EXAMINING THE CO-EVOLUTION OF KNOWLEDGE AND EVENT MEMORY

Angela B. Nelson

Submitted to the faculty of the University Graduate School

in partial fulfillment of the requirements

for the degree

Doctor of Philosophy

in the Department of Psychological & Brain Sciences

and the Department of Cognitive Science

Indiana University

August 2009

Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Doctoral Committee

_____

Richard M. Shiffrin, Ph.D.

_____

Robert L. Goldstone, Ph.D.

_____

Thomas A. Busey, Ph.D.

_____

Karin H. James, Ph.D.

August 13th, 2009

Dedication

For my Mom. Without her love, encouragement, and support I never would have made it through graduate school.

Acknowledgments

First and foremost, I would like to acknowledge my advisor, Rich Shiffrin, who has been a wonderful mentor. He has devoted countless hours to the cultivation of this research and to helping me grow as a scientist, and I am truly thankful for both. I would like to thank Tom Busey for his willingness to provide insight, programming help, and resources for the neural portion of this project, as well as my committee members Rob Goldstone and Karin James for their guidance and useful discussions. Also, I want to acknowledge my lab-mates who have provided a great deal of assistance over the years in the form of discussions, programming advice, and help running subjects (especially George Kachergis).

Angela Nelson

Examining the Co-evolution of Knowledge and Event Memory

In this paper, we present 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 describe two experiments which provide data to inform our theory: these studies involve the development of new knowledge, and then testing in tasks involving episodic memory, retrieval from knowledge, and perception. Both studies train subjects to differential degrees on novel items (Chinese characters), then test the subjects in three transfer tasks: pseudo-lexical decision (or global recognition), episodic recognition memory, and forced choice perceptual identification. Experiment 1 uses a visual search training task, and found significant effects of training frequency in all transfer tasks. The structure of the visual search task however introduced differences in similarity between items of varying frequency, thus, Experiment 2 was conducted to isolate the effects of "pure" frequency. This was accomplished through the use of a same/different judgment training task, which eliminated contextual and similarity differences between items of differing frequencies. The results of the transfer tasks in experiment 2 still showed significant effects of training frequency, indicating a substantial role of pure frequency or raw exposure differences in the transfer task effects. The SARKAE model is presented, and shown to account for the effects of both experiments.

Contents

Examining the Co-Evolution of Knowledge and Event Memory

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.

Another example of the interaction that occurs between knowledge and event memory is the finding that semantic memory, or "gist" memory, can be retained while the specifics of an event or episode are forgotten, shown in the classic studies of Bransford and Franks (1971). Perhaps the most parsimonious interpretation of such findings posits incorporation in a stored event trace of general knowledge extracted from memory, and filling in gaps in information retrieved from an event trace by incorporating knowledge during retrieval. Perhaps the clearest demonstration of such an effect is recent research by Hemmer and Steyvers (2009a, b): They obtained ratings of environmental base rates for sizes of fruits and vegetables; other participants viewed objects, some novel and some fruits and vegetables. Later size judgments were distorted in ways consistent with the degree of prior knowledge and information in the base rates. In related research (personal report) base rates were obtained for objects likely to be found in kitchens; other participants saw kitchen scenes and tried to recall the contents. Recall was a mixture of event memory and intrusions from knowledge, well modeled in terms of the base rates. Closely related research by Brainerd, Wright, Reyna, and Payne (2002) also shows the interaction of knowledge and event memory in recall. Other research by Brainerd and Reyna (e.g. 1990; Brainerd Reyna, & Mojardin, 1999) investigate the use of gist memory in addition to more veridical memory ('verbatim') in recognition. The details of their model aside, the findings quite strongly show that storage and retrieval processes in recognition include a significant component due to knowledge. Brainerd and Reyna and their group have published many related studies pointing to the importance of these effects in children's memory retrieval.

Other developmental psychology studies provide additional insight into the co-evolution of event memory and knowledge. Rovee-Collier (1997) showed that whereas it has been previously thought that implicit and explicit memory systems develop at different rates in infants

during the first year of life, the two systems actually develop simultaneously. At the least such results are consistent with the view that the two systems develop conjointly and together. Such views provide support for a major theme of the present modeling, by which event memory and knowledge (sometimes referred to by implicit and explicit memory, or episodic and semantic memory, though these terms are used with different connotations in different contexts) lie along a continuum of memory traces, with event traces at one end of the continuum, for example existing 'alone' the first time an event is encountered (e.g. first encounter with a word), to developed knowledge at the other end of the continuum (e.g. adult knowledge of a common word). This view will be explicated in some detail as we lay out our theory in later sections of the article.

Modeling the Co-Development of Event Memory and Knowledge

There are many models of the storage and retrieval of event memories (e.g. – Atkinson & Shiffrin, 1968, Raaijmakers & Shiffrin, 1981, Shiffrin & Steyvers, 1997, Mueller & Shiffrin, 2006, Howard & Kahana, 2002, Grossberg, 1978, Anderson, 1983, Murdock, 1982). There are also models that explain retrieval of semantic memories (Quillian, 1967). A few models attempt to explain aspects of the way events produce knowledge (e.g. McClelland & Rumelhart, 1981, McClelland & Elman, 1986, Jones, Kintsch, & Mewhort, 2006), but these models primarily are aimed to explain memory and learning when knowledge has already formed (to various degrees). 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. This co-occurrence matrix accrues knowledge over time, represented as the number of co-occurrences observed for each feature pair. The REM-II model (which is a considerable elaboration of the REM model of Shiffrin and Steyvers, 1997)) describes the interaction between episodic memory and semantic memory, and accounts for phenomena such as polysemy and connotation effects. 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 article. As we shall see, 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.

The most 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, in which a single event trace may be stored initially, and gradually gains added information as similar events occur during experience, until a rich knowledge trace results. Of course if one only looks at the ends of this continuum, a single event trace compared to a mature knowledge

trace, these can appear quite different in their effects on storage and retrieval, as seen in a variety of dissociations (Jacoby & Dallas, 1981, Neely, 1989).

In SARKAE (as in REM-II), 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 from the physical properties of the item or concept being learned, as well as features drawn from the context surrounding the item during learning, modified and governed by attentional focus. Such features include other events nearby in time and attention, and the various components of internal and external context that numerous investigators have discussed for many years (Estes, 1955, Godden & Baddeley, 1975, Klein, Shiffrin, & Criss, 2007). Thus for example 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 no one event stands out and the knowledge seems context free.

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. The 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).

Most of these mechanisms that produce storage of event memory and knowledge 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 punctate, 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).

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 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. In recent years we have found it particularly useful to characterize the match of 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, 1980, 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. Much of this article will cover the theory in detail, but one cannot sensibly construct a theory without data to constrain it. Thus we present first studies that will inform the theory. These studies involve the development of new knowledge, and then testing in tasks involving episodic memory, retrieval from knowledge, and perception. The studies focus on the effects of differential experience

during training, because such effects are omnipresent in cognition, and the mechanisms for such effects are presently an issue under investigation in the field. The studies reported here are of course insufficiently rich to allow detailed theory specification, so even our simplified theory incorporates key concepts that we view essential due to prior research, due to conceptual coherence, or due to previous good results in applications of the REM theory (a theory that has been shown to give good accounts of memory, priming, and knowledge retrieval; see for example Shiffrin & Steyvers, 1997, Schooler, Shiffrin, & Raaijmakers, 2001, Wagenmakers, Steyvers, Raaijmakers, Shiffrin, van Rijn, & Zeelenberg, 2004).

It is wise to warn the reader that the model will be applied to five quite disparate tasks. Although we endeavor to incorporate the same processes in the models for the different tasks, whenever such common processes are appropriate, different tasks of course require models that are individually tailored to them, using task specific assumptions. Even in simplified form, when task specific assumptions are laid our in sufficient detail to allow quantitative modeling, the result will be a fairly elaborate model with high complexity. We will try to highlight those processes that are fundamental to the SARKAE approach and the goal of explicating the co-evolution of event memory and knowledge, and distinguish those key assumptions from the sometimes ad hoc and simplified assumptions needed to model the individual tasks.

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. Thus we will vary presentation frequency in the present studies. In particular, novel characters are trained to

different degrees for an extended period and then tested in tasks requiring perception, or retrieval of event memories or knowledge.

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', or WFE, especially when found in recognition memory (Glanzer & Adams, 1985, Glanzer & Adams, 1990, Kinsbourne & George, 1974). In episodic recognition memory tasks, words that occur rarely in the environment are recognized *better* than words that occur frequently in the environment. This effect has been called one of the regularities of recognition memory (Glanzer, Adams, Iverson, and Kim, 1993). Word frequency has also been shown to have effects on recall performance (high frequency words are recalled better (Gregg, 1976)), and perceptual tasks such as lexical decision and perceptual identification (forced choice, etc.). In lexical decision, high frequency (HF) words are identified both more accurately and more quickly than LF words (Becker, 1979, Rubenstein, Garfield, & Millikan, 1970, Scarborough, Cortese, & Scarborough, 1977). Perceptual ID shows a more complex pattern of results, with the general findings in two alternative forced choice showing that HF targets are better identified, and both HF and LF targets are better identified when paired with a LF foil (Wagenmakers, Zeelenberg, & Raaijmakers, 2000).

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 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. Several previous studies have used training to examine the effects of experience on memory and perception. Maddox and Estes (1997) trained subjects on letter and number strings using a memory task. The frequency of presentation of the stimuli in the memory task was varied such that the strings were familiarized to varying degrees. This training phase was followed by an episodic recognition memory task. The results of this study indicated that both hits (correctly responding "old" to a studied item) and false alarms (incorrectly responding "old" to an unstudied item) increased as a function of familiarity (as measured by training

exposure). A training study by Reder, Angstadt, Cary, Erickson, and Ayers (2002) also found differences in post-training memory performance due to training frequency. Their study used pseudowords as the stimuli, and trained the subjects on the pseudowords to different degrees using a free recall task. The subjects were tested several times throughout the training period, and the results showed that early in the training increased familiarity resulted in increased hits and false alarms (replicating the results of Maddox and Estes (1997)). However, later on in training when recognition was tested again, the results showed a mirror effect: more hits and fewer false alarms occurred for low frequency trained pseudowords compared to high frequency.

These studies provide valuable background for our research, but are not quite ideal as a basis for theory development. For one thing, the letter and number strings used by Maddox and Estes (1997) and the pseudowords used by Reder et al. (2002) were only partially novel, and are related to a good deal of alphanumeric existing knowledge. Previous studies have shown that in addition to the effects of the frequency of the entire word, the frequency of single letters, such as those used in the letter and number strings, can affect recognition memory (Malmberg, Steyvers, Stephens, and Shiffrin, 2002). Pseudowords also contain parts of words, and bigrams and trigrams that differ in frequency in the language, factors known to affect performance in lexical decision (Rice & Robison, 1975). These stimuli could therefore produce differing performance due to differential interference based on bigram/trigram frequency, and even meaning, to the extent that a pseudoword reminds the viewer of a word or words in the lexicon. In order to better control such factors, 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).

In a study by Nelson and Steyvers (2004), subjects were trained on Chinese characters for seven sessions. A recognition memory task was used for both training and testing, but produced results that were difficult to interpret. It could well be that use of the same task for training and testing produced interactions between the two phases of the study that obscured the underlying processes. Related concerns could be raised about the studies by Maddox and Estes (1997) and Reder (2002). The studies reported in this article therefore use a training task that is as different as possible from the subsequent transfer tasks. 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.

Experiment 1, Part One: Training

Method

Subjects. Eight people, recruited through an email advertisement, participated in the experiment for monetary compensation. All participants reported no prior experience with Chinese characters.

Apparatus. All tasks were displayed on Samsung SyncMaster 700NF 17" flatscreen CRT monitors, and responses were collected through keyboard presses. Experiments were run using the programs Authorware and MATLAB. Participants were seated in dark booths with ventilation fans that greatly reduced ambient noise.

Procedure. The visual search task required the participants to judge, as quickly as possible without making more than a few errors, whether a single Chinese character presented just before a display was present in a subsequent display of two or four Chinese characters. A varied mapping procedure was used, so that targets on some trials were foils on others, and vice versa. Each trial was initiated by a key press, which was followed by a fixation cross for 500 milliseconds. The cross was followed by a target character presented centrally for 1000 ms; the target was then replaced by a blank screen for 500 milliseconds. Then a display of either two or four characters appeared and remained until a response was made. The characters each subtended about 3.5 degrees visual angle vertically and 2.9 degrees horizontally. For the display size of four, the characters were positioned evenly in each quadrant of the screen, in a square pattern, with a separation of about 4.3 degrees visual angle. For display size two, the characters were randomly placed in two of the four possible positions. The procedure is illustrated in Figure 1 with two sample sequences: a) display size two with target present; b) display size four with target absent. Half the trials used display size two, and half of each type had target present. There were a total of 640 trials per session, and 12 visual search sessions were completed by each subject, over the course of roughly two weeks.

Design and Stimuli. The occurrence of characters were permuted so that some occurred more often than others: There were four frequency conditions, with different characters occurring in a ratio of 1::3::9::27. These same ratios held for occurrence of a character as target or foil: In each session, for every occurrence of a character as a target, it was also present five times as a foil. For each participant, a set of 32 characters was selected randomly from a pool of approximately 200 characters. In order to keep the complexity of the characters similar, all characters were composed of seven strokes or less. Figure 2 shows a sample of eight characters.

From the 32 characters for a given participant, eight were randomly assigned to each frequency condition. The foils for each trial were of mixed frequency. The permutation was arranged in a block of 160 trials, and there were 4 blocks in each session. Thus there were 640 trials per session, and 7680 trials per participant at the end of training.

Results

The principal measure used to analyze learning over training sessions was the slope of the search function, calculated separately for present and absent trials. Slope was defined as half the difference between response time for display sizes of two and four. Figure 3a shows mean slope per session, averaged over the 8 subjects, as a function of session number. The slopes show a decrease over training, beginning at approximately 100 ms/item and dropping to 60 ms/item for present trials, and falling from 220 ms/item to 150 ms/item for absent trials. Figure 3b shows the estimated zero intercept of the search function, defined (for present trials) as the mean response time to a present trial of display size (4 or 2) minus (4 or 2) times the present slope (the result when size 4 vs. size 2 was used was averaged). The intercept for absent trials was calculated the same way. Like the slopes, the intercepts showed improvement over training; approximately, the positive intercept dropped from 700 ms to 475 ms, and the absent intercept dropped from 550 ms to 400 ms. The intercept is usually taken to include various perceptual, encoding, decision making, and motor response components that may be independent of display size, and therefore might not demonstrate character learning. The slope is usually taken to reflect processing time per character in a serial or limited capacity search, and is a better measure of character learning.

When separated into frequency groups, the slope patterns are similar, but the high frequency items show a more pronounced and clear decrease, likely because there are many

more data points for these items. It might be expected under some learning models that search time per character would vary with training frequency. Under other models this would be a less clear prediction, because analysis time for a given display character might depend on the alternative characters that were, or could have been, present, so that search time would reflect overall character learning for the entire set. The data did not exhibit clear slope or intercept differences by frequency, but the very limited amount of data at the lowest frequencies make meaningful inferences difficult, so these data are not shown here (they are presented in Appendix A).

Discussion

Training produced a clear and pronounced improvement in search rate and search intercept, providing evidence for the development of knowledge about the initially novel Chinese characters. Although the search results do not demonstrate frequency differences, they do not rule these out. In any event findings of differential results by frequency in the transfer tasks will be sufficient to prove that training did indeed produce frequency effects.

The present search results can be compared to those found in the study by Shiffrin and Lightfoot (1997). That study did not vary frequency, but instead carefully controlled features of each stimulus, because the aim was exploration of perceptual learning. There were just three simple and spatially distinct features per stimulus (three line segments pointing inwards from the periphery of a rectangle). Further, no one of these features by itself could produce successful search, because targets and foils always shared exactly one feature (a conjunction search was required on every trial). That study showed a reduction of search slope over training from around 270 ms per stimulus to about 90 ms per stimulus, interpreted as a shift from initial sequential

consideration of each of the three features for each display stimulus, to eventual consideration of each entire character in one search step. This perceptual unitization was verified in a wide variety of subsequent transfer tasks. In contrast, the featural composition of the present Chinese characters was not controlled, and by inspection is obviously quite complex. Furthermore, the randomization of stimuli across participants and conditions insures that any feature specific effects would differ across participants, and probably disappear in the average results. What learning consists of in the present task is an open question. The task requires that on each trial the target be distinguished from all foils. Because any of the training characters can be foils on any trial, it seems likely that the participant will try to identify and learn a feature or feature combination that will uniquely identify each character in the sense that it will distinguish that character from all the others. (Note that such a requirement, being based on the composition of the entire set of characters, could result in a reduction or elimination of frequency differences in training). The distinguishing feature combination could be a quite simple component of a character, or as complex as the entire character itself, depending in good part on the character set for a given participant. Because pre-existing feature combinations in each participant's visual knowledge might serve in some cases as distinguishing features, and because conjunction search is not required by the present task, full perceptual unitization, and its attendant large slope changes, might not be expected. Thus it is not surprising that the changes in slope with training in the present task are smaller than that seen in Lightfoot and Shiffrin (1997).

Experiment 1, Part Two: Post-Training Tasks

Following the training on the visual search task, the subjects completed three post-training tasks: episodic recognition, pseudo-lexical decision, and forced choice perceptual identification.

Episodic Recognition

Method

Subjects. All eight subjects who were trained on the characters completed this task shortly after their final training session.

Design and Procedure. This 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

Performance on the episodic recognition task was measured in terms of the hit rate (probability of correctly identifying a studied item as 'old') and false alarm rate (probability of incorrectly identifying a non-studied item as 'old'). Performance of individual subjects as well as performance averaged over all subjects was analyzed. All subjects showed better performance for low frequency trained characters than for high frequency trained characters. The average performance also produced a mirror pattern: more hits and fewer false alarms for low frequency items (see Figure 4). A contrast analysis showed that there was a significant negative relationship between frequency and hit rate, and a marginally significant positive relationship between frequency and false alarm rates. Characters of zero frequency produced performance intermediate between the levels for trained characters (see Appendix A for detailed description of analysis and statistics).

Discussion

In episodic recognition tasks using words as stimuli, it is reliably found that low frequency items produce better performance, and that a mirror effect occurs: low frequency words produce more hits and fewer false alarms than high frequency words. Many theories have been proposed to explain both the frequency effects and the mirror pattern, and include such factors as attention (Glanzer & Adams, 1990), context (Sikstrom, 2001), a dual-process of familiarity and recollection (Reder et al., 2000), and the Bayesian-based retrieval models of Shiffrin and Steyvers (1997) and McClelland and Chappell (1998). The present results do not clearly distinguish the competing theories, nor was that their aim. They do show that frequency

and the difference in amount of exposure, and/or the correlated character context (the character context mainly consisting of that trial's foils and perhaps the previous trial's target) was enough to produce the classic word frequency effects: The low frequency trained characters produced better performance, and a mirror pattern also occurred. Further discussion will be deferred until the presentation of the SARKAE model and theory in following sections.

It is commonly found when words are used as stimuli that non-words or very low frequency words (effectively non-words for many participants) do not fall on the same function as other words. Typically such words exhibit performance intermediate between low and high frequency words (e.g. – Estes & Maddox, 2002). Our data for untrained characters show a similar pattern. It is perhaps unsurprising that untrained items, whether words or our characters, would produce performance inconsistent with the trend for trained characters. Untrained items do not have a knowledge trace, and therefore might be processed with a different set of mechanisms than items that do. For example, at both study and test the features attended and used might be restricted to low level physical characteristics, and/or features extracted from knowledge from traces of trained items that are similar. Processing of trained items with knowledge traces would undoubtedly use the contents of those knowledge traces at both study and test.

Whatever the processes at play, the replication of the recognition patterns found for words increases the likelihood that these processes are similar for the two types of stimuli. To the degree that this is so, one can discount explanations for the word data that rely on other factors than exposure frequency and the word context that is correlated with frequency.

Pseudo-Lexical Decision

Method

Subjects. All eight subjects who were trained on the characters completed this task shortly after their final training session.

Design and Procedure. Subjects viewed one list, which contained all 32 trained characters, 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. The placement of the characters in the list was randomized. Subjects were presented with a single character on the screen, and were asked to decide as quickly as possible whether they had ever seen that character during any of the previous training sessions. Responses were made by pressing either the 'v' or 'm' button on the keyboard.

Terminology. Lexical decision tasks require distinguishing words from non-words. Our task involves Chinese characters rather than words, hence the prefix.

Results

Response times and accuracy were analyzed over the four separate frequency groups, as well as the new items. For response times, the analysis showed that the higher frequency characters produced significantly faster response times than the lower frequency items (see Table 1). The average response time for new items was 820 milliseconds. Accuracy also differed by frequency group; Table 1 shows that the accuracy of response was higher for the high frequency items than the low frequency items. New items showed an accuracy level of approximately 93 percent. A contrast analysis showed that there was a significant positive relationship between frequency and accuracy, and a significant negative relationship between frequency and response

time (see Appendix A). Separate analyses broken down by test position of the same character were somewhat noisy for accuracy, but showed a decrease in response time for later tests. This data is given in Appendix A.

Discussion

The results from the pseudo-lexical decision task indicate that the degree of experience with a character, and/or the character context that is correlated with frequency in our tasks, is responsible for the pseudo-lexical decision frequency results. These findings align with lexical decision results for words in previous studies (Becker, 1979, Rubenstein, Garfield, & Millikan, 1970, Scarborough, Cortese, & Scarborough, 1977). Thus our frequency findings for Chinese characters are not due to any of those other factors correlated with word frequency. One could also reverse this inference and argue that it is likely that the word results are mostly due to the same basic processes that produced the present results. One way to understand the processes involved in pseudo-lexical decision (and perhaps lexical decision as well) will be laid out in the modeling sections of this paper.

Forced Choice Perceptual Identification

Method

Subjects. Six out of the eight trained subjects completed the forced choice perceptual identification. The task was administered several months after completion of the initial training, so the subjects completed three sessions of re-training on the characters prior to the task, using the same visual search task as was used in the previous training sessions. The slope and intercept

of the search function was measured to assure that the subjects were at the same level of performance as when they had completed the previous tasks.

Design and Procedure. This task consisted of five lists, each with 50 two-alternative forced choice trials. The first list was used to adjust the length of target presentation to a 75 percent correct threshold, using the Best-PEST algorithm (Lieberman & Pentland, 1982). The average length of target presentation was 67.8 ms. Each subject's individual threshold presentation speed was used for the four test lists. Throughout the task, every combination of foil and target frequency was tested (frequencies 0, 2, 6, 18, 54), creating a total of 25 conditions.

For each trial, subjects viewed a target character presented briefly in the center of the screen which was immediately covered by a mask stimulus. The mask consisted of a jumbled mix of Chinese character pieces. After the mask, the subjects were presented with two choice characters: one on the right side of the screen, the other on the left. The subjects were asked to choose which of the two characters matched the target character that had been presented immediately prior. These two characters stayed on the screen until a decision was made, and the correct answer was always one of the choices. Subjects completed one block of 50 speed adjustment trials and four blocks of 50 trials at their established presentation speed. Only data from the last four blocks were analyzed.

Results

The proportion of correct responses was measured for each condition of target frequency and foil frequency. The results showed that when target frequency increased (averaged over all foil frequency conditions), performance increased. The same was true for foil frequency: when the frequency of the foil increased (averaged over all target frequency conditions), the

probability of responding correctly increased (see Figure 5). Both of these effects were marginally significant (see Appendix A).

Discussion

The first portion of these findings agrees with what is found in word frequency literature: when the frequency of the target word increases, performance generally increases (Broadbent, 1967). However, the second portion of our findings is slightly harder to explain: when the foil is higher frequency, performance is also increased. When words are used for this type of task, the frequency of the foils produces a much more complex pattern of results, and usually a high frequency foil will hinder rather than help performance (e.g. Wagenmakers et al., 2000). Why words and the Chinese characters in the present task show a different pattern for foil frequency is not clear. Regardless, one explanation for the present findings for foils posits a negative inference: Suppose the participant or the cognitive system takes into account the fact that high frequency targets are easier to perceive correctly. If so, and if nothing or almost nothing is perceived on a given trial, then it would make sense to guess that a low frequency choice had been presented (on the 'reasoning' that a higher frequency target would have been seen). This idea will be elaborated in the modeling discussion that follows.

A Version of the SARKAE Model Applied to Experiment 1

The visual search task used for training varied character frequency, but the randomization of trials and foils insured that higher frequency characters 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. Following

Experiment 2 we will present the model SARKAE that posits separate and at least somewhat independent roles for both character-context and pure frequency. However, following Experiment 1 we decided to test the adequacy of a model with only one factor. Because Adelman, Brown, and Quesada (2006) proposed that only word diversity, not word frequency per se, was responsible for most word frequency effects, we decided to see if such a model could be used to fit our findings. It was the perhaps surprising success of this approach that led to the formulation of Experiment 2. That study showed the need for both factors, and the model was subsequently elaborated. It nonetheless may aid understanding to make a few brief remarks at this point about the single factor model.

Both the one and two factor versions of SARKAE are conceptually similar to the REM-II model of Mueller and Shiffrin (2006) and in certain respects to the TCM model of Howard and Kahana (2002). Each item is represented as a vector that has a fixed number of features, and a fixed number of possible values per feature. The feature values do not have a frequency associated with them before training. The features that make up an item are a combination of many low-level features, and one high-level feature (these features represent only the item itself but not context; context features are a separate set). The high-level feature of an item is unique to that item; no two items share the same high-level feature. The high-level feature represents the distinguishing feature that we have argued earlier is learned in order to accomplish visual search efficiently.

Using this representation, the knowledge trace is built up over training in the following fashion: Each time a character is presented as a target, features are added to its knowledge trace from four possible sources: 1) the actual target item features; 2) the features of the foils from that trial; 3) the features of the previous target item; 4) guesses based on base rates in the knowledge

base. Sources 2) and 3) represent the nearby character context for a given target character. Because high frequency (HF) characters occur more often, they build up more counts in the vector representations (see Figure 6). In addition, because high frequency (HF) characters make up most the character-context for any character of any frequency, it is the feature values of HF characters that dominate the contents of the developing knowledge traces. This fact changes the inter-character similarity between characters of different frequencies. As one way to illustrate this, Figure 7 shows the normalized inner product between the knowledge traces of two characters of stipulated frequencies after training is complete. This figure shows that traces of higher frequency have become more similar to each other. This similarity effect is not due to the differing number of counts in the various traces, because the comparisons are normalized for total counts. Thus it is the similarity of the patterns of counts across features that produces the similarity findings.

How did such similarity effects come about? Context for a given character during training consisted of the most part of features gathered from the surrounding foil items in the visual search task. These foils tended to be high frequency, so the features gained by a character's knowledge trace tended to be drawn from other high frequency items. In addition, a high frequency item occurred more often and hence in a more diverse set of contexts. Together these factors produced frequency dependent similarity effects.

In the application of SARKAE to Experiment 1, it was this similarity effect that was represented in knowledge at the end of training that produced good predictions for the various post-training transfer tasks, even though no role was posited for frequency per se (i.e. the total number of counts in a trace). One might ask how it is possible to implement a SARKAE like model in such a way that the number of counts in a knowledge trace plays no direct role. This

was accomplished by making both the probes of memory, and the retrieved knowledge traces to which the probes are compared, of similar 'density': In particular, retrieved knowledge traces are formed by sampling at most one feature value per feature, even when a feature and its values has multiple counts. Although the one factor SARKAE model fit the results of Experiment 1, the confounding of frequency and character context made inference about causal mechanisms uncertain, and hence led to the design of Experiment 2. The results of that study demonstrated the need for frequency per se to be taken into account, leading to the full model that will be presented later.

It is useful to note here another important component of the one-factor SARKAE model that will be changed in the full model: The one factor model applied to episodic recognition memory assumed that the test probe only accesses the traces of characters presented on the list. Higher frequency items were predicted to produce lower performance because the list traces were more similar to higher frequency probes, thereby producing more confusions. The design of Experiment 2 eliminated frequency-dependent within-list similarity differences, so this explanation of frequency effects in recognition would not be tenable. As shall be seen, frequency effects were indeed found in episodic recognition in Experiment 2, so the recognition model had to be augmented.

One other note is useful at this point. The sampling scheme mentioned in the paragraph two back has another important property—it enables us to calculate likelihood-ratios for each trace-to-probe comparison in a consistent and coherent manner, resulting in a set of nice properties true of the REM approach. Thus we will retain this sampling scheme in only slightly altered form in the full SARKAE model.

Experiment 2: Eliminating Character-Context Effects of Training[1]

Experiment 2 used a training paradigm not involving visual search. Participants are trained using a same vs. different judgment task: A character is presented briefly twice in succession, and half the time the two presentations vary 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. Experiment 2 used the same variation of training frequency as Experiment 1. 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.

[1] Between Experiments 1 and 2 an attempt was made to reduce character-context effects in a study that used a visual search paradigm. This study used visual search training, but items of a given frequency always occurred with foils of that same frequency. This manipulation was not sufficient to remove frequency dependent similarity effects, according to the simulations of the model. It became clear that the paradigm of visual search made it difficult to remove all context effects, thus leading to the paradigm introduced in Experiment 2. This experiment is detailed in Appendix B.

Training: Same/Different Judgments

Methods

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

Apparatus. All tasks were displayed on Samsung SyncMaster 700NF 17" flatscreen CRT monitors, and responses were collected through keyboard presses. Experiments were run using MATLAB. Participants were seated in dark booths with ventilation fans that greatly reduced ambient noise.

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. 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.

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 8 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.

Discussion

The increases in performance during training are sufficient to show that something about the characters is being learned. One could question whether more is being learned for higher frequency characters, but the staircase algorithm does not adjust separately for different

frequencies, so it is impossible to say for sure. However, analysis of visual search training in experiment 1did not show any significant differences in training results based on frequency, suggesting that the same may be true here. One might also question whether the nature of the present task causes more shallow character representations to develop than develops due to visual search, possibly reducing the importance of frequency variations. However, both questions become moot if the transfer tasks show frequency effects (they will), so this issue will not be discussed further here.

Post-Training Tasks

Following the training on the visual search task, the subjects completed three post-training tasks: pseudo-lexical decision, episodic recognition, and forced-choice perceptual identification. Testing was again carried out 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

Method

Subjects. All seven subjects who were trained on the characters completed this task shortly after their final training session (within approximately 2-3 days), and again approximately 6 weeks after their final training session.

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. The placement of the characters in the list was randomized. Subjects were presented with a single character on the screen, and were asked to decide as quickly as possible whether they had ever seen that character during any of the previous training sessions. Responses were made by pressing either the 'v' or 'm' button on the keyboard.

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 9. A contrast analysis showed that there was a significant positive relationship between frequency and accuracy, and a significant negative relationship between frequency and response time (see Appendix A).

Response time and accuracy were measured again 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 positive relationship between accuracy and frequency, and a significant negative relationship between response time and frequency (see Figure 9). 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 (see Appendix A for statistics.)

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. An explanation will be laid out in the later exposition of the SARKAE model.

Episodic Recognition

Method

Subjects. All seven subjects who were trained on the characters completed this task shortly after training, immediately following the lexical decision task described above. This task was also completed approximately 6 weeks after the completion of training.

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 10. Similar to post-training results discussed in the previous experiment, when tested shortly after the completion of training, false alarms significantly increased as frequency increased (panel A). There was also a marginally significant decrease in d' due to frequency. The hit rate analysis however showed no significant effect of frequency (see Appendix A for statistics).

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 10. 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 significantly larger immediately after training (see Appendix A for details).

Discussion

When tested shortly after the completion of training, the results in the episodic recognition task are similar to results found in Experiment 1 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 experiment 1, experiment 2 did not show a significant effect of frequency on hit rates. However, previous work using normative word frequency manipulations in this task has shown that the effect of frequency on false alarm rates is much more robust than the effect on hit rates, which only surfaces a portion of the time (Criss & Shiffrin, 2004b).

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 Experiment 1. 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 presented next provides an explicit role for frequency per se (especially to explain pseudo-lexical decision findings) and an elaborated model for recognition.

The SARKAE Model: Co-Development, Event Memory, Knowledge Retrieval, Perception

It is easy in the following descriptions to lose sight of the central theme of the modeling effort: Knowledge develops through a process of accumulation due to re-occurrences of similar events. Each event produces an event trace, but also addition to the developing knowledge trace. This process is instantiated by either of our training methods, with the result that event traces and knowledge traces form for the Chinese characters. The two training methods produce different knowledge traces with a different inter-character structure. The knowledge traces that result at the end of training then form the backbone upon which the processes of the transfer tasks operate, for storage, perception, and retrieval. The ways in which this knowledge forms and is then used in the recognition, pseudo-lexical decision, and forced-choice perceptual identification tasks is laid out in the sections that follow.

The results of Experiment 2 demonstrate that frequency per se plays a large role in event memory, knowledge retrieval, and perception, even when training has controlled context so that similarity between characters should not be dependent on frequency of character training. This finding led us to produce a version of the SARKAE model includes separate roles for contextual diversity (operating in our model through the similarity structure of the knowledge traces) and frequency per se. There are many (potential) ways by which frequency could influence storage and retrieval. It is unlikely that the present results are sufficient to force convergence on any one such method, and we therefore chose the simplest implementation that seemed likely to be successful: The gathering of features from a knowledge trace (whenever knowledge is accessed) is assumed to be more accurate for more highly developed traces (i.e. – traces of high frequency items). The resultant model is applied to the results of both studies.

General Assumptions

Novel characters are represented by a vector of 20 character features, each with eight values, each feature having a ‘one’ for its assigned value and ‘zero’ elsewhere. These assignments are initialized randomly for each character. These represent the ‘physical features’ of a character prior to learning and training. General situational context (incorporating list context) is assumed to be represented by 30 features, each with eight values. In any reasonable model the context values should change as training continues, and as the shift is made from training to transfer tasks. However, introducing an explicit model of context change would add considerable complexity to the model, and is not needed for the present applications. Instead some simple approximations can be used to capture the critical aspects that are needed. Thus we simplify by not explicitly modeling the development of a context portion of the lexicon during training. One adjustment to this simple assumption is later added to model the effects of delay in Experiment 2. Like item features, each context feature has a single ‘one’ in a specified value and ‘zero’ elsewhere. For a given simulated participant the marked values are assigned randomly. Of course noise in storage will assure that different traces do not have identical values, as described in the following.

Both a character’s episodic trace and its knowledge trace will be constructed with these 20+30 features, but the values stored will not necessarily match the originally specified values, because storage is incomplete and error prone. Each event trace will have some features with all zeros (incomplete storage) and some features with a single value marked with a one. The knowledge trace accumulates features and feature values over time, so that given enough training all features would come to have all values marked with at least one ‘one’, and the number of counts for any value would increase as training continues. This is illustrated in Figure 6, albeit

with fewer features and feature values than in the actual simulation. The figure shows an simplified example character representation (with item and context features) and two example knowledge traces, one for a character after relatively few occurrences, and the other for a character presented relatively many times. For any given presentation of a character, some character features may have values stored and some may not. If a value is stored, the value may be copied from that feature's value in any of several sources -- the character's representation, the character's knowledge trace, other characters nearby in time and space, or base rates in the knowledge base. Similarly, for any given episode, a context value can be stored or not. If stored, there are two possibilities for the source -- the set of specified context values, and the base rates in knowledge (given our simplifying context assumptions, the storage of context features is only explicitly modeled in the episodic recognition task).

In addition to the low-level item features, each item also has one high-level feature, which distinguishes it from all other items. Although not strictly needed for Experiment 2, given the change in training paradigm, it does no harm to use such a feature for both studies. In addition, the parameter which governs the attention given this high-level feature is tuned appropriately to training type: when training consists of visual search, this high level feature is given more attention. Considering the nature of visual search, the process of forming such a high-level feature for novel items such as Chinese characters is likely very complex. However, given the lengthy period of training it seemed reasonable to assume that the relevant high level feature was formed early in training. We therefore simplified by assuming that the high level feature is formed instantly and is present from the outset of training. Because the high-level feature has a unique value for every Chinese character, it is given as many values as characters given training, a number of values that is larger than that for each of the other features in the

simulation. The rules for storing feature values during training for the low-level and high-level features are the same with one exception: Because participants are likely to give more attention to the high-level distinguishing feature identifying a target character than other features, we assume the high-level feature is more likely to be stored in the knowledge trace and episodic trace of the character presented as a target on that trial.

As far as possible we made the assumptions governing storage of features and feature values, and the construction of vectors with which to probe memory, consistent across experiments, and across training and transfer tasks. However, the differences among tasks, and between storage and retrieval, require some slight differences in these assumptions, so they are described in detail in the following exposition, in association with each setting.

We apply the model to two different training tasks (visual search, and same-different character matching, three immediate post-training transfer tasks representing different tests of retrieval (event memory, knowledge retrieval, perception), and two six week delayed transfer tasks. These tasks come close to spanning the domain of cognition, so it is a non-trivial exercise to model all these with the same assumptions let alone the same parameters. When possible we keep parameter values constant across conditions for which it seems reasonable that the same values should apply. We will present the model in segments, as it applies to each task in turn, introducing the parameters during the discussion. As we present estimated parameters we will indicate which have been estimated for multiple tasks or conditions, and which estimated separately.

Training

When an item is presented as a target during training, either in visual search or the same/different judgment task, in principle two types of storage occur: An event trace is stored, and the same types of information stored in the event trace are stored in that item's developing knowledge trace. We are not modeling event storage in detail, so the following storage rules define only the way that information is added to the developing knowledge trace. Upon presentation of an item as a target during training (in either experiment), each of its features have a value stored (i.e. added to the counts already in knowledge for that item) with probability *u*. With probability *1-u* no value is stored for that feature for that presentation.

For the search task, if a value is stored, the source of the feature value is the physical feature value of the target with probability *ti*, the physical feature value of a randomly chosen foil with probability *td*, and the physical feature value of the previous target item with probability *tp*. The values of *ti*, *td*, and *tp* sum to one. In the same/different judgment task, *td* is set to zero, because there are no foils; the parameters *ti* and *tp* are then renormalized to sum to one. Whatever the source, the feature value is copied correctly from that source with probability *c*. If not copied correctly, the value stored is chosen randomly in proportion to the base rate for that feature and feature value in the knowledge base[2]. Both high-level and low-level features are stored in this fashion, but we let the *u* parameter be estimated separately for high-level features, on the theory that such features might be given more attention. This difference between *u* for low level features and *u* for high level features is estimated separately for the two experiments, as the type of training (visual search vs. same/different judgment) is almost sure to influence the

[2]We simplify by using the base rate assessed at the end of training, but this assumption makes little if any difference for the present applications.

attention given the high level feature of an item. These assumptions implement our conceptual view that storage is incomplete and noisy, and incorporates features from the nearby spatial and temporal surround of an item.

The storage of physical feature values is implemented trial by trial, as described above. Conceptually something similar could be done for context features, but the end result would be a noisy uniform distribution across features and values, so the storage of context in knowledge is not simulated on a trial by trial basis. Though context as stored in knowledge is never utilized in the current simulations, we assume that the representation of context in knowledge built over training forms a uniform distribution.

These storage rules produce quite different knowledge structures for the two experiments. For the visual search task a similarity structure develops in which high frequency items grow to be more similar to all other items and most similar to each other (see Figure 7). In the matching task, there is no character context to bias storage toward experimental neighbors of high frequency. Thus the similarity structure no longer varies systematically with frequency. Figure 11 shows the dot product between normalized knowledge traces following training, and shows no systematic variation with frequency – all similarities are approximately equal.

Transfer Tasks after Training

Following training, the knowledge traces are used in the retrieval and/or storage processes for the various transfer tasks. The knowledge traces contribute in a variety of ways as outlined for each task next. Due to the change of knowledge across the experiments, some

parameters are estimated separately for the two tasks, and a few assumptions are adjusted as indicated.

Recognition Memory

Within-list Confusions and Extra-list Confusions: Item and Context Noise

Before turning to the way SARKAE implements recognition memory, it is useful to discuss in more general terms the factors that affect recognition performance and the influence of frequency. This provides useful background for the model.

Previous experiments in recognition memory have shown that items that have been presented more recently but were not on the list that is currently being tested are more likely to produce a false alarm (Criss & Shiffrin, 2004a). That is, the more recently that an unstudied item has been seen, the more likely a subject will incorrectly call that item "old" when presented with it on the test list. The experiment by Criss and Shiffrin (2004a) showed this through manipulating an item's presentation on previous study lists; however, it follows that the recency of an item's presentation in any similar circumstance (not just another study list) should have the same effect (see Shiffrin and Steyvers, 1997, and Dennis & Humphreys, 2001). The preliminary model applied to Experiment 1 assumed activations of present-list traces only, and as we shall see, this approach is inadequate. In previous research (Shiffrin and Steyvers, 1997) it was pointed out that an approach that involved activations of every trace in memory event traces would run into trouble, so that there would likely be a probabilistic cutoff based on degree of match to the probe, and only event traces exceeding the cutoff would take part in the recognition decision. In the present setting we assume for simplicity that when recognition testing occurs, traces of characters from training sessions (or previous test lists) that are different from the test

character will not be activated. The reasoning is that such traces will differ in both character features and context features, and thereby fall below the cutoff for activation. Thus when character A is tested SARKAE assumes activation of all event traces from the list, whether they were due to study of A or other characters (these match well in the many context features), and also activation of a portion of A traces from the training sessions (these match well in character features). Presumably the activated traces come mostly from recent training sessions, but because we are not specifically modeling context change during training, we simply estimate the number activated. However, in each training session the number of event traces of a character is linearly related to its frequency of training. Thus we assume the numbers of such extra-list traces activated will be in proportion to the training frequency.

The model we use assumes decisions are based on summed activations across all activated traces. The extra activations of higher frequency items from training sessions will cause confusions that lower performance for such items. This factor will operate in both experiments. Having said this, it is critical to note that it is not the activation of these extra list traces that causes confusions, but rather the variability of encoding associated with those traces. The reason is simple: Both high frequency foils and high frequency targets have the same number of extra list traces in memory.

To understand this it is useful to consider the situation in terms of signal detection theory. Although SARKAE codes activations as likelihood ratios, it is easier to follow the following reasoning if we imagine activations are based simply on activation strength that increases monotonically with probe-to-trace match. First consider the distribution of the sum of such activations for targets, and the distribution of the sum of such activations for foils, when extra list traces are not activated, or ignored. We would have two overlapping distributions with targets

having higher average familiarity (targets have a list trace that usually matches well). Now suppose we add activation of extra list traces. Both targets and foils have such traces, and hence the average familiarity rises equally for both—the two distributions shift upward together, but the difference between the means of the distributions does not change (see Figure 12). By itself this shift in means would not produce a change in performance (measured say by d', or the overlap of the two distributions). However, the extra-list traces that are activated are not all equal – the storage process is very noisy, and these traces vary considerably from each other. Hence the familiarity they contribute adds noise, and causes the distributions to have higher variance. Higher variance causes more overlap, and lowers performance.

The final piece of the story is the effect of test item frequency. The amount of added variability rises with the number of extra-list traces activated. This number is higher for higher frequency items, causing more variance and lower performance. This effect of frequency due to extra-list traces is similar for both experiments. However an extra effect of frequency occurs in Experiment 1: In that study test items of higher frequency better match list traces (due to higher similarity induced by visual search training). Thus SARKAE predicts a greater effect of frequency for Experiment 1 than Experiment 2, the difference depending on the parameters chosen.

To summarize, a test character activates: 1) list traces (mostly of different characters, but half the time including a self-trace) and 2) extra-list self-traces in numbers reflecting frequency. For Experiment 1 the list traces are differentially activated by test items of different frequency, due to similarity differences induced by training. For Experiment 2 the list traces are activated similarly by test characters of different frequency. For both experiments, more extra-list self-traces are activated by high frequency test characters. Higher frequency test items produce worse

performance because the additional extra-list trace activations and (for Experiment 1) the higher strength of list-trace activations, cause increased noise in the summed activations. Dennis and Humphreys (2001) among others term the errors caused by confusions with extra-list traces context-noise (because one misjudges the context of study) and the errors caused by confusions with list traces item noise (because one misjudges the item information).

This discussion of recognition has been intentionally simplified for the sake of exposition: It has assumed that activations are measured by strengths. However, in SARKAE (as in many other models including REM) activations are actually measured in terms of likelihood ratios: The ratio of the probability that a trace originated from storage of the test probe to the probability that the trace originated from storage of some other character. This way of measuring activations helps to explain the mirror effect that is observed in our studies and most studies. In Figure 12, higher frequency items have summed activation distributions that are larger (shifted further to the right). If a single criterion were adopted independent of frequency then higher frequency test items would be given both higher hit rates and higher false alarm rates. To explain the observed mirror effects it would therefore be necessary to place the decision criteria for different frequency items at different positions, roughly where the distributions for items of a given frequency cross. However, the measurement in terms of likelihood ratios, when done appropriately, co-centers the distributions for items of different frequency, as illustrated in Appendix A. Co-centering allows a single criterion to be used for items of all frequencies, and due to the co-centering a mirror effect is predicted.

Many previous applications of the REM model assumed only item-noise, and it was easy to derive likelihood ratios in terms of the model parameters. Furthermore, due to the Bayesian inspired analysis, the distributions of summed likelihood ratios for different frequencies were co-

centered. Dennis and Humphreys (2001) assumed only context noise, and again derived likelihood ratios in terms of the model parameters so that co-centering was predicted. However, the present model is considerably more complex than either of these models, both in the way that features are recruited and stored, and in the assumption that there is both item noise (list traces) and context noise (extra-list traces). This makes it difficult to derive the likelihood ratios in terms of the model parameters, but as we shall see, a method is available to simulate the appropriate likelihood ratios, thereby producing co-centering and the mirror effect.

Modeling Recognition Memory in SARKAE

Episodic recognition is modeled in two phases: the study phase and the test phase.

Recognition Study.

When an item is presented for study, an episodic trace is formed for that item, using the following steps (see Figure 13): First, for every feature (both item and context) there is a probability *u* that something is encoded for that feature (the high-level item feature has a higher probability u_h of being encoded). If something is encoded for a given item feature, the value to be encoded is gathered probabilistically from one of three sources: from the study item's physical features with probability *si*, from the study item's lexical entry with probability *sl*, and from the previous study item's physical features with probability *sp*. Once the source of the feature value is decided, with probability *c* the value is copied correctly into the episodic trace. If the value is being gathered from the physical item features or previous study item features, then *c* does not depend on frequency. If however, the value is gathered from the item's lexical entry, then *c* depends on the total number of counts in the lexical entry being contacted. The

value of *c* varies linearly with the number of counts in the lexicon: the intercept (*ci)* and slope (*cs*) of this function are parameters of the model. If the value is not copied correctly, then a random value is chosen in proportion to the base rates in knowledge for the values for that feature.

For context features, the storage process is simpler. For a given study list, there is one context vector that represents that list. When an item is presented for study, for every context feature, with probability *u* the feature is encoded; if it is encoded, then the value in the list context vector is copied correctly into the episodic trace with probability *c*, otherwise is chosen randomly according to base rate. Since we are encoding context from the current study situation (which, unlike a lexical trace, is identical for high and low frequency items), the value of *c* during context storage is not dependent on frequency.

Recognition Test: Construction of a Memory Probe.

For each item on the test list, the following steps ensue: First, a test probe is constructed. Because the test item is available during probe construction, we assume one feature value is encoded for each and every item and context feature ($u = 1.0$), and we assume that this value is copied correctly from its given source. The source of the encoded physical feature value varies using the same probabilities as during study: The value copied into the test probe is taken from the physical features of the test item with probability *si*, from the test item's lexical entry with probability *sl*, and from the previous test item with probability *sp*. The source of the encoded context feature value is always the current context vector; and since the probability of encoding and the probability of correct copy are both 1.0 in test probe construction, the context portion of the test probe is always identical to the current context vector.

Recognition Test: Retrieval of Event Traces**.**

Once the test probe has been constructed, it is compared in parallel to all of the episodic traces in memory from the previously studied list, as well as any extra-list traces of the test item that are similar enough in context to the current study list to be activated. We could have assumed and explicitly modeled regular context change during training, and again until recognition testing, but this would greatly complicate the model for little gain. Instead we approximate the intended situation.

In conceptual terms, we theorize that each individual prior event involving the current test item (presentations during training, etc.), whether it occurred a long or a short period before the current test, has the potential to create confusion for the recognition decision. While it is possible that all previous events may contribute to this confusion, it is more likely that only those events which possess a similar context to the current test context will actually influence the recognition judgment. One possible method for selecting traces to include in the set of extra-list intrusions would be to simply use a criterion method: every event trace in memory is compared to the current context and only those with a given level of context match are included as extra-list traces. However, given that we do not explicitly keep record of every event that is stored during training, we must use a method that approximates this process.

In order to accomplish this approximation, we first establish a set of prior event traces to be probabilistically included in the intrusion set. The size of this initial set is directly proportional to the frequency of the item being tested: a high frequency item will have many more recent previous presentations than a low frequency item, and thus the inclusion of extra-list traces follows the same ratio. Specifically, for a given frequency, there are a base number of

possible extra-list traces (1, 3, 9, or 27). This base number is then multiplied by a multiplier parameter, *em*, to determine the size of the set of possible extra-list traces. Each trace in the initial set has a certain probability *ep* of being included in the actual set of intrusions. The total number of extra-list traces included in a given test trial is therefore determined by a combination of test item frequency and parameters *em* and *ep*. The number of extra-list intrusions is of course lower when there is a longer delay between training and recognition testing, because it is assumed that context changes during the interval and fewer training session traces exceed the theoretical activation threshold. This is reflected in the model through a change in the *em* and *ep* parameters (see Table 2).

The mechanisms above determine the number of extra-list traces which are included, now we must specify how these traces are constructed. Each extra-list trace represents a previous occurrence of the current test item. We do not keep records of every training event; however, the model includes a mechanism for the construction of an episodic trace. Therefore, we simply create *n* separate extra-list traces (the quantity *n* is determined by the parameters *em* and *ep* in the previous paragraph) using the same process for storing item features as is used to store item features in the episodic recognition study phase (described above). This creates *n* distinct item feature traces. Next, the context features must be added to these traces. Since these traces come from prior events, and we do not know what the context was at the time of their storage, we must approximate. We assume that any trace that is included as an extra-list intrusion possesses a context that is similar to the current context to a certain degree; therefore, the generating (or true) context vector of each extra-list trace shares10 out of 30 features with the current context. This generating or true extra-list context vector is then stored into each extra-list trace in the same noisy and incomplete fashion that occurs during the episodic recognition study

phase. The end result is a set of extra-list event traces that were each separately created (in a noisy, error-prone fashion) from the item features of the current test item and a context vector that shares one third of its features with the current context.

Recognition Test: Comparing Probe to Trace.

The test probe is compared to each trace (list traces and activated extra-list traces) individually, feature by feature. The value for a given feature might be missing in a trace, and if so, that feature is ignored. If the trace has a value it can either match the probe value or mismatch the probe value. The number of matching values (*km*) and the number of mismatching values (*kq*) are counted. Each match increases the likelihood ratio that the trace is that of the probe rather than some other item, and each mismatch decreases this likelihood ratio. The likelihood ratio for the whole trace is obtained by multiplying all the match and mismatch ratios for that trace, as in Equation 1.

$$\lambda = \left[\frac{P(m|s)}{P(m|d)}\right]^{km} \left[\frac{P(nm|s)}{P(nm|d)}\right]^{kq} \quad 1$$

In this equation the ratio for a matching feature is given in the first bracket, and for a mismatching feature in the second bracket. In the brackets, s denotes a trace generated by the item being tested, and d indicates a trace generated by some other item. The superscripts are the number of matching and mismatching features respectively. For simpler models (like REM) the terms in brackets can be derived in terms of the model parameters. However, the complex rules we posit for construction of the event traces and the test probe make it difficult to derive these probabilities analytically for the present model. Therefore they are instead estimated through a

simulation technique (see Appendix D for details regarding this technique.) Once estimated for a given set of parameters, the two ratios are fixed and used to calculate the likelihood ratios for all activated traces.

The recognition decision is based directly on the average of the likelihood ratios. The average is termed the odds ratio Φ for old over new, and is shown in Equation 2.

$$\Phi = \frac{1}{n}\sum_{i=1}^{n} \lambda_i. \qquad 2$$

In the REM model an 'optimal' Bayesian decision is made by using a decision criterion of 1.0: If Φ is greater than 1.0, then the item is called "old", if not, then the item is called "new". In the present model it is difficult to calculate an optimal Bayesian decision criterion, so we let the criterion be a parameter (but this decision criterion is the same for test items of all frequencies).

Parameter Estimation and SARKAE Predictions for Recognition

Using this model, parameters (some jointly and some separately) were estimated for the recognition results for Experiments 1 and 2. Table 2 gives the parameters and their estimated values for the two experiments and all tasks. Whenever a value is identical across tasks or conditions, it indicates that that parameter was forced to have the same value in those cases. Parameter estimation was carried out by inspection, and the results we show are not necessarily the best possible but are sufficient to show that the approach captures at least the qualitative patterns seen in the data. Figure 14 shows that the model produced patterns of predictions similar qualitatively to the observed data, predictions that are not too deviant quantitatively: LF items produced better performance than HF items, in the form of both more hits and fewer false

alarms for experiment 1 (panel A), and fewer false alarms but no hit rate advantage in experiment 2 (panel B). To obtain these predictions the odds criterion was set at 0.50 for Experiment 1, and 0.25 for Experiment 2, and the number of estimated extra list traces activated for test characters of frequency x were 0.50x. The model was also fit to the six-week delay data, a condition for which many fewer activations of extra list traces are expected. The estimated number of activated extra list traces for characters of frequency x dropped to 0.01x. Given the drop in extra list traces activated, the predicted frequency effects diminished correspondingly, as shown in Figure 15.

It is important to remember that SARKAE chooses values from knowledge in a way that favors higher frequency items (parameter *c* is higher for high frequency entries). This factor by itself would produce a high frequency advantage. Such an advantage will indeed be seen for the other tasks. However the advantage in recognition is clearly outweighed by the factors discussed above that favor low frequency characters. Other parameterization factors of note include changes in the encoding parameter *u*. First, the difference between the encoding parameter for low-level features (*u*) and high-level features (*uh*) is larger in experiment 1 than experiment 2. That is, in experiment 1, the attention advantage given to high-level features is larger. This difference is attributed to the training paradigms used in each experiment; in experiment 1, the visual search task induced a strong reliance on the high-level unique feature to search for a target character among distracters. While we still assume that a high-level feature is constructed during the same/different judgment task in experiment 2, the nature of the task makes the use of this feature much less necessary. This differing usage of the high-level feature during training is reflected in the levels of the *uh* parameter used in the episodic recognition task.

Furthermore, the *u* parameter for experiment 2 changes based on whether the test was administered shortly after training or after a 6 week delay. Simply put, after a 6 week delay, encoding is no longer as efficient as it was immediately following training: this decrease in efficiency is represented with a decrease in the *u* parameter (see Table 2). The drop in the encoding parameter allows the model to produce the decline in performance after delay that occurs in the data.

A final topic of note is the characterization of performance for untrained (or zero-frequency) items in the episodic recognition task. The behavioral data for untrained items (hit and false alarm rates) for experiment 1 is shown in Figure 4, for experiment 2 in Figure 10. However, these data points were not included in the model. The data for these zero frequency items follows the same general pattern that has been found to occur for very low frequency words: the hit and false alarm rates for these unknown or untrained items do not necessarily fit the same function as known or trained items (e.g. – Estes & Maddox, 2002). What the SARKAE model would predict for untrained items is unclear. The knowledge trace of an item plays a large role both in constructing an episodic trace during study and making a recognition decision during testing of that item. An untrained item would presumably not possess a knowledge trace, making it unclear what mechanism should be used in the existing framework. Of course, machinery could be added to the model that would make it possible to predict performance of these untrained items without accessing knowledge. However, given that these are only two data points and that the overall goal of our model is to account for items that are in knowledge at least to some degree, we did not add special processes to the model to account for the untrained data.

Discussion

The SARKAE model assumes both item noise (activation of list traces) and context noise (activation of extra-list traces of the test character). For words, Dennis and Humphreys (2001) have argued that all interference is due to context noise, in distinct contrast to many recognition models which ignore context noise and employ only item noise. Even for words, almost certainly both play a role (Criss & Shiffrin, 2004a), although the degree to which each controls performance varies with task. In the application of SARKAE to the present studies using Chinese characters, recognition in Experiment 1 is frequency dependent due to both context and item noise, but recognition in Experiment 2 is frequency dependent only due to context noise. The latter assumption is validated by the Experiment 2 delayed testing results, showing a reduction in the frequency effect. This result would be expected if the number of extra-list activated traces from the training sessions drops as six weeks pass between training and recognition study and test. By the same reasoning we can infer that item noise plays a major role in limiting performance in the six week delayed condition. Performance in the six week delay conditions is slightly lower than that for immediate recognition. One might expect that performance would improve if one source of noise is removed. I.e. If activations of traces from the training sessions largely stop after delay, then context noise is greatly removed, and the main source of noise must be activations of list traces. This reasoning does raise a question: Why does performance decrease? Certainly the decrease in extra-list intrusions following a delay will decrease the familiarity of test items. However, this decrease should be equal for high and low frequency items, resulting in no change in the qualitative pattern of results. Why then does performance (d') decrease with delay? SARKAE explains this drop in performance as stemming from a loss of encoding efficiency with time (represented by a decrease in the encoding

parameter – see discussion above). The delay results also tend to rule out a variety of alternative theories of frequency effects in recognition in which list study produces differential strength of storage of characters of different frequencies. If so, it is not clear why delay should reduce the frequency effect (especially since the delay results for pseudo-lexical decision show very strong frequency effects, verifying that the frequency differences in the knowledge traces have not 'decayed' over the six weeks).

Pseudo-lexical Decision

Pseudo-lexical decision requires the participant to judge whether a test item has ever been studied, or instead is novel. We assume this task is carried out by reference to the knowledge traces. The general idea is that a test probe activates all 32 knowledge traces. Each knowledge trace contributes a likelihood ratio. These are averaged and a decision is based on the average (the odds ratio). Because the data in this task consist of both response time and accuracy, the model must be extended to predict the time to respond. We do so by assuming that features are gradually extracted starting from presentation of the test character, and accumulate in the test probe. At each point in time the current probe is used to produce an odds ratio. A positive response is given if and when the odds ratio reaches a positive boundary, and a negative response is given if and when the odds ratio reaches a negative boundary. The effect of frequency is fairly direct: The gathering of features from a knowledge trace is more accurate for more highly developed traces (i.e. – traces of high frequency items).

The SARKAE Model for Pseudo-lexical Decision

The model outlined above is a dependent random-walk or diffusion. The features extracted at each time point are not necessarily accurate, though on average a known test character will produce a feature matching its knowledge trace, and an unknown character will produce a feature that matches any given trace randomly. The noise in feature extraction produces odds ratios at successive time points that sometimes moves toward one boundary and sometimes toward the other. Of course, on average, test of a known character causes movement toward the positive boundary, and test of an unknown character causes movement toward the negative boundary. The drift rate of this diffusion is not stationary, however, because we assume that features accumulate in the probe. Thus, as time passes, the odds ratio becomes less noisy, and eventually the odds ratio moves steadily toward the correct boundary. If the participant would wait sufficiently long before responding (by placing the boundaries far apart) then no errors would be made. However, the instructions require responding as 'rapidly and accurately' as possible. This dual emphasis causes placement of the boundaries fairly close together, at a point that balances the conflicting demands of speed and accuracy.

Specifically, when a test item is presented, at each time step there is a probability u^* that each not-yet-encoded item feature is encoded in the percept[3]. If a feature is encoded, with probability c^* the feature value is copied correctly from the physical features of the presented item; with probability $1-c^*$ the value is chosen randomly in proportion to base rate in the knowledge base.

[3] The parameter u^* in lexical decision is conceptually identical to the u parameter used in episodic recognition. However, since it is estimated separately and used for a different task, we will denote it with the *. The same * notation is used for the c, cs, and ci parameters.

In this task, we ignore context features and they do not play a role. Previous research has shown that context features do join the probe even though the match between current context and stored context is irrelevant to the decision. In fact this assumption has been used in previous publications (Schooler et al, 2001) to explain long-term priming of knowledge retrieval. However, our tasks do not manipulate priming, so adding this assumption here would not change the qualitative pattern of predictions. Thus we exclude context features for simplicity.

At each time step, following feature encoding, the current percept (i.e. probe) is compared to the 32 traces in the knowledge base. Each trace contributes a likelihood ratio, and the average of these is an odds ratio on which the decision is based. The probe has at most one value per feature, but each knowledge trace generally has a distribution of values, and often multiple counts per value. To enable calculation of a likelihood ratio for each trace, a sample of feature values is taken from the knowledge trace: With probability $c^*(n)$ a feature value will be copied from the distribution in the knowledge trace, in proportion to the counts in the trace for that feature. The value of $c^*(n)$ is based on and rises with the total counts n in the knowledge trace, summed over all features. A simple linear function is used, with two parameters cs^* and ci^* representing the slope and intercept respectively. The copying is proportional and hence on average the sampled vale for each feature accurately reflects the contents in the trace. The result is a sample with at most one value per feature. The process of comparing the probe to the trace sample is then the same as that used in recognition: Matches and mismatches are counted, and a likelihood ratio calculated (see Equation 1).

To be precise, a likelihood ratio is computed for each trace, based on the number of matches and mismatches between the percept and the trace's sample, as given in Equation 1: As was true in recognition, the complexity of the sampling rules makes it hard to derive the ratios in

the parentheses in terms of the model's parameters, so again a simulation method is used to derive the ratios (see Appendix D). Once simulated for a given set of parameters, the ratios are fixed and used for all calculations and all time points. After the percept is compared to each knowledge trace sample, the average of the likelihood ratios is the odds ratio given in equation 2.

Whenever the odds ratio becomes greater than an "old" criterion, then an old response is given. Whenever the odds ratio drops below a "new" criterion, then a new response is given. If the odds ratio is between these two boundaries, then a new sample of features is taken from the presented character and the percept tends to grow. Features already in the percept do not have their value replaced, but values for not-yet-encoded features can be added at each time step. This new percept is then used to compute a revised odds ratio. This process repeats until the odds ratio passes one of the boundaries, and a response is made.

Parameter Estimation and SARKAE Predictions for Pseudo-lexical Decision

Parameters were estimated (some jointly and some separately) from the data for pseudo-lexical decision from Experiment 1, Experiment 2 with no delay, and Experiment 2 after delay. Table 2 gives the parameters and their estimated values for the two experiments and all tasks. Whenever a value is identical across tasks or conditions, it indicates that that parameter was forced to have the same value in those cases. Parameter estimation was carried out by inspection, and the results we show are not necessarily the best possible but are sufficient to show that the approach captures at least the qualitative patterns seen in the data.

In all three cases, the HF items were recognized both more quickly and more accurately than the LF items, as shown in Figure 16. When a studied character is tested, consider its

knowledge trace. Because the probability of copying correctly into the lexical sample is higher for higher frequency characters, at each time step the likelihood ratio for the test character's knowledge trace will tend to be higher. Hence the diffusion will move more quickly on average toward the correct (studied) boundary. When the test character is new, no knowledge trace will match, and each trace will tend to contribute a negative trending likelihood ratio. One might suggest that, instead of high frequency lexical traces being sampled more accurately, high frequency test items should create a more complete percept. However, this would require making assumptions about the processing of high frequency physical item features that our model does not include. In the lexical decision task, it is the *item* features that are being processed and compared to a sample of all item's *lexical* features to check for a match. If the lexical entry for a specific (say high frequency) item was contacted to influence the creation of a percept, this defeats the purpose of lexical decision: it gives you an answer as to whether the item is in the lexicon. It is possible that the high level feature of high frequency items is processed more efficiently in this task. However, if that were the case here, such a processing advantage should be added to the episodic recognition task as well. For simplicity therefore, and because it is not needed in this task, such a process is not implemented.

Discussion

The simple observed patterns of data, in which response time drops and accuracy rises with character frequency, could undoubtedly be modeled in many ways. We attempted to construct a simple model that was consistent with the assumptions made for the other tasks, but make no claim that these assumptions are necessarily better than other possibilities. The intricacies of calculations aside, the model is conceptually straightforward: Starting with

presentation of the test character, a percept gradually fills out with feature values as time passes. At each time step this percept is compared to all knowledge traces and an odds ratio calculated by averaging the trace likelihood ratios. At the time step when this odds ratio passes one of two boundaries a response is made. The dependence on frequency is due to the fact that the trace comparisons are more accurate for the higher frequency traces.

In future research we might try to test the detailed model assumptions by fitting the response time distributions, even for individual participants. The present studies were designed in such a way that testing was of limited duration, because we did not wish testing to undermine the frequency manipulations through learning during testing. The limited data available were not sufficiently accurate to merit fitting response time distributions.

Perceptual Identification in Experiment 1

Two alternative perceptual identification could conceivably be modeled by simple visual matching: Some physical features from the flashed stimulus could be extracted and matched directly to the two following choices. Much research has shown that this model is inadequate, particularly when the flashed stimulus is masked, and even more so when the mask is a pattern mask. Many studies and findings show that the probe consists not just of the flashed character but a variety of features from the contextual surround (e.g. Huber, Shiffrin, Lyle, & Ruys, 2001), and that the collection of probe features is used to access the various traces in the knowledge base (e.g. Ratcliff & McKoon, 1997, Schooler et al., 2001) producing features that are then used in comparison to the two choices. Evidence calculations and inferences can also be complex, as demonstrated by short-term priming in which there are various forms of discounting of features that could have been been present due to the presence of the primes (e.g. Huber et al., 2001). In

our present studies, we do not use short-term priming, and do not vary most of the variables manipulated in prior studies, and therefore use a simplified model.

The model is predicated upon prior research showing that the use of a pattern post mask tends to inhibit the use of low level features in forced choice decision making. Sanborn, Malmberg, and Shiffrin (2004; based in part on prior research by Huber et al., 2001) demonstrated this by showing that neither the case not the color of a flashed word matters when making a choice between two choice words, one of which matches the case or color, and the other of which does not. Instead the choice word's spelling is critical, even when the spelling differs by only one letter. Sanborn et al showed that this result held when a pattern mask was used, a mask containing features confusable with the target stimulus (e.g. multiple colors for colored stimuli, and letter-like fragments for stimuli varying in case). For more primitive masks (e.g. uncolored or pixel noise) lower level features seemed to be available and were used. The present task used jumbled fragments of Chinese characters, surely a pattern mask, and very likely to reduce the use of physical features in the forced choice decision. For this reason we adopted a very simple model in which the decision is based solely on the high level feature and its value.

Given this to be the case, we need only model the extraction of a high level feature value, and its comparison to the two choices. Let the probability of extraction of a correct high level feature from the flash be c**(f), where f denoted the target character's frequency[4]. The dependence upon frequency is simple: The higher the frequency of the flashed character, the greater the probability that the high-level feature of the item is extracted. This was instantiated

[4] The ** notation represents that while this *c* parameter is conceptually identical to the *c* parameter used in episodic recognition and the *c** parameter used in lexical decision, it is estimated separately here to apply only to the forced choice task.

with a two parameter linear function (see Equation 3), where *cs*** and *ci*** represent the slope and intercept of the function respectively.

$$c^{**}(f) = (cs^{**} \times f) + ci^{**} \quad (3)$$

Although it seems reasonable to assume that the high level feature extracted could sometimes be the wrong one, such an assumption would not significantly change the predictions for our present task because frequency dependent guessing is assumed to take place if no extraction occurs (see just below). This biased guessing introduces sufficient noise and frequency bias into the model. We therefore again simplified by assuming that if a feature is extracted, the value would be correct.

When a feature value is extracted the decision is simple: Select the choice having that feature (in which case the choice is correct). However, if the high-level feature is not extracted from the flash, then a biased guess is made. The guess will be based on the frequencies of the two choices. We express the guess in terms of the probabilities of choosing the target (Equation 4):

$$p(\textit{guess target}) = \frac{1 - c^{*}(f_{target})}{(1 - c^{*}(f_{target})) + (1 - c^{*}(f_{foil}))} \quad (4)$$

This equation posits that guessing a choice will drop as its frequency rises. The idea is that feature extraction is higher for higher frequency targets, so the failure to extract implies that the target probably was of lower frequency.

Parameter Estimation and SARKAE Predictions for Forced-Choice Identification

The parameter values estimated were $cs^{**} = 0.075$, and $ci^{**} = 0.60$. The dependence upon frequency is well predicted by the model. The increase in predicted performance for high frequency targets is due to increased extraction of information from the flash: i.e. frequency dependent perception. It would have been possible to model this frequency dependence with direct reference to the counts in various knowledge traces, but such a model would have introduced much additional complexity without significantly changing the qualitative predictions, so was omitted. The rise in predictive performance for foils of higher frequency was due to the assumption of inference in guessing when information was not extracted from the flash. There was an assumed bias to guess the low frequency alternative, presumably based on the inference that something would have likely been extracted had a higher frequency character been flashed. The predictions and he experimental data are shown in figure 17.

General Discussion

The SARKAE model fits within a framework designed to provide first steps toward a theory for the co-evolution of event memory and knowledge. There are many caveats: The experiments are very simple and only manipulated a few variables, chiefly training frequency. The way that the general theoretical framework was instantiated to deal with the training tasks and the three transfer tasks was not only highly simplified but utilized assumptions that were in many cases somewhat arbitrary. There are surely many alternative implementations that would be consistent with the general approach that would also fit the patterns of results. These caveats aside, the research project presented here is non-trivial on both conceptual and empirical

grounds. Conceptually, it is important to have a framework by which knowledge grows from events, and by which knowledge informs the coding of events. The 'knowledge' formed in the present studies is of course rather simple and unstructured in comparison to most knowledge we form in the real world, but it is not hard to imagine how the approach could be augmented to deal with the formation of much more richly structured knowledge, using the same basic building blocks. Empirically, it may be the case that relatively few variables were manipulated in the studies, but it is nonetheless rather unusual to carry out in the same empirical framework training, event memory, perception, and knowledge retrieval, using very different tasks that come close to spanning those in the field of cognition. Such an approach is more typical of applications of cognitive architectures (e.g. SOAR, ACT-R, EPIC) but those architectures focus upon established knowledge, whereas our focus is on the mechanisms by which event memory and knowledge co-evolve, and inform each other in the process.

Our two studies provide evidence concerning the effects upon episodic memory, knowledge retrieval, and perception of event frequency during the establishment of knowledge. The results point to two effects of frequency: one due to an item's temporal and spatial context when context is correlated with an item's frequency (such correlation almost always occurring in the real environment) and one due to frequency per se (i.e. the increased strength of an event's knowledge trace when frequency increases). Such a finding would surprise few scientists or laypersons but the ongoing debate concerning word frequency gives the findings some importance. It is interesting that we found in our initial modeling efforts (aimed at the results of Experiment 1) that most if not all effects of frequency could be explained by a model utilizing only event context. The empirical dissociation of context and frequency in Experiment 2 was needed to demonstrate the existence of two roles for frequency.

Frequency manipulations aside, the main focus of this article is a theory providing plausible mechanisms by which knowledge grows from events, and knowledge informs the coding and retrieval of both events and knowledge itself. The approach we have pursued, termed SARKAE, grows from and is largely consistent with a prior history of memory modeling starting with the Atkinson and Shiffrin (1968) model, continuing through the SAM models of Raaijmakers and Shiffrin (1980, 1981) and Gillund and Shiffrin (1984) and continuing through the REM modeling of Shiffrin and Steyvers (1997) and subsequent applications of that Bayesian-inspired approach to perception and knowledge retrieval (e.g. Huber et al., 2001; Schooler et al., 2001, Wagenmakers et al., 2004). However even accepting the basic conceptual underpinnings of the theory, SARKAE is surely not a unique way to implement the concepts. We could imagine other approaches, for example using composite and distributed neural nets. We note however, that many neural net approaches instantiate knowledge traces for words in a lexicon of separate traces. In addition, we have greatly simplified the general conceptual theory for the purposes of modeling the present experimental results, and in the process of simplification and approximation, a number of fairly arbitrary assumptions were made when other simplifying assumptions would almost surely have served just as well. Thus we regard the present implementation as a proof of concept of the general approach rather than a commitment to the detailed assumptions. As the theory is applied to more and more data from a variety of studies, we would hope a convergence on the detailed assumptions would take place.

