Psychological Review
2013, Vol. 120, No. 2, 356–394

0033-295X/13/$12.00 DOI: 10.1037/a0032020

The Co-Evolution of Knowledge and Event Memory

Angela B. Nelson and Richard M. Shiffrin
Indiana University

We present a theoretical framework and a simplified simulation model for the co-evolution of knowledge and event memory, both termed SARKAE (Storing and Retrieving Knowledge and Events). Knowledge is formed through the accrual of individual events, a process that operates in tandem with the storage of individual event memories. In 2 studies, new knowledge about Chinese characters is trained over several weeks, different characters receiving differential training, followed by tests of episodic recognition memory, pseudo-lexical decision, and forced-choice perceptual identification. The large effects of training frequency in both studies demonstrated an important role of pure frequency in addition to differential context and differential similarity. The SARKAE theory provides a framework within which models for various tasks can be developed; we illustrate the way this could operate, and we make the verbal descriptions of the theory more precise with a simplified simulation model applied to the results.

Keywords: memory, knowledge, events, frequency effects, storage, retrieval

The processes involved in the accumulation of knowledge and the formation of event memories are interdependent: Knowledge grows from events experienced, and events are coded in terms of existing knowledge. This is not only an old idea but seems self-evidently true. It is a structural component of many neural network/connectionist models (discussed later). Memory theorists have, however, focused mainly on either event memory (*episodic memory*, as Tulving, 1972, termed it) or semantic memory (access to knowledge). This article therefore builds on traditional memory theorizing, but looks at the larger picture, and illustrates (one way) event memories and knowledge interact at all phases of storage, coding, and retrieval. It does so with both a general theoretical framework and a simplified simulation model.

The theory is informed by several studies; each study builds new knowledge in an extended training task and then tests the effects of that knowledge in three transfer tasks spanning the field of cognition. The focus of these initial studies is differential experience: Individual items are given widely differing amounts of training. In these tasks, initially unknown Chinese characters are trained for several weeks so that the cognitive system comes to develop a simple form of perceptual knowledge. One training task, visual search, builds knowledge about both the individual characters and the context in which those characters are trained. The other training task, character matching, eliminates variation of the training context. Frequency of character training was varied because frequency of event occurrence produces some of the most reliable effects upon performance in almost all cognitive and behavioral tasks, and according to the theory is the basis for knowledge formation. In addition, the cause of frequency effects has heretofore been a matter for debate.

To set the stage and motivate the studies and analyses to follow, we lay out here in brief a few key elements of the theoretical framework and models. A central goal is the presentation of a coherent conceptual framework within which models can and hopefully will be developed. An extremely simplified simulation model is presented and fit to the experimental results, but its purpose is neither to verify a task model by detailed quantitative fitting nor to demonstrate one model quantitatively superior to another, but rather to make precise the core concepts of the approach and to show how the basic framework can be applied to quite disparate tasks. In the General Discussion, we take up the ways in which it would be necessary to form more realistic, albeit more complex, models of the present tasks and a few others.

The theory is termed SARKAE, an acronym for "Storing and Retrieving Knowledge and Events." It is assumed that there are (mostly separate) memory traces for events. What are events is a complex matter, not yet explored very well by the field. We discuss this matter at the end of the article, but for the simple studies used as a basis for present theory development, it does no harm and it is convenient to define events in terms of the stimuli presented on each successive "trial." The traces stored for such events tend to be weak, imprecise, inaccurate, and impoverished. Consider the first few times an event occurs, at the onset of knowledge development. The first occurrence produces an event trace. A later repetition of an event produces a new event trace. There are three possibilities: (a) a new event trace is stored in addition to the previous one; (b) the new event retrieves the earlier trace, and if the two are similar enough, the new event and the previous trace are combined into a single new trace; (c) both (a) and (b) could occur. When an augmentation occurs and a trace combines information from the present event with a similar previous event, the resultant trace gains in strength, accuracy, and precision, in comparison to what would have been stored in an

This article was published Online First March 4, 2013.

Angela B. Nelson and Richard M. Shiffrin, Department of Psychological and Brain Sciences, Indiana University.

This research was supported by Air Force Office of Scientific Research Grant FA9550-09-1-0178 to Richard M. Shiffrin.

Correspondence concerning this article should be addressed to Richard M. Shiffrin, Department of Psychological and Brain Sciences, Indiana University, Room 350, 1101 East 10th Street, Bloomington, IN 47405. E-mail: shiffrin@indiana.edu

independent event trace. It is this augmentation process that gradually causes the development of knowledge traces. Once a knowledge trace of some strength has formed (as happens for words in one's language), it will generally be retrieved by another occurrence of a similar event (such as another encounter with that word), and in such cases both a new event trace and an augmentation of the knowledge trace will take place. In our model, therefore, there is no hard and fast boundary between event traces and knowledge traces: There is a continuum of traces from generally weak and impoverished event traces at one end to very rich and developed knowledge at the other end. Knowledge traces of words are usually termed lexical traces, but there are knowledge traces for all types of events at all levels of abstraction and complexity (e.g., a golf swing, reading, poker playing).

When an event is encountered, it is always encoded with reference to retrieved knowledge. For example, when the visual form "cat," the auditory expression of that word, or a picture of a cat is encountered, a cascade of retrieval processes occurs that among other things will retrieve the lexical knowledge trace and various aspects of the meaning of "cat." This encoding information plus various types of context will then be stored (incompletely and not entirely accurately) as a new event trace. In addition, the developing knowledge trace will gain (some part of) this information, particularly including the context that is specific to the current event and not yet represented in the knowledge trace. The kinds of information accessed, encoded, and stored are surely determined by a variety of implicit and explicit attention processes, but these are not the main concern of this article and are therefore only mentioned incidentally in the exposition. In SARKAE, both event traces and knowledge traces are represented as vectors of counts of feature values (e.g., "red") organized by feature types (e.g., "color"). As knowledge traces grow richer, the counts keep rising.

Because developed knowledge contains elements of all the contexts associated with the events that produced the knowledge trace, no one context stands out, and the knowledge appears to the person retrieving the knowledge to be context free: We "know" but do not associate the knowledge with a single life event.

The theory is applied with a very simplified computational model fit to the data; however, the purpose is not to produce a fleshed out model for each of the tasks but rather to make precise the way a theory like the present one can be applied. The model is applied to five very different tasks: visual search or physical form matching (the two training tasks), pseudo-lexical decision (the knowledge retrieval transfer task), episodic recognition memory (the event memory transfer task), and two-alternative forced-choice perceptual identification (the perception transfer task). The studies had limitations on the amount of testing that could be carried out without distorting the manipulations of training frequency; this constraint produced insufficient data to make it reasonable to apply and test a state-of-the-art quantitative model of each task. The patterns of data were nonetheless informative, and we shall see that our simplified simulation was able to capture the qualitative trends in the data. In designing the simulation, we were most interested in commonalities: Different tasks necessitate some differences in assumptions, but all the tasks have many elements in common and surely utilize many similar processes. To give just one or two examples, a stimulus is presented and used as a probe of knowledge in long-term memory; information is retrieved from knowledge to build a representation in short-term memory; and the information in short-term memory is used to store an event trace and add to knowledge. Therefore, we kept the models as conceptually consistent as possible and kept parameters and their estimated values constant across the various applications, whenever such constancy was sensible.

In the first portion of the article, we describe some of the background that gives rise to the theory and provide a summary of SARKAE's main assumptions and processes. Experience is the basis for the formation of knowledge, so we then review some of the relevant research on the effects of frequency of training. In the second portion of the article, we report two training studies in which novel knowledge is formed for Chinese characters that are trained to differing degrees; training is followed by tests of event memory and knowledge retrieval. Both studies revealed strong frequency effects. According to the theory, the visual search training used in the first study could have induced frequency effects by making traces of high frequency characters more similar to each other (because they co-occurred during training). The second study eliminated this co-occurrence factor by training with character self-matching. Frequency effects were found nonetheless, leading us to incorporate a role for "pure frequency" in the theory. The next portion of the article makes the theory's assumptions more precise by instantiating them in a simulation model and showing that such model captures the major trends in the data. Lastly, in the General Discussion, we flesh out the theory and describe how one might model other sorts of data from the present tasks and how the theory might be used to form models for other tasks.

Background

The distinction between event memories and knowledge is an old one, but most present day researchers refer to the distinction laid out by Tulving (1972). He termed event memories *episodic* and distinguished them from *semantic* memories. Tulving's ideas have of course evolved and become more complex over the years, especially in light of many findings of cognitive deficits caused by various brain abnormalities, and in light of studies using brain measurements such as positron emission tomography (PET) and functional magnetic resonance imaging (fMRI). One update of his views was presented in summary form (Tulving, 1993) and focused on different sorts of awareness: *noetic* awareness of the contents of episodes or semantic knowledge, *autonoetic* awareness of personal participation in the retrieved memory (episodic), and *anoetic* awareness of procedural knowledge. The present theory finds it more useful to divide memories into just two categories, with procedural knowledge part of knowledge generally. Perhaps most relevant for present purposes is that Tulving's recent view treats episodic memory as an extension of semantic memory rather a separate system. One could say without too much distortion that the present treatment caches out this view in considerable detail.

Of course, the main theme of the present theorizing goes beyond distinguishing events and knowledge to positing the way to two co-evolve. The interactions are intertwined at every phase of encoding, storing, and retrieval. 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 retrieving effectively from memory

(REM) model's account of priming (Shiffrin & Steyvers, 1997): Study of an event produces an event trace but also adds information to an existing trace that is called to mind. For words the trace called to mind will be the lexical trace, and the added information produces priming. For relatively novel stimuli, the trace called to mind would be a previous event trace, and the added information begins the process of forming a knowledge trace. These interdependent processes create a feedback loop in which knowledge and episodic memory formation, and retrieval of both, develop jointly over lifelong learning.

Studies of memory and perception, particularly priming studies over the last 30 years, have provided strong support for the interdependence of event memory and knowledge during retrieval (e.g., Jacoby & Dallas, 1981; Roediger & Challis, 1992; for some summaries, see Jacoby, 1991; Roediger & McDermott, 1993). Priming studies typically present a word (the "event"); a later apparently unrelated test presents the same word for a test of knowledge retrieval, with the result that knowledge retrieval is affected by the earlier event presentation. Our prior modeling (e.g., Malmberg & Shiffrin, 2005; Schooler, Shiffrin, & Raaijmakers, 2001; Shiffrin & Steyvers, 1997; Wagenmakers et al., 2003) and SARKAE account for these effects through a process in which the lexical trace for a given word is augmented by study of that word (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 its lexical representation. When that word is later presented in a task requiring retrieval from knowledge (such as naming, perceptual identification, lexical decision), the context tends to be similar to that at study, increasing the match of the probe cues to the lexical trace, enhancing and/or biasing retrieval and predicting the priming results. This can be described as an effect of experience upon perception: The inclusion in knowledge of information such as current context affects 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). A parsimonious interpretation of such findings posits storage of an event trace that incorporates general knowledge extracted from long-term memory, and retrieval that is partially a matter of recovering features from the stored trace, and partially a matter of inference and reconstruction that uses general knowledge. Perhaps the clearest demonstration of such an effect and process is recent research by Hemmer and Steyvers (2009a, 2009b): 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; Hemmer and Steyvers, 2009a, 2009b), 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. 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 (1990; also see Brainerd, Reyna, & Mojardin, 1999) investigates 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.

Developmental studies of memory in infants provide additional insight into the co-evolution of event memory and knowledge, although the relation of that literature to our present theory is complex. Theories in the developmental literature sometimes follow current theory applied to adults and posit a division of memory into implicit and explicit memory systems. These terms are not precisely defined and seem to mean different things in the hands of different theorists, increasing the difficulty of mapping our present theory onto this binary dichotomy. Rovee-Collier (1997) has several studies showing that very young infants (starting as young as 3 months of age) can and do learn associations between, for example, two puppets. In her understanding of the explicit/implicit distinction, these studies show that the two systems develop together. On the other hand, researchers such as Newcombe have studies (e.g., Lloyd, Newcombe, & Doydum, 2009) showing that other types of associations involving relating event context to situational context (presumably associations that require hippocampal involvement in adults) are much slower to develop in infancy. She therefore concludes that early event storage is implicit rather than explicit. Regardless of the mapping of the results onto a binary categorization of memory systems, the results are consistent with the view that event memory and knowledge develop conjointly and together, although (in our terms) the content of event memory may change over early development. We note that some of the difficulty in assessing early memory development resides in differing views of the role of the medial temporal lobes and the hippocampus in forming, storing, and retrieving memories, and helping the formation of knowledge.

In our view, the extant literature provides support for a major theme of the present modeling, by which event memory and knowledge are best viewed "dualistically": On the one hand, there are very good reasons to distinguish event memories (episodic) and knowledge, both functionally and neurally; this view is fostered by our focus on the ends of a continuum—very recent event memories ("where did I park my car this morning?") or developed knowledge ("who was our first president?"). On the other hand, knowledge surely develops from experience so memory traces must lie along a continuum, with event traces at one end of the continuum, 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., full lexical knowledge). The issue is complicated by evidence showing a transition from traces formed and stored with the help of the hippocampus and adjacent regions to traces stored cortically elsewhere. This transition is presently an area of intense research, behaviorally, neutrally, and chemically, and is not settled. Our views of the co-development of event memory and knowledge are explicated in some detail as we lay out our theory in later sections of the article. In the General Discussion, we return to the way our present model maps onto the ways in which theorists sometimes divide memory systems into two systems, and although we do not in this article explore the neural substrates of memory, we also briefly discuss the ways in

which brain measurements of hippocampal involvement in memory storage relate to both our theory and system dichotomies.

Our theory, in common with many others, adopts a second dualism: a distinction between short-term active memory and long-term passive memory, an idea going back to the first days of psychology (explicated quantitatively by Atkinson & Shiffrin, 1968, for example). This distinction should not be confused with that between event traces and knowledge traces, both of which are formed with the use of short-term memory, are represented in long-term memory, and produce information in short-term memory during retrieval.

Modeling the Co-Development of Event Memory and Knowledge

There are many models of the storage and retrieval of event memories, and some models of the addition to existing knowledge of information from recent events (e.g., J. R. Anderson, 1983; Atkinson & Shiffrin, 1968; Howard & Kahana, 2002; Mueller & Shiffrin, 2006; Murdock, 1982; Raaijmakers & Shiffrin, 1981; Shiffrin & Steyvers, 1997). There are also models that explain retrieval of semantic memories (e.g., Quillian, 1967). A few models attempt to explain aspects of the way events produce knowledge, especially for aspects of the role played by words in language (e.g., Jones, Kintsch, & Mewhort, 2006; McClelland & Elman, 1986; McClelland & Rumelhart, 1981), though the focus of such models is more on language than memory per se.

Different goals aside, many neural net and connectionist models (Rogers & McClelland, 2004, is one well developed example) directly link presentation of new events, development of long-term knowledge, and retrieval. In a typical treatment, a new input (the current event) is presented in the form of a feature vector to a set of first stage units. These are connected in feedforward manner to a (usually) smaller number of hidden units that are in turn connected to further sets of units, and eventually to a set of output units. The output units produce a pattern that deviates from a desired output, and the connections in the system are adjusted (with the use of error backpropagation) so that the output units come closer to the outcome desired. This process continues as new inputs occur, some of which may be repetitions (very similar inputs) to previous inputs. Partly because the various intermediate layers of units are smaller in number, and partly due to the learning rules, the connections between units come to encode abstractions (information condensations) of the information distinguishing different groups of similar inputs from each other. A new input produces a pattern on the output units and this is a typical form of retrieval. The main point is that a system like this combines events (the inputs) to form knowledge (the connection weights), and the encoded knowledge then responds to a retrieval probe (a new input). In this sense, neural net modelers pursue similar aims as the present theory.

However, there are a myriad of systems and models of this sort each of whose structures, representations, goals, and data to be predicted differ from each other as well as the present theory. Even if a direct comparison were possible, it would not be clear which theory and which studies and data sets to use for comparison. Thus, we simply note the existence of the neural network approach to the co-evolution of event memory and knowledge, but we do not attempt to compare and contrast generally. However, given that such systems have been used to produce richly structured knowledge, we return briefly to this issue in the General Discussion.

The present exposition is long for a journal article but is far too short to produce a quantitative set of models for memory tasks involving storage and retrieval in event memory and knowledge, and their interaction. Thus, we limit our goals to a description of the general framework, coupled with a very simplified simulation that produces qualitative predictions for a few critically important findings from the present studies. A longer term goal is the development of increasingly accurate and sophisticated models of particular tasks that (hopefully) will be consistent with the present framework. Pointers to some of the ways that this larger goal can be accomplished are taken up in the General Discussion.

The REM model (Shiffrin & Steyvers, 1997, 1998) provided a preliminary hint of the way SARKAE could deal with both event memory and knowledge. Those articles were aimed at event recognition and presented a model whose assumptions were simplified enough to allow a mathematical derivation of predictions from a Bayesian inspired theory. That article included a few paragraphs indicating how addition of information to lexical traces could explain long term priming, a process that might also be the basis for the formation of a lexicon. The idea that addition of information to a lexical trace could explain long term priming was fleshed out in Schooler et al. (2001). Although these earlier articles hinted at the present development, the scope was extremely limited, many of the implementation assumptions differed from the present treatment, and most important, those articles did not deal seriously with the recurrent flow of information between knowledge and event memory that is the main theme of this article.

Earlier modeling that more directly led to the present development, and dealt explicitly with the co-evolution of the two systems, was seen in the REM-II model, created by Mueller and Shiffrin (2006). The main focus of this research was the development of knowledge traces. It departed from the common approach of representing traces by a vector of feature values by instead representing knowledge traces as an accumulation of the co-occurrence of features: Features that are present in an episodic event were 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 describes the interaction between episodic memory and semantic memory, and accounts for phenomena such as polysemy and connotation effects. In this article, we revert to a vector of features values primarily because this representation is simpler: There is much to be said for the REM-II approach, and it could well be extended to larger groups of feature co-occurrences, but such a system becomes not only complex but also too powerful to test, able to explain almost any result without enhancing our understanding of cognition. The present treatment and simulation considers quite simple kinds of features, but the general SARKAE theory allows a "feature" to be any established knowledge trace, of any complexity (e.g., a favorite TV program could be a feature).

A fundamental storage assumption in SARKAE allows both event memories and knowledge to develop in concert: Each storage episode produces both (1) an event trace and (2) additional information added to traces in memory that are brought to mind due to similarity to the present event. The trace brought to mind can be a previous event trace (the basis for the start of knowledge

accumulation) or a developing or mature knowledge trace, or both. There is no fundamental functional distinction between the representation of event traces and knowledge traces in this view (possible neural distinctions are discussed shortly). Instead, there is a continuum: Traces are stored initially for each single event; some of these are retrieved (when a similar new event occurs), gain additional information, and are re-stored. As this process continues over successive occurrences of similar events, a rich knowledge trace results. 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).

It is worth a brief segue to discuss an (apparently) alternative view in which there are separate systems for event memory and knowledge. In one version, (some kinds of) event memories are stored initially in the medial temporal lobe (MTL)/hippocampus and gradually transferred into more permanent memory traces elsewhere in the cortex. This hypothesis is compatible with and somewhat orthogonal to the present proposal, in the sense that the (more) permanent cortical traces would include a continuum of traces from individual event traces to knowledge traces. Another version of the alternative approach would posit that each event occurrence would result in an event trace (presumably involving the hippocampus) and separately would result in a trace of a different qualitative character (presumably involving a cortical storage route separate from the hippocampus). In such a view, subsequent events would build knowledge by adding to the cortical representation rather than the initial event trace. This issue may someday be resolved through neuro-cognitive research, but the difference is rather subtle from a behavioral perspective, so we focus solely on the SARKAE approach in which knowledge grows from event traces.

In SARKAE, accumulation of knowledge about an item or concept (e.g., for words, its lexical entry) includes features of the surrounding context that is present at the time of learning. Specifically, knowledge traces develop during learning by storing features that come both from the physical properties of the item or concept being learned, and also from the context surrounding the item during learning; both types of storage are modified and governed by attentional focus. These context features arise from other (attended) events nearby in time and the environment, and from the various components of internal and external context that numerous investigators have discussed for many years (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, information about the spatial relations and layout, 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 features of the events that accumulate to form the knowledge. A mature knowledge trace includes features of numerous events, so the features of a specific episode tend to be swamped in the accumulation of features of many episodes. When a specific event is retrieved, it is through access to an event trace rather than access to a knowledge trace. Thus, a knowledge trace in most instances seems to be context free. What are retrievable from a mature knowledge trace are features that are consistent across many episodes, such as the spelling, pronunciation, and meaning of a word.

This view of context leads back to the issue of feature representation, and the issue of the formation of new features with which to represent an item. Suppose we hear a word, or see a random dot pattern, for the first time. The features used to code such an item might be low level physical features (phonemes and dot arrangements), supplemented by features of the surrounding context (the observed animal with the long neck might be a feature, or features, associated with the first hearing of the word "giraffe"). In general, it is also possible for new features to emerge over time, often consisting of re-combinations of existing features. Thus, if many dot patterns observed are generated as distortions of a prototype pattern with a central square of four dots, the re-occurrence over events of this partial pattern might be noticed and become a new feature used to encode such patterns. Of course this new feature is actually a new trace on its own (both an event trace and the start of a knowledge trace). In general, features of any item trace are probably best thought of as relatively unitized other traces in knowledge. In our present simulations of such a system, we simplify greatly by fixing the total number of potential features and feature values (i.e., the number of vector slots) and leaving certain vector positions empty until they are filled. These ideas are elaborated later in the article.

There is a dualism between the formation of knowledge and the coding of event traces, because event traces are formed on the basis of current knowledge. 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 event 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 a considerable degree, the features chosen are modified by attentional focus, so that, for example, a current focus on danger might lead to features from one meaning of "gun," while a focus on racing might lead to features from another meaning of "gun." 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).

We have been highlighting mechanisms that produce storage of event memory and knowledge. Storage depends heavily on retrieval (certainly from knowledge, and often from event memory) and retrieval produces storage (of an event trace and in knowl-

edge), so their separate discussion should not be allowed to obscure their close interrelation. Furthermore, the mechanisms operating in retrieval and storage have many similarities that will be obvious from the following discussion.

We adopt the generally accepted view that retrieval is cue dependent and based on similarity of the retrieval probe to the traces in memory (e.g., Tulving & Thompson, 1973). The generation of such a probe cue can be discrete, 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 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, both comprised of features already extracted from knowledge, 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 these processes. Following the approach in the REM model, we assume that storage and retrieval are probabilistic, incomplete, and error prone. When errors are made, it is natural to assume they are based on information in the knowledge base, and not completely random. Thus, errors in retrieving and storing features are assumed to be relevant and consistent, in the sense that they are feature values for the feature in question (a "blue" color feature might be retrieved or stored as "green," 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 knowledge traces. It would be unworkable and likely unreasonable to calculate explicitly the match to each of the essentially uncountable traces in memory. Thus, we assume that there is a probabilistic cutoff, only traces sufficiently similar to the probe becoming activated and participating in subsequent retrieval operations.

Similarity is a fairly vague term and needs to be defined more carefully. We assume that the relation of memory probe to trace can be characterized by an "activation strength," used to define the set of traces that exceed the threshold for activation and to govern subsequent retrieval. This activation strength is defined as a relative measure: In our Bayesian-inspired approach, the activation strength of a trace is a likelihood ratio; the numerator expresses the probability that the probe and cue were generated from the same event, and the denominator expresses the probability that the two were generated by different events. Both numerator and denominator are calculated on the basis of the features that match or mismatch between probe and trace. High strengths depend on having both a high ratio of matching to mismatching features and also a high total number of features. These likelihood ratios occupy the theoretical niche played by "strengths of activation" in various other theories (such as SAM; Raaijmakers & 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. The latter portions of this article cover the theory in detail, but a theory described verbally, even in great detail, will inevitably be interpreted and applied differently by different readers for different tasks. Therefore, to make the basic tenets of the theory more precise, we carry out studies that explore the development of new knowledge from events, and the ways in which that knowledge is used in event and knowledge retrieval, and fit the data with a simplified simulation model consistent with the theory. The goal of the quantitative modeling is not the usual one of delving deeply into processes and mechanisms, but rather to make precise the basic elements of the theory. In fact, the data are rather straightforward and limited in extent and are not suitable for the former goal. Thus, aspects of the general theory are based on key concepts that are rooted in data from prior research, that are chosen in order to obtain conceptual coherence, or are copied from previous useful concepts in applications of the REM theory (a theory that has been shown to give good accounts of memory, priming, and knowledge retrieval; see, e.g., Schooler et al., 2001; Shiffrin & Steyvers, 1997; Wagenmakers et al., 2004). The simplified simulation comes close to assuming the minimum needed to handle the present data, but nonetheless illustrates the co-evolution theme of the present article.

The studies do serve another and different purpose by answering a fundamental question about the way that events produce knowledge: What is the role of event frequency? The answer provides a starting point for the theory development. In particular, the studies use an extended period of training to foster the development of new knowledge, and then use transfer tasks to explore the effects of training upon episodic memory, retrieval from knowledge, and perception. The effects of differential experience during training are omnipresent in cognition, and the mechanisms for such effects are presently an issue under investigation in the field.

It is important to be aware that the simulation model will be applied to five quite disparate tasks. This allows us to highlight the co-evolution processes that are common to all, and otherwise add the minimum assumptions demanded by task differences. Had we tried to produce a simulation containing a fully fleshed out state-of-art model for each task, the result would be very complex, and the theme of this article would be lost in a forest of details.

Role of Experience and Frequency in Cognition

If one hopes to develop a theory in which events accumulate to form knowledge, 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 vary presentation frequency in the present studies. In order to control the total experienced frequency, we train novel characters (Chinese characters). Different characters are given substantially different amounts of training, over many days. Following training, these characters are tested in an episodic memory task (storage and retrieval of recent events), a perception task (identifying briefly flashed characters), and a knowledge retrieval task (pseudo-lexical decision: was the test character trained?).

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, defined as normative occur-

rence in the environment. Estimates of these frequencies are computed from various databases of (typically) textual materials. Words differing in frequency produce different results in a wide variety of tasks, and show these differences quite consistently. When found in episodic recognition memory, the pattern of differences is termed the *word frequency effect* (Glanzer & Adams, 1985, 1990; Kinsbourne & George, 1974). In these tasks words that occur rarely in the environment are recognized *better* than words that occur frequently in the environment, a consistent finding that has been called one of the regularities of recognition memory (Glanzer, Adams, Iverson, & Kim, 1993). This advantage for lower frequency is however the exception: In most tasks higher frequency benefits performance. In episodic recall, high frequency (HF) words are recalled better (Gregg, 1976), and in perceptual tasks such as lexical decision and perceptual identification (forced choice, etc.), high frequency improves response speed and accuracy. In lexical decision, HF words are identified both more accurately and more quickly than low frequency (LF) words (Becker, 1979; Rubenstein, Garfield, & Millikan, 1970; Scarborough, Cortese, & Scarborough, 1977). Perceptual identification shows a more complex pattern of results: Generally, in two alternative forced-choice studies, 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 three large corpora of texts that vary in both 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. 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 & 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, & Shiffrin, 2002). Pseudowords also contain parts of words as well as bigrams and trigrams that differ in frequency in the language, factors known to affect performance in lexical decision (Rice & Robinson, 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 selected 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 et al. (2002). It is of course the case in actual experience that the training and testing of knowledge occur in similar tasks, but inferences about underlying processes are more difficult when this is the case.

The studies reported in this article therefore use training tasks that are 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. The second study had participants compare a character to itself, looking for slight physical changes. 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.

For both studies, note that only a limited amount of data was collected from each participant in the three transfer tasks. Learning of course continues during testing in the transfer tasks. The transfer designs required equal use of stimuli trained at different frequency levels. Thus, as testing continues, there would be an inevitable dilution of the frequency effects that were one of the research goals. Testing was limited to minimize such dilution effects.

The results of the training tasks and transfer tasks are given in figures, and the pattern of results and their statistical significance are described in the text. The exact results and details of the statistical analyses are given in Appendix A. Although the presentation of the model and its parameters occurs later, the figures also give the model predictions.

Experiment 1: Visual Search Training

Training Task

Method.

Participants. Eight people, recruited through an e-mail 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-in. (43.18-cm) 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 ms. The cross was followed by a target character presented centrally for 1,000 ms; the target was then replaced by a blank screen for 500 ms. Then a display of either two or four characters appeared and remained until a response was made. The characters each subtended about 3.5° visual angle vertically and 2.9° 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° 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, and (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 each subject completed 12 visual search sessions, over the course of roughly 3 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 four blocks in each session.

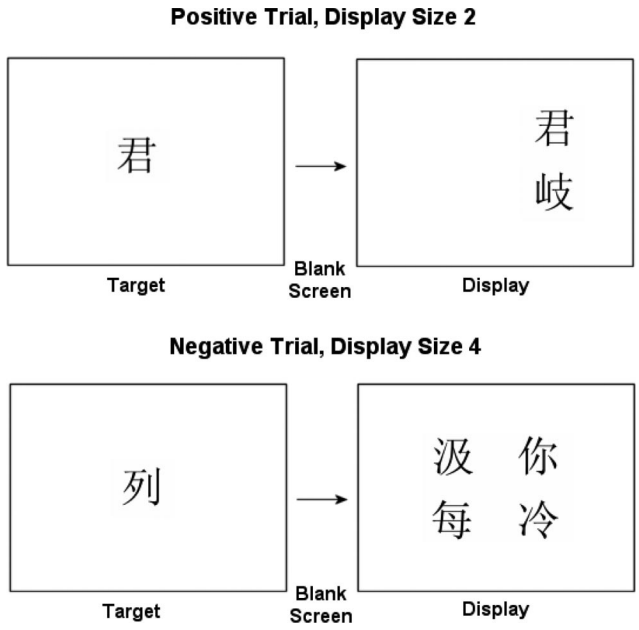

Figure 1. Example visual search trials.

Thus, there were 640 trials per session, and 7,680 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 times for display sizes of two and four. Figure 3A shows mean slope per session, averaged over the eight 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, although, as might be expected, the pattern of results became quite a bit noisier for characters of lower frequency. 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

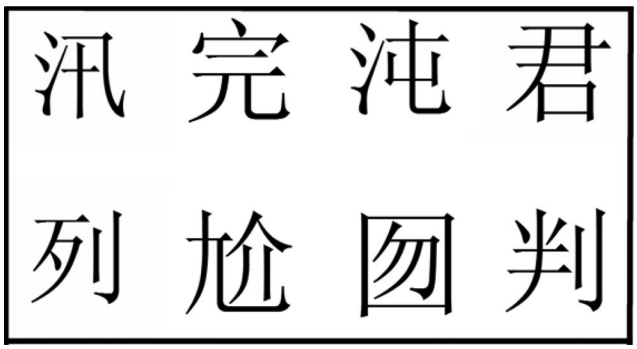

Figure 2. Sample of characters used.

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.

Discussion. Training produced a clear and pronounced improvement in search rate and search intercept. The slope decreases extended over a longer period of time than the intercept decreases, consistent with the view that they reflect different processes. 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 results provide evidence for the development of knowledge about the initially novel Chinese characters. Furthermore, the slope reduction is not due to the automatic direction of attention to targets. As demonstrated by Schneider and Shiffrin (1977; Shiffrin & Schneider, 1977; and verified in many studies since), use of a varied mapping procedure prevents the learning of automatic attraction of attention to targets—such learning takes place in a consistent mapping procedure.

Instead, the slope reduction is likely due to one of two closely related factors: the increasing integration of the features of each character, or the identification of a feature combination that is unique to each character. These factors were demonstrated 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 inward 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.

Thus, a good part of the learning seen in the present task is likely due to perceptual unitization, but a quantitative prediction would not be possible because the feature composition of the present Chinese characters was not controlled, and indeed varied with character sets that differed for each participant. By inspection the feature overlap appears quite complex, so that some characters and sets might allow search for a single distinctive feature (once identified), while other characters and sets might require search for a conjunction of features. Thus, part of the learning might involve discovery of distinctive features and other parts of the learning might involve perceptual unitization of feature combinations. Whichever way distinctive features are produced, the participants are likely to find those that are unique for their entire character set: 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 relative to all 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. Finally, note that some learning that could lower slopes could occur after perceptual unitization is complete, if the efficiency of such search improves with training. For these various reasons, quantitative predictions concerning the degree of slope changes are not possible, but we can conclude that learning has occurred.

Post-Training Tasks

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

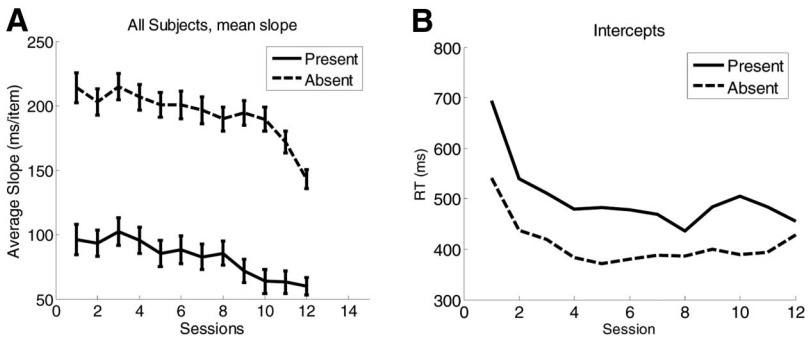

Figure 3. Panel A: Slope of the search function over training. Error bars represent the standard error of the mean. Panel B: Intercept of the search function over training. RT = response time.

Pseudo-Lexical Decision

Method.

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

Participants. All eight subjects who were trained on the characters completed this task shortly (2–3 days) 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 three 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 decreased and accuracy increased, as frequency increased (see Figure 4, solid black lines with inverted triangle markers; for reference, new items had a mean response time of 820 ms and accuracy of .93). Analysis showed that higher frequency characters produced significantly faster response times and higher accuracy than lower frequency items. 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.

Discussion. The pseudo-lexical decision results show that the degree of experience with a character, and/or the character context that is correlated with frequency in our tasks, produces decreases in response times and increases in accuracy. These findings align with lexical decision results for words in previous studies (Becker, 1979; Rubenstein et al., 1970; Scarborough et al., 1977). The present frequency results are due to the factors we introduced in visual search training (differences in experience and differences in the character context for characters with different experience) and are not due to the many other factors that are correlated with word frequency. It seems reasonable to reverse the logic and infer that a major component of the word frequency effect is due to those same factors. One way to understand the processes involved in pseudo-lexical decision (and perhaps lexical decision as well) is laid out in the modeling sections of this article.

Episodic Recognition

Method.

Participants. 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 1,000 ms, 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 to 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 5, left panel; also see Table 1, top). 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.

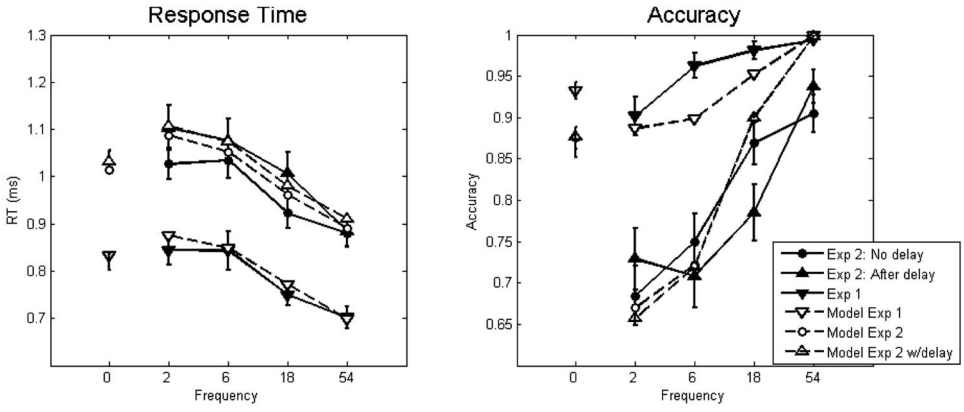

Figure 4. Pseudo-lexical decision: Observed data and simulated data for Experiment 1, Experiment 2 immediate test, and Experiment 2 delayed test. Response time (RT) is given in the left panel, and accuracy is given in the right panel. Error bars represent the standard error of the mean.

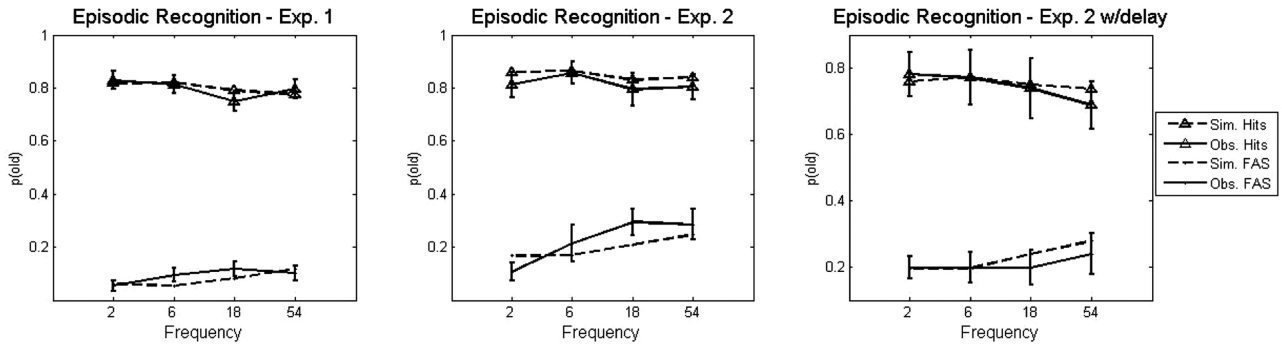

Figure 5. Episodic recognition: Observed (Obs.) data and simulated (Sim.) data for Experiment 1 (left panel), Experiment 2 (center panel), and Experiment 2 after a 6-week delay (right panel). Error bars represent the standard error of the mean. FAS = false alarms.

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 (Sikström, 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, and this was not 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: better performance for lower frequency items, and a mirror pattern. Further discussion is deferred until the presentation of the SARKAE model and theory in following sections.

Table 1
Episodic Recognition Results

Item type	P(H)	P(FA)	*d'*
Experiment 1			
Novel items	0.807	0.142	1.94
Frequency 2 items	0.828	0.055	2.54
Frequency 6 items	0.813	0.094	2.21
Frequency 18 items	0.750	0.117	1.86
Frequency 54 items	0.797	0.102	2.10
Experiment 2 immediate test			
Novel items	0.680	0.097	1.77
Frequency 2 items	0.813	0.107	2.13
Frequency 6 items	0.857	0.214	1.86
Frequency 18 items	0.795	0.295	1.36
Frequency 54 items	0.804	0.286	1.42
Experiment 2 delayed test			
Novel items	0.707	0.129	1.68
Frequency 2 items	0.781	0.198	1.62
Frequency 6 items	0.771	0.198	1.59
Frequency 18 items	0.740	0.198	1.49
Frequency 54 items	0.688	0.240	1.20

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. In our study, it may be especially important that new characters contain some new features, not previously encountered. These could be noticed at both study and test, and improve performance beyond that expected for characters encoded with features that are familiar and shared among several items (this idea is used in the SARKAE model for lexical decision).

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.

Forced-Choice Perceptual Identification

Method.

Participants. Six out of the eight trained subjects completed the forced-choice perceptual identification. The task was administered approximately 3 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% correct threshold, using the Best-PEST (parameter estimation by sequential testing) 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 6). Both of these effects were marginally significant.

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 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 is based on the use of 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 is elaborated in the modeling discussion that follows.

Characterizing Differential Experience in SARKAE

The frequency of character presentation was varied in Experiment 1. Over repetitions, the SARKAE model accumulates feature counts in a developing knowledge trace. The number of counts could possibly be one cause of the observed frequency effects, not only for retrieval from knowledge but for storage and retrieval of event traces (to the degree that event storage and retrieval depend in part on access to knowledge). However, the randomization of targets and foils over the trials of visual search 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. In fact, Adelman et al. (2006) proposed that a word's contextual diversity, not word frequency by itself, was responsible for most word frequency effects, and this factor could be another cause of our frequency effects.

There is no real controversy about the existence of context effects in memory storage and retrieval—they are omnipresent in cognition at every level of analysis. In the present discussion, context includes the general situational context but more importantly also the context of items that may be co-occurring in the physical and mental environment. This approach is closely related to that in the temporal context model (e.g., Howard & Kahana, 2002; they use the nearby item context to explain, among other findings, the tendency for freely recalled items studied together to be output together, to a degree determined by the presentation

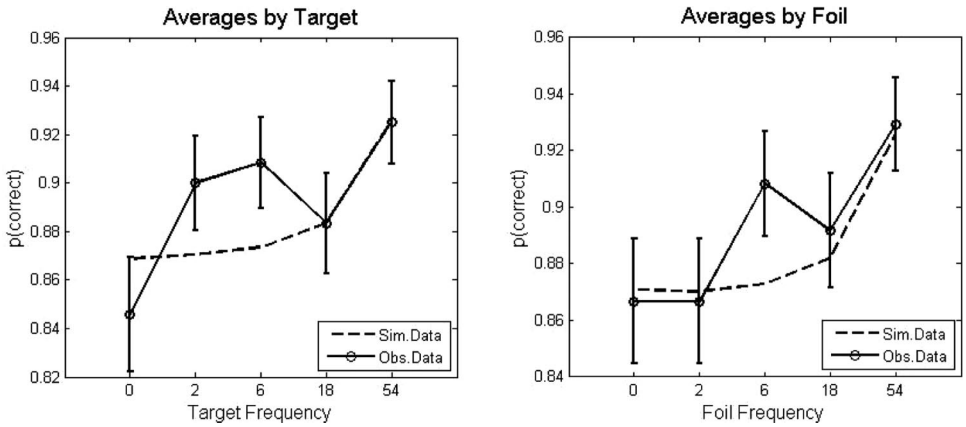

Figure 6. Two alternative forced choice: Observed (Obs.) data and simulated (Sim.) data, averaged by target frequency (left panel) and by foil frequency (right panel). Error bars represent the standard error of the mean.

separation of the items). One natural way to introduce nearby item context would be to allow an "event" to encompass a group of several nearby items. However, suppose for simplicity we limit an event to storage of a single item (e.g., the target on a visual search trial), as is done in the SARKAE simulation. The effects of nearby items (such as the foils on a trial) are then introduced by including storage of their features.

Because knowledge traces are formed from events, they accumulate information about the types of items that are near the various event occurrences. Thus, the presentation of a Chinese character as a target causes that character's knowledge trace to gain information not only about that character's features but also about those from the trial's foils, and possibly from the target on previous trial. Because high frequency items tended to co-occur in our study, their knowledge traces grew to include features of each other. The feature vectors for the knowledge traces that develop for higher frequency characters therefore come to overlap more: More generally, the similarity between two knowledge traces is higher to the degree that their training frequencies are higher. As one way to illustrate this, we analyze the similarity of the vectors representing knowledge traces for characters of differing training frequencies. There are many ways to characterize the similarity of two vectors. We have tried several and all produce the same result. Figure 7A shows representative results for one type of normalized inner product: Each feature value is divided by the sum of values for that feature. This is done for every feature. Then a simple inner product is calculated: If the resultant counts in the two traces are (a1, a2, . . ., aN) and (b1, b2, . . ., bN), the measure depicted in Figure 7 is the sum of (ai)(bi). This calculation insures that increased similarity is not due to larger counts per se, but rather the similarity of the pattern of counts cross the vectors. We give the results for one simulation of the SARKAE model, but the pattern is true quite generally. Figure 7A shows that traces of higher frequency have become more similar to each other, in that the patterning of the counts across the vector is more similar. According to SARKAE (and probably any model taking context into account), the similarity structure of knowledge should co-vary with presentation frequency to the degree that the co-occurrence of items of differing frequency is correlated.

We next assessed the possibility that the data from Experiment 1 could be fit with a version of SARKAE that did not include any explicit role for frequency other than the change in similarity. This model (could be made to) fit most of the findings, but left open the possibility that pure frequency might play a role as well. Experiment 2 was designed to remove the confound between frequency and context. To take a peek ahead, it demonstrated that an adequate model requires also a role for "pure" frequency, so both the general SARKAE theory and the simplified simulation will include both.

The Experiment 2 results had another important implication that is worth mentioning briefly here because it caused us to employ the more complex version of our previous event recognition model. The REM model of Shiffrin and Steyvers (1997) allowed for activation by the test item of both list traces and pre-list and pre-experimental traces, but the basic model worked well with activation of list traces only. The basic model predicted higher performance for higher frequency words 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, but effects of training frequency effects were found nonetheless, ruling out the basic model. The full model predicts effects of training experience due to activation of event traces from the training sessions. SARKAE and the simulation therefore employed the full recognition model.

Experiment 2: Eliminating Character-Context Effects of Training

Experiment 2 switched training from visual search to same-different character matching: 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 matching task) between every two experimental character judgments. This extremely high frequency character was not subsequently used in the post-training

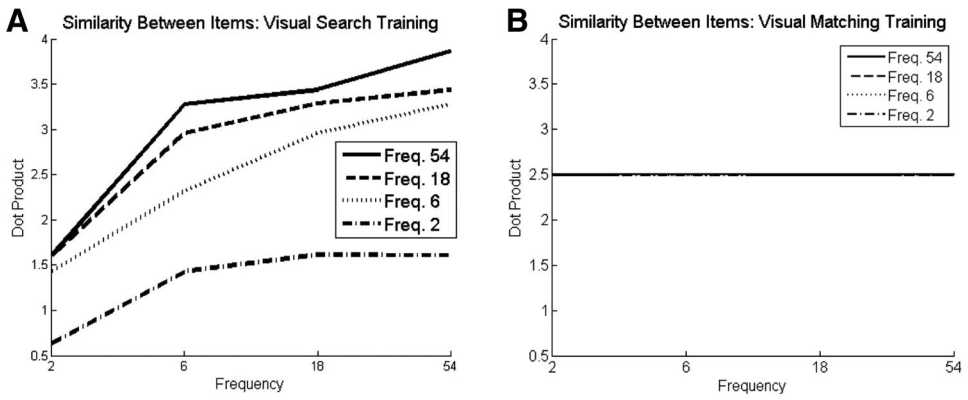

Figure 7. SARKAE (Storing and Retrieving Knowledge and Events): Similarity between normalized lexical entries (measured by dot product) after completion of training for visual search training (Panel A) and visual matching training (Panel B). Freq. = Frequency.

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]

Training Task

Method.

Participants. Seven people, recruited through an e-mail 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-in. (43.18-cm) 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, eight 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 seven 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 × 4.3 degrees. Specifically, each trial proceeded as follows: A character was presented on the screen for 500 ms, a white screen was then shown for 250 ms, the character was presented again for 500 ms, and then text appeared instructing the subject to respond "identical" or "different" by pressing the "v" or "m" key, respectively. This text stayed on the screen until the subject made a response. Once a response was given, the subject was given feedback (correct or incorrect) that remained on the screen for 1,000 ms, and was immediately followed by the next trial.

The two presentations of the character in each trial were either identical or varied slightly in size, rotation, or contrast of the character. If a trial contained a variation, only one dimension varied. 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. For example, in the case of rotation, 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 three per week. There were a total of 1,060 trials for Sessions 1–11, and 1,140 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. Is such learning frequency dependent? Frequency differences in acquisition of visual search were not significant in Experiment 1, suggesting that the same could have been true here. In fact, one might speculate that character matching produces more shallow character representations than visual search, possibly reducing further the impact of frequency variations. Unfortunately, the present design does not allow this question to be answered because the staircase algorithm does not adjust separately for different frequencies. In any event, the issue becomes moot given that the transfer tasks show frequency effects.

Post-Training Tasks

Following the training on the character matching task (approximately 2 days later), the subjects completed three post-training tasks: pseudo-lexical decision, episodic recognition, and forced-choice perceptual identification. In addition, for Experiment 2, post-training testing was carried out again 6 weeks after training. A programming error, discovered after the transfer tasks were initially analyzed, caused the forced-choice data to be very noisy and essentially uninformative. These results are therefore neither reported nor analyzed. Also, because the forced-choice results were not useful for the initial transfer tasks, forced-choice testing was omitted for the delayed testing at 6 weeks.

Pseudo-Lexical Decision

Method.

Participants. 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 three times

[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.

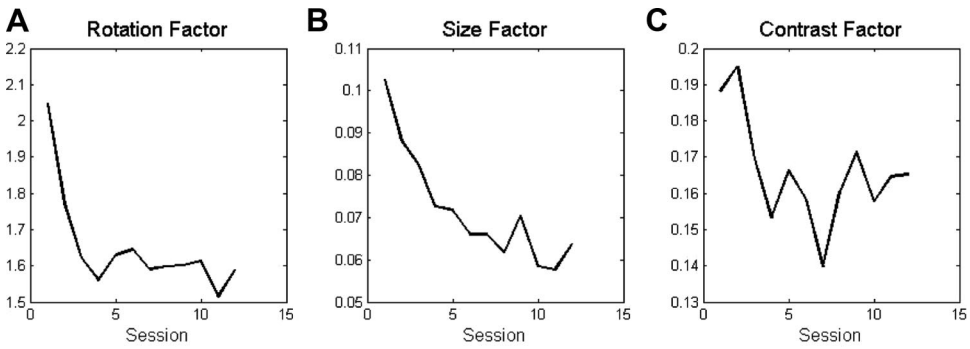

Figure 8. Average change in rotation (Panel A), size (Panel B), and contrast (Panel C) needed to obtain 75% accuracy as a function of training session. Rotation factor is measured in degrees, size factor is measured in percentage size difference, and contrast factor is measured in percentage contrast difference.

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 accuracy and response time results for both new and trained items are shown in Figure 4 for Experiment 2 immediate testing (solid black lines with circle markers) and delayed testing (solid black lines with triangle markers), along with the results for Experiment 1 (solid black lines with inverted triangle markers). In all cases, 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, and the patterns were not significantly different for any of these testing situations.

Discussion. The lexical decision results showed strong effects of training frequency on speed and accuracy of decision. The magnitude of the effects was also quite large, in the general neighborhood of that from Experiment 1. Furthermore, this frequency effect showed little signs of reduction over 6 weeks. This study eliminated character-context during training, so the results cannot be due to that factor. The obvious explanation is based on the stronger knowledge traces that result from more repetitions during training. The details are laid out in the later exposition of SARKAE.

It is noteworthy that accuracy is well below ceiling, response time is much slower than response times to a simpler task (such as onset of a flash of light), and both accuracy and response time are strongly dependent on frequency, yet all of these effects survive largely intact over 6 weeks of delay. Such results contrast quite strongly with episodic memory tasks that perhaps start with similar levels of accuracy (and possibly response time) when tests immediately follow study, but fall off sharply as the delay until test increases. Pseudo-lexical decision in the present setting is a kind of episodic memory task, given that one has to judge whether the test character had been studied in the initial few weeks of training. Of course the pseudo-lexical decision task does not require a discrimination based on context, so the failure to see forgetting could reflect a more general principle that much of forgetting is due to changes in context between study and test, and that such forgetting will be observed whenever context discrimination is integral to the task.

Episodic Recognition

Method.

Participants. 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 1,000 ms, 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 carried out shortly after the completion of training 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 5 (middle panel; also see Table 1, center). Similar to the findings from Experiment 1, false alarms significantly increased as frequency increased. There was also a marginally significant decrease in d' due to frequency. The hit rate analysis however showed no significant effect of frequency. Novel items in this study showed a bias to respond "new" compared with trained items.

Six of the seven subjects were tested again following a 6-week delay. The results of the delayed test are shown in the right panel

of Figure 5 (also see Table 1, bottom). After 6 weeks, one might expect some lapse in the efficiency of encoding and some increase in the variability of encoding of these relatively novel Chinese characters, resulting in a decrease in overall performance. Conversely, one might expect a delay to reduce activation and retrieval of the irrelevant training session traces, thereby boosting performance. According to SARKAE (discussed in the model description that follows), if the criterion used for a recognition judgment about items on a just studied list does not get adjusted according to the delay since training, the decrease in access to irrelevant traces will result in a decrease in both hits and false alarm rates, leading to a general decrease in "old" responses without a large change in d'.

Of primary interest for theory was the effect of frequency upon recognition study and test 6 weeks after initial training. Statistical analyses showed no significant effect of frequency on hit rates, false alarm rates, or d': Although the trends in the data were in the same direction as the other recognition findings, the noise in the data made it impossible to infer the presence of such patterns. Comparisons across immediate and delayed conditions 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 delayed tests of novel items again showed a bias to respond "new," compared with trained items.

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).

After a 6-week delay, unlike the pseudo-lexical decision task that produced large and essentially unchanged frequency effects, the d' and false alarm rate effects were reduced and possibly absent. The existence of frequency effects in recognition in Experiment 2 and their reduction with delay have important implications for recognition modeling. It is not uncommon to use a simplified model for recognition by assuming probes activate only traces of items from the study list (this approach was used in early versions of SARKAE applied to Experiment 1). In such a model, poorer performance for high frequency test items is due to increased confusions with traces of list items, because those traces are more similar to the high frequency test probes. The present design should have eliminated such similarity differences. In addition, if similarity differences are not present in knowledge traces, then performance differences caused by within-list confusions should not decrease over a 6-week delay, because the event traces are stored on the basis of list study at both immediate and delayed testing. Thus, to model recognition, we use an augmented model (similar in certain respects to previous models by Criss & Shiffrin, 2004a; Dennis & Humphrey, 2001; Shiffrin & Steyvers, 1997) that assumes confusions are produced also by the activation of event traces from the (recent) training sessions. If context changes over 6 weeks, then the activation of training session traces will be reduced, and the magnitude of frequency effects caused by such activation will be reduced.

The data showed a slight lowering of performance after delay, and at first glance could be thought surprising, given that the 6-week delay is not between the study list and test, but between the training sessions and both study and test after the delay. A simple application of the model would not predict this result: Activation of training session traces adds noise that reduces overall performance, so that a decrease in such activation caused by a 6-week delay would reduce this noise and, if the only factor affected by delay, would increase performance. Of course, many other factors are likely affected by delay. For example, the choice of features to attend and encode during list study, and to use as a probe at test, might be less optimal than immediately following training. In the simulation, we implemented this idea by allowing a lower encoding probability after delay than for immediate testing. It is very likely that another factor that was not incorporated in the simulation produces lowered performance after delay: There is considerable research, some quite recent (Criss, Malmberg, & Shiffrin, 2011; Malmberg, Criss, Gangwani, & Shiffrin, 2012; Murdock & Anderson, 1975), showing that traces are stored during testing as well as during study. Such traces would harm recognition performance more and more as they accumulate. We do not have sufficient data to look at test position effects, so that the performance shown and modeled, for either immediate or delayed testing, is in effect the average performance midway through testing. However, when comparing immediate to delayed testing, there is a difference: The test traces from immediate testing will tend to be activated during delayed testing, adding noise and reducing performance compared with immediate testing. In any event, whatever the mechanisms at work, the fact that delayed performance is slightly lower than immediate performance indicates that the factors harming delayed recognition outweigh the factor improving recognition.

The Experiment 2 results showed that new items, both targets and foils, tended to elicit a higher probability of responding new than trained items. It is possible that new test items lead participants to notice the presence of new features, causing a bias to respond new for a given level of global familiarity. There are a large number of complex issues regarding the coding of novel characters, given they do not have knowledge traces. These issues are taken up in the General Discussion.

Because Experiment 2 eliminated the interaction of character context with frequency, the correlation of similarity differences with frequency differences cannot be used to explain the pattern of frequency findings for lexical decision and episodic recognition. Thus, we introduce an explicit role for frequency per se, as described in the next section.

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

We next build a simplified simulation of the SARKAE theory and use it to predict the qualitative pattern of results from the two studies. The simulation is quantitative nonetheless, thereby making precise the assumptions in a way not possible with verbal descriptions. The simplification of the simulation is intentional, because building a model of all the relevant factors at work in five quite different tasks would bury the essential and common elements of

SARKAE in a forest of detail. Furthermore, SARKAE is intended to be an evolving theory rather than a finished product. The General Discussion takes up more realistic extensions of the theory, additional applications, and new results requiring additional components. Here, we give the "barebones" of the simulation (simulation details are given in Appendix B), with justification and alternatives reserved for the General Discussion.

SARKAE: Summary and Implementation of Simulation

Representation. Memory is organized into separate traces, each represented as a vector of feature values, values being integer counts. The traces are classified into event traces and knowledge traces, though there is a continuum from one to the other, because knowledge traces form through an accumulation over many events. At any point during training an event occurs and may cause retrieval of one or more previously stored traces based on similarity of the two vectors. Each selected trace can gain new information (additional counts) from the current event. It is assumed that at most one previous event trace, and at most one developing knowledge trace, can be selected. Thus, there is only one knowledge trace that develops for each class of event (in the present case, a Chinese character).

The values in a trace vector are grouped into features (e.g., "size," "color," "orientation") with the values specifying the kind of feature (e.g., "huge," "red," "45°"). Event traces are sparse, incomplete, and contain some values in error. Knowledge traces develop over experience as relevant events occur, and become replete with many feature value counts. As such traces accumulate information a kind of law of large numbers takes place: For the kind of information that remains consistent over events, the noise comes to be dominated by the "true" values so that the distribution of feature value counts for a consistent feature has a mode at or near the true value.

The simulation assumes that each trace vector, whether a knowledge trace or event trace, has 432 vector positions, each position encoding a count of values for some feature; 0 represents no values stored, 1 represents a single value stored, 2 represents two values stored (which must have come from two different events), and so on. The vector is illustrated schematically in Figure 9. The vector is divided into three parts. The first part consists of 160 positions encoding content features (including content from items nearby in time, space, and thought). These values are organized into eight feature values for each of 20 features. Because the Chinese characters in the studies are largely represented by physical features (such as shape features), we often refer to the content features as physical features. The second part consists of 32 feature values that encode the value for a single high level feature. The "true" representation of a given character consists of initial values assigned to the physical features and the high level feature; the 20 content values are assigned randomly (so there is random content-feature overlap between the representations of different characters).

Each of the 32 characters is assigned a different high level feature value, because we assume that the high level feature is chosen to make each character distinguishable from all the others. Such a distinguishing feature is useful and perhaps necessary for the training task of visual search. The physical matching required in Experiment 2 does not require the use of a distinguishing feature, but it does no harm to include such a feature in the traces formed for that study, and it maintains consistency in the assumptions across the two studies. 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 third part of a trace vector consists of 240 context feature values, organized into 30 features each with eight values. These values represent general list and environmental context, including both external context (e.g., room setting, furniture, temperature, etc.) and internal context (e.g., mood, thought processes about the study and other matters, etc.). At the start of training, one context value is chosen at random for each of the 30 context features.

Retrieval probe. SARKAE assumes cue dependent retrieval. In any task or setting, access to event or knowledge traces is carried out with use of a retrieval probe. The construction and use of a retrieval probe occurs both during storage and testing. For characters presented above threshold, a retrieval probe is formed from various sources of information including the presented character, nearby characters, current context, and the knowledge trace of the presented character. For characters presented at perceptual threshold, the probe will consist of low level perceptual features. In tasks and settings in which retrieval is initiated with a test cue, the retrieval probe evolves dynamically over time as features are extracted from the test stimulus. These retrieval dynamics are used in the present simulation of pseudo-lexical decision in order to produce response time predictions. However, because we analyze only accuracy of responding for event recognition and perceptual identification our simulation for these tasks is simplified by allowing the retrieval probe to evolve to a final asymptotic state that is then used in comparisons with the memory traces. In the General Discussion, we take up models that use the dynamics of activations for these tasks. The details of probe construction are given in the sections to follow, but we note in advance that the role of pure frequency is built into the construction of the retrieval probe.

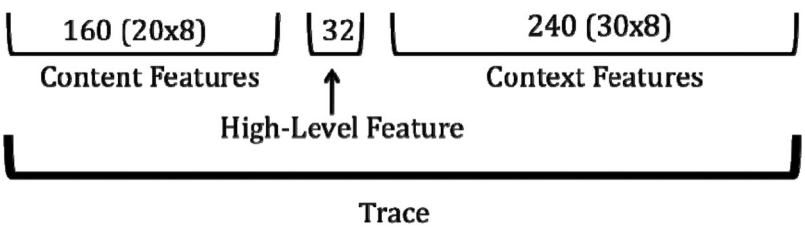

Figure 9. Schematic example of a trace vector in SARKAE (Storing and Retrieving Knowledge and Events).

Calculation of likelihood ratios. At essentially every stage of processing in storage, training, study, and retrieval, probe cues are used to activate memory traces. In our Bayesian-inspired approach, such activation is characterized for each trace, whether an event trace or knowledge trace, as a likelihood ratio: the probability that the trace in question matches the probe, divided by the probability that it does not. The calculation of such a likelihood ratio could reasonably be based on the number of counts, the pattern of counts, or both. We decided for all traces to use the pattern of counts. Event traces have one count per value, and the calculation uses the numbers of matching and mismatching feature values. Knowledge traces have multiple counts, so a single sample value is taken from the knowledge trace for each feature, the sample being taken in proportion to the counts in that trace for the different values. The resultant vector of values is matched to the retrieval probe, in the same manner as for event traces. If, for example, we would take a given knowledge trace and exactly double every feature value count, the likelihood ratio resulting from matching to a fixed retrieval probe would not on average change.

The calculation of the likelihood ratio is based on the matches and mismatches of feature values for features that have a value in both probe and trace (this approach borrowing from the REM model). The value for a given feature might be missing in a trace or probe, and if so, that feature is ignored (a more elaborate and probably better conceived approach would take missing values into account; such an approach is taken up in the General Discussion). If there is a value for a feature in both probe and trace, then the values can either match or mismatch. The number of matching values (k_m) and the number of mismatching values (k_q) 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.

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

In this equation, the ratio for a matching feature (*m*) is given in the first bracket, and for a mismatching feature (*nm*) 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. Therefore they are instead estimated through a simulation technique (see Appendix C for details). 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 trace likelihood ratios are the basis for all retrieval in all tasks, as described in the following sections.

It should be noted that training frequency affects the likelihood ratios in a variety of direct and indirect ways. First, there are more event traces when there is more training, obviously producing more likelihood ratios. In addition, knowledge traces have a decreased number of missing values as training frequency increases. Event traces are also stored partly on the basis of retrieval from knowledge, thus indirectly producing frequency-dependent effects inherited from knowledge retrieval, and retrieval probe construction depends in part on retrieval from knowledge and thus will reflect frequency differences encoded in knowledge traces. Finally, the accuracy of knowledge retrieval is dependent on pure frequency (as described shortly).

Context change. Situational context changes in complex and poorly understood ways. It certainly does not change uniformly as time passes. For the present tasks, it is nonetheless possible to make plausible and simple assumptions and incorporate these in the simulation. After each trial in each training session, the values representing current context are altered: Each context value is replaced with probability .01 by a uniformly chosen value (the same value for both studies). Between each training session, between the last training session and the first transfer session, and between successive transfer sessions, this context change process is carried out N_c times (estimated to be $N_c = 20$). For the delayed transfer tasks in Experiment 2, the context change process is carried out an additional 36 N_c times (representing the 6 weeks that have passed—this value was set arbitrarily). The values chosen capture the idea that context change is slow but accumulative within session, substantial between sessions, and then quite a bit larger during a several week delay.

SARKAE Simulation: Modeling of Training

Each presentation of a character (whether in training, or study or testing during transfer tasks) produces an event trace, with at most one value per feature, and also adds feature values to the vector representing its knowledge trace. The features stored or added are sampled from several sources and are sometimes stored incorrectly in which case the value stored is proportional to the knowledge base rates (obtained by summing across knowledge traces for that feature and its values). The sources of features for both experiments are the physical character features, current context (context gradually changes over the course of training), the character studied on the previous study trial, and character features extracted from the character's knowledge trace. For Experiment 1, additional sources are the foils on the current trial. At the end of training there are for each character many incomplete and error prone event traces, and a knowledge trace that has accumulated feature values over training. Due to accumulation over training, the counts in the knowledge trace for each content/physical and high level feature tend to be largest for the feature value in the representation of that character. The values for context features are more diffusely spread out because context changes over the course of training.

The knowledge traces that develop during training are involved in both storage and retrieval in many ways that will be described in the sections to follow. The event traces that are stored during training have a more limited use, because they do not get activated and participate in pseudo-lexical decision or forced-choice perceptual identification—only knowledge trace activation determines performance in our modeling of knowledge retrieval. The event traces formed in the training sessions do play a role in the later recognition task: They are activated by a recognition test probe to the degree that they incorporate context feature values that match those in the probe. Such activation is used to explain recognition

frequency effects in Experiment 2. Because context keeps changing, the activated training session event traces tend to be those most recently stored. It is slow to simulate the storage of every event trace for 2 weeks of training and cumbersome to do so when only the recent traces play a significant role. Thus, in our simulations we allowed the knowledge traces to evolve over all training sessions, but the simulation only stored event traces for the last training session.

Event storage and knowledge development. For both content and context features, and for both event and knowledge traces, the feature values stored will not necessarily match the actual values: A value might not be stored, and if stored might be stored incorrectly. We assume for simplicity that event traces do not have multiple counts for a given feature, so each event trace stored will have some features with all zeros (incomplete storage) and will have some features with a single value marked with a one, but the value marked might not match that in the event. As remarked above, the knowledge trace accumulates feature values; thus, for content and high level features, the mode of the values gradually comes to approximate the true value of each feature. This scheme is illustrated in Figure 10, albeit with fewer features and feature values than in the actual simulation. The figure shows a simplified set of features and feature values representing a presented character (Row 1), an event trace stored for that presentation (Row 2), the knowledge trace that includes storage due to all previous event occurrences (including the present one) when a large amount of training has taken place (Row 3), and when only a small amount of training has taken place (Row 4).

Storing feature values during training. When an event occurs there will be a process of storing features of the event in both an event trace (possibly more than one event trace) and one (or more) knowledge trace. Our simulation assumes each presentation produces storage in one event trace and one corresponding knowledge trace. Many factors affect which features are stored, and with what probability and timing, factors such as attention, coding strategy, and presentation time. In our training sessions, for either Experiment 1 or 2, the task and timing remain constant throughout, so we ignore these factors and simplify the simulation as described in the following sections. Because we let the vector representations of all trained characters (event and knowledge) have the same "slots" for features and feature values, we can specify storage by letting each vector position have a chance at storage (in parallel). For each position, we first choose a source for storage, and then choose whether to store, and if something is stored, what to store.

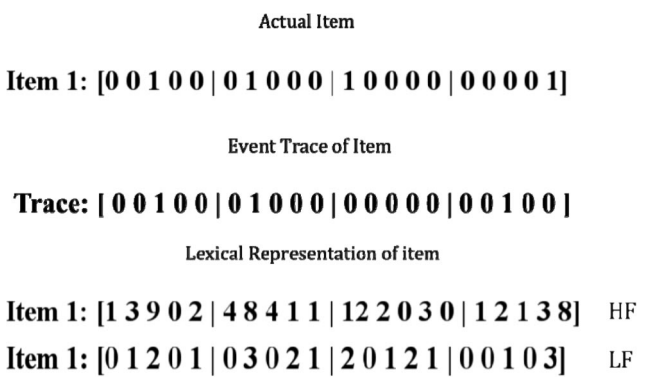

Figure 10. Item representation and lexical entries in SARKAE (Storing and Retrieving Knowledge and Events). HF = high frequency; LF = low frequency.

Sources of features for storage. An important component of SARKAE is the assumption that knowledge develops by incorporating feature values of other items in the temporal, spatial, and mental neighborhood. We instantiate this idea by stipulating probabilities for the source for storage of feature values: For context features the only source of feature values is the vector of current context values. For physical features the source rules vary with experiment. For the search task (Experiment 1), the source of the feature value is the (high or low) physical feature value of the target with probability s_p, the physical feature value of the previous target item with probability s_t, a physical feature chosen from the character's knowledge trace (proportional to the counts there) with probability s_k, and the physical feature value of a randomly chosen foil with probability $1 - s_p - s_t - s_k$ (the first three "estimated" to be .57, .14, and .14, respectively). For Experiment 2, there are no foils, so these three estimates were simply renormalized by dividing by their sum, thereby making them add to 1.0. For Experiment 2, note that the previous target item is the same fixed character that appears every other trial. These parameters determine sources for storage in both knowledge and event traces, but are applied independently.

Note on terminology. In SARKAE, one source is nearby items, but these are not termed "context"; that term is reserved for generalized context not dependent on the particular items studied. Confusion is possible because several researchers have used context to include features of other nearby items (e.g., as in the temporal context model of Howard & Kahana, 2002). This difference in terminology is not by itself substantive, although of course the detailed instantiations of the different models differ.

Storage probabilities. Assume a source for storage of a feature value has been selected. If the value is for a character's low- or high-level feature, some value (see below) is stored with probability u_k ("estimated" to be 0.5); with probability $1 - u_k$, no value is stored for that feature;[2] if the value is for a context feature, the storage probability is u_c, ("estimated" to be 0.1), and no value is stored with probability $1 - u_c$. For storage in the knowledge trace, the probabilities of adding a feature value were set to the same probabilities, but applied independently.[3]

Correct or incorrect storage of a feature value. Assuming that a feature value is stored, there is a probability that its value it will be copied correctly from the selected source. For all sources but the knowledge trace, this probability, c, is the same value ("estimated" to be .8). When the value is not copied correctly, the value stored is chosen in accord with the base rates for that feature (in proportion to the current summed counts for the values of that feature in the knowledge base). If a feature has no values at all in

[2] Conceptually, it seems appropriate to let the rate of storage differ for high- and low-level features, and differ between the two experiments. However, when we did allow such variation, the parameter estimates and the fit did not differ very much.

[3] Because decisions in the knowledge retrieval tasks (pseudo-lexical decision and perceptual identification) are based only on activation of knowledge traces, and decisions in the event retrieval task (episodic recognition) are based solely on activation of event traces, the assumption of independence is probably not consequential in the present setting.

the knowledge base (which happens very early in training), then a value is chosen uniformly randomly.

Experiment 2 demonstrated that SARKAE must have a role for "pure frequency." There are many ways to let SARKAE include a role for pure frequency. We chose to assume that the dependence on trace richness of extraction of information and of subsequent storage is reflected in the probability of copying a feature value correctly when the source of the feature is the knowledge trace (discussion of alternatives is deferred until the General Discussion). Because trace richness is represented by the total counts, n, in a character's knowledge trace, we let the copy-correct probability, p_{cc} rise with n: $p_{cc} = 1 - \exp(-\alpha n)$.[4]

Similarity structures formed in the two experiments. These storage rules, when used to develop knowledge traces, produce different knowledge structures for the two experiments. Figure 7A shows that training with visual search (Experiment 1) causes knowledge traces of higher frequency characters to grow more similar to each other. Figure 7B shows that training with character matching (Experiment 2) does not cause similarity to vary with training frequency.

SARKAE Simulation of the Three Transfer Tasks

There are three immediate post-training transfer tasks representing retrieval of different types: event memory, knowledge, and perception. In Experiment 2, retrieval of event memory and knowledge are repeated at 6-week delay. These tasks come close to spanning the domain of memory, so it is a non-trivial exercise to produce a simple simulation model that will utilize the same processes and assumptions, and most of the same parameter values (for conditions of "like" kinds). The generation of such a simulation is nonetheless useful as a means of making concrete the core processes of SARKAE.

Episodic recognition.

Characters with and without knowledge traces. Recognition study and test lists included some untrained characters without knowledge traces. Such characters would in principle be treated in ways analogous to trained characters. However, there are several respects in which such characters would require special assumptions. First, such characters would very likely contain new features not in the set of trained characters; such new features are in fact used in modeling pseudo-lexical decision, when untrained characters are presented. Second, new characters have no corresponding knowledge trace, raising issues about the extraction of features when the "source" is knowledge: Presumably similar knowledge traces provide such features, but specifying just how would involve much additional machinery. Although assumptions about novel feature use and extraction of features from similar knowledge traces could be introduced, doing so would add much complexity just to predict what is in effect one data point: recognition performance for novel characters. We therefore decided to leave the recognition simulation model in its simpler form, predicting only recognition performance for trained characters. Some issues regarding novel items, and items with incipient but weak knowledge traces, are taken up in the General Discussion.

List study. Event traces are stored for each character on the study list, just as during training, with two exceptions: (1) The source of feature storage is not allowed to be visual search foils or prior trial visual search targets, since these do not exist during list study for recognition. (2) Features for novel characters are restricted to the same set as for trained characters (if novel features had been included there would have been no noticeable change in the predictions for trained items). Recognition instructions emphasize storing for memory, so the probability that a feature will have a feature value stored is allowed to be higher than during training. Furthermore, we allowed the value of this study-list storage probability, u_s, to differ for Experiment 1, Experiment 2, and Experiment 2 delayed ("estimated" values were .85, .75, and .55, respectively).

For study of words in a list, there is a long history of research (e.g., Atkinson & Shiffrin, 1968) showing that storage occurs due to joint rehearsal and joint coding of groups of words, usually words in close temporal proximity. This sort of rehearsal is likely less pronounced for the Chinese characters in the present studies, but it would still be reasonable to allow the source of features for encoding character j to include a nearby character, particularly the previous one. However, because the data collected were insufficiently rich to allow sequential analyses, we did not include this possibility in the simulations (and we did not include such a possibility in the process of constructing a test probe, discussed next).

Test probe construction. The test list includes targets (on the study list) and foils (not on the study list). For each test character, a memory probe is constructed as follows: Because the test character is available during probe construction, we assume that one feature value is encoded for each and every character and context feature. For content features, the encoded value is copied correctly from its given source (physical features and the knowledge trace). The parameters determining the source of the values are identical to the probabilities applying during list study, renormalizing as needed. The source of the encoded context feature value is always the current context vector (so the context portion of the test probe is always identical to the current context vector).

Probe to trace matching. In the general model, the test probe is compared to all traces sufficiently strong and sufficiently similar to the probe. For our simulation, this process was simplified: The probe is compared in parallel to two sets of event traces: all event traces of characters studied during list presentation, and event traces from the last training session that have at least M_C context feature matches out of the 30 possible. The criterion M_C is a parameter that determines the amount of confusion from the traces in the last training session.[5] The result of matching is a likelihood ratio for each final-training-session trace that passes the activation threshold, and each list trace. The calculation of the likelihood ratio was described in Equation 1.

[4] This function was chosen fairly arbitrarily and was used in all our simulation efforts; it is not unlikely that some other another monotonically increasing function is and will eventually prove to be a better choice.

[5] A better justified activation criterion would be based on both (1) the number of feature values in a trace and (2) the trace likelihood ratio, and applied to all extra-list and list event traces. However, the present simplification should produce similar predictions because it excludes a few extra-list traces that would have just managed to get activated and includes a few list traces that would not have exceeded activation threshold, but including or excluding traces that match poorly does little to reduce performance and tends only to change the best placement of the decision criterion. The sum of likelihood ratios used to decide is dominated by traces that match well.

Recognition decision. The average of the likelihood ratios is the odds in favor of the test character being from the list, and is therefore used to produce a response (as in the REM model). The odds ratio Φ for old over new is given in Equation 2.

$$\Phi = \frac{1}{n}\sum_{i=1}^{n} \lambda_i \quad (2)$$

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 *R*, independent of item frequency ("estimated" to be .5, .4. and .5 for Experiment 1, Experiment 2, and Experiment 2 delayed, respectively).[6]

SARKAE predictions for recognition. Figure 5 gives the simulation predictions (and data) for recognition for the two experiments and the delayed test. The predictions and data are close both qualitatively and quantitatively. Given the substantial noise in the results caused by the small amounts of data collected, it would be unwise to ask for additional precision. Both data and predictions show that LF items produce 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). At 6-week delay, predicted performance is slightly lower than immediately after training, and predicted frequency effects are smaller, both roughly consistent with the data.

Pseudo-lexical decision. Pseudo-lexical decision requires the participant to judge whether a test item has ever been studied, or instead is novel. In principle a response could be based on summed familiarity to event traces, knowledge traces, or both. However, because the knowledge traces are far more complete and accurate than individual event traces they provide a much better basis for response. We therefore assume lexical decision is based on a comparison of the test character to all 32 knowledge traces. Each knowledge trace contributes a likelihood ratio and these are averaged to produce a decision statistic.

If measures of accuracy were the sole goal of lexical decision, one could simply use this statistic to make a decision, but this is not the case: The participant is asked to respond as quickly as possible while maintaining high accuracy, conflicting goals that can be modeled best in the context of a joint model of accuracy and response time. SARKAE is therefore extended by assuming that probe features are gradually extracted after presentation of the test character, these extracted probe features accumulating as time passes. At each point in time the current probe is compared to the 32 knowledge traces. Each comparison results in a likelihood ratio, and these are averaged to produce an odds for "old"/"new."

Because there is noise in the process by which the probe features accumulate, and noise in the process of comparing the probe to each knowledge trace, the odds moves somewhat randomly up and down as time passes, though generally rising for old test characters and generally falling for new test characters—this is known as a non-homogenous random walk or diffusion.

New features for novel characters. If novel characters are assumed to have the same features and the same distribution of feature overlap to all other items as trained characters, then the knowledge traces for the trained characters tend to have a fairly large similarity to novel test characters, and the pseudo-lexical performance is poor. To predict good performance for novel characters, it is useful to make the reasonable assumption that novel characters have some novel physical features. To introduce this idea into the simulation, it is simplest to assume that these novel features will occupy some of the 20 physical features in the probe, displacing other features, and to assume that the value of these novel features will not find a match in any of the 32 knowledge traces. The result will be to reduce matching for a novel test item, and thereby improve the accuracy and speed of responding. We let a parameter *nf* be the proportion of features in a novel item probe that are new; for Experiment 1, *nf* was estimated to be 0.6, and for Experiment 2, immediate and delayed *nf* was estimated to be 0.35.

Context features. In principle, the probe should contain context features. In fact, previous research has shown that context features do join the probe: The addition of context features to knowledge traces caused by an item presentation 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 the simulation simplifies by assuming that the probe contains only physical character features.[7]

Probe construction. The probe accumulates physical features extracted from the test character. At each time step after presentation of the test character, for each not-yet-encoded physical character feature, there is a probability u_s (estimated to be .75) that that character's actual physical feature value is encoded in the percept/probe. If encoded, the probability of encoding the value correctly is the standard value (.8) and otherwise it is encoded with a value in proportion to the base rates for that feature. We assume that the value of any feature encoded, whether correct or incorrect, is kept as a part of the accumulating probe until a response occurs; thus, the probe gains features as time passes, but does not revise previously encoded features. The features that join the probe do not do so differentially as a function of trace frequency because the knowledge trace is yet to be contacted.

Comparing the probe to knowledge traces. At each time *t*, the probe is compared to each knowledge trace. As usual, we assume one sample is made from the values for a given feature in a particular knowledge trace, in proportion to that feature's counts in that trace. This occurs for each feature. We assume that richer knowledge traces are easier to contact and hence let the probability of correctly copying that sampled feature be frequency dependent, based on the same equation and parameter used for storage in training: $1 - \exp(-\alpha n)$, where n is the total counts in the trace. The value selected given an incorrect copy is simply proportional to the base rates for that feature's values in knowledge. Then at each sample moment, the incomplete probe vector is compared in parallel to all knowledge traces, each trace being represented by

[6] Using an *R* of 1.0 for all three gave good qualitative fits, and was not far worse qualitatively. This insensitivity to *R* is related to the extreme skewing of the distribution of odds.

[7] In an expanded model, context features would be present from the start of the feature extraction process, producing a baseline starting level for activation. One method for adjusting for this baseline level would be an upward adjustment of both response thresholds.

one value per feature. The likelihood ratio for a given trace for the probe used at time t is given by Equation 1.[8] The odds at time t that the test character is a trained character is the average likelihood ratio over the 32 knowledge traces produced by a probe at time t.

Timing of the random walk. The accumulation of feature values into the probe is simulated as a sequence of discrete steps, each step assumed (somewhat arbitrarily) to occur every 100 ms. Note that the sample from each knowledge trace is made independently from one sample time to the next. This fact and the noise in constructing the probe both make the odds fluctuate from one sample to the next, although the value does ted to stabilize somewhat as the probe becomes more complete. Because the value of the odds fluctuates over time, we can think of the odds as making a type of random walk. This random walk does not start until after an initial mean residual time (MRT; "estimated" to be 590, 780, and 800 ms for Experiment 1, Experiment 2, and Experiment 2 delayed, respectively). The residual time was first subtracted from all response times (RTs) in a given condition, and the model then fit to the data to produce model RTs and accuracy—the *MRT* was then added to these RT values to produce the predicted response times.

The random walk decision process. A positive response is given if and when the odds reaches or exceeds an "old" boundary, R_O ("estimated" to be 0.6, 0.75, and 0.75 for Experiment 1, Experiment 2, Experiment 2 delayed, respectively), and a negative response is given if and when the odds ratio reaches or drops below a "new" boundary R_N ("estimated" to be 0.30, 0.25, and 0.25 for Experiment 1, Experiment 2, Experiment 2 delayed, respectively). Higher frequency "old" characters produce faster and more accurate responses because the gathering of features from a knowledge trace is more accurate for more highly developed knowledge traces.

SARKAE predictions for pseudo-lexical decision. The SARKAE simulation predicts, as shown in Figure 4, that in Experiment 1, Experiment 2, and Experiment 2 delayed, the HF items are recognized both more quickly and more accurately than the LF items. The level of performance and the effects of frequency are not much altered by delay. These predictions match the data fairly well.

Forced-choice perceptual recognition. Naively, one might imagine that two-alternative perceptual-identification could 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. For one thing, many studies 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 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 simulate with a very simplified model.

The model is predicated upon prior research showing that the use of a pattern mask immediately following the flashed target 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 nor 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. (2004) showed that this result held when a pattern mask was used, which is defined as 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 reduced the use of physical features in the forced-choice decision. For this reason, the simulation bases the decision solely on the high level feature and its value.

High-level feature extraction. The probability of extracting a high level feature from the flashed character is frequency dependent, based on a simple linear function: $e^H(\mathrm{n}) = b + m\mathrm{n}$, where n is the total number of counts in the trace, and b and m are parameters of the linear function. If a high level feature is extracted, the probability of copying its value correctly is frequency dependent in the same way as assumed heretofore: The probability of copying the value correctly, $c^H(\mathrm{n})$, is assumed to be a function of n, the total counts in the trace: $c^H(\mathrm{n}) = 1 - exp(-\alpha\mathrm{n})$. If copying is incorrect, a value is chosen in proportion to base rates for the feature. The product $e^H(\mathrm{n})c^H(\mathrm{n})$ gives the probability, $p^H(\mathrm{n})$, of extracting from the flash the higher level feature for the flashed character.[9]

Discounting when guessing. If a high level feature value is extracted and matches one of the choices, that choice is selected. If there is no match, or if nothing is extracted, then a sophisticated guess is made. The idea is that feature extraction is more accurate for higher frequency targets, so the failure to extract implies that the target probably was of lower frequency. If n1 and n2 are the number of feature counts in the knowledge traces for choices 1 and 2, the probability of guessing 1 is as follows: $[1 - p^H(\mathrm{n1})]/\{[1 - p^H(\mathrm{n1})] + [1 - p^H(\mathrm{n2})]\}$.

SARKAE predictions for forced-choice identification. The dependence upon frequency found in the data is predicted by the model, as shown in Figure 6. The predicted increase in performance for higher frequency targets is due to a higher probability of extracting the correct value of the high level feature. The predicted increase in performance for foils of higher frequency is due to the bias to choose the lower frequency choice when guessing. This

[8] As was true in recognition, the complexity of the sampling rules makes it hard to derive analytically the feature ratio terms in Equation 1 that produce the likelihood ratios, so again a simulation method is used to derive them, as described in Appendix C. Once simulated for a given set of parameters, the ratios are fixed and used for all calculations and all time points.

[9] We simplify the simulation by ignoring the possibilities of contacting the knowledge trace of some other character, and of choosing a feature value from the base rates that happens to match the correct feature value.

bias is based on the inference that something would have likely been extracted had a higher frequency character been flashed.

General Discussion

Simplified Simulation Versus Theoretical Framework

The simulation of SARKAE presented in the preceding section was not intended as a deeply developed model of any one of our tasks, let alone other tasks, and instead was presented to make concrete the way that the SARKAE approach could be applied to quite distinct tasks of both event and knowledge retrieval. The SARKAE theoretical framework is meant to guide theory development as models are incorporated for new tasks and extended for existing tasks, especially in light of new data. Thus, much of this general discussion focuses on the ways SARKAE could be extended in more realistic ways and could be applied to more tasks.

Frequency Representation and Effects

Because our studies had frequency as a primary focus, and because frequency has many and complex roles in the model, it is worth reviewing these here. When an event is encountered the first time, we assume it is encoded as an incomplete, impoverished, and error prone event trace. For simplicity, we assume each feature value has at most one count, and that extra study time fills empty slots rather than adds counts to a given slot. This idea is extended to the next several repetitions. However, knowledge traces develop by accumulation of information from event traces and these do accumulate counts in given feature value slots. Thus, a better conceptualized model would allow count accumulation at every stage of storage of both event traces and knowledge traces. Implicitly, we are therefore assuming such count accumulation occurs seldom enough to be ignored in typical episodic tasks.[10]

In most tasks and real world settings, the accumulating information in a knowledge trace will incorporate the statistics of the featural context of the repeated events (including other nearby events, and environmental context), thus naturally producing a major role for context as a basis for frequency effects. However, our Experiment 2 went as far as possible in reducing such context variation over repetitions, and the studies nonetheless showed frequency effects. We therefore included in SARKAE a role for "pure frequency": Knowledge trace richness was characterized in terms of increasing number on feature value slots that have at least one count, and in terms of increasing counts in given slots. For retrieval of knowledge traces richness had two effects: (1) whenever a feature value is extracted from a knowledge trace the probability the value will be encoded correctly, as opposed to a random choice in accord with base rates, is assumed to increase with total trace richness; (2) when time to process is limited, as in perceptual identification when the target is flashed briefly and masked, the probability of extracting any feature value (correct or incorrect) is assumed to rise with trace richness.[11] Both factors should increase monotonically with trace richness (the present data do not allow the form of the functions to be inferred, so the simulation model adopted simple functions that were arbitrarily chosen—see Appendix B).

The dependence upon trace "richness," and the representation in terms of knowledge trace counts and completeness, and in terms of number of event traces, adds a great deal of structure to the simple account of frequency effects in the REM model. The REM account depended only on what could be described as differential distinctiveness of event traces. That idea is incorporated in the present approach, albeit in a different form, but the present approach is far more fleshed out, and considers more factors.

It is of course the case that 2 weeks of training of Chinese characters falls well short of the lifetime of training given to relatively common words. Thus, we must be careful in generalizing the conclusions about the role of pure frequency from the present studies to words. It could be the case that the relative strength of the pure frequency factor drops as training continues to extremely high levels. Thus, pure frequency might predict a large difference between, say, 10 and 100 presentations, but not much difference between, say, 1,000 and 10,000 presentations. The present studies do not speak to this issue.

Consider next that a first stage when any item is presented for study or test is retrieval from knowledge. Knowledge trace richness will therefore affect encoding in episodic tasks as well. Such effects are indirect but nonetheless important. For example, in event recognition the encoding of higher frequency items at study and test will be less error prone. Beyond these effects is the obvious factor that repetitions generally produce more event traces. (We say "generally" because a repetition may sometimes cause accumulation of information in a previous trace rather than formation of a second trace; this process causes a repeated trace to become less similar to others, a process we have termed differentiation.) The additional event traces tend to have similar content, but differing contexts (both in terms of other nearby events, and in terms of environmental context, because environmental context tends to drift over time). Episodic retrieval uses memory probes with both content and context, and the context tends to be that present at the time of retrieval. This produces a tendency to discount older event traces, and favors retrieval of recent events over older ones. Nonetheless, the existence of extra traces due to event repetitions will interfere with episodic retrieval of other items/events, in both tasks like recognition that rely mainly on familiarity judgments and tasks like recall that relay on trace sampling and recovery, to the extent that the traces share content or context similarity.

Finally, we note that storage and retrieval may vary with pure frequency in ways beyond those specified by SARKAE. However, processes that improve storage and retrieval (of event and knowledge traces) as pure frequency rises must be relatively weak, partly because recognition is harmed by higher frequency, partly because environmental frequencies vary by many orders of magnitude, and partly because perception studies show perception (sometimes) tends to be driven more by bottom up than top down factors (as in Pelli, Farell, & Moore's, 2003, studies, which are discussed soon).

[10] Whether this simplifying assumption should be adjusted for studies varying study time, spacing of repetitions, and type of encoding is an open question, and might depend on the degree to which repeated items tend to form a single richer trace, rather than multiple separate event traces.

[11] In principle, the extraction probability should also rise with trace richness early in the processing of any presented stimulus, but this feature is not included in our simulations of recognition or lexical decision because it would not change the predictions significantly.

Matching of Probe to Trace, and the Nature of Activation

In the simulation, the activation of a trace by a probe is a likelihood ratio calculated on the basis of matches and mismatches of probe feature values to trace feature values, but only when there exists a non-zero value on a given feature in both probe and trace. If such a corresponding value is present in one but missing in the other, no evidence is calculated. In principle, the absence of a value is usually diagnostic, and SARKAE therefore needs to be augmented. Suppose, for example, that an event trace is stored for a word studied in a list, and contains word content and list context. Suppose a dot pattern is presented at test. The dot features will not match the word features, so will be ignored, but the context features in the probe and trace will match, producing a large likelihood ratio for match. We have recently developed a more sensible matching formula that infers negative evidence (1) when the probe has a value in a feature for which the trace has no value, or (2) when the trace has a value in a feature for which the probe has no value. The formula is described in Cox and Shiffrin (2012) and is given in Appendix D. Note that our treatment of pseudo-lexical decision assumed new feature values for novel characters; using the new formula, one could just as well assume new features and accomplish the same result.

The simulation of recognition and SARKAE generally assume that only reasonably similar event traces are activated and take part in determining the value of familiarity. The simulation implemented this idea crudely by assuming all list traces and all traces in the last training session were in the relevant set of traces. A threshold for similarity would be a better-justified basis for activation. Whether this threshold would need to be adjusted for different conditions (say for foils differing in similarity to targets) remains an open question.

Representation and Features

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. Yet, some broadening of the present vector representation is probably needed, particularly to deal with configurality. Because our studies were rather simple, we needed to take only one step toward configurality by assuming a single "high level" distinguishing feature for the Chinese characters. Mueller and Shiffrin (2006) broadened the representation to encompass all binary feature combinations. The system proved able to account for a number of findings in cognition (Mueller & Shiffrin, 2006, 2007), but at the cost of increasing the ability of the model to explain most results.

Decision Criteria for Recognition

An advantage of using likelihood ratios to represent activations (in SARKAE and in other similar models) is that a fixed or close-to-fixed recognition criterion can be used despite variations in factors like the frequency of test characters. In REM, for example, a criterion of 1.0 did an excellent job of producing qualitatively correct predictions across variations in item strength, list strength, and list length (Shiffrin & Steyvers, 1997). The present model is more complex, so the appropriate criterion no longer lies at 1.0, but use of 1.0 nevertheless produces decent predictions from the SARKAE simulation.

Our studies here, and indeed most studies in the literature, use a class of stimuli that are relatively homogenous (e.g., Chinese characters in our studies, or words in many other studies). We have carried out recognition experiments in which the stimulus classes differ widely (e.g., words, faces, Chinese characters, vacation scenes, snowflakes, dot patterns) and are seen and tested once each. Familiarity can be expected to vary widely across these classes. The studies (Cox & Shiffrin, 2011, 2012) make it clear that responses are made appropriately for different classes of stimuli. For example, two-alternative force choice recognition testing gives similar performance whether the two choices are from one class (e.g., both words or blobs) or from different classes (e.g., a word vs. a blob). Standard signal detection models would have trouble with such results, given they seem to require different criteria for each class but there is no opportunity to learn such criteria. A possible solution is rooted in an alternative approach based on the shape of the dynamic profile of activation, discussed next.

The Dynamics of Retrieval

The need to model response times for pseudo-lexical decision required us to model the dynamics of activation and decision. Logically, all tests evolve dynamically, and in principle response times could be measured and modeled in other tasks, particularly including event recognition. The SARKAE simulation for event recognition did not utilize dynamic assumptions, in good part because the data were too limited to allow reliable estimates of response time statistics and therefore a dynamical model would have added much complexity for little purpose.

Cox and Shiffrin (2011, 2012) have been developing dynamical models for event recognition, and using them to predict both accuracy and response times. The model assumes that features are extracted from the recognition test stimulus over time. The probe contains context features at the start of evidence accumulation, and context features are gradually extracted from the test stimulus and added to the growing probe. Familiarity thus changes over time, generally rising when targets are tested, and dropping when foils are tested. In order to deal with different familiarity levels and profiles for differing item types (words, faces, visual objects, blobs, dot patterns, and so forth) the model bases decisions not on the level of familiarity itself but rather on the accumulated changes in familiarity as time passes. There are thresholds for responding that do not depend on the type of item tested. The response choice depends on the threshold reached, and the response time is determined when that threshold is passed. The resultant model has the nice property that it predicts data that Jacoby and colleagues had interpreted as evidence for a *fluency heuristic* (e.g., Jacoby & Whitehouse, 1989). The present model has the machinery to calculate familiarity for a probe at a given moment in time, so it would be reasonable to extend the model dynamically, in the way suggested by Cox and Shiffrin.

Stages of Processing

Detailed models of perceptual processing almost always assume stages of processing. Such an approach is not in debate, though there remain many questions about the roles of top-down feedback processes. Just one of numerous examples showing the importance of stages (albeit in this case minimizing the role of top-down processes) is research by Pelli and colleagues on visual word perception. For example, Pelli et al. (2003) showed that word length does not improve word perception at threshold to the degree one would expect from an ideal observer model, by which threshold should drop strongly with word length. Instead, the threshold for words of differing lengths seemed determined largely by the threshold for individual letters, leading to a bottom-up stage model in which letter perception is a prior stage to word perception.

The need for stages of processing is clear, and these should be incorporated in SARKAE as the theory is further developed. We explicitly modeled pseudo-lexical decision as a dynamic process, and in the previous section discussed modeling event recognition dynamically, but the current forms of these dynamic models at most incorporate stages of processing in implicit fashion. Consider visually presented words. Notwithstanding top down feedback, it is very likely that processing begins with primitive features such as line edges, shape, and color; at a later point in time, processing will occur for higher order features such as letters; at that time or even later, the lexicon will be contacted and meaning features retrieved, and so on. These stages likely overlap, and likely have recurrence. The timing of these stages could well vary as well.

The picture is further complicated in the SARKAE framework because the stages in accessing knowledge produce features that are added (over time) to any probe of event memory, a process that then evolves over time in accord with the changing probe. The data from the present studies do not require a model with stages of processing, especially because the recognition testing was too minimal to produce reliable response times, so such modeling was not included in the simulation. The augmentation of SARKAE with stages of processing is left for future development.

Factors That Determine Recognition Performance

We have discussed several factors that affect (harm) recognition performance, including activation of non-target traces from the study list (termed *item noise* by Dennis & Humphreys, 2001) and activation of traces of the test item itself that were stored on the basis of study events other than the list presentation (termed *context noise* by Dennis & Humphreys, 2001). The distinction between context noise and item noise is actually not a sharp one in the SARKAE framework because the event trace of an item contains features of other items that are co-rehearsed. Thus, according to a context noise model, the activation of an item's list trace will act partially as if the co-rehearsed item had been activated. The distinction then becomes a subtle one in which the critical issue is whether a tested item activates traces of other items that were never co-rehearsed. In any event, SARKAE assumes that activation of all types of traces is a matter of degree depending on similarity to the retrieval probe.

Another important factor is the successive storage of the event traces of the successive items tested (given that essentially all studies use multiple tests): The context for the n-th test following a given study list is surely very similar to that for the traces of the prior $n - 1$ tests, leading to high likelihood ratios for those test traces. Since context likely changes between list study and the subsequent test sequence, the traces of list items will contribute smaller likelihood ratios. This would have two implications: Recognition performance should drop as testing proceeds following study of a given list, and the size of this effect should be larger than the effect of list length. Several recent studies demonstrate these effects and assertions (see Criss et al., 2011; Malmberg et al., 2012). We have not included this factor in our present simulations largely because the studies produced too little data to analyze the effects of test position. Had test traces been included in the simulations, predicted performance would have been lower, so our listed parameter estimates are surely a little too low. In addition, if storage of test traces had been included in the simulation, then that would have added another factor lowering predicted performance after a 2-week delay in Experiment 2. However, the qualitative pattern of the present predictions should be unchanged if test trace storage had been included.

In SARKAE, the way that extra and larger trace activations operate to harm performance is a technically tricky matter. At first glance, it appears that lowered performance is caused by the increased variance produced by extra and stronger traces. However, the recognition decision is based on the average likelihood ratio, and the average requires division by the number of activated traces. These factors together produce effects that turn out to be controlled largely by the skewing of the distribution of the average likelihood ratio. Shiffrin and Steyvers (1997) discussed how this works (albeit in a different version of the model, and one simpler in many ways). This picture is further complicated when one wants to understand the predicted effects of training frequency. The effects of training frequency depend on (a) number of training session event traces, (b) knowledge trace access whose accuracy depends on trace richness (and therefore affects both storage during list study and retrieval at test), and (c) changes in similarity of knowledge traces to each other that are induced by the visual search paradigm in Experiment 1. The interaction of these various factors makes it hard to gain an intuitive understanding of the basis for the model's predictions. After writing several long and convoluted verbal explanations, none of which clarified matters very much, we have decided to omit these and let the simulated predictions speak for themselves. It suffices to note that activation of traces other than those due to list study of the test item, whether those traces are from the list or from personal history and experience, and whether those traces are of other items (similar in content and context) or traces of the test item itself (in other but similar contexts), will reduce recognition performance. The harm caused by such activations is, however, largely caused by the few very similar traces (i.e., the tail of the similarity distribution) rather the average familiarity of the activated traces.

High Level Distinctive Features

The process by which one or more unique features are identified and learned are an interesting subject of research. For example, Shiffrin and Lightfoot (1997) showed that training gradually causes the separate features of a novel object to cohere into a unified whole, so that search that starts by dealing with several features of a given stimulus changes to the point where a given character can be dealt with as a single feature. The end result

would be a whole character code that could serve as a unique distinguishing feature. That study used simple three-feature stimuli but search was difficult because a conjunction of features was needed to identify a target. The difficulty was high enough that serial terminating search continued to be used throughout 30 sessions of training. The training did, however, produce considerable learning: search at the outset of training operated by joint operation of two sequential comparisons—sequential comparisons of features within each object (to determine whether a given object is a target) and sequential comparisons across the display objects. As training proceeded, the features defining a given object gradually coalesced into a single whole-character code that uniquely identified each object. Search therefore gradually switched (over many sessions) from sequential comparisons of features to sequential comparisons of whole objects. Because the time needed for a feature comparison was similar to that needed for an object comparison, search speed increased by a factor of three as training continued.

For the case of visual search for quite complex Chinese characters, the process of finding and learning unique features would likely be more complex, but not necessarily difficult, because some feature of each character already coded by our visual system might well be a unique identifier. For this reason, and because the learning of unique character codes is not the goal of the present investigations, we simply assumed that such a high level feature is present from the start of training. This simplifying assumption is unlikely to distort the modeling effort because training went on for a few weeks, and the distinguishing features were very likely learned early in training. The main function served by such a high level feature is its use in the two-alternative forced-choice (2AFC) perception task (we have described earlier why this feature alone governs choices). Unique identifying features would not be required for the matching tasks in Experiment 2, though there is no reason why they could not be noticed, learned and utilized. Given that there is no direct evidence from Experiment 2 concerning this possibility (partly because there were no 2AFC data from that study), we simply carried over the Experiment 1 assumption and assumed that a unique feature was available for the outset of training.

In the general framework, we intend "feature" to include references to other traces in the knowledge base: A knowledge trace develops based on a set of events, but then can itself be a feature stored in other event and knowledge traces.

Modeling Recall

Recall tasks are a major class of memory paradigms that we did not explore in our studies and hence did not model with SARKAE. The model we have in mind for recall essentially adopts that of the SAM model (Raaijmakers & Shiffrin, 1980, 1981). The general idea is that recall uses a probe that does not "contain" the answer being sought. Such a probe and subsequent changes to that probe are used in what can be and often is an extended search of memory. Each step of the search consists of sampling a trace, recovering the contents to the degree possible, assessing the validity of the information recovered (akin to the "recognize" part of the "generate and recognize" heuristic) and outputting the desired response, continuing the search, or ending the search unsuccessfully. The trace sampling process begins with trace activation as in SARKAE, likelihood ratios being the result for all activated traces. Sampling from these traces is made in proportion to the strength of the likelihood ratio for a given trace. Recovery of information from a trace is based in part on the value of the likelihood ratio for that trace (see the next paragraph). A decision whether the trace contains the information sought is based on the recovered information. Decisions whether to continue sampling and whether to change the probe cue are based both on the information recovered and the history of the search to that moment.

Performance in both cued recall and free recall is generally higher for higher frequency words, especially for knowledge tasks, but also for event memory. Cued recall studies generally show performance is higher for targets of higher frequency (probably due to the higher probability of "recovery"; e.g., Gillund & Shiffrin, 1984), but the effects of cue frequency vary from study to study. The factors in SARKAE that are frequency dependent must also operate in event recall. Any factors that increase the likelihood ratio between the probe and a given trace will affect recall in two primary ways. First, an increase in the ratio of the likelihood ratio for a given trace to the sum of likelihood ratios across all activated traces will increase the probability of sampling that trace. Second, recovery of information from the sampled trace will be higher for a trace with a higher likelihood ratio. The recovery process is frequency dependent for two reasons. A first stage of recovery will depend on extraction of features from the sampled trace, and more features will generally be extracted when the likelihood ratio is higher for that trace. However, the extracted features by themselves will typically be few in number, and error prone, and therefore insufficient to govern decisions and produce responses. Thus, a second stage of recovery will compare the sampled features to knowledge, in the hope of inferring what items are encoded in the trace. The process of comparing the trace features to knowledge traces will produce strong frequency effects in the recovery process (for the same reasons such frequency effects are expected in knowledge retrieval, such as lexical decision).

Recollection and Familiarity

SARKAE posits that a probe of memory has dual consequences: The cue itself engenders a feeling of familiarity (in the form of an average trace likelihood ratio), and the cue starts a process of sampling traces from memory (in proportion to the trace likelihood ratios), thereby allowing recall to occur. Thus, it is a logical and conceptual necessity in the SARKAE framework that recognition testing would involve both familiarity and recall. Demonstrating that this is the case is however no easy matter. In the thousands of studies that collect only old/new judgments, and the lesser number that (also) collect confidence judgments, the data are relatively sparse, making definitive conclusions difficult. Wixted and Mickes (2010) have a recent review of this research and conclude that a continuous version of the unequal variance signal detection model can account well for both familiarity and recollection. The adequacy of this model may reflect the relative paucity of accuracy data.

SARKAE requires both familiarity and recall processes to operate in recognition, but implicit in its architecture is a complex inter-relation of these two components. Not only SARKAE but almost any plausible model would have to predict a positive correlation between recognition and recall processes, because any

factor that makes a trace stronger and more accurate would lead to improvement of both familiarity-based and recall-based performance. The complexities arise in the details. If the task is such that decisions based on familiarity and recall are highly correlated when recall occurs, then a strategy might be adopted to ignore recall and base all decisions on familiarity. Many recognition models adopt this position either explicitly or implicitly. However, it might well be the case that decisions are based on the joint outcome of both processes. Some recognition tasks might even make recall the preferred basis for decisions, if foils are so similar to targets that targets and foils produce similar distributions of familiarity (e.g., when plural and single word forms must be discriminated—see Malmberg, Holden, & Shiffrin, 2004). If familiarity and recall are used jointly, then one needs to consider the different information that is used in the two cases. Event recall is characterized as a series of samples from memory. Recall sampling is characterized as proportional sampling from the likelihood ratios whose average is "familiarity." Suppose a recognition test results in some information being recovered from a sampled trace. There is not at present data that would allow a determination of the way that such recalled information is combined with familiarity. It might be simplest to assume that a positive decision based on recovered information from a recall sample dominates familiarity when the two differ. However, cases when the two sources differ might be low in number, making the various possible models indistinguishable. A number of investigators pursuing this issue have asked for two recognition judgments, one often termed *know* (presumably aligned with familiarity) and the other termed *remember* (presumably aligned with recall). Unfortunately, it turns out to be the case that most such data are consistent with the view that *remember* judgments are simply familiarity judgments with a higher criterion.

It may be that more definitive inferences can be achieved with studies analyzing other forms of data such as response times and measurements of neural activity. However, simply obtaining such measures will not necessarily provide easy answers. Consider response times: One might think that recall response times would sometimes be longer than those produced by familiarity judgments, due to extended sampling. However, such an inference requires strong knowledge of the shape of the slow tail of the distribution of familiarity judgments, and this knowledge would likely be model dependent. In addition, typical recognition tasks might well cause a curtailing of the recall process to just a single sample; if so, the comparisons of the response time distributions produced by a familiarity process and a single cycle of the recall search process would surely be dependent on assumptions of the models used for each.

The inference difficulties are only increased when considering dynamic models for recognition judgments. For example, in our new dynamic model for recognition familiarity judgments, familiarity will change over time, but there is no evidence that tells us when during this time the sampling process begins and ends. Given that successful recovery will usually occur for "target traces" whose familiarity tends to rise over time, and given that recovery is posited to depend on the value of the likelihood ratio for the sampled trace, success in sampling will be higher if the (first) sample occurs relatively late in the dynamic process of familiarity change. This could make it reasonable to assume as a first approximation that the timing of recognition responses is determined by the familiarity decision process even if the decision is sometimes based on recall.

In summary, we believe and the SARKAE approach assumes that recognition testing will usually involve both familiarity and recall components, but very precise and clever studies combined with careful modeling will likely be needed to assess the relative contributions of each.

Context

Features from the general surround of an event are added to both event traces and knowledge. Such context features include general internal states, the environmental surround, and other nearby events (similar to the ideas posited in the temporal context model of Howard & Kahana, 2002). One consequence of incorporating context into event traces is an increasing complexity of the structure of knowledge; another is the tendency for knowledge traces formed from events nearby in time and space to become increasingly similar to each other, and thereby to reflect the co-occurrence statistics of the environment. Both context due to nearby events (often varied experimentally) and general context are important. 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).

The SARKAE simulation modeling of context and content storage during event study is very simple and bypasses many of the complexities that are associated with different storage strategies and different allocations of attention. A detailed treatment of these matters is outside the scope of this article, but we believe that storage of a given element of context will be important and affect performance to the degree that that element is (1) attended and (2) integrated with other aspects of the event. These elements should also be important in encoding context for a probe of memory. In many situations, a given element of environmental context, such as aspects of the room environment, will not be attended or coded explicitly, and storage will be incidental. Changing that context element between study and test would therefore produce small effects, and those observed might be characterized as changes of "bias." However, when instructions or task lead to a binding of context to content, then larger effects and changes in accuracy can be expected (Murnane & Phelps, 1993, 1994, 1995; Murnane, Phelps, & Malmberg, 1999).

One type of context that may be very important but not much studied heretofore is "task context." An example is found in recent research by Annis, Malmberg, Criss and Shiffrin (in press): Following study of a word list, successive tests of words for recognition produce increasing output interference (Criss et al., 2011), but interpolating tests of words for lexical decision produces little if any interference for event recognition (even though those same words are stored, as demonstrated in a separate and later event recognition test). The simplest interpretation holds that the task context is a major part of the attended and stored information during testing. Then a recognition probe for item presence on the study list would contain recognition task context that would be quite dissimilar from the lexical decision task context in the lexical decision test traces. The lexical decision test traces would then not be activated and would not interfere.

Incidental context nonetheless plays in role in storage and retrieval, even when not in the focus of attention. In earlier writings (using the REM modeling framework), we suggested and obtained evidence for the "one-shot-of-context" hypothesis. The basic idea is that the explicit focus of storage is on content (particularly semantic content in the case of words). In contrast, storage of generic list context occurs incidentally at the time of presentation and only lasts for a short period of time. Thus, extended study does not add to context storage but only adds to content storage. Evidence for this view was obtained in several studies (Malmberg & Shiffrin, 2005). The SARKAE modeling did not explicitly incorporate this hypothesis because the studies we carried out did not vary study time and did not explore the attendant issues. Nonetheless, the hypothesis is very important for understanding of certain types of dissociations (e.g., massed study time affecting event memory but not knowledge retrieval). Dissociations are the subject of the next section.

Dissociations

In the eyes of many theorists, dissociations provide a major part of the rationale for separate event traces and knowledge traces. In this context, dissociations refer to an experimental variable that produces one pattern of outcomes when event memory is tested, but a different pattern when knowledge is tested (e.g., Roediger, Weldon, & Challis, 1989). To take another example than study time, "depth" of processing of a word is often manipulated (say, judgments of pleasantness, "deep," vs. judgments of word length, "shallow"); deeper processing produces better event memory but does not produce better knowledge retrieval (e.g., naming times may be the same for deep and shallow processing). The REM model has been used previously to explain dissociations (e.g., Malmberg & Shiffrin, 2005; Shiffrin & Steyvers, 1997); SARKAE predicts dissociations in conceptually similar fashion although the details differ from the REM account. In SARKAE, study of an event adds feature values to a new event trace and also adds (some of the same) values to the relevant knowledge trace. The effects of this addition are quite different, however, because for event traces, each feature value occurs once (if at all) and in this sense all such values are equivalent. However, the knowledge trace already has numerous feature value counts, so adding one more can have a negligible effect: If we know that fire engines are red, study of a red fire engine may add one more count to a large number of counts in the fire engine knowledge trace, but this will not change the knowledge trace appreciatively. On the other hand, novel or relatively novel information added to knowledge can produce a noticeable effect. For example, suppose one sees a green fire truck, adding "green" for the first time to the "fire truck" knowledge trace. If one is later asked to name a fire truck presented in green, the match of the color green in the test probe to the added green feature in the knowledge trace would be expected to speed responding. This example is artificial; more commonly context will produce the new information added to knowledge. Context generally keeps changing, so a new encounter with an instance of an established knowledge trace (say, study of a word in a study list) will add new context to that knowledge trace. The novel context features added to the knowledge trace can produce additional matching when (some of) those same new context feature values are used in probing knowledge a short time later. The additional matching can produce significant priming. In other cases, when the addition to knowledge is only a negligible increment, then priming will not occur. These are cases in which dissociations can be expected. To return to our examples, consider the account of Malmberg and Shiffrin (2005). They provide evidence that context features are stored in the first second or two of study, in somewhat automatic fashion, but not thereafter (they term this the *one shot of context* hypothesis). Suppose, therefore, that both "deep" and "shallow" processing of a word produce equal context storage in the word's event trace and the word's knowledge trace. As a result, "depth" of processing will not produce a difference in knowledge access due to context. What depth of processing will do is cause more storage (in both event and knowledge traces) of semantic feature values, because the depth manipulation will change the amount of elaborative semantic processing that takes place. The extra semantic information due to deep processing will improve event memory. However, the extra semantic processing will add more semantic features to a huge number of those features that already exist in the knowledge trace, the effect being a negligible change in knowledge access.

The preceding discussion is based on the assumption that the test item has a corresponding knowledge trace. What will occur when there is no corresponding knowledge trace, or such a trace is in early stages of formation, is largely unexplored territory. Some of the key issues are discussed next.

Weak or Missing Knowledge Traces

There are many unknowns concerning test items that have no corresponding knowledge trace, or a corresponding knowledge trace that is in early stages of development (such a trace lies in descriptive limbo, not clearly best described as either event or knowledge). When a presented item has a missing or weak corresponding knowledge trace, then it is reasonable to assume that the item will be encoded in terms of (1) lower level features that are represented in knowledge (e.g., for a random geometric line shape: lines, junctions, edges), and (2) one (or more) knowledge trace(s) that are retrieved based on similarity to the presented item. Thus, a visual "blob" might be novel and have no corresponding knowledge trace but be encoded in terms of textures, and in terms of other knowledge traces such as those for "cloud" or "ink stain"; a pseudoword like "event" might be encoded in terms of orthography and phonology, and also cause retrieval of the phonetically similar lexical trace "event." For event trace access in such cases, the features that result from the access to knowledge, whichever traces are the source of those features, will join the evolving probe of event traces. Beyond these reasonable assumptions, there is at present little empirical or theoretical basis for modeling in greater detail. In access to knowledge, how strongly does a weakly developed corresponding knowledge trace compete with similar but different knowledge traces? Should this process of knowledge access be modeled as the first sample of our general model for recall? Suppose the task involves activation of event traces: Should the incipient knowledge trace be treated as just another event trace, albeit one with more filled in values, and perhaps one with multiple counts per value? What sorts of priming and dissociations could be expected in such cases? These sorts of questions are good ones for future research. At the moment, however, it would really be a matter of guesswork for us to attempt to formulate a model for

items with missing knowledge traces (this was a good part of our decision not to model episodic recognition of untrained characters).

Forgetting From Long-Term Memory

In the following discussion, complexities due to possible differences in retention due to storage in the medial temporal lobes versus cortex are ignored—this issue will be raised briefly in the following section.

As a general rule, retrieval is better to the degree that there is a better match between the retrieval probe cue(s) and the trace(s) sought in memory. This principle applies both to event traces and knowledge traces. It is an old concept that was termed *encoding specificity* and was highlighted by Tulving and Thompson (1973). It has been a key principle in most theories of memory, including SARKAE and its predecessors such as SAM and REM. Forgetting is then expected to the degree that there is a reduction of the match of probe cue to stored trace. Thus, for example, given the expectation that context generally changes over time, and given that present context will either explicitly or implicitly be part of both the event trace and the probe, the match of probe to trace determined in part by context feature matching will decrease over time and produce forgetting.

In SARKAE (and other similar models), this general principal must be refined because the absolute value of the match to the desired trace is only one factor governing retrieval, another being this strength relative to the sum of the matches to other traces. For recognition, for example, the decision statistic is based on familiarity defined as an average likelihood ratio; the average is determined in part by the match of the probe to other traces. The more other traces are similar, the more likely the test item will be recognized, and conversely. This factor is important in predicting and explaining "false memory" effects (e.g., Brainerd & Reyna, 2002; Deese, 1959; Roediger & McDermott, 1995) and interference effects (Mensink & Raaijmakers, 1988). Turning to recall, note that one stage in the search process is the sampling of a trace in proportion to its likelihood ratio, again a relative measure (one closely related to that critical for recognition). However, once a trace is sampled, then recovery will depend mostly on the match (likelihood ratio) of the probe to the sampled trace.

These principles are fairly commonly used to explain forgetting, although some other factors could also play a role, such as trace degradation or retrieval inhibition (RI; e.g., M. C. Anderson, Bjork, & Bjork, 1994). RI is said to occur when there are several traces competing for retrieval to a probe; when one "wins" the others are inhibited. There has been much debate concerning RI, with some researchers claiming much of forgetting is due to its action, and others claiming the findings supporting RI can be explained by the factors discussed earlier in this section (e.g., Raaijmakers & Jacob, 2013). We have not included RI in our simulation and suspect it plays at most a small role in forgetting, but note that it could be incorporated in the general SARKAE theory without much distortion.

The discussion in the last few paragraphs considers forgetting as a unitary phenomenon, not distinguishing the type of materials or features retrievable at different time points. However, there is good reason to think that certain high level information is retained better than more superficial types of information. This idea has been formalized most carefully by Brainerd and Reyna (e.g., Brainerd et al., 1999; and used to explain developmental differences in false memory—e.g., Brainerd & Reyna, 2007), who distinguished "gist" information from verbatim information and who argued for longer survival of "gist." These ideas also accord with everyday subjective experience—the day after a research talk, we remember many details, but as time passes, we gradually seem to lose these details and perhaps eventually retain only a sense of whether the talk was "good" or "bad."

Brainerd and Reyna have explored recognition judgments of verbatim versus gist matches, obtaining parameter estimates in their quantitative model suggesting longer retention of gist. What could explain selective retention of high level "gist"? There are a number of possibilities, and these might differ somewhat for recognition and recall. One factor likely involves probe encoding. Each time an item is presented the current context could help govern the choice of features for encoding, but the superficial features might vary more as time passes than the core "meaning" features. For example the word "apple" might be encoded with a mental image of a green apple if one had consumed a green apple a few hours previously; a few days later, the encoding is likely to be red. In both cases, the core meaning of apple may be the same. By itself this factor would simply produce forgetting, but use of a probe that emphasizes the higher level gist features would tend to produce better retrieval than a probe that emphasizes superficial features, due to better matching to the stored trace. Such a factor is easy to incorporate into SARKAE. In recall, the factor of recovery from a sampled trace comes into play: We have assumed recovery improves as a function of the overall match of the retrieval probe to the sampled event trace, but this is likely too simple a view. It seems likely that what is recovered from a sampled trace will depend on the internal structure of the information in the trace, and that superficial detail may be more recoverable when the probe context matches the stored context well, and less recoverable when the context shifts. On the other hand, gist information in the trace may be recovered well due to matches to similar information in the probe even when context has changed. Such ideas are not yet incorporated in SARKAE and must be left for future studies and implementation.

A third possibility is based on the idea that "gist" features are more important and stored more strongly than "verbatim" features. The present simulation allows such strength to be represented by a higher probability of storing one kind of feature than another. This approach would require an elaborated representation of event traces, since the simulation of SARKAE assumed at most one count per feature. Of course it makes perfect sense, and accords with the representation of knowledge traces, to allow features that are given more coding and attention to be stored with more than a single count. If so, then matching of probe to event traces could be simulated in the way that this is done for knowledge traces: One trace feature value is sampled per feature class, in proportion to the counts in that feature class. It is a very complex issue whether a mere increase in gist counts would lead to selective retention of gist. It could be that proportional sampling of counts would have to be replaced by sampling that selectively chooses stronger feature values as delay increases and trace likelihood ratio decreases. For recall tasks and settings, all the above factors could be playing a role, but in addition, recovery of features from a sampled trace might selectively favor stronger feature values when the trace

likelihood ratio drops. The usual way to arrange this is through a non-linearity. For example, the lower a trace's overall likelihood ratio, the higher could be the feature-value-count threshold for accurate recovery of that feature value. These various possibilities have to be left for future exploration.

Neural Implementations of SARKAE

The role of the hippocampus and the medial temporal lobes (MTL) in encoding and establishing memories has been a topic of immense research interest, becoming especially prominent with Brenda Milner's research in the middle of the last century (e.g., Milner, 1970). It is clear that this region need not be intact to allow normal cognitive functioning in the present, but must be working properly to allow encoding of event configurations, and retrieval of those events once attention has been diverted for a few minutes. This superficial summary obscures the many issues that are continuing targets of research, such as the nature of events for which the MTL is uniquely required (as opposed to direct cortical storage), the time course and nature of transfer to cortical storage, the lability of MTL encoding, the degree to which MTL traces and cortical traces can be independent, and much more. This article is not the place for discussions of the many issues, but there are a few points worth raising.

One involves a puzzle that is essentially unanswered in all memory models including SARKAE: If retrieval is cue dependent, why do not the knowledge traces that strongly match the probe dominate retrieval, and prevent retrieval of the much weaker event traces? In some earlier writings, the second author speculated that the knowledge trace activation that occurs quickly might be temporarily inhibited, thereby allowing the weaker event traces to be accessed. A different idea is worth considering, that recent events stored in the MTL regions can be accessed to some degree independently of traces in the other brain regions (even when some start has been made at transferring those events to cortical traces). In such a view, knowledge access is largely from cortical regions and episodic access would be from MTL. This is of course too simple an idea. Given an intact MTL, event or episodic traces will gradually lead to the formation of a corresponding event trace in cortex, and event trace access might be some combination of retrieval from both. One might also be led to form a distinction between relatively short term event memories stored in MTL and much longer term event memories stored in cortical regions. This distinction is not often highlighted, because laboratory studies of event memories tend to focus almost exclusively on quite recent events whose recall might well be subserved by MTL ("what was on the just studied list?") but not events in the distant past ("what did you do on the day after your twelfth birthday?").

It is quite common to describe memory in terms of two "systems." Let us put aside the obvious proviso that our memory is extraordinarily complex and any binary division would be a gross oversimplification. A division of memory into a small number of systems can nonetheless be helpful as an aid to understanding and as a guide to ongoing and future research. One common division is described as implicit versus explicit (or other similar terms), and this binary division is at least partly motivated by the different memory functions for MTL versus other brain regions. Such a division is however unsurprisingly characterized differently in the hands of different theorists, making it difficult for us to relate the idea to our model. In addition, the division is often stated not in precise model terms, but in a fashion that leaves open many possibilities for interpretation. We have decided therefore to describe SARKAE as precisely as we can, and leave interpretation of the relation to various binary system frameworks in the hands of the theorists who either proposed or use them.

One interesting and very recent set of research findings on memory storage, and on the building of knowledge, is termed *reconsolidation* (e.g., Schiller et al., 2010). It seems clear that the formation of knowledge requires additions to existing memories—this concept lies at the core of the present theory. Further, it seems clear that additions to an existing memory require accessing, or calling to mind, that memory. This is termed *reminding* in reconsolidation research. The importance of the reconsolidation research lies in the idea that such reminding may place the recovered memory in a plastic state, and that the augmented memory would then require re-storage in order to be again available. Chemical intervention to prevent such re-storage seems to eliminate the memory (or its access). It seems likely to us that this finding and process, if further confirmed by ongoing research, would be limited to new memories still primarily stored in MTL.

There are many other issues that relate to the neural correlates of the SARKAE approach. 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, and Crommelinck (2002) utilized training of a novel set of objects (greebles). When faces were tested in upright or inverted fashion (inverted being "novel"), subjects showed an N170 effect: delayed and enhanced N170 for inverted versus upright faces. Prior to training, this inversion was not seen for greebles, but following 2 weeks of training, the N170 (at least in the left hemisphere) was delayed and enhanced for inverted versus upright greebles. Furthermore, James and Atwood (2009) 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.

The Nature of Events

Related to these discussions is the concept of "event." For simplified experimental tasks, such as the study and test of a list of words or characters, it is possible to think of events as separate and discrete sets of information corresponding to the separate task elements. This approach is obviously inadequate when experience is continuous and complex—for example, what are the "events" when we are carrying out a conversation while walking to school, or playing a tennis match, or reading a book? With few exceptions, the field has not explored the issue (see Zacks & Tversky, 2001). It seems likely and perhaps necessary that events exist at many levels of abstraction, overlap in complex and structured ways, and contain information with different temporal spatial extent. For example, events occurring while reading a book could include a word, a paragraph, a casual story element, a character, a character plot interaction, and the book itself, among many other possibilities. Surely the representation of events must involve the way our

cognitive system uses attention to parse experience, but the way in which this might operate goes well beyond the scope of this article. If one does adopt such an elaborated idea of "event," then an "association" probably ought to be elaborated similarly, perhaps as a hierarchy with an association event subsuming two item events. If the concept of event is generalized in such a way, then it would be natural to assume that knowledge traces could be structured similarly, given that knowledge traces form through event accumulation.

Separate Traces

The use of a representation in which traces are "separate" is a convenient heuristic and an aid to understanding. The last 30 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," given the extraordinary complexities of mind and brain, but each aids understanding in different ways. Some sorts of knowledge appear quite discrete, such as words in our lexicon, lending themselves more naturally to separate representations (indeed, many models using composite and distributed representations for word events nonetheless instantiate word knowledge with a lexicon of separate traces). Other forms of knowledge are far more continuous, such as "the actions of playing tennis," and may lend themselves better to composite and distributed representations. Our choice of separate traces is useful for behavioral modeling, but we certainly do not mean to argue for purely separate neural representations for events or for knowledge, something that is almost certainly impossible.

Structured Knowledge

Our studies used novel stimuli with no designed structure of their features. The knowledge traces that resulted from training had relatively little structure. According to SARKAE, the trial structure in Experiment 1 caused the similarity between knowledge traces to rise with their frequency, but that structure is highly impoverished compared to knowledge generally. Many neural net models are designed to produce sophisticated structure as their weights adjust over time to continued inputs. This is typically caused by the decisions about the sizes of the banks of nodes and their connectivity. For example, when feedforward connections are forced through a "choke point," the system will end up performing a kind of discriminative learning, separating groups of inputs into classes defined by similarity. There are far too many types of neural net models to try to characterize them as a group, but this sort of discriminative learning can and does produce sophisticated structure in the resultant sets of connection weights (a good example is found in the modeling of Rogers & McClelland, 2004).

In its simplest form, SARKAE is limited in its ability to form structured knowledge because it simply aggregates features of events. However, the present description of SARKAE has left out any serious discussion of attention and short-term memory. All storage of events and knowledge goes through active memory (variously called short term memory or working memory) where a variety of attention processes operate on the then contents of active memory. This article is not the place to describe these attention processes, but in the SARKAE framework, they are a major source of the eventual structure of knowledge. Discriminative learning will occur for example when attention is focused on the features in active memory that are most useful for separating classes of events important for the given task. We admit that this approach is just a promissory note at this point, because we have not yet attempted to apply the model to tasks in which highly structured knowledge forms. In addition, even if the approach is carried out, it is not clear that it will produce structured knowledge in as elegant a fashion as many neural net models.

Summary and Final Remarks

The research presented here is important on both conceptual and empirical grounds. Conceptually, it provides a traditional memory framework and model that shows how knowledge grows from events, and how knowledge informs the coding of events. Empirically, it provides some closure about the ways that frequency of occurrence affects storage and retrieval of event memory and knowledge: We demonstrated separate roles for the contexts in which items of different frequency appear, and frequency per se (represented as trace "richness"). Although few variables were manipulated, the studies are useful because it is unusual to combine in one study a variety of tasks: training (different types in the two studies), event memory, perception, and knowledge retrieval, thus coming close to spanning memory studies. Such an approach is more typical of applications of cognitive architectures (e.g., SOAR, e.g., Laird, 2012; ACT-R, e.g., Anderson, 1993; EPIC, e.g., Meyer & Kieras, 1997), but those architectures tend to 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.

SARKAE is obviously not the only way to implement the conceptual ideas presented here, but is a natural outgrowth of the second author's prior research, starting with Atkinson and Shiffrin's (1968) model, continuing through the SAM models of Raaijmakers and Shiffrin (1980, 1981) and Gillund and Shiffrin (1984) and continuing further 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).

According to SARKAE, the key to the storage of events and the development of knowledge is the assumption that an event causes two types of memory storage: An event is stored as an incomplete and error prone trace, including both content and context information. The event also causes addition of the same sorts of information (again in incomplete and error prone fashion) to an already stored trace that is brought to mind 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. When the trace brought to mind is an already rich knowledge trace, the addition of new event information makes only a

small change, but nevertheless is sufficient to produce long term priming.

The converse of storage in knowledge is coding by knowledge: Whenever an event is encountered, and whenever a probe of memory is formed or constructed, the knowledge base is consulted and relevant information retrieved: Most of the way an event is coded beyond infancy is based on learned features in the knowledge base.

In conclusion, the SARKAE model presented in this article provides one principled way of thinking about the co-evolution and interactive nature of human knowledge, event memory, and perceptual systems. This theory, and others of a similar character, might join other recent developments that focus research increasingly on the ways that cognitive, behavioral, and neural systems evolve together and interrelate 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,* 814–823. doi:10.1111/j.1467-9280.2006.01787.x

Anderson, J. R. (1983). A spreading activation theory of memory. *Journal of Verbal and Learning Behavior, 22,* 261–295. doi:10.1016/S0022-5371(83)90201-3

Anderson, J. R. (1993). *Rules of the mind*. Hillsdale, NJ: Erlbaum.

Anderson, M. C., Bjork, R. A., & Bjork, E. L. (1994). Remembering can cause forgetting: Retrieval dynamics in long-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 1063–1087. doi:10.1037/0278-7393.20.5.1063

Annis, J., Malmberg, K. J., Criss, A. H., & Shiffrin, R. M. (in press). Sources of interference in recognition testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

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). doi:10.1016/S0079-7421(08)60422-3

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. doi:10.1037/0096-1523.5.2.252

Brainerd, C. J., & Reyna, V. F. (1990). Gist is the grist: Fuzzy-trace theory and the new intuitionism. *Developmental Review, 10,* 3–47. doi:10.1016/0273-2297(90)90003-M

Brainerd, C. J., & Reyna, V. F. (2002). Fuzzy-trace theory and false memory. *Current Directions in Psychological Science, 11,* 164–169. doi:10.1111/1467-8721.00192

Brainerd, C. J., & Reyna, V. F. (2007). Explaining developmental reversals in false memory. *Psychological Science, 18,* 442–448. doi:10.1111/j.1467-9280.2007.01919.x

Brainerd, C. J., Reyna, V. F., & Mojardin, A. H. (1999). Conjoint recognition. *Psychological Review, 106,* 160–179. doi:10.1037/0033-295X.106.1.160

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. doi:10.1006/jmla.2001.2796

Bransford, J. D., & Franks, J. J. (1971). The abstraction of linguistic ideas. *Cognitive Psychology, 2,* 331–350. doi:10.1016/0010-0285(71)90019-3

Broadbent, D. E. (1967). Word frequency effect and response bias. *Psychological Review, 74,* 1–15. doi:10.1037/h0024206

Cox, G. E., & Shiffrin, R. M. (2011). Criterion setting and the dynamics of recognition memory. In L. Carlson, C. Hölscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 114–119). Austin, TX: Cognitive Science Society.

Cox, G. E., & Shiffrin, R. M. (2012). Criterion setting and the dynamics of recognition memory. *Topics in Cognitive Science, 4,* 135–150. doi:10.1111/j.1756-8765.2011.01177.x

Criss, A. H., Malmberg, K. J., & Shiffrin, R. M. (2011). Output interference in recognition memory testing. *Journal of Memory and Language, 64,* 316–326. doi:10.1016/j.jml.2011.02.003

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. doi:10.1037/0033-295X.111.3.800

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. doi:10.1037/0278-7393.30.4.778

Deese, J. (1959). On the prediction of occurrence of particular verbal intrusions in immediate recall. *Journal of Experimental Psychology, 58,* 17–22. doi:10.1037/h0046671

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. doi:10.1037/0033-295X.108.2.452

Estes, W. K. (1955). Statistical theory of spontaneous recovery and regression. *Psychological Review, 62,* 145–154. doi:10.1037/h0048509

Estes, W. K., & Maddox, W. T. (2002). On the processes underlying stimulus-familiarity effects in recognition of words and nonwords. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28,* 1003–1018. doi:10.1037/0278-7393.28.6.1003

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review, 91,* 1–67. doi:10.1037/0033-295X.91.1.1

Glanzer, M., & Adams, J. (1985). The mirror effect in recognition memory. *Memory & Cognition, 13,* 8–20. doi:10.3758/BF03198438

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. doi:10.1037/0278-7393.16.1.5

Glanzer, M., Adams, J. K., Iverson, G. J., & Kim, K. (1993). The regularities of recognition memory. *Psychological Review, 100,* 546–567. doi:10.1037/0033-295X.100.3.546

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. doi:10.1111/j.2044-8295.1975.tb01468.x

Gregg, V. H. (1976). Word frequency, recognition and recall. In J. Brown (Ed.), *Recall and recognition* (pp. 183–216). London, England: Wiley.

Hemmer, P., & Steyvers, M. (2009a). A Bayesian account of reconstructive memory. *Topics in Cognitive Science, 1,* 189–202. doi:10.1111/j.1756-8765.2008.01010.x

Hemmer, P., & Steyvers, M. (2009b). Integrating episodic memories and prior knowledge at multiple levels of abstraction. *Psychonomic Bulletin & Review, 16,* 80–87. doi:10.3758/PBR.16.1.80

Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology, 46,* 269–299. doi:10.1006/jmps.2001.1388

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. doi:10.1037/0033-295X.108.1.149

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language, 30,* 513–541. doi:10.1016/0749-596X(91)90025-F

Jacoby, L. L., & Dallas, M. (1981). On the relationship between autobiographical memory and perceptual learning. *Journal of Experimental Psychology: General, 110,* 306–340. doi:10.1037/0096-3445.110.3.306

Jacoby, L. L., & Whitehouse, K. (1989). An illusion of memory: False recognition influenced by unconscious perception. *Journal of Experimental Psychology: General, 118,* 126–135. doi:10.1037/0096-3445.118.2.126

James, K. H., & Atwood, T. P. (2009). The role of sensorimotor learning in the perception of letter-like forms: Tracking the causes of neural specialization for letters. *Cognitive Neuropsychology, 26,* 91–110. doi:10.1080/02643290802425914

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. doi:10.1016/j.jml.2006.07.003

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. doi:10.1016/S0022-5371(74)80031-9

Klein, K. A., Shiffrin, R. M., & 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* (pp. 171–190). New York, NY: Psychology Press.

Laird, J. E. (2012). *The Soar cognitive architecture*. Cambridge, MA: MIT Press.

Lieberman, H., & Pentland, A. (1982). Microcomputer-based estimation of psychophysical thresholds: The Best PEST. *Behavior Research Methods & Instrumentation, 14,* 21–25. doi:10.3758/BF03202110

Lloyd, M. E., Newcombe, N. S., & Doydum, A. (2009). Memory binding in early childhood: Evidence for a retrieval deficit. *Child Development, 80,* 1321–1328. doi:10.1111/j.1467-8624.2009.01353.x

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. doi:10.1037/0278-7393.23.3.539

Malmberg, K. J., Criss, A. H., Gangwani, T. H., & Shiffrin, R. M. (2012). Overcoming the negative consequences of interference that results from recognition memory testing. *Psychological Science, 23,* 115–119. doi:10.1177/0956797611430692

Malmberg, K. J., Holden, J. E., & Shiffrin, R. M. (2004). Modeling the effects of repetitions, similarity, and normative word frequency on old-new recognition and judgments of frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30,* 319–331. doi:10.1037/0278-7393.30.2.319

Malmberg, K. J., & Shiffrin, R. M. (2005). The "one-shot" hypothesis for context storage. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 322–336. doi:10.1037/0278-7393.31.2.322

Malmberg, K. J., Steyvers, M., Stephens, J. D., & Shiffrin, R. M. (2002). Feature frequency effects in recognition memory. *Memory & Cognition, 30,* 607–613. doi:10.3758/BF03194962

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. doi:10.1037/0033-295X.105.4.734-760

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology, 18,* 1–86. doi:10.1016/0010-0285(86)90015-0

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of the effect of context in perception: I. An account of basic findings. *Psychological Review, 88,* 375–407. doi:10.1037/0033-295X.88.5.375

Mensink, G.-J. M., & Raaijmakers, J. G. W. (1988). A model for interference and forgetting. *Psychological Review, 95,* 434–455. doi:10.1037/0033-295X.95.4.434

Meyer, D. E., & Kieras, D. E. (1997). A computational theory of executive cognitive processes and multiple-task performance: Part 1. Basic mechanisms. *Psychological Review, 104,* 3–65. doi:10.1037/0033-295X.104.1.3

Milner, B. (1970). Memory and the medial temporal regions of the brain. In K. H. Priobram & D. E. Broadbent (Eds.), *Biology of memory* (pp. 29–50). New York, NY: Academic Press.

Mueller, S. T., & Shiffrin, R. M. (2006, June). *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.

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. doi:10.1037/0033-295X.89.6.609

Murdock, B. B., & Anderson, R. E. (1975). Encoding, storage and retrieval of item information. In R. L. Solso (Ed.), *Theories in cognitive psychology: The Loyola Symposium* (pp. 145–194). Hillsdale, NJ: Erlbaum.

Murnane, K., & Phelps, M. P. (1993). A global activation approach to the effect of changes in environmental context on recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 882–894. doi:10.1037/0278-7393.19.4.882

Murnane, K., & Phelps, M. P. (1994). When does a different environmental context make a difference in recognition? A global activation model. *Memory & Cognition, 22,* 584–590. doi:10.3758/BF03198397

Murnane, K., & Phelps, M. P. (1995). Effects of changes in relative cue strength on context-dependent recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 158–172. doi:10.1037/0278-7393.21.1.158

Murnane, K., Phelps, M. P., & Malmberg, K. (1999). Context-dependent recognition memory: The ICE theory. *Journal of Experimental Psychology: General, 128,* 403–415. doi:10.1037/0096-3445.128.4.403

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, NJ: Erlbaum.

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

Pelli, D. G., Farell, B., & Moore, D. C. (2003, June 12). The remarkable inefficiency of word recognition. *Nature, 423,* 752–756. doi:10.1038/nature01516

Quillian, M. R. (1967). Word concepts: A theory and simulation of some basic semantic capabilities. *Behavioral Science, 12,* 410–430. doi:10.1002/bs.3830120511

Raaijmakers, J. G. W., & Jacob, E. (2013). Rethinking inhibition theory: On the problematic status of the inhibition theory for forgetting. *Journal of Memory and Language, 68,* 98–122. doi:10.1016/j.jml.2012.10.002

Raaijmakers, J. G. W., & Shiffrin, R. M. (1980). SAM: A theory of probabilistic search of associative memory. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 14, pp. 207–262). doi:10.1016/S0079-7421(08)60162-0

Raaijmakers, J. G. W., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review, 88,* 93–134. doi:10.1037/0033-295X.88.2.93

Ratcliff, R., & McKoon, G. (1997). A counter model for implicit priming in perceptual word identification. *Psychological Review, 104,* 319–343. doi:10.1037/0033-295X.104.2.319

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. doi:10.1037/0278-7393.28.1.138

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. doi:10.1037/0278-7393.26.2.294

Rice, G. A., & Robinson, D. O. (1975). The role of bigram frequency in the perception of words and nonwords. *Memory & Cognition, 3,* 513–518. doi:10.3758/BF03197523

Roediger, H. L., & Challis, B. H. (1992). Effects of exact repetition and conceptual repetition on free recall and primed word fragment completion. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 3–14. doi:10.1037/0278-7393.18.1.3

Roediger, H. L., & McDermott, K. B. (1993). Implicit memory in normal human subjects. In F. Boller & J. Grafman (Eds.), *Handbook of neuropsychology* (Vol. 8, pp. 63–131). Amsterdam, the Netherlands: Elsevier.

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words notpresented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 803–814. doi:10.1037/0278-7393.21.4.803

Roediger, H. L., Weldon, M. S., & Challis, B. H. (1989). Explaining dissociations between implicit and explicit measures of retention: A processing account. In H. L. Roediger & F. I. M. Craik (Eds.), *Varieties of memory and consciousness: Essays in honour of Endel Tulving* (pp. 3–41). Hillsdale, NJ: Erlbaum.

Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach*. Cambridge, MA: MIT Press.

Rossion, B., Gauthier, I., Goffaux, V., Tarr, M. J., & Crommelinck, M. (2002). Expertise training with novel objects leads to left-lateralized facelike electrophysiological responses. *Psychological Science, 13,* 250–257. doi:10.1111/1467-9280.00446

Rovee-Collier, C. (1997). Dissociations in infant memory: Rethinking the development of implicit and explicit memory. *Psychological Review, 104,* 467–498. doi:10.1037/0033-295X.104.3.467

Rubenstein, H., Garfield, L., & Millikan, J. A. (1970). Homographic entries in the internal lexicon. *Journal of Verbal Learning and Verbal Behavior, 9,* 487–494. doi:10.1016/S0022-5371(70)80091-3

Sanborn, A. N., Malmberg, K. J., & Shiffrin, R. M. (2004). High-level effects of masking on perceptual identification. *Vision Research, 44,* 1427–1436. doi:10.1016/j.visres.2004.01.004

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. doi:10.1037/0096-1523.3.1.1

Schiller, D., Monfils, M., Raio, C. M., Johnson, D., LeDoux, J. E., & Phelps, E. A. (2010, January 7). Preventing the return of fear in humans using reconsolidation update mechanisms. *Nature, 463,* 49–53. doi:10.1038/nature08637

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review, 84,* 1–66. doi:10.1037/0033-295X.84.1.1

Schooler, L. J., Shiffrin, R. M., & Raaijmakers, J. G. W. (2001). A Bayesian model for implicit effects in perceptual identification. *Psychological Review, 108,* 257–272. doi:10.1037/0033-295X.108.1.257

Shiffrin, R. M., & Lightfoot, N. (1997). Perceptual learning of alphanumeric-like characters. *The Psychology of Learning and Motivation, 36,* 45–81. doi:10.1016/S0079-7421(08)60281-9

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. *Psychological Review, 84,* 127–190. doi:10.1037/0033-295X.84.2.127

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM—Retrieving effectively from memory. *Psychonomic Bulletin & Review, 4,* 145–166. doi:10.3758/BF03209391

Shiffrin, R. M., & Steyvers, M. (1998). The effectiveness of retrieval from memory. In M. Oaksford & N. Chater (Eds.), *Rational models of cognition* (pp. 73–95). Oxford, England: Oxford University Press.

Sikström, S. (2001). The variance theory of the mirror effect in recognition memory. *Psychