounding context features (taken from other characters presented in close temporal proximity) are stored into the knowledge representation. In a more general sense, the knowledge trace that represents the concept of "table" will include information about the physical properties of various types of tables, information about the contents of events that involved tables (e.g. forks, dinners, conversations, replacing light bulbs), information about thoughts and feelings experienced at tables, and information about other events that occurred in the nearby temporal surround of table events (e.g. dropping of a milk bottle when removing it from the refrigerator). These features include context specific events themselves, such as the breakfast event in a given morning. Knowledge development is therefore built upon the events that accumulate to form the knowledge. Of course a mature knowledge trace includes features of numerous events, so a specific episode tends to be swamped in the accumulation of many episodes and tends not to be retrieved (from the knowledge trace—it can be retrieved as an episodic trace). Thus a knowledge trace in most instances seems to be context free. What do come to be retrievable from a mature knowledge trace are features that are consistent across many episodes, such as the spelling, pronunciation and meaning of a word.

Conversely, the formation of episodic memory traces is determined by prior knowledge and experience. Although certain very primitive features of experience might not depend upon learning and experience (e.g. a loud sound), most features of events are encodings based on prior learning (e.g. encoding and storing a table feature as 'dinner'). The model therefore creates episodic traces by choosing features of events from knowledge. Such features come from several sources: some are directly related to the central defining elements of the event such as the physical features of which it is composed (e.g. table physical features) and the central organizing concept (e.g. dinner); some come from other knowledge traces that are brought to mind during encoding of the event (e.g. the illness one encountered when eating breakfast last Sunday, or one's commitment to a new diet); some come from features of other nearby events still in short-term memory at the time of the present event. To some degree, the features chosen are modified by attentional focus. In terms of the experiment discussed in this paper, an episodic memory consists of a combination of physical features of the studied item, features drawn from the knowledge trace of that item, and features drawn from other items in close temporal proximity. One key concept is the perhaps non-controversial idea that the features comprising an event representation in short-term memory, and thereafter the stored event trace, are recruited from knowledge (e.g. one's prior experience and knowledge regarding tables will influence the formation of an event trace concerning a physically present table).

We have been highlighting mechanisms that produce storage of event memory and knowledge. Very similar mechanisms also occur in retrieval. We adopt the generally accepted view that retrieval is cue dependent, and based on similarity of the retrieval probe to the traces in memory. The generation of such a probe cue can be clearly defined, as when one is asked: "What is the capital of South Dakota"? In other cases retrieval seems more continuous and automatic, as when information moving through short-term

memory acts as retrieval cues to bring other associations to mind. However, because modeling continuous retrieval is quite complex, we will treat all retrieval in terms of discrete retrieval operations occurring one at a time, each based on some defined set of retrieval cues. The features that comprise such a retrieval cue are generated with the same processes that generate features for storage: They come from the query (if there is one), or from feature sets presently in short-term memory and attentional focus, and include features from the contextual surround at the time (internal and external context, and nearby events). More specifically, in the modeling of our experimental results, the retrieval cue consists of a combination of physical features of the test item, features drawn from the knowledge trace of the test item, and features taken from other items in close temporal proximity to the test item.

An absolutely essential component of storage and retrieval is noise in the processes. Following the approach in the REM model, we assume that both storage and retrieval are probabilistic, incomplete and error prone. When errors are made, it is natural to assume they are based on information in the knowledge base, and not completely random. Thus errors in retrieving and storing features are assumed to be relevant and consistent, in the sense that they are feature values for the feature in question (a 'blue' color feature might be retrieved or stored as 'red', but not as 'wet') and occur in proportion to the base rates of such values in knowledge.

When a cue is used to probe memory, it is compared in parallel to the event traces (and/or knowledge traces) in parallel. It would be unworkable and likely unreasonable to explicitly consider the match to each of the essentially uncountable traces in memory. Thus we assume that there is a probabilistic cutoff, only traces sufficiently similar to the probe becoming activated and participating in subsequent retrieval operations.

Similarity plays a role in both storage and retrieval, but we define similarity operations in such a way that similarity is measured as a relative construct: For both storage and retrieval a process based on similarity is defined as similarity of a given match compared to the similarity of matches that could have but did not occur. Thus in recent years we have characterized the match of a probe to an activated trace as a likelihood ratio: The numerator expresses the probability that the probe and cue were generated from the same event, and the denominator the probability that the two were generated by different events. These likelihood ratios occupy the theoretical niche played by 'strengths of activation' in various other theories (such as SAM; Raaijmakers and Shiffrin, 1981).

This brief summary of some of the central tenets of SARKAE provides hints concerning the theory, but is only the barest scaffolding upon which the model is constructed. When the full detailed processes are implemented, the model produces predictions that fit the results of the various post-training tasks from both the initial visual search training experiment as well as the no-context experiment

described in this paper. We cannot fully describe the modeling processes and results here due to space; however the aim of this discussion is not to focus on quantifiable model fits, but instead to convey the basics of the theory that inspired both the experiments described in this paper and the subsequent model development. The SARKAE model provides plausible mechanisms by which knowledge grows from events, and knowledge informs the coding and retrieval of both events and knowledge itself. Based on this theory, or others of a similar character, we hope that future research developments will not focus so strongly on differences among systems as upon the ways they grow together, in highly dependent fashion.

References

Adelman, J. S., Brown, G. D. A., & Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. Psychological *Science, 17,* 815-824.

Anderson, J. R. (1983). A Spreading Activation Theory of Memory. *Journal of Verbal and Learning Behavior, 22*, 261-295.

Glanzer, M., & Adams, J. (1985). The Mirror Effect in Recognition Memory. *Memory & Cognition, 13*, 8-20.

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

McClelland J. L., & Rumelhart, D. E. (1981). An interactive activation model of the effect of context in perception, Part I. An account of basic findings. *Psychological Review, 88,* 375-407.

Mueller, S. T., & Shiffrin, R. M. (2006). REM-II: A Model of the developmental co-evolution of episodic memory and semantic knowledge. *Paper presented at the International Conference on Learning and Development (ICDL), Bloomington, IN, June 2006.*

Raaijmakers, J. G. W., & Shiffrin, R. M. (1981). Search of Associative Memory. *Psychological Review, 88,* 93-134.

Ratcliff, R., & McKoon, G. (1997). A Counter Model for Implicit Priming in Perceptual Word Identification. *Psychological Review, 104*, 319-343.

Reder, L. M., Angstadt, P., Cary, M., Erickson, M. A., & Ayers, M. S. (2002). A Reexamination of Stimulus-Frequency Effects in Recognition: Two Mirrors for Low- and High-Frequency Pseudowords. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 28,* 138-152.

Schooler, L. J., Shiffrin, R. M., and Raaijmakers, J. G. W. (2001). A Bayesian Model for Implicit Effects in Perceptual Identification. *Psychological Review, 108,* 257-272.

Shiffrin, R.M., & Lightfoot, N. (1997). Perceptual Learning of Alphanumeric-like Characters. *The Psychology of Learning and Motivation, 36*, 45-81.

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