What then are the key conceptual bases that the present theory places on the table? An event causes two types of memory storage, with storage processes that are essentially the same: One type is storage of an event trace (an episodic memory trace). Such a trace is identified by the general contextual surround, including other events in the nearby temporal/spatial vicinity

(presumably to a degree governed by attention, although attentive processes are not a focus of the present article), and a variety of internal features (e.g. one's thoughts, goals, and emotions) and environmental features (internalized coding of the features of the environment --e.g. place and time). In general the formation of such a trace is incomplete and error prone. The second type is addition of the same sorts of information (again in incomplete and error prone fashion) to an already stored trace that is brought to mind (brought to mind most often by similarity of features to the present event). The already stored trace can be an earlier event trace-- this mechanism allows the development of knowledge traces. As addition continues to occur over successive events (for example spaced repetitions of a word in different contexts) information accumulates and the resultant trace becomes richer and more complete. The information in common to the various events comes to dominate the developing knowledge trace, and the information that varies over events, such as each event's context is present but inconsistent. Thus retrieval from a knowledge trace that is relatively mature brings the consistent information to mind, but not the inconsistent information, and knowledge gradually becomes seems to introspection to become context free. Of course the various event contexts are indeed present, as demonstrated by long term priming: E.g. study of a word adds local context to its knowledge trace, and later knowledge retrieval (e.g. word naming) in a similar context is enhanced by the partial match of the current context to the stored context.

The process described above gives the bare bones of the process by which events come to form knowledge. We have made no attempt to capture the complexity of real knowledge formation, involving attentive and control processes, and the addition of various forms of knowledge from elsewhere in the developing knowledge base. In principal the mechanisms to

incorporate more complex storage can be added to the model, but such awaits further research and development.

A key aspect of these storage processes for both event storage and knowledge storage is the idea that features from the general surround of an event (perhaps sampled from the current contents of short-term memory) are added to both. Thus storage of an event generally includes samples of features from both general internal and environmental context, and other nearby events. The same types of features are encoded into the developing knowledge traces. Thus the structure of knowledge becomes quite complex and the similarity of knowledge traces to each other comes to reflect the co-occurrence statistics in the environment. In our Experiment 1 high frequency Chinese characters trended to co-occur in search trials, so our model predicted that the knowledge traces of high frequency characters grew similar to each other. The storage of nearby events of course takes place in event storage, and this process is similar to that posited in the Temporal Context Model (Howard & Kahana, 2002). In studies of free recall, for example, retrieval of one word tends to be followed (and preceded) by retrieval of a word that was studied in the former word's temporal vicinity, following a temporal gradient. Of course experimentally varied context, such as the words in a list, are just one type of context that is stored in event and knowledge traces. A variety of studies have been carried out in which the environmental context is changed from study to test, and these changes can have substantial effects, especially in recall tests (e.g. Godden & Baddeley, 1975, Smith, Glenberg, & Bjork, 1978). Murnane, Phelps, and Malmberg (1999) have discussed two ways that such context can be added to event and knowledge storage. In one case, termed associated context, the context joins the stored trace, but is not integrated with the other information. In the other case, termed integrated context, the contextual information is integrated into the representation. The difference can be probed

experimentally, and assessed, as shown first in fairly convincing fashion in the studies of Murnane et al. (1999).

The processes of storage are of course only one half the story of co-evolution. Whenever an event is encountered, and whenever a probe of memory is formed or constructed, the knowledge base is consulted and relevant information retrieved. The features found there are used to construct the coding of the event or memory. In fact most of the way an event is coded beyond infancy is based on learned features in the knowledge base. We use feature here in a general way that includes references to other traces in the knowledge base. That is, a trace that develops based on any set of events can be used as a feature stored in other knowledge traces. Most knowledge stored in memory can be used as examples of these conceptions, but the knowledge trace for the word 'table' is of course learned, and the features in that trace include reference to many other knowledge traces, such as 'fork', 'legs', 'surface', 'plate', 'flat' and so forth. Thus events are accumulated to form knowledge, but the features of events are coded with reference to the then current state of the knowledge traces. This recurrence is what leads us to refer to the co-evolution of events and knowledge.

The use of a representation in which traces are 'separate' is likely a convenient heuristic and an aid to understanding rather than a necessity. The last thirty years especially have seen substantial progress in development of neural net models in which numerous elements or nodes are linked by weights in networks, and various forms of events and knowledge are encoded by the weights. Given high enough dimensionality of the network, and/or recurrence in the network, many different types of information can be encoded in mixed fashion in the weights. This approach is particularly compelling for descriptions at the level of neural processes. For descriptions at the behavioral level, separate traces can provide a better avenue for

understanding. Neither approach is ‘right’, each aiding understanding in different ways (this use of ‘right’ is not meant in any technical sense, given that both approaches are certainly far too simplistic and far from anything resembling the actual processes of cognition and brain). In addition, some sorts of knowledge appear much more punctuate, such as words in our lexicon, lending themselves more naturally to separate representations, whereas other forms of knowledge are far more continuous, such as ‘the actions of playing tennis’, and lend themselves better to composite and distributed representations.

The way in which we represent knowledge, as vectors of feature values, is of course impoverished. Yet one must be careful not to enrich the representation too far, lest the theory become capable of explaining everything, and predicting nothing. Such a concern would apply if one broadens the concept of feature to include all possible combinations of existing features, as attractive as such an idea appears conceptually. Along similar lines, we call attention again to the earlier mention of the likely need to have features of one knowledge trace include (reference to) other developed knowledge traces as features. We did not implement this idea in the present modeling, but had we done so, the resultant model would be far too powerful unless strongly constrained. Such constraints would most sensibly arise in a theory specifying access to other knowledge traces and rules for incorporating reference to such. A particular concern with the present approach, utilizing a vector of primitive feature values, is the failure to take into account configurality: One might have trouble encoding a striped barber shop pole by separately listing primitive features (black, white, tall, etc.) rather than their conjunction (unless features are allowed to include conjunctions). The REM-II model developed by Mueller and Shiffrin (2006) therefore broadened the representation to encompass all binary feature combinations. The representation of events and knowledge traces then became a matrix of co-occurrence counts.

The authors showed the power of this approach to explain a number of findings in cognition (Mueller & Shiffrin, 2006, 2007), but this simple augmentation of the theory enriches it to the point where it lies at the limits of testability.

The feature vector representation we adopted for present purposes, knowing it was surely oversimplified, nonetheless is instructive and serves to illustrate the approach. We are of course not alone in coding features in vectors. There are numerous examples. One early example is the TODAM model proposed by Murdock (1982). It used vector representations, albeit in the context of a composite distributed model. Items in this model were represented as vectors, associations as convolutions of two item vectors (forming an association vector), and memory as a single composite vector summing all item vectors and association vectors. Other vector based models have encountered similar restrictions. Another early example was the MINERVA 2 model (Hintzman, 1984). It used a vector representation, but kept the traces separate. Associations in MINERVA 2 were concatenated vectors, back to back. In most neural net models the nodes could be considered features (and the weights connecting the nodes to be the 'strengths' positioning a memory in composite multidimensional weight space). However, given the nodes are typically arranged in layers or modules, it is not very natural in those cases to consider the nodes as a single vector

The vector representation we have adopted in this article here would likely run into difficulties should one wish to deal with associations and other forms of configural information. The present experiments were focused on the learning about individual characters, rather than associations. Hence the SARKAE model as presented here adopts this focus and provides no assumptions about the way to represent associations. Given the way that SARKAE operates, in which a given item trace includes features from the contextual and item surround of that item,

including nearby items, it might seem natural to encode an association as another trace of the same sort, but in which the information about the two items being associated is roughly equally balanced. This approach would not work well using the present vector representation: The features of the two items being associated would get mixed up in the simple feature vector. One way to handle associations and other sorts of configural information would involve expanding the present SARKAE representation to become a matrix of co-occurrence counts, as in REM-II. Mueller and Shiffrin (2006, 2007) provided examples that (indirectly) showed how such a representation might encode associations. For example, a representation of an association consisting of a novel combination of Chinese characters, each of which is new, might treat the combination as a single set of correlated features. For associations between items already existing in the knowledge base, it is conceivable that one could form a new trace matrix in which each existing item is a submatrix of highly correlated feature counts, and an association represented as smaller counts connecting the features of these submatrices. This approach strikes us as awkward. A better approach is probably that mentioned earlier, in which a feature can be a reference to a previously encoded item trace. Then the association trace could be a matrix with high counts for the co-occurrence of these two item-reference features. Such an approach could possibly be implemented in a vector representation, but it would fit more naturally into the framework of REM-II.

Our intent in this article is not to present a large scale model that could be applied quantitatively to the results of a large number of experimental results spanning the memory and learning literature. Although that is one of our long run goals, much modeling exploration and further empirical research would be needed before such a goal could be realized. Our aim here is far more modest: to lay out the bare bones of possible mechanisms by which event memory and

knowledge can co-evolve. A small number of conceptual assumptions seemed particularly important to underlie such a process, and we therefore implemented SARKAE in such a way that certain key assumptions were retained, but many other elements of the general theory were radically oversimplified. As the approach is gradually fleshed out in future research, more realistic assumptions will be incorporated, but at the inevitable cost of increasing complexity, and increased difficulty of understanding which of the many assumptions are critical for predicting which phenomena. This state of affairs is one good reason for starting with an simplified model that highlights just some of the critical assumptions: One can extrapolate to understand which factors produce what behaviors when the more complex models follow later.

Another direction in which the present research project ought to be extended is exploration of the neural correlates of the co-evolution that SARKAE implements (we have started moving in this direction with a pilot study using EEG). Previous studies show that training of novel objects (even for periods as short as one hour) produce measurable neurological changes. For example, a study by Rossion, Gauthier, Goffaux, Tarr, & Crommelinck (2002) utilized training of a novel set of objects (greebles). When faces are tested in upright or inverted fashion (inverted being 'novel') subjects showed an N170 effect: delayed and enhanced N170 for inverted vs. upright faces. Prior to training this inversion was not seen for greebles, but following two weeks of training, the N170 (at least in the left hemisphere) was delayed and enhanced for inverted vs. upright greebles. Furthermore, James & Atwood (2008) found that training on pseudoletters (letter-like stimuli) can produce activation in areas known to be involved in letter processing. Presumably, these pseudoletters are not receiving higher level feedback (as they have no linguistic association), and yet they are showing expertise effects

similar to roman letters. Such studies provide one way to use neural measures to link established knowledge with the learning of new knowledge.

In conclusion, the SARKAE model presented in this article provides a principled way of thinking about the co-evolution and interactive nature of human knowledge, event memory, and perceptual systems. 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.

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In R. W. Spence & J. T. Spence (Eds.), *The psychology of learning and motivation,* (Vol. 2, pp. 89-195). New York: Academic.

Becker, C.A. (1979). Semantic Context and Word Frequency Effects in Visual Word Recognition. *Journal of Experimental Psychology: Human Perception and Performance, 5*, 252-259.

Brainerd, C. J., & Reyna, V. F. (1990). Gist Is the Grist: Fuzzy-Trace Theory and the New Intuitionism. *Developmental Review, 10,* 3-47.

Brainerd, C. J., Reyna, V. F., & Mojardin, A. H. (1999). Conjoint recognition. *Psychological Review, 106*, 160-179.

Brainerd, C. J., Wright, R., Reyna, V. F., & Payne, D. G. (2002). Dual-Retrieval Processes in Free and Associative Recall. *Journal of Memory and Language, 46,* 120-152.

Bransford, J. D., & Franks, J. J. (1971). The Abstraction of Linguistic Ideas. *Cognitive Psychology, 2*, 331-350.

Broadbent, D.E. (1967). Word frequency effect and response bias. *Psychological Review, 74*, 1-15.

Criss, A. H., & Shiffrin, R. M. (2004a). Context Noise and Item Noise Jointly Determine Recognition Memory: A Comment on Dennis and Humphreys (2001). *Psychological Review,111,* 800-807.

Criss, A. H., & Shiffrin, R. M. (2004b). Interactions Between Study Task, Study Time, and the Low-Frequency Hit Rate Advantage in Recognition Memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*, 778-786.

Dennis, S., & Humphreys, M. S. (2001). The role of context in episodic recognition: The bind cue decide model of episodic memory. *Psychological Review, 108,* 452-478.

Estes, W. K. (1955). Statistical theory of spontaneous recovery and regression. *Psychological Review,* 62, 145–154.

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

Glanzer, M., & Adams, J. (1990). The Mirror Effect in Recognition Memory: Data and Theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 5-16.

Glanzer, M., Adams, J. K., Iverson, G. J., & Kim, K. (1993). The regularities of recognition memory. *Psychological Review, 100,* 546-567.

Godden, D. R., & Baddeley, A. D. (1975). Context-dependent memory in two natural environments: On land and under water. *British Journal of Psychology, 66,* 325-331.

Gregg, V. H. (1976). Word frequency, recognition and recall. In J. Brown (Ed.), *Recall and recognition.* London: Wiley.

Grossberg, S. (1978). Behavioral contrast in short-term memory: serial binary memory models or parallel continuous memory models? *Journal of Mathematical Psychology*, 17, 199–219.

Hemmer, P., Steyvers, M. (2009a). Integrating Episodic Memories and Prior Knowledge at Multiple Levels of Abstraction. *Psychonomic Bulletin & Review, 16(1), 80-87.*

Hemmer, P. & Steyvers, M. (2009b). A Bayesian Account of Reconstructive Memory. *Topics in Cognitive Science, 1,* 189-202.

Hintzman, D. L. (1984). MINERVA 2: A simulation model of human memory. *Behavior Research Methods, Instruments, and Computers, 16,* 96-101.

Hintzman, D. L. (1988). Judgments of Frequency and Recognition Memory in a Multiple-Trace Memory Model. *Psychological Review, 95*, 528-551.

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

Huber, D. E., Shiffrin, R. M., Lyle, K. B., & Ruys, K. I. (2001). Perception and preference in short-term word priming. *Psychological Review,* 108, 149-182.

Jacoby, L. L, & Dallas, M. (1981). On the Relationship Between Autobiographical Memory and Perceptual Learning. *Journal of Experimental Psychology: General, 110,* 306-340.

Jones, M. N., Kintsch, W., & Mewhort, D. J. K. (2006). High-dimensional semantic space accounts of priming. *Journal of Memory and Language, 55,* 534-552.

Kinsbourne M., & George, J. (1974). The mechanism of the word-frequency effect on recognition memory. *Journal of Verbal Learning and Verbal Behavior, 13,* 63-69.

Klein, K.A., Shiffrin, R.M., and Criss, A.H. (2007). Putting context in context. In J. S. Nairne (Ed.), *The Foundations of Remembering: Essays in Honor of Henry L. Roediger III*. New York: Psychology Press.

Lieberman, H., & Pentland, A. (1982). Microcomputer-based estimation of psychophysical thresholds: The Best PEST. *Behavior Research Methods & Instrumentation, 14,* 21-25.

Maddox, W.T., & Estes, W.K. (1997). Direct and indirect stimulus-frequency effects in recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23,* 539-559.

Malmberg, K., Steyvers, M., Stephens, J., & Shiffrin, R. (2002). Feature frequency effects in recognition memory. *Memory & Cognition, 30,* 607-613.

McClelland, J. L, & Chappell, M. (1998). Familiarity Breeds Differentiation: A Subjective-Likelihood Approach to the Effects of Experience in Recognition Memory. *Psychological Review, 105*, 724-760.

McClelland, J. L, & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology, 18,* 1-86.

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.*

Mueller, S. T., & Shiffrin, R. M. (2007). Incorporating connotation of meaning into models of semantic representation: An application to text corpus analysis. *In D. S. McNamara & J. G. Trafton (Eds.), Proceedings of the 29th Annual Cognitive Science Society (pp. 64-70). Austin, TX: Cognitive Science Society.*

Murdock, B. B. (1982). A theory for the storage and retrieval of item and associative information. *Psychological Review, 89,* 609-626.

Murdock, B. B. (1983). A Distributed Memory Model for Serial-Order Information, *Psychological Review, 90,* 316-338.

Murnane, K., Phelps, M. P., & Malmberg, K. (1999). Context-dependent recognition memory: the ICE theory. *Journal of Experimental Psychology: General, 128*, 403-415.

Neely, J. H. (1989). Experimental dissociations and the episodic/semantic memory distinction. In H. L. Roediger, & E I. M. Craik (Eds.), *Varieties of memory and consciousness: Essays in honour of Endel Tulving* (pp. 229-270). Hillsdale, N J: Erlbaum.

Nelson, A., & Steyvers, M. (2004). *Memory for Chinese Characters.* Unpublished manuscript, University of California, Irvine.

Quillian, M. R. (1967). Word concepts: A theory and simulation of some basic semantic capabilities. *Behavioral Science, 12*, 410-430.

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

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

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, and Cognition, 28*, 138-152.

Reder, L.M., Nhouyvanisvong, A., Schunn, C.D., Ayers, M.S., Angstadt, P., & Hiraki, K. (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.

Rice, G. A., & Robinson, D. O. (1975). The role of bigram frequency in the perception of words and nonwords. *Memory and Cognition, 3*, 513-518.

Rovee-Collier, C. (1997). Dissociations in Infant Memory: Rethinking the Development of Implicit and Explicit Memory. *Psychological Review, 104*, 467-498.

Rubenstein, H., Garfield, L., & Millikan, J.A. (1970). Homographic entries in the internal lexicon. *Journal of Verbal Learning and Verbal Behavior, 9*, 487-494.

Sanborn, A. N., Malmberg, K. J., & Shiffrin, R. M. (2004). High-level effects of masking on perceptual identification. *Vision Research, 44*, 1427-1436.

Scarborough, D.L., Cortese, C., & Scarborough, H.S. (1977). Frequency and repetition effects in lexical memory. *Journal of Experimental Psychology: Human Perception and Performance, 3*, 1-17.

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.

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

Sikstrom, S. (2001). The variance theory of the mirror effect in recognition memory. *Psychonomic Bulletin & Review, 8*, 408-438.

Smith, S. M., Glenberg, A., & Bjork, R. A. (1978). Environmental context and human memory. *Memory & Cognition, 6*, 342-353.

Wagenmakers, E. J., Steyvers, M., Raaijmakers, J. G. W., Shiffrin, R. M., van Rijn, H., & Zeelenberg, R. (2004). A model for evidence accumulation in the lexical decision task. *Cognitive Psychology, 48*, 332-367.

Wagenmakers, E., Zeelenberg, R., Raaijmakers, J. (2000). Testing the counter model for perceptual identification: Effects of repetition priming and word frequency. *Psychonomic Bulletin & Review, 7*, 662-667.

Tables

Table 1 – Mean response times and probability correct for Pseudo-Lexical Decision

	Freq 2	Freq 6	Freq 18	Freq 54	New
Response Time	845 ms	843 ms	750 ms	702 ms	820 ms
Probability correct	.902	.962	.981	.994	.932

Table 2 – Parameter values for model fits of Episodic Recognition and Lexical Decision

Parameter	Episodic Recognition			Lexical Decision		
	Exp. 1	Exp. 2: No Delay	Exp. 2: After Delay	Exp. 1	Exp. 2: No Delay	Exp. 2: After Delay
ti	.2	.2	.2	.2	.2	.2
td	.05	0	0	.05	0	0
tp	.05	.01	.01	.05	.01	.01
u	.6	.8	.4	.7	.6	.5
uh	.8	.85	.45	.9	.65	.55
c	.8	.65	.65	.55	.55	.55
si	.9	.75	.75	-	-	-
sl	.05	.2	.2	-	-	-
sp	.05	.05	.05	-	-	-
odds	.5	.35	.75	0,1	1,8	1,8
ci	.75	.65	.65	.55	.55	.55
cs	.000005	.000007	.000007	.005	.005	.005
timestep	-	-	-	.35	.675	.675
em	.8	.5	.001	-	-	-
ep	2	1	1	-	-	-

Figure Captions

1. Example Visual Search Trials
2. Sample of characters used
3. Panel A: Slope of the search function over training, Panel B: Intercept of Search function over training
4. Episodic Recognition performance: Hits are shown in blue, false alarms in red.
5. Forced Choice Perceptual Identification performance
6. Item representation and lexical entries in SARKAE
7. SARKAE: Similarity between normalized lexical entries (measured by dot product) after completion of training
8. Average 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.
9. 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.
10. 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.
11. Similarity between items following training with no context, measured by dot product of normalized lexical entries.
12. Signal Detection Theory demonstration of changes in hit and false alarm rates due to frequency. The solid lines represent the familiarity distributions for foils (solid red distribution on the left) and target items (solid blue distribution on the right), ignoring extra-list traces. The effect of adding extra-list traces is to shift the distributions to the right (higher familiarity); the dashed red line represents foils with extra-list traces, the dashed blue line represents targets with extra-list traces. The addition of extra-list traces also makes the distributions broader (reducing performance). In this simplified account both the amount of shift and the amount of broadening would be higher for HF items due to the greater number of extra list traces activated. However the Bayesian approach and the use of likelihood ratios for activations can center the distributions so a single criterion can be used for both HF and LF cases (for either the cases with or without extra-list traces). The solid vertical lines represent representative decision criteria (the one on the left for the case without extra-list traces, and the one on the right for the case with extra-list traces (without taking into account the extra shift for HF items).
13. Modeling process for episodic recognition task: study phase
14. Model and data for episodic recognition task for experiment 1 (panel A) and experiment 2 (panel B).
15. Model and data for episodic recognition in experiment 2 after 6-week delay.
16. Lexical Decision: Observed data and simulated data for Experiment 1 (black lines), Experiment 2, immediate (blue lines) and Experiment 2, delayed (red lines). Response time is given in the left panel, and accuracy is given in the right panel.
17. Two Alternative Forced Choice: Observed data and simulated data, averaged by target frequency (left panel) and by foil frequency (right panel).

Figures

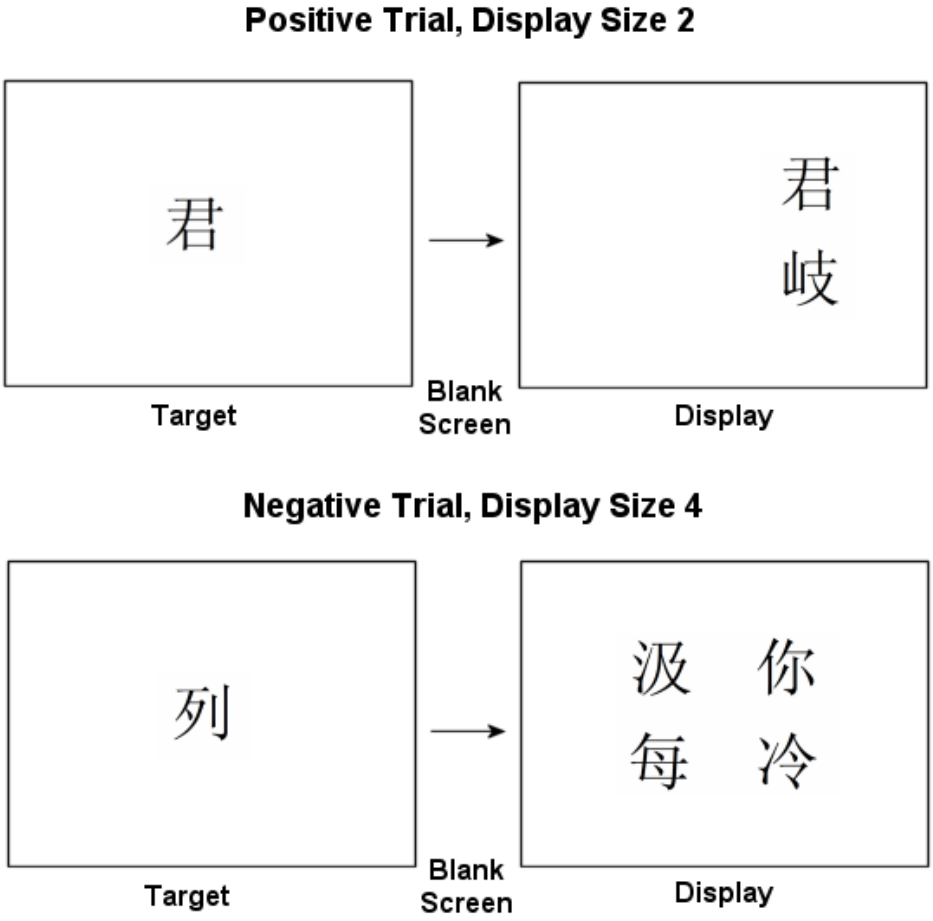


Figure 1

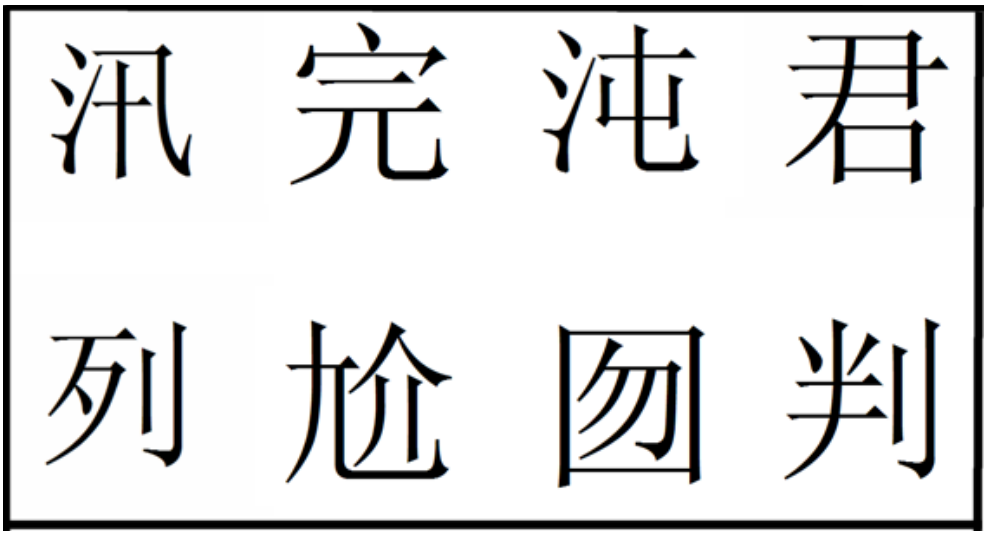


Figure 2

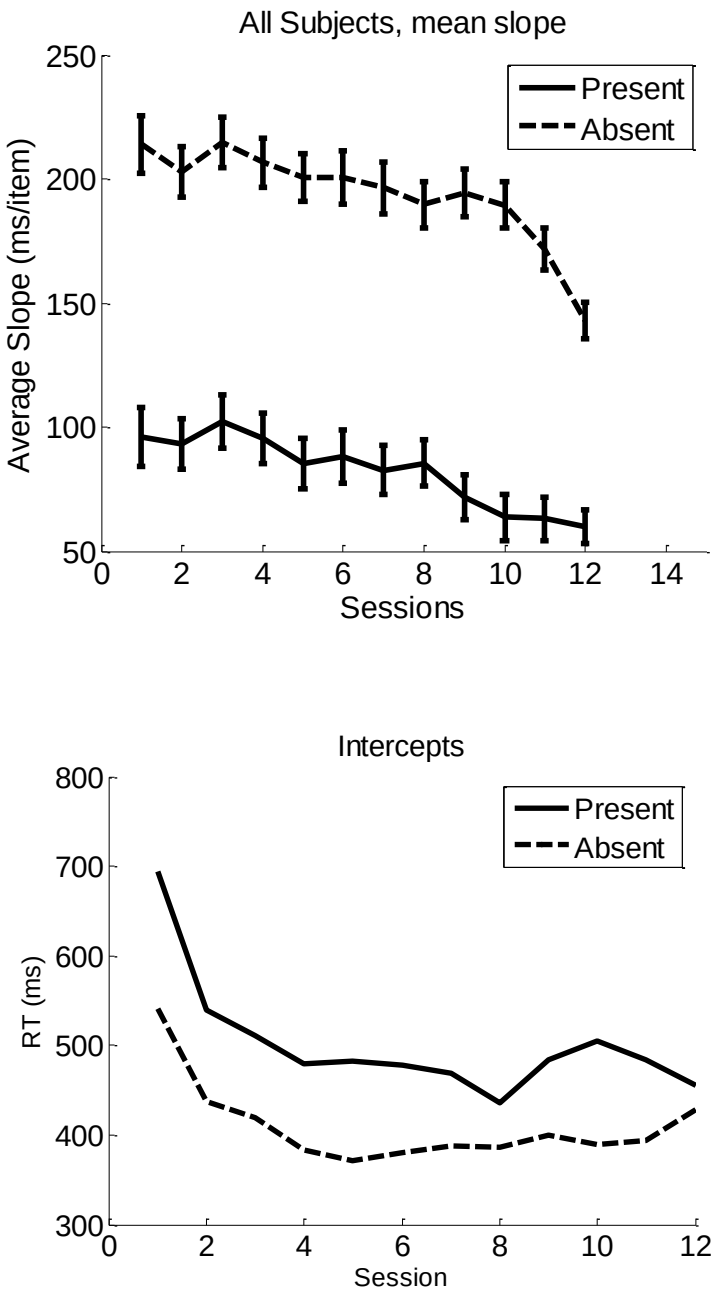

All Subjects, mean slope
Present
Absent
Average Slope (ms/item)
250
200
150
100
50
0
2
4
6
8
10
12
14
Sessions
Intercepts
Present
Absent
RT (ms)
800
700
600
500
400
300
0
2
4
6
8
10
12
Session

Figure 3

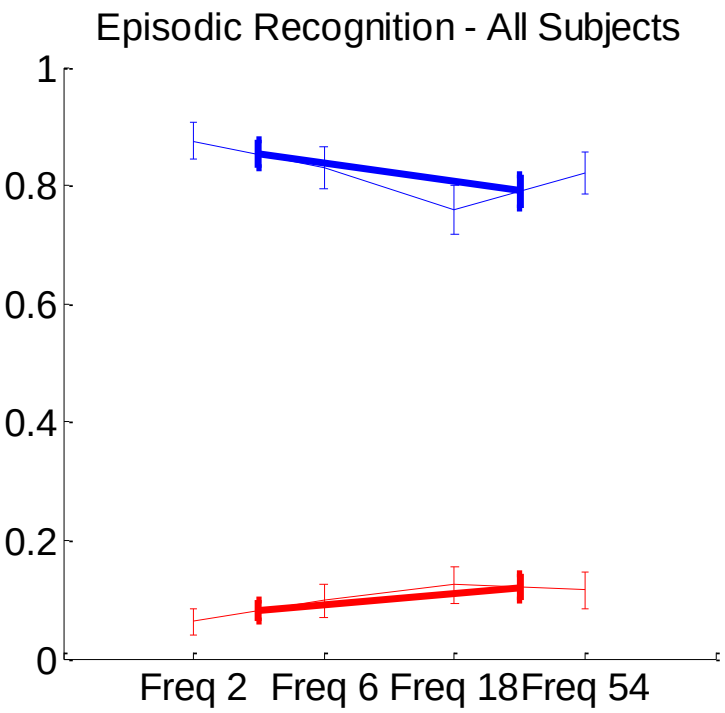

Episodic Recognition - All Subjects
1
0.8
0.6
0.4
0.2
0
Freq 2
Freq 6
Freq 18
Freq 54

Figure 4

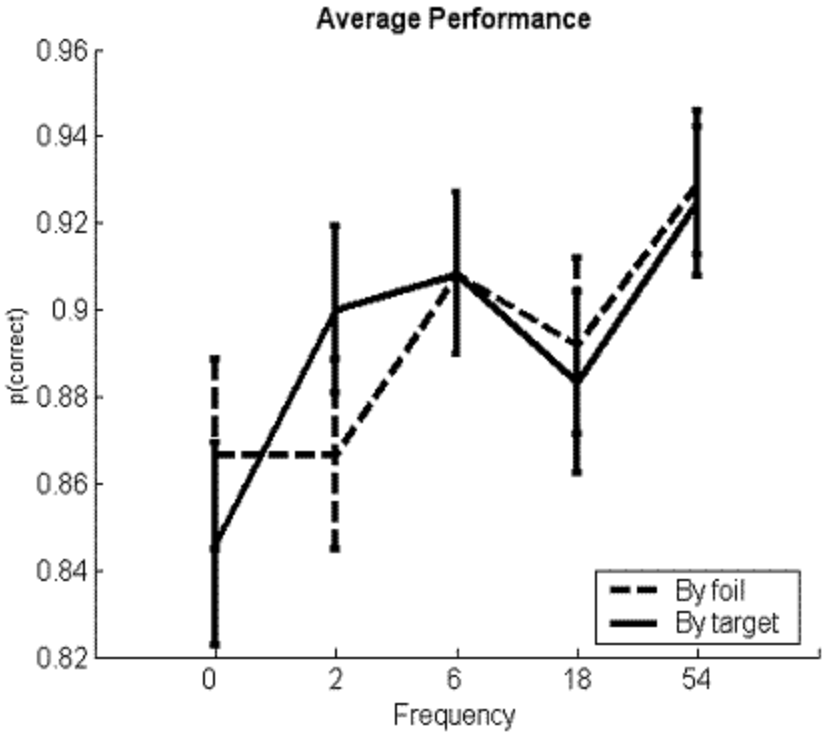

Average Performance
0.96
0.94
0.92
0.9
0.88
0.86
0.84
0.82
p(correct)
0
2
6
18
54
Frequency
By foil
By target

Figure 5

Actual Item

Item 1: [0 0 1 0 0 | 0 1 0 0 0 | 1 0 0 0 0 | 0 0 0 0 1]

Lexical Representation of item

Item 1: [1 3 9 0 2 | 4 8 4 1 1 | 12 2 0 3 0 | 1 2 1 3 8] HF

Item 1: [0 1 2 0 1 | 0 3 0 2 1 | 2 0 1 2 1 | 0 0 1 0 3] LF

Figure 6

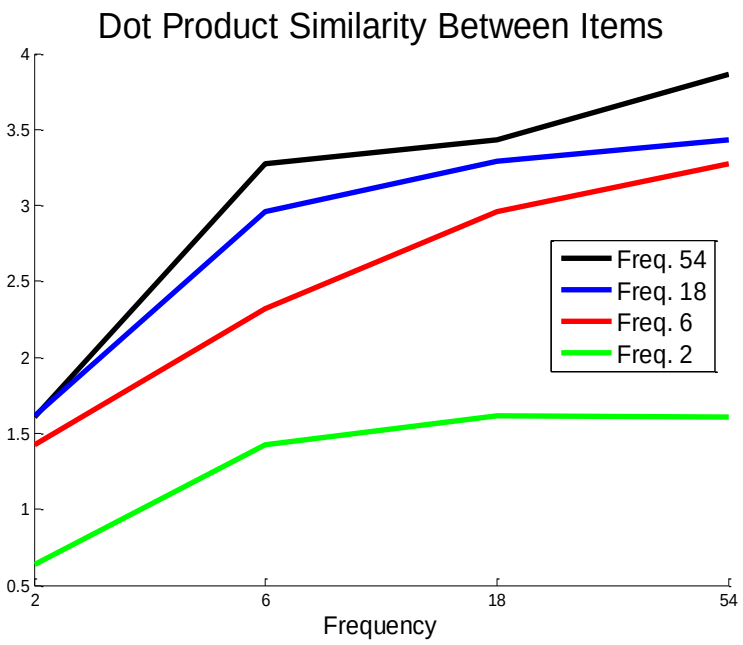

Dot Product Similarity Between Items
4
3.5
3
2.5
2
1.5
1
0.5
2
6
18
54
Frequency
Freq. 54
Freq. 18
Freq. 6
Freq. 2

Figure 7

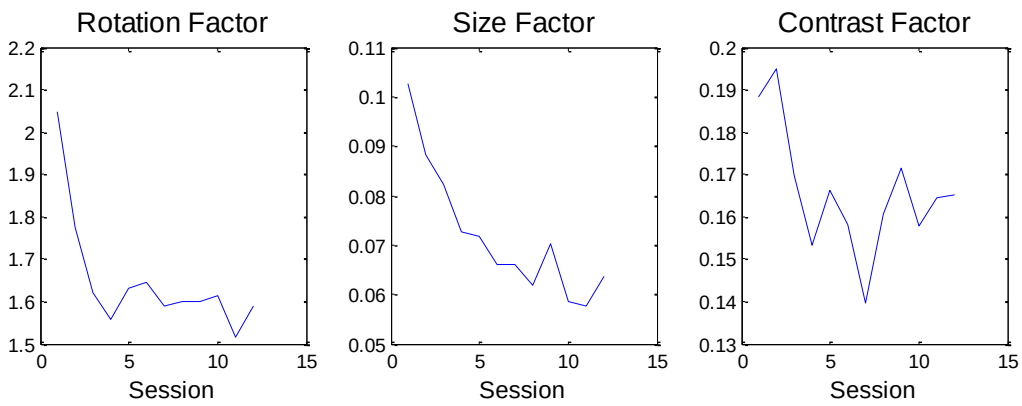

Rotation Factor
2.2
2.1
2
1.9
1.8
1.7
1.6
1.5
0
5
10
15
Session
Size Factor
0.11
0.1
0.09
0.08
0.07
0.06
0.05
0
5
10
15
Session
Contrast Factor
0.2
0.19
0.18
0.17
0.16
0.15
0.14
0.13
0
5
10
15
Session

Figure 8

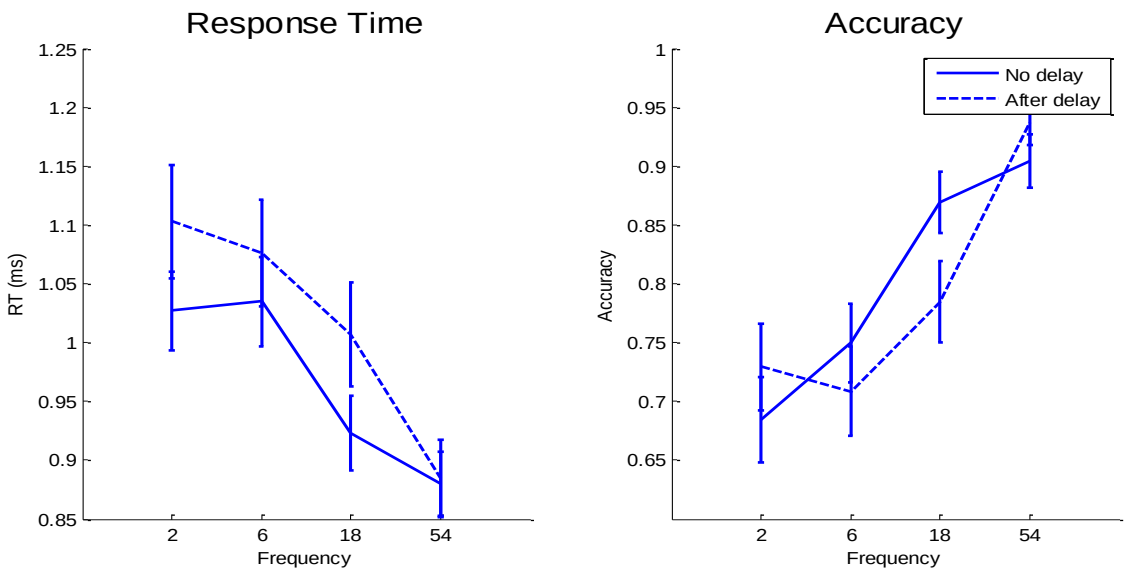
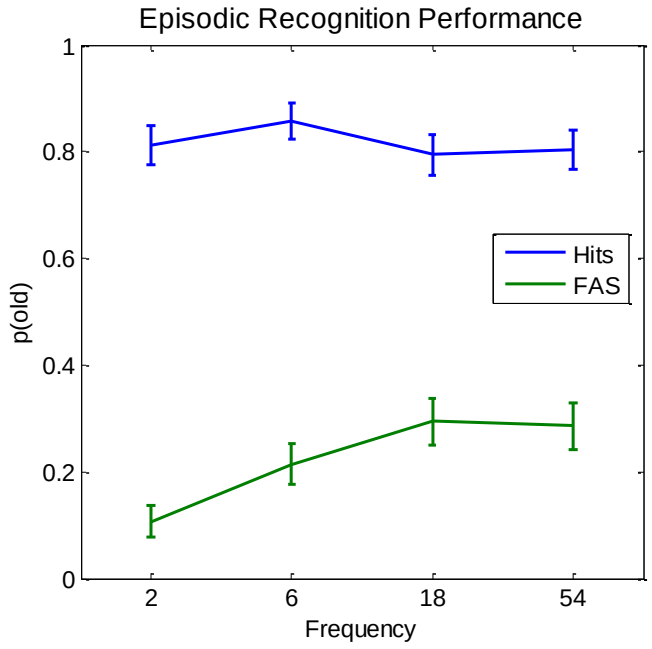

Response Time
RT (ms)
1.25
1.2
1.15
1.1
1.05
1
0.95
0.9
0.85
2
6
18
54
Frequency
Accuracy
Accuracy
1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
No delay
After delay
2
6
18
54
Frequency

Figure 9


Episodic Recognition Performance
1
0.8
0.6
0.4
0.2
0
p(old)
Hits
FAS
2
6
18
54
Frequency

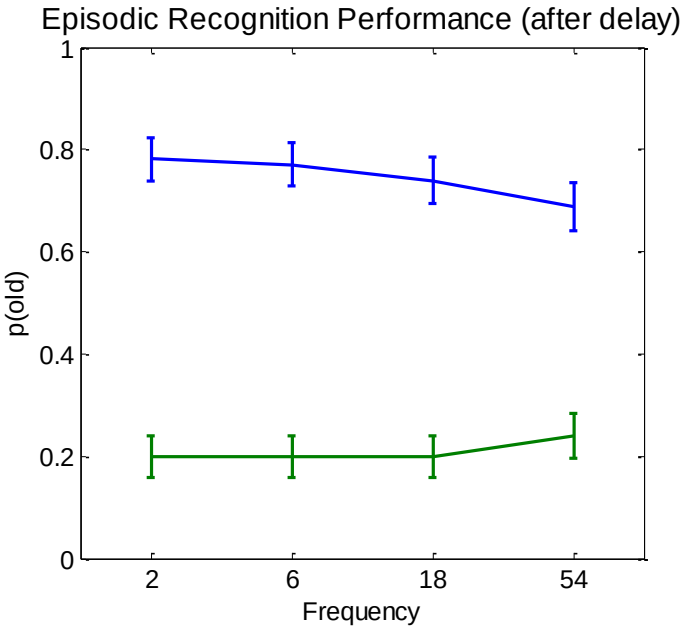

Episodic Recognition Performance (after delay)
1
0.8
0.6
0.4
0.2
0
p(old)
2
6
18
54
Frequency

Figure 10

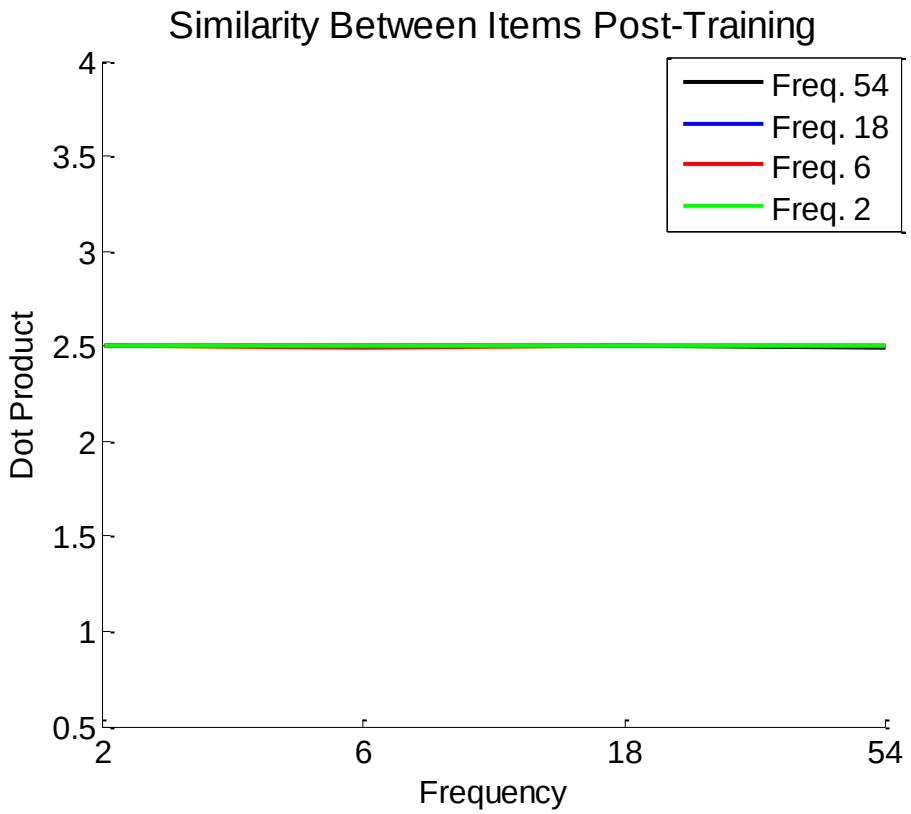

Similarity Between Items Post-Training
Freq. 54
Freq. 18
Freq. 6
Freq. 2
Dot Product
4
3.5
3
2.5
2
1.5
1
0.5
2
6
18
54
Frequency

Figure 11

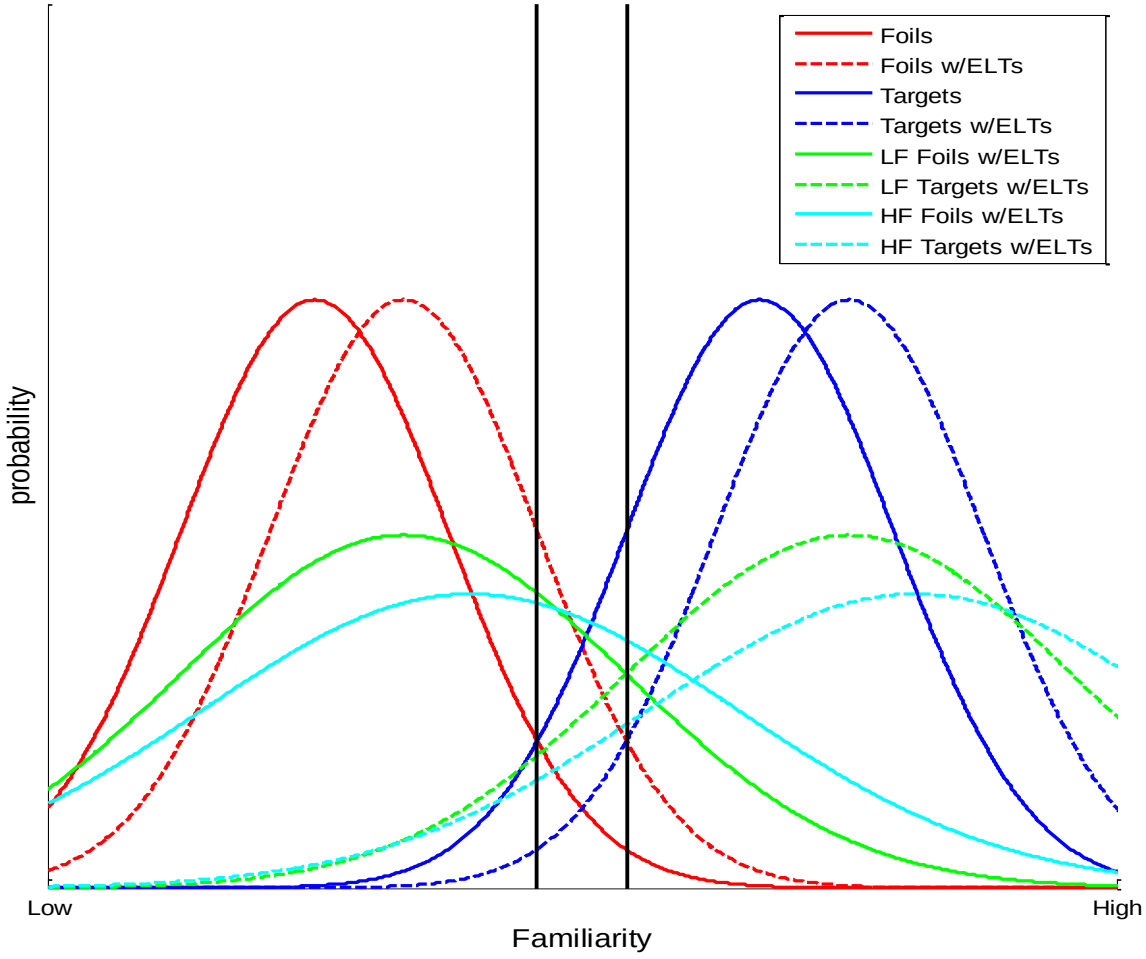

Foils
Foils w/ELTs
Targets
Targets w/ELTs
LF Foils w/ELTs
LF Targets w/ELTs
HF Foils w/ELTs
HF Targets w/ELTs
probability
Low
High
Familiarity

Figure 12

Episodic Recognition: Study Phase

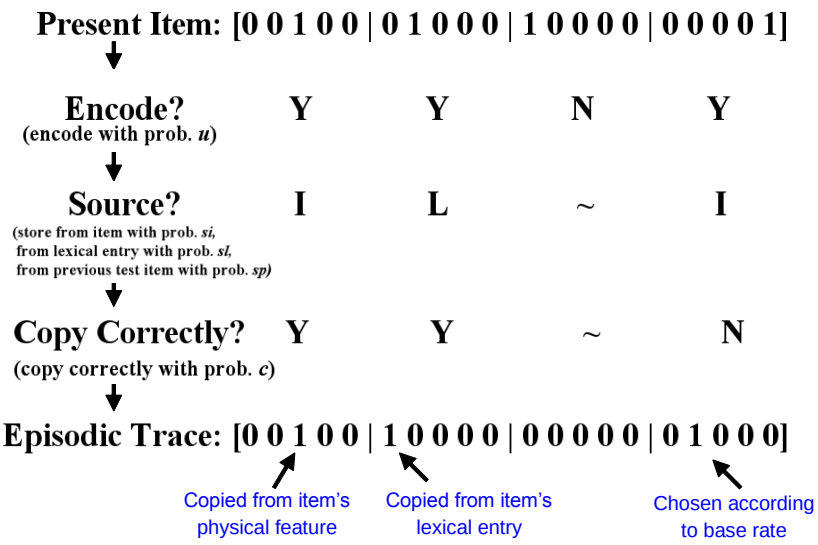


Figure 13

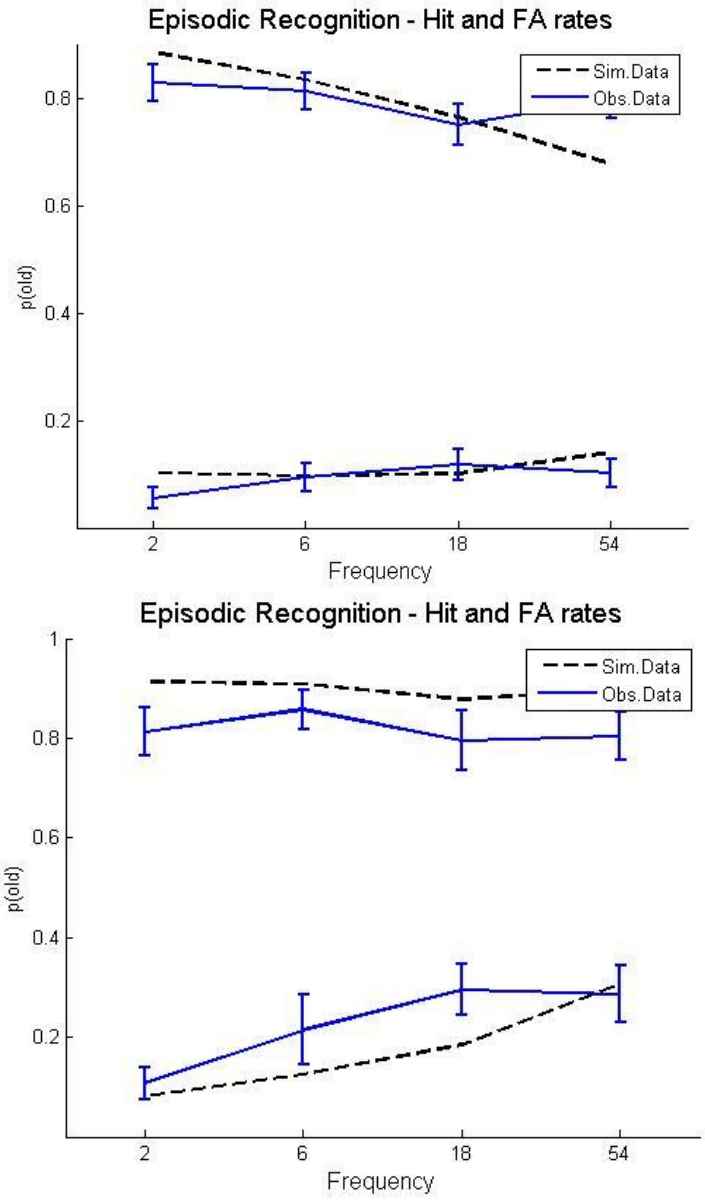

Episodic Recognition - Hit and FA rates
Sim.Data
Obs.Data
p(old)
0.8
0.6
0.4
0.2
2
6
18
54
Frequency
Episodic Recognition - Hit and FA rates
Sim.Data
Obs.Data
p(old)
1
0.8
0.6
0.4
0.2
2
6
18
54
Frequency

Figure 14

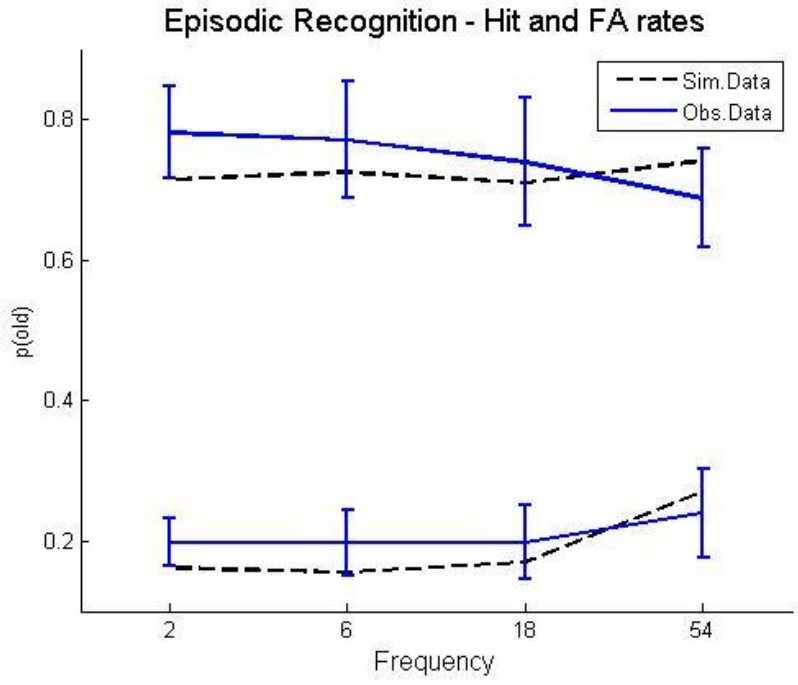

Episodic Recognition - Hit and FA rates
Sim.Data
Obs.Data
0.8
0.6
0.4
0.2
p(old)
2
6
18
54
Frequency

Figure 15

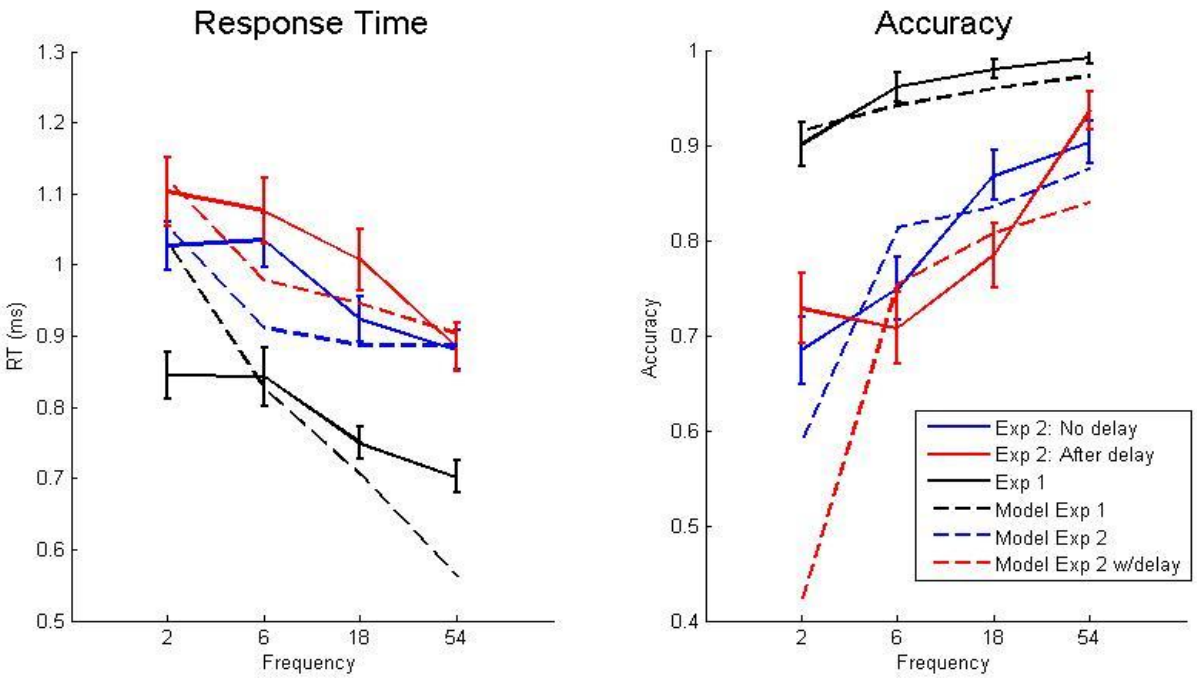

Response Time
RT (ms)
Frequency
Accuracy
Accuracy
Frequency
Exp 2: No delay
Exp 2: After delay
Exp 1
Model Exp 1
Model Exp 2
Model Exp 2 w/delay

Figure 16

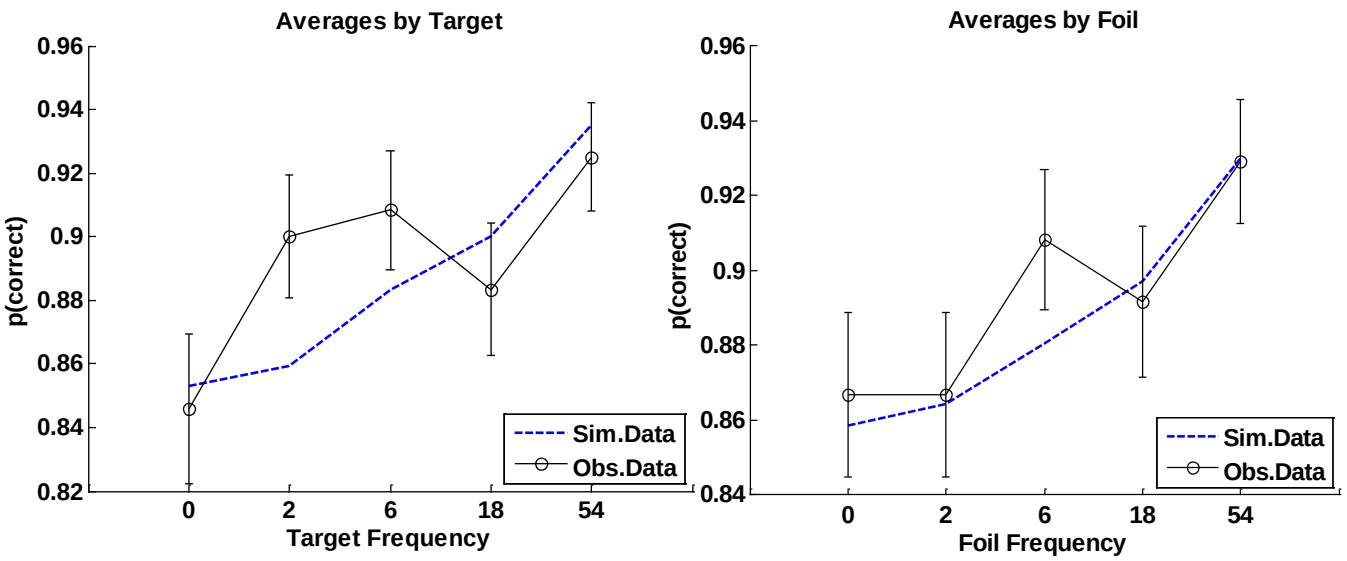

Averages by Target
p(correct)
0.96
0.94
0.92
0.9
0.88
0.86
0.84
0.82
0
2
6
18
54
Target Frequency
Sim.Data
Obs.Data
Averages by Foil
p(correct)
0.96
0.94
0.92
0.9
0.88
0.86
0.84
0
2
6
18
54
Foil Frequency
Sim.Data
Obs.Data

Figure 17

Appendix A – Additional Figures & Statistics

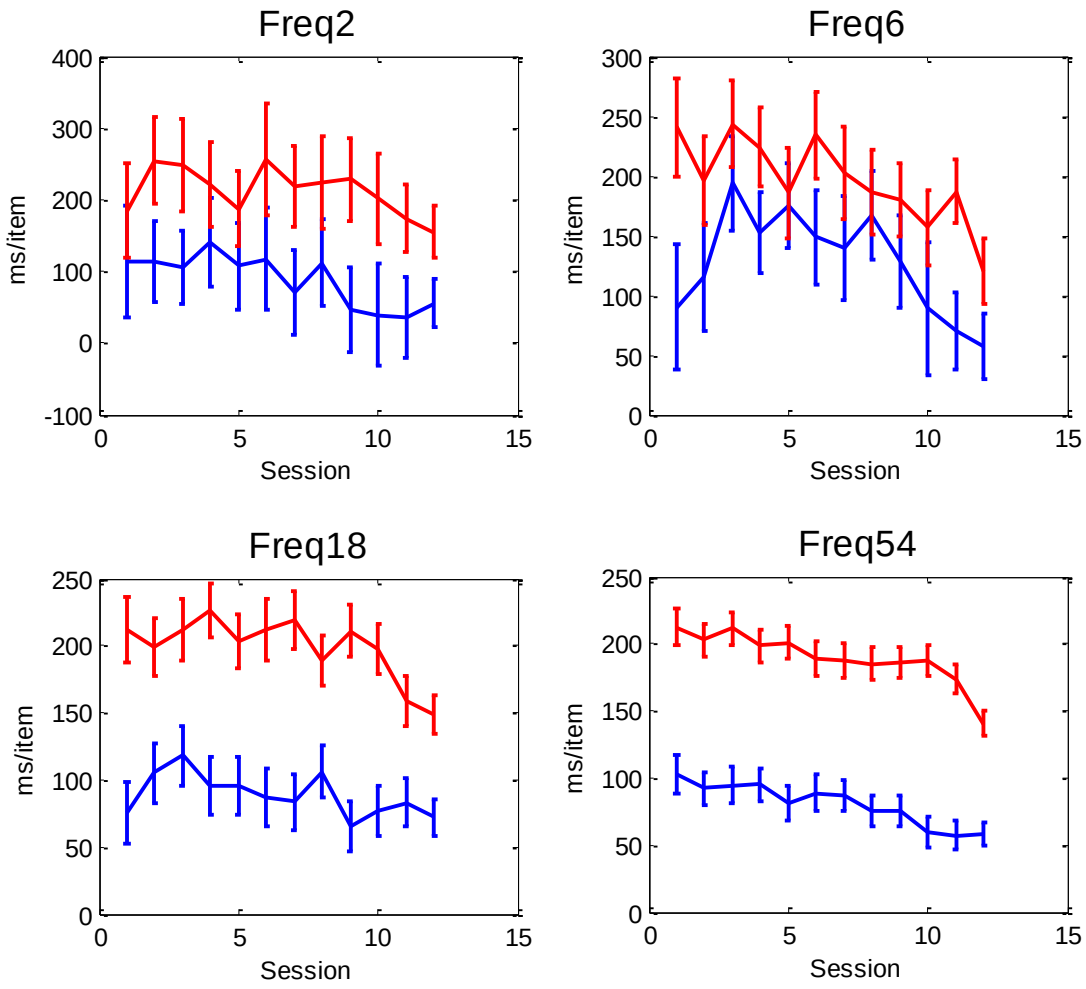


Figure A1 – Experiment 1, slope as function of training session, separated by item frequency.

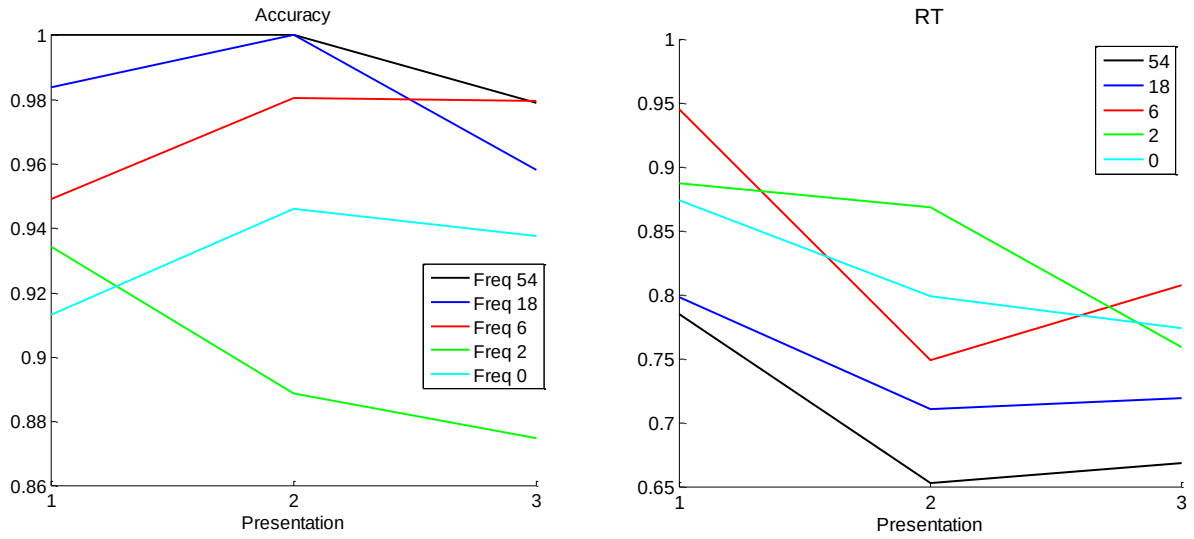


Figure A2 – Accuracy (panel A) and Response time (panel B) for Lexical Decision (Exp. 1) separated by 1st., 2nd, and 3rd presentation of each item.

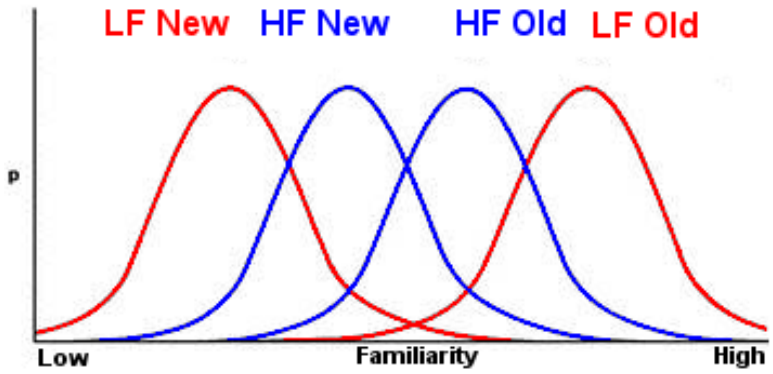


Figure A3 – Co-centered distributions of likelihood ratios for low frequency and high frequency studied (old) and unstudied (new) items.

Experiment 1: Analysis & Statistics

Episodic Recognition

A contrast analysis was conducted to examine the significance of the effects. For each subject, a dot product was taken of the vector of hit rates by frequency and the contrast vector [-

3,-1, 1, 3]. A dot product was also computed for false alarm rates and the contrast vector. A t-test was performed on the hit dot-products and false alarm rate dot-products to test whether they were significantly different than 0. The hit rate analysis showed a significant negative relationship between frequency and hit rates ($t(7) = -2.32$, $p= .059$), and the false alarm rate analysis showed a marginally significant positive relationship between frequency and false alarm rates ($t(7) = 1.96$, $p=.097$).

Lexical Decision

The significance of the trends in the behavioral data was assessed by conducting contrast analyses. For each subject, two dot products were computed: one between the vector of accuracy rates and the contrast vector [-3,-1, 1, 3], and another between the vector of response times and the same contrast vector. A t-test was used to examine whether the accuracy dot products were significantly different than zero, and found that there was a significant positive relationship between frequency and accuracy ($t(7) = 2.27$, $p=.057$). The same analysis was applied to response times, and found a significant negative relationship between frequency and response time ($t(7) = -2.30$, $p=.055$).

Forced Choice Perceptual Identification

A contrast analysis was run to examine the significance of the effect of target frequency, and of foil frequency. For each subject, two dot products were computed: one between the vector of accuracy rates by target frequency and the contrast vector [-3,-1, 0, 1, 3], and another between the vector of accuracy rates by foil frequency and the same contrast vector. A t-test was used to examine whether the dot products were significantly different than zero. The results

showed that the increase in performance due to target frequency was marginally significant ($t(5) = 1.94$, $p = .11$), as was the increase due to foil frequency ($t(5) = 1.73$, $p = .14$).

Experiment 2: Analyses and Statistics

Lexical Decision

A contrast analysis was used to look for consistent trends of frequency on accuracy and response time. For each subject, two dot products were computed: one between the vector of accuracy rates and the contrast vector [-3,-1, 1, 3], and another between the vector of response times and the same contrast vector. A t-test was used to examine whether the accuracy dot products were significantly different than zero, and found that there was a significant positive relationship between frequency and accuracy ($t(6) = 2.90$, $p=.03$). The same analysis was applied to response times, and found a significant negative relationship between frequency and response time ($t(6) = -2.97$, $p=.03$).

A linear regression analysis was also run on each subjects response times and accuracy. Each analysis produced a value β representing the slope of the best fitting regression line. The β values for accuracy (one β from each subject) were then tested for significance using a t-test. The analysis showed that the slopes of the regressions on accuracy were significantly greater than zero ($t(6) = 2.59$, $p=.04$). An analysis of the slopes from the regressions on response time were shown to be significantly less than zero ($t(6) = -2.45$, $p=.05$). These two results are in agreement with the contrast analyses above.

In addition to testing for consistent trends in each subject's data, a linear regression analysis was also used to analyze trends in the averaged data. The results or this analysis of

group means showed a (non-significant) negative effect of frequency for response time (β = -.002, r2=.054, p=.23), and a significant positive effect of frequency on accuracy (β = .004, r2= .147, p<.05).

Response time and accuracy were measured again approximately 6 weeks after the previous test session. The results followed the same pattern as they did 6 weeks prior: there was a significant positive relationship between accuracy and frequency for both the contrast analysis ($t(5) = 2.44$, $p = .059$) and individual regression b analysis ($t(5) = 2.54$, $p = .05$), and a significant negative relationship between response time and frequency for both the contrast analysis ($t(5) = -2.36$, $p = .06$) and individual regression b analysis ($t(5) = -2.45$, $p=.058$). Furthermore, a contrast analysis comparing the results of the delayed test to the immediate test showed that there was no significant decrease in the magnitude of the effects, either for accuracy ($t(5) = 1.14$, $p = .31$) or for response time ($t(5) = .51$, $p = .63$).

Episodic Recognition

A contrast analysis was performed to test whether consistent effects of frequency were present in each subject's data. For each subject, a dot product was taken of the vector of hit rates by frequency and the contrast vector [-3,-1, 1, 3]. A dot product was also computed for false alarm rates and the contrast vector. A t-test was performed on the hit dot-products and false alarm rate dot-products to test whether they were significantly different than 0. The hit rate analysis showed no significant difference from zero ($t(6) = -.387$, $p= .71$), but the false alarm rate analysis showed a significant positive relationship between frequency and false alarm rates ($t(6) = 3.19$, $p=.02$).

A linear regression analysis was also run to examine trends in the group data. This analysis showed a marginally significant positive relationship between frequency and false alarm rates (r^2 = .113, p=.08). There was no significant correlation between hit rates and frequency (r^2 =.008, p=.66).

The results were also examined by analyzing effects of frequency on d'. A contrast analysis was conducted to look for consistent trends over subjects, and found a marginally significant decrease in d' due to increased frequency (t(6) = -1.86, p=.11). A linear regression on the group d' data did not show a significant relationship (r^2 =.053, p=.24).

Six of the seven subjects were tested again following a six week delay. Linear regression found no significant relationship between hit rates and frequency (r^2 = .041, p=.34), or false alarm rates and frequency (r^2 = .023, p= .48). Furthermore, a contrast analysis showed that there was no significant effect of frequency on hit rates (t(5) = -1.12, p = .31) or on false alarm rates (t(5) = .605, p=.57). Analysis of d' after delay found no significant effect in the contrast analysis (t(5) =-.989, p = .37) or in the linear regression analysis (r^2 = .017, p=.54). A contrast analysis was also used to examine the change in magnitude for the delayed test vs. immediate test. This analysis showed that there was no significant difference in the magnitude of the hit rate effect (t(5) = .30, p = .77), but there was a marginally significant change in the false alarm rate effect (t(5) = 2.11, p = .09).

Lastly, a t-test (of non-paired samples) was conducted examining the change in magnitude of effects from experiment 1 to experiment 2. This analysis found that there was no significant difference in the magnitude of the hit rate effect (t(13) = -.27, p = .79), but there was a significant change in the false alarm rate effect (t(13) = -2.24, p = .04).

Appendix B – Consistent Context Experiment

As discussed above, previous results by Adelman et al. (2006), in combination with the predictions of the SARKAE model in its original form, have called into question the effect of frequency in the absence of contextual diversity variability. The aim of the current experiment is therefore to remove the differences in contextual diversity that were previously associated with frequency; that is, during training on novel items, contextual diversity will be held constant while frequency of occurrence is varied. Following this controlled or consistent context training, the same post-training tasks as used in experiment 1 of this paper will be conducted to assess the effects of frequency in the absence of contextual diversity.

Training: Visual Search

Method

Subjects. Six people, recruited through an email advertisement, participated in the experiment for monetary compensation. All participants reported no prior experience with Chinese characters.

Apparatus. All tasks were displayed on Samsung SyncMaster 700NF 17" flatscreen CRT monitors, and responses were collected through keyboard presses. Experiments were run using the programs Authorware and MATLAB. Participants were seated in dark booths with ventilation fans that greatly reduced ambient noise.

Design and Stimuli. The occurrence of the characters in the visual search task was manipulated to produce four frequency conditions which varied in a ratio of 2::6::18::54. For every two presentations of a character as a target, it was also present 5 times throughout the session as a foil. 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. A varied mapping approach was used for the visual search task, meaning that the character was present throughout the task as both a target and a foil.

In order to keep the context consistent for each frequency group, the foils for each trial were randomly drawn only from characters that belonged to the same frequency category as the target item.

Procedure. Each trial was initiated by a key press by the participant, which was followed by a fixation cross. Next, a target character was presented centrally for 1000 ms followed by a blank screen for 500 ms. A display of either two or four characters then appeared and remained until a response was made. For the display size of four, the characters were positioned evenly in each quadrant of the screen, and for display size two, the characters were randomly placed in two of the four possible positions. The task of the participants was to respond as quickly as possible whether or not the target character was present in the display set. The target was present on half of the trials. There were a total of 640 trials per session, and 10 training sessions were completed by each subject.

Results

The results of training were examined by measuring response time during the visual search task, and deriving the slope and intercept of the search function. The slopes were calculated by subtracting the response time from visual search of a display of 2 items from the response time to a display of 4 items and dividing by 2. This was done separately for target

present and target absent trials; the mean slope of the 6 subjects is shown as a function of training session in Figure B1 (panel A) along with the mean intercept over session (panel B).

A linear regression was run separately on present slope, absent slope, present intercept, and absent intercept, and found that all decreases except absent slope were significant ($p<.05$). The results of the regression analyses are shown in table A1. A contrast analysis was also run on the data from each individual subject. A dot product was computed for each subject multiplying that subject's slopes/intercepts for sessions 1-10 by the vector [-9 -7 -5 -3 -1 1 3 5 7 9]. A t-test was then run on the resulting values to determine if they were significantly different from zero. The results of these analyses found a marginally significant negative effect of session on present slope ($t(5) = -1.83$, $p=.13$), but no significant effect of session on absent slope. The contrast analysis also showed a marginally significant negative effect of session on both present intercept ($t(5) = -2.37$, $p = .06$) and absent intercept ($t(5) = -2.28$, $p = .07$).

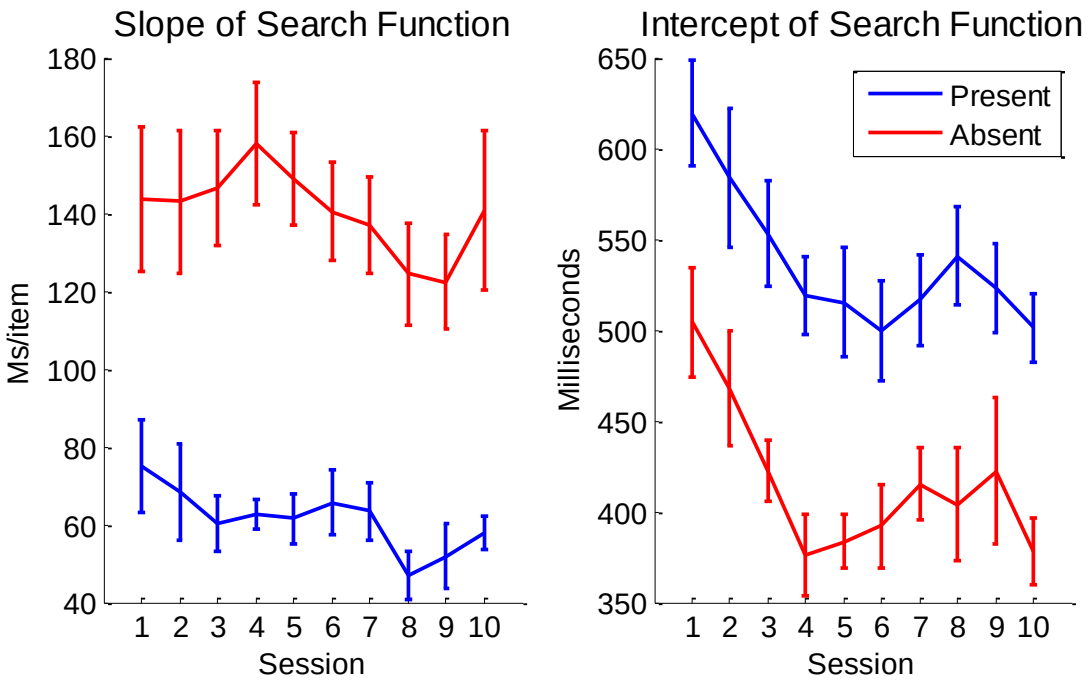


Figure B1 – Slope (panel A) and Intercept (panel B) of Search Function in the Visual Search task over training. The blue line shows results for target present trials, the red line for target absent trials.

Table B1 – Results of regression analyses on training data

Data:	β_0	β_1	r^2	F	p
Present Slope	72.37	-2.01	.087	5.50	.023
Absent Slope	152.25	-2.14	.030	1.78	.188
Present Intercept	589.27	-9.50	.145	9.84	.003
Absent Intercept	463.74	-8.61	.125	8.30	.006

Discussion

Consistent with the results found by Shiffrin and Lightfoot (1997) in a similar training study, the analysis of the current experiment showed a significant decrease in search rate over training session. This result supports the hypothesis that the subjects are learning the characters throughout the training sessions. Shiffrin and Lightfoot interpreted the decrease in slope over training as resulting from the subjects forming a unitized representation of the stimuli. However, since the Chinese character stimuli were not controlled in the same way that the stimuli in that study were, it is unclear how processing is changing over training. However, given that the main objective of the current study is to examine how processing in the post-training tasks is influenced by contextual effects during training, it is enough to show that subjects were

exhibiting learning through in increase in search efficiency, and thus have built up knowledge about the characters of varying frequencies to differing degrees.

Post-Training Tasks

Following the training on the visual search task, the subjects completed three post-training tasks: pseudo-lexical decision, episodic recognition, and forced choice perceptual identification. All subjects completed the tasks in the order that they are presented here.

Pseudo-Lexical Decision

Method

Subjects. All six subjects who were trained on the characters completed this task shortly after their final training session.

Design and Procedure. Subjects viewed one list, which contained all 32 trained characters, 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. The placement of the characters in the list was randomized. Subjects were presented with a single character on the screen, and were asked to decide as quickly as possible whether they had ever seen that character during any of the previous training sessions. Responses were made by pressing either the 'v' or 'm' button on the keyboard.

Results

Response times and accuracy were analyzed for each of the four frequency groups, as well as for new items. The results for trained items are shown in Figure B2. A contrast analysis

was used to test the significance of the effect of frequency on accuracy and response time. For each subject, two dot products were computed: one between the vector of accuracy rates and the contrast vector [-3,-1, 1, 3], and another between the vector of response times and the same contrast vector. A t-test was used to examine whether the resulting dot-products were significantly different from zero. The analysis of accuracy showed that there was a marginally significant positive relationship between frequency and accuracy ($t(5) = 1.93$, $p=.11$), and the analysis of response times found a significant negative relationship between frequency and response time ($t(5) = -3.12$, $p=.03$).

A linear regression analysis was also run on each subjects response times and accuracy. Each analysis produced a value b representing the slope of the best fitting regression line. The β values for accuracy (one β from each subject) and the β values for response time were then tested for significance using a t-test. The analysis found that the positive effect of frequency on accuracy again showed only marginal significance ($t(5) = 2.12$, $p=.09$), but the negative effect of frequency on response time was shown to be significant ($t(5) = -4.49$, $p=.01$). These two results are in agreement with the contrast analyses above.

To examine the effects of frequency in the grouped data (mean of all subjects), a linear regression analysis was conducted. This analysis produced a significant relationship between frequency and accuracy ($r^2 = .173$, $p=.04$), but did not show a significant correlation between frequency and response time ($r^2 = .049$, $p=.30$). However, this result should be interpreted cautiously, as it treats each data point as if it were independent. In reality, the data points at each frequency level for a given subject are highly dependent.

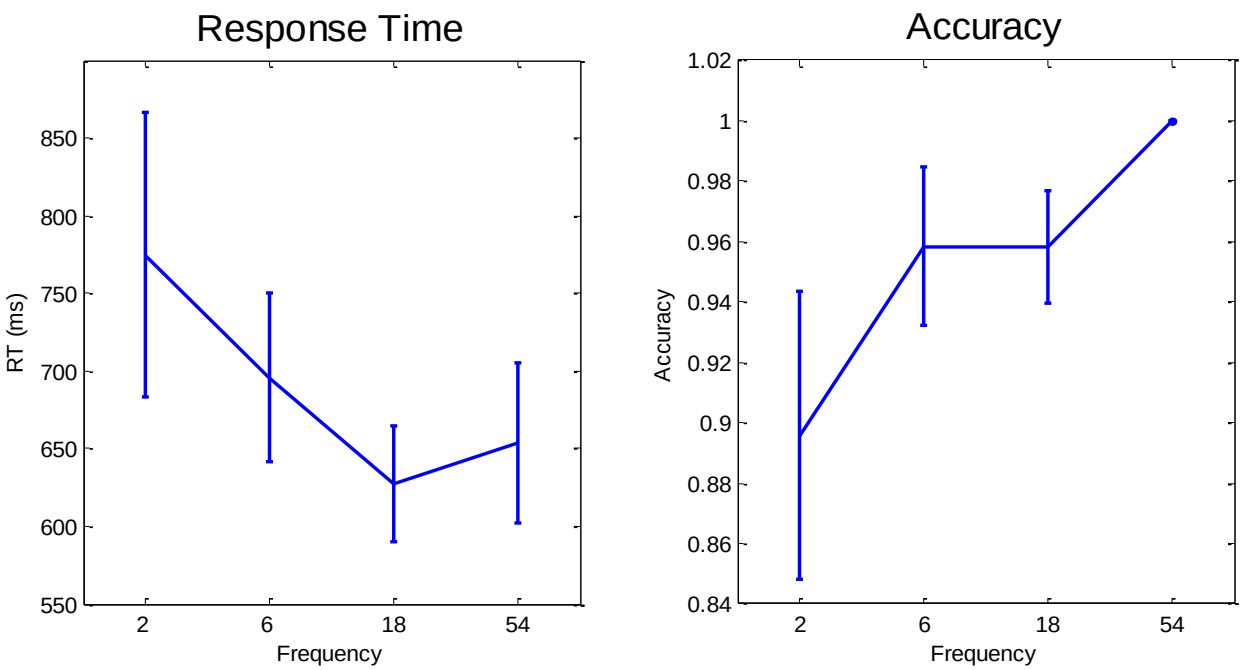


Figure B2 – Pseudo-Lexical Decision Results. Panel A shows the mean response time in milliseconds for items from each frequency group (for correct responses only). Panel B shows the mean accuracy for items from each frequency group.

Discussion

The results of this post-training task are consistent with both normative word frequency findings as well as the results of experiment 1 above. Even though the effect of frequency on accuracy was only marginally significant in the contrast and individual regression analyses, the data shows a clear trend of faster response times and higher accuracy for more highly trained items. In terms of the current contextual manipulation, it appears that equating the diversity of the surrounding context of items did not disrupt the frequency results found in this lexical decision task. This suggests that the strength of a knowledge trace itself, not its overlap or

similarity with other items in knowledge, is the principal source of the frequency effects that occur in this task.

Episodic Recognition

Method

Subjects. All six subjects who were trained on the characters completed this task shortly after their final training session.

Design and Procedure. This 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 effects of frequency in this task were investigated by examining the hit rates (probability of correctly identifying a studied item as old) and false alarm rates (probability of

incorrectly identifying an unstudied item as old). The hit rates and false alarm rates (averaged over all subjects) are shown in figure B3.

A contrast analysis was performed to test whether consistent effects of frequency were present in each subject's data. For each subject, a dot product was taken of the vector of hit rates by frequency and the contrast vector [-3,-1, 1, 3]. A dot product was also computed for false alarm rates and the contrast vector. A t-test was performed on the hit rate dot-products and false alarm rate dot-products to test whether they were significantly different than 0. The hit rate analysis showed no significant difference from zero ($t(5) = 1.53$, $p = .19$), but the false alarm rate analysis showed a significant positive relationship between frequency and false alarm rates ($t(5) = 2.88$, $p = .03$).

A linear regression analysis of group averaged data found that neither was significant (hit rates: $\beta = .0012$, $r^2 = .016$, $p = .56$; false alarm rates: $\beta = .002$, $r^2 = .065$, $p = .23$).

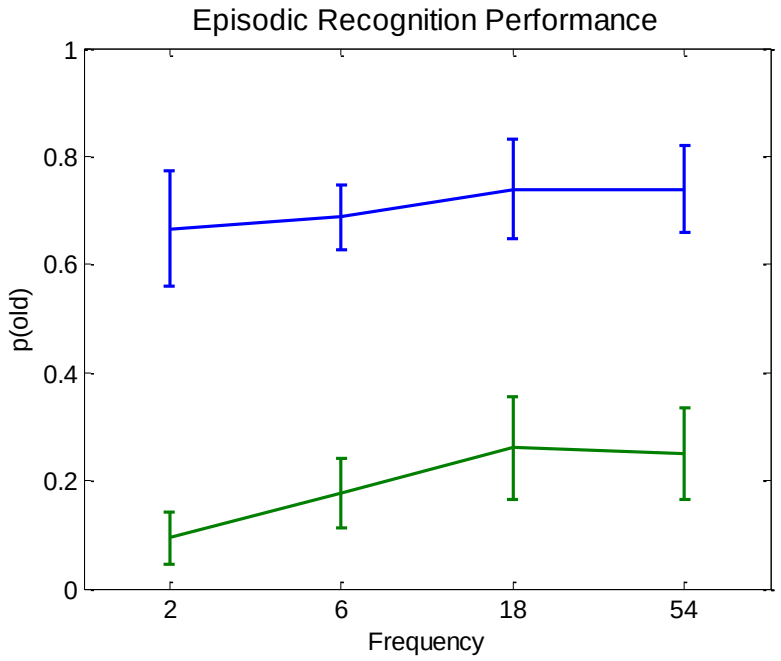


Figure B3 – Mean hit (blue line) and false alarm (green line) rates for each frequency group.

The results from this task were examined further through d-prime analyses. A contrast analysis as described above was performed, and found that there was not a significant decrease in d-prime due to frequency ($t(5) = -1.18$, $p = .29$). A regression analysis of the averaged data also did not find a significant relationship between frequency and d-prime ($r^2 = .062$, $p>.05$).

Discussion

In episodic recognition tasks using words as stimuli, it is reliably found that low frequency items produce better performance, and that a mirror effect occurs: low frequency words produce more hits and fewer false alarms than high frequency words. Many theories have been proposed to explain the mirror pattern, and include such factors as attention (Glanzer & Adams, 1990), context (Sikstrom, 2001), and a dual-process of familiarity and recollection (Reder et al., 2000). Furthermore, the results of experiment 1 showed that exposure differences that occurred during training produced an effect like that found for words: low frequency trained items produced better performance in episodic recognition, in the form of both more hits and fewer false alarms.

The results found in the current experiment indicate that differences in the diversity of context (as defined by the distractor items which surround an item during training) do not significantly alter the effects of frequency on false alarm rates in an episodic recognition task. However, the mirror pattern is not found for this training paradigm. In fact, the slope found by the linear regression analysis of hit rates was positive, contradicting the decrease in hit rates for

high frequency items found in previous studies. The data also did not show a significant effect of frequency on d-prime as is usually found for episodic recognition.

There are two possible explanations for the combination of similarities and differences to previous results found in the current data. First, it may be that the contextual manipulation did not have an effect, and disappearance of the hit rate effect is due to chance. Previous studies have shown that the hit rate effect is less robust than the false alarm rate effect in this task (Criss & Shiffrin, 2004b), so it is possible that it failed to occur by chance. The other possible explanation is that the contextual manipulation disrupted the hit rate but not the false alarm rate for the same reason: the hit rate effect is less robust, and therefore was eliminated due to the consistent context training even though the false alarm rate effect was not.

Forced Choice Perceptual Identification

Method

Subjects. All six subjects completed the forced choice perceptual identification shortly after their final training session.

Design and Procedure. This task consisted of five lists, each with 50 two-alternative forced choice trials. The first list was used to adjust the length of presentation to a 75 percent correct threshold, using the Best-PEST algorithm (Lieberman & Pentland, 1982). This adjusted presentation speed was restricted to the range of between 20 and 80 milliseconds. Each subject's individual threshold presentation speed was used for the four test lists. Throughout the task, every combination of foil and target frequency was tested (frequencies 0, 2, 6, 18, 54), creating a total of 25 conditions.

For each trial, subjects viewed a character flashed briefly on the screen, and were then presented with two characters. The subjects were asked to choose which of the two characters matched the character that had been briefly presented just prior. These two characters stayed on the screen until a decision was made, and the correct answer was always one of the choices. Subjects completed one block of 50 speed adjustment trials and four blocks of 50 trials at their established presentation speed. Only data from the last four blocks were analyzed.

Results

The proportion of correct responses was measured for each condition of target frequency and foil frequency. Each combination of target and foil frequency was tested, making a total of 25 conditions. These 25 conditions were collapsed separately across target frequency and foil frequency for analysis (i.e. – the mean at target frequency = 0 is the mean across all foil frequencies of trials where the target frequency was 0). Figure B4 shows the mean performance as a function of target frequency (panel A) and as a function of foil frequency (panel B). The two plots each contain data from all 25 conditions but are organized differently: according to target frequency in panel A, and according to foil frequency in panel B.

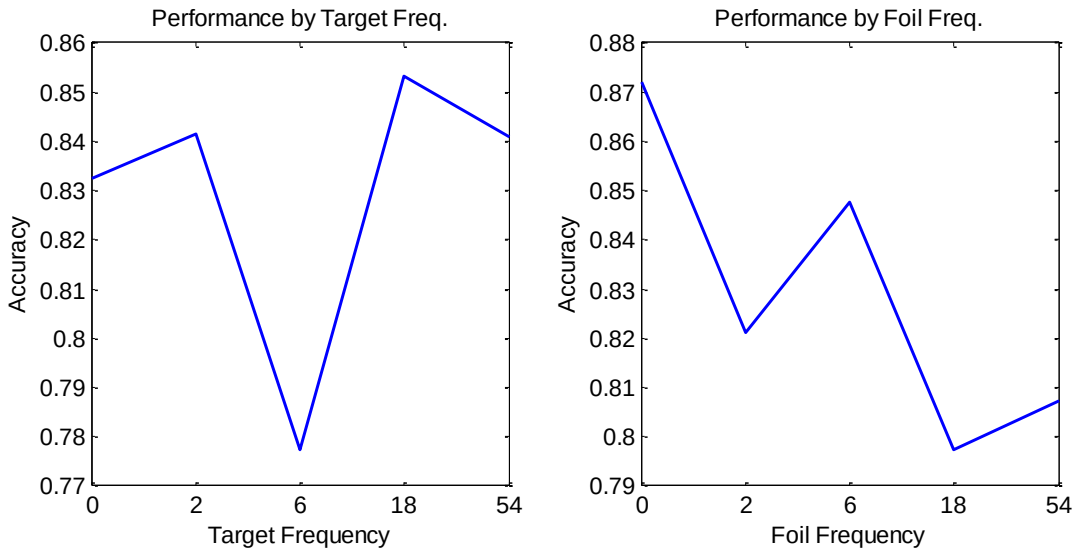


Figure B4 – Forced choice perceptual identification performance, mean of all subjects as a function of target frequency (panel A) and foil frequency (panel B).

A contrast analysis was performed on the data to find trends consistent over subjects. The results of the analysis showed no significant effect of target frequency ($t(5) = -.03$, $p = .98$), and a marginally significant negative effect of foil frequency ($t(5) = -1.81$, $p = .13$).

Considering the fact that the visual search training always presented items of the same frequency category together, it follows that discriminating between items of the same frequency category may produce effects that are lost in the collapsed analyses above. As such, performance on trials with equal target and foil frequency was examined more closely. A contrast analysis of pairs of all frequencies (new items included as "0" frequency) did not show a significant effect of frequency ($t(5) = .208$, $p = .84$). However, the curves tended to form a u-shaped pattern (see Figure B5), with performance decreasing from high to low frequency but then jumping up again for pairs of new items. A contrast analysis was therefore performed on

just the trained items, and found a significant increase in performance as frequency increased ($t(5) = 2.55$, $p = .05$).

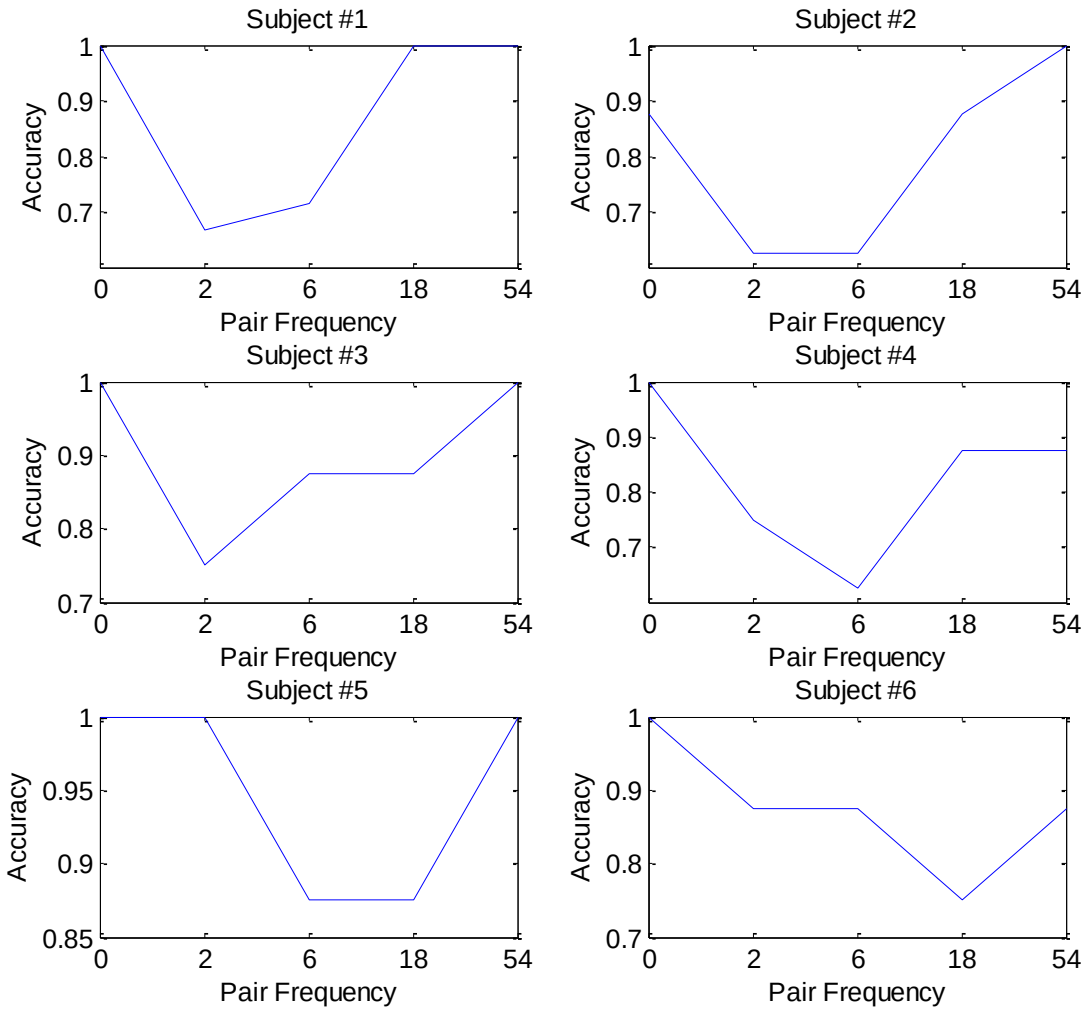


Figure B5 – Forced choice perceptual identification performance for each subject when the target and foil pairs were of equal frequency.

Performance was also examined as a function of frequency when the target and foil came from different frequency categories (i.e. – when the pairs of equal frequency examined above

were removed from the data). A contrast analysis was first run on all pairs of unequal frequency, organized by foil frequency and then by target frequency. This analysis showed no significant effect of target frequency on performance ($t(5) = -0.03$, $p = 0.97$), and again only a marginal negative effect of foil frequency ($t(5) = -1.87$, $p = 0.12$). However, given the result above indicating that performance for new items differed from trained items, this analysis was performed again using only pairs of trained items of unequal frequency. The results showed no significant effect of target frequency ($t(5) = -0.49$, $p = 0.64$), but a significant negative effect of foil frequency ($t(5) = -2.61$, $p = 0.05$). Figure B6 shows performance of unequal trained pairs by target frequency (panel A) and foil frequency (panel B).

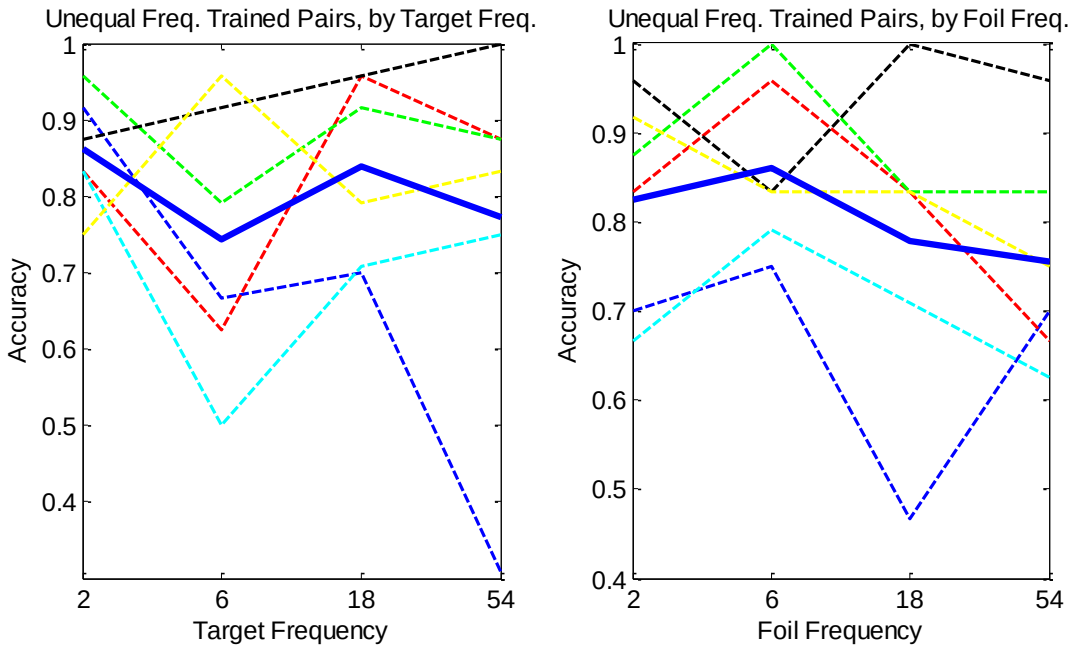


Figure B6 – Forced Choice Perceptual Identification Performance for pairs of trained items of unequal frequencies. Panel A shows performance as a function of target frequency, panel B shows performance as a function of foil frequency. Both panels show individual subject data with dashed lines, and the mean of all subjects with a solid bold line.

Discussion

The results of this portion are important because, as did the results of the episodic recognition task, they show a large divergence from the previous findings. Whereas experiment 1 found a marginally significant ($p = .14$) positive relationship between performance and foil frequency, the results of the current study showed a marginally significant negative effect of frequency. Similarly, experiment 1 found a marginally significant ($p = .11$) positive relationship between target frequency and performance, and the current study showed absolutely no overall effect of target frequency on performance. Clearly something about the consistent context in the current training paradigm altered the processing or strategies used in this perceptual task.

Further evidence for the influence of the consistent context present in the training paradigm comes from the analyses of same-frequency pairs. The current experiment found a significant increase in performance due to higher training frequency (when new items were excluded) for pairs of equally trained items. This indicates that as training progresses, subjects become better at distinguishing between items of a given training group, which makes sense given that this is the distinction they are required to make during the visual search training task. The jump in performance for pairs of two new items is slightly harder to explain. One possibility is that since these items are new, subjects are forced to rely on a sense of what they thought they saw and make a guess (which is usually quite accurate), as opposed to trying to fill in the gaps in their perception with what they know about the target, which could lead to inaccuracies for low frequency trained items. This U-shaped pattern of same frequency pairs, while interesting, is a result that needs to be supplemented with further data before any solid theoretical conclusions can be reached.

The analyses of performance on pairs of unequal frequency produced basically the same findings as the overall analyses: no effect of target frequency, and a marginally significant negative effect of foil frequency on performance. This leads to an interesting question regarding the positive frequency effect for equal frequency pairs: is this effect due to the target frequency, the foil frequency, or perhaps an interaction? The training paradigm would suggest that it is the interaction of the two, or more directly put, the ability of the subjects to distinguish between the two items, that leads to the increase in performance as frequency increases. This increase in distinguishing ability as a result of further training is able to overcome the negative effect that foil frequency has on performance in the case of unequal frequency pairs, and produce instead the increase in performance shown for equal frequency pairs.

To put these results in modeling terms, if low-level features were used for this task, then the relationship between frequency and performance for pairs of equal frequency should be negative, not positive. However, the SARKAE model posits that the high level feature, not the low level features, is what is extracted from the flash and used to make perceptual identification decisions. Our result of a positive relationship between frequency and pairs of equal frequency therefore is consistent with the SARKAE model, for the following reasons. If during training, an item from a given frequency class need only be distinguished from other items in that same class, it is logical for the system to create a high level feature which differentiates the item from other members of its class. Therefore, when two items of the same frequency are the two choices in the perceptual identification task, higher frequency improves performance, presumably because the high level feature in high frequency items is more fully developed. Frequency does not improve performance in perceptual identification for pairs of unequal frequency because the high

level feature was developed to distinguish an item only from other items in its class, and thus may not distinguish it well from items of a different class.

General Discussion & Modeling

The previous experiment (described in experiment 1) used a variable training paradigm that, in terms of the model, led to differences in contextual diversity between high and low frequency items. These contextual variations created differences in the similarity structures of high and low frequency items, which led to the model producing frequency effects matching those found in the behavioral data. The Consistent Context experiment was designed in an attempt to equate the diversity in lexical representation for high and low frequency items.

The SARKAE model was run under the consistent context training paradigm, using the same processes and parameters as were used in simulating the original variable context experiment (see experiment 1 for a description of the original model formulation). Following training, similarity between items of varying frequency was analyzed by taking the dot product of the normalized lexical entries. Figure B7 shows the similarity between items of a given frequency (each individual line is one frequency category) in comparison to other items of each frequency group (plotted on the x-axis). The results of this analysis showed that as frequency increases, similarity tends to increase (as is the case for variable context). However, there is also a spike in similarity when items are being compared to other items in their own frequency category (i.e. – when a frequency "6" item is being compared to other frequency "6" items). This increase in familiarity is due to the fact that during training a target item and its surrounding distractor items always had identical frequencies (i.e. – if the target was a frequency "6" item, then the distractors were also frequency "6" items). Therefore, the majority of the surrounding

contextual information that was stored into the target's lexical representation was gathered from items that belonged to the same frequency category, causing items in each category to become more similar to one another.

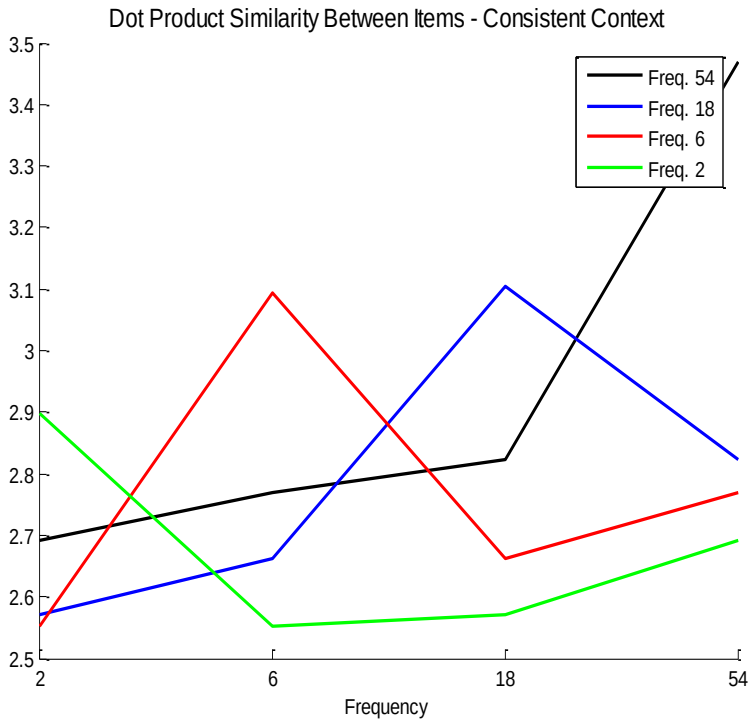


Figure B7 – Similarity between items following training, measured by the dot product of their normalized lexical entries.

The model in its original form is thus highly sensitive to the contextual differences present during training. It is these training differences that produce differences in similarity between items, leading to the frequency effects in the post-training tasks. The experiment just discussed shows that keeping context relatively consistent between frequency groups does change the frequency effects slightly, however, confounds created by this design (such as the development of the HL feature mentioned in the forced-choice discussion and the similarity structure of items discussed above) make it difficult to interpret the effect of context. Therefore, in order to provide a better picture of the true effects of context, instead of keeping context

consistent, frequency effects should be examined in the absence of context, as is the focus of experiment 2.

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.

Criss, A. H., & Shiffrin, R. M. (2004b). Interactions Between Study Task, Study Time, and the Low-Frequency Hit Rate Advantage in Recognition Memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*, 778-786.

Glanzer, M., & Adams, J. (1990). The Mirror Effect in Recognition Memory: Data and Theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*, 5-16.

Lieberman, H., & Pentland, A. (1982). Microcomputer-based estimation of psychophysical thresholds: The Best PEST. *Behavior Research Methods & Instrumentation, 14*, 21-25.

Reder, L.M., Nhouyvanisvong, A., Schunn, C.D., Ayers, M.S., Angstadt, P., & Hiraki, K. (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.

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

Sikstrom, S. (2001). The variance theory of the mirror effect in recognition memory. *Psychonomic Bulletin & Review, 8*, 408-438.

Appendix C – Forced-Choice Perceptual Identification results of Experiment 2

Method

Subjects. All seven subjects completed the forced choice perceptual identification shortly after their final training session. This task was not completed again during the re-test session after the 6 week delay.

Design and Procedure. This task consisted of five lists, each with 50 two-alternative forced choice trials. The first list was used to adjust the length of presentation to a 75 percent correct threshold, using the Best-PEST algorithm (Lieberman & Pentland, 1982). This adjusted presentation speed was restricted to the range of between 20 and 80 milliseconds. Each subject's individual threshold presentation speed was used for the four test lists. Throughout the task, every combination of foil and target frequency was tested (frequencies 0, 2, 6, 18, 54), creating a total of 25 conditions.

For each trial, subjects viewed a character flashed briefly on the screen, and were then presented with two characters. The subjects were asked to choose which of the two characters matched the character that had been briefly presented just prior. These two characters stayed on the screen until a decision was made, and the correct answer was always one of the choices. Subjects completed one block of 50 speed adjustment trials and four blocks of 50 trials at their established presentation speed. Only data from the last four blocks were analyzed.

Results

Contrast analyses were performed as described in the forced choice analysis section of the experiment detailed in Appendix B. Although theoretically speaking pairs of equal frequency in this task should be no different than pairs of unequal frequency (given the current

no-context training paradigm) these analyses were included for the sake of comparison and thoroughness. In the 10 contrasts that were analyzed, only one was marginally significant: the positive relationship between performance and target frequency ($t(6) = 2.08$, $p = 0.08$). Performance broken down by target frequency is shown in panel A of Figure C1, as is performance by foil frequency (panel B, no significant pattern). The results of all contrast analyses run are shown in Table C1.

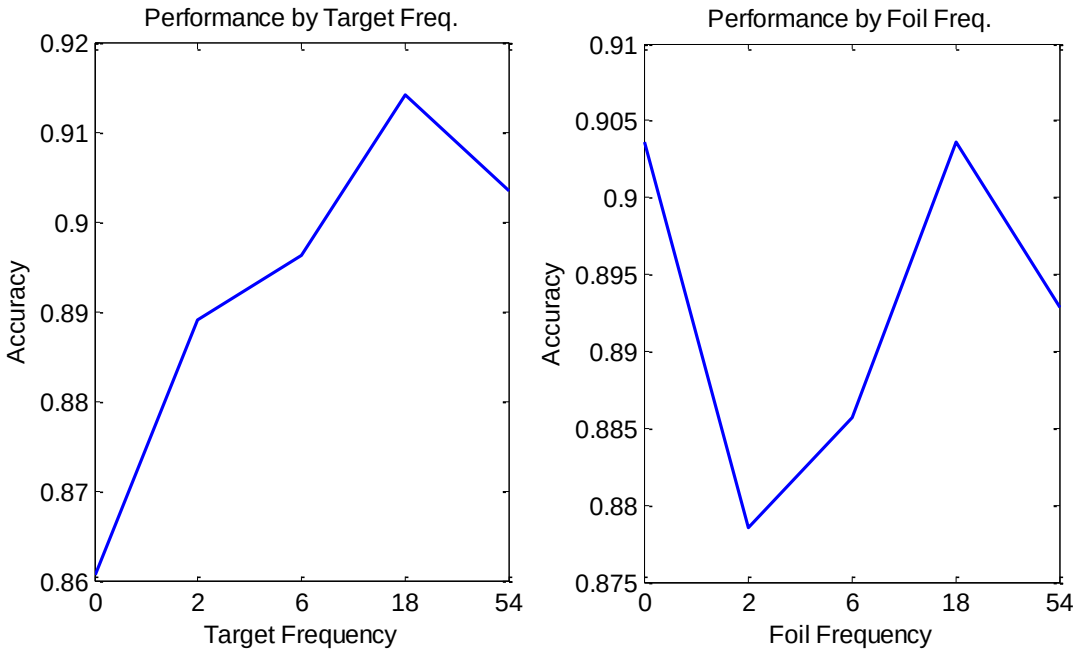


Figure C1 - Forced Choice Performance: as a function of target frequency (panel A) and foil frequency (panel B).

Table C1 – Contrast Analyses for Forced Choice Perceptual Identification

Contrast Type:	**df**	**t**	**p**
All items by target frequency	6	2.08	0.08
All items by foil frequency	6	-0.05	0.96
Trained items by target frequency	6	0.89	0.41
Trained items by foil frequency	6	0.50	0.63
All items, pairs of equal frequency	6	1.60	0.16
Trained items, pairs of equal frequency	6	0.69	0.52
All items, pairs of unequal frequency, by target frequency	6	1.22	0.27
All items, pairs of unequal frequency, by foil frequency	6	-0.58	0.58
Trained items, pairs of unequal frequency, by target frequency	6	0.65	0.54
Trained items, pairs of unequal frequency, by foil frequency	6	-0.27	0.79

Discussion

The results from this task failed to show a major effect of frequency, and were not consistent with previous experiments. However, caution should be taken when interpreting these results. A large influence on performance is the item presentation speed that is chosen for each individual subject. Due to possible computer malfunctions, the presentation speed may not have been adjusted correctly for some subjects. When individual subject data was examined, it showed that while some subjects seemed to be placed very near the appropriate accuracy (75%),

other subjects showed almost perfect performance. Clearly, the presentation speed either was not adjusted correctly, or if it was, the computer did not display the items for the correct amount of time. As such, the results from this portion of the task will not be addressed in the modeling section, and the subjects were not re-tested on this task after the 6 week delay. In the future, it would be advantageous to repeat the experiment with corrected timing and display, and include in the model an explanation for the results obtained. However, given the time constraints of the current project, it is not possible to include such a repetition in this paper.

References

Lieberman, H., & Pentland, A. (1982). Microcomputer-based estimation of psychophysical thresholds: The Best PEST. *Behavior Research Methods & Instrumentation, 14*, 21-25.

Appendix D – Probability estimation through simulation

The complex construction of the traces, percepts, and test probes used in this model make it necessary to find an alternative to the traditional Bayesian equations used in REM based models (Shiffrin & Steyvers, 1997). The technique used in the modeling in this paper is an estimation of the probabilities of matches and mismatches for a given set of parameters through simulation. The idea is to run the model through a large number of simulations of the task being modeled, but instead of calculating likelihood ratios to simulate data, to simply count the number of matches and mismatches that occur when an episodic trace or a percept is being compared to its own trace and when it is being compared to a different trace. After a large number of simulations, estimated probabilities of match and mismatch can be produced, given that an item is being compared to its own trace vs. being compared to a trace that is not its own.

Before being put to use, this method was tested using the basic REM model. Since the correct equations for the REM model are easily derived, the simulated estimates can be compared to the results produced by the Bayesian equations to check the accuracy of the estimates. The test was run (using the episodic recognition task) as follows: Over many simulations, a count was kept of how many times each test value occurred with each trace value when a target item was being compared with its own trace, as well as when a target item was being compared with other traces. The estimation was calculated for a given set of parameters, and the results from the simulation were compared to the results of the Bayesian formulas using those same parameters. The results showed that the simulation technique produced accurate estimates of the probabilities of match and mismatch given the item was being compared to its own trace or some other trace (see Figure D1).

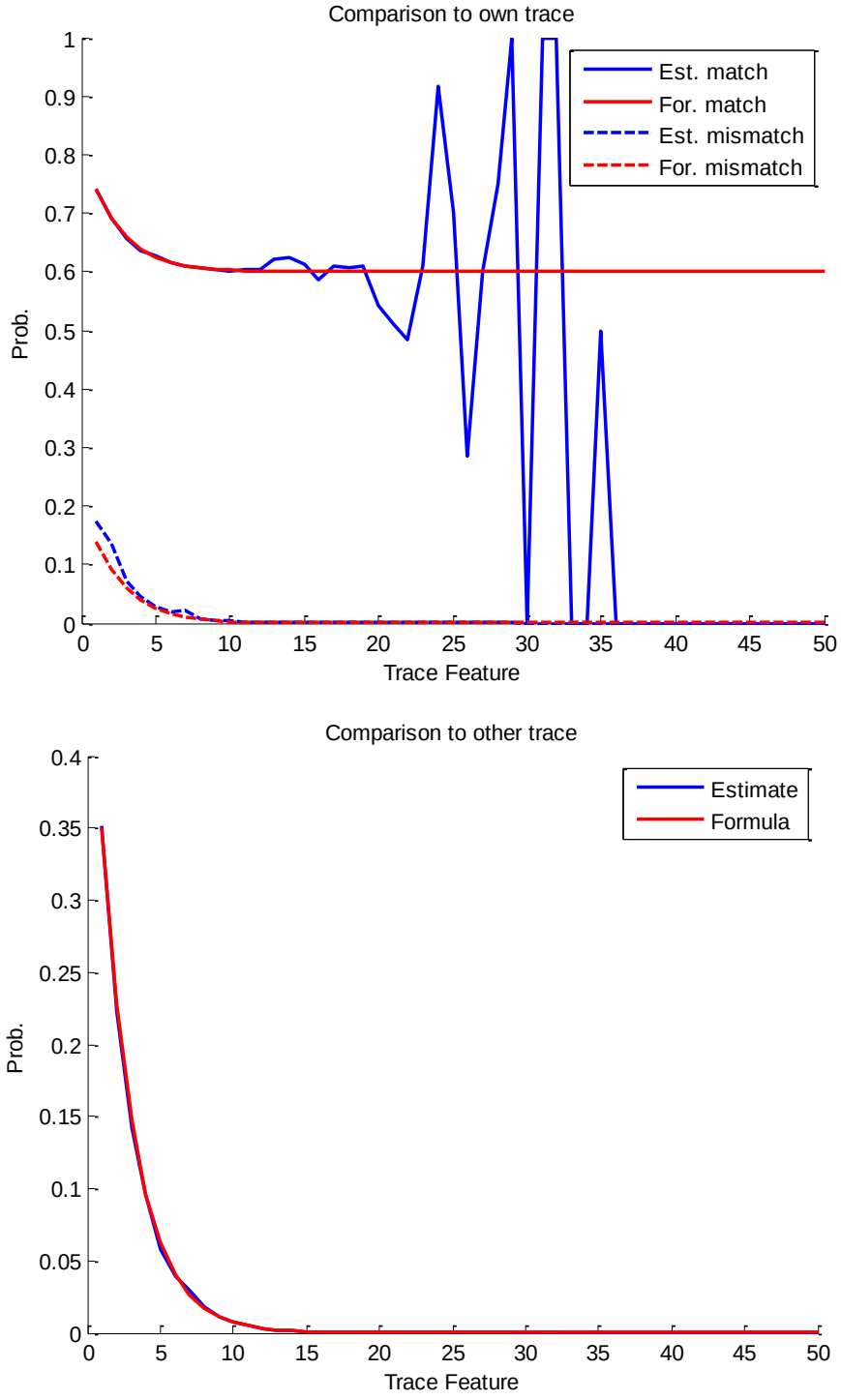


Figure D1 – Estimated probabilities compared to probabilities calculated by formula when an item is compared to its own trace (panel A) or some other trace (panel B). Note that the estimated probabilities in Panel A differ from the formula probabilities only for large feature values, due to the extremely low instance of these very low frequency values.

References

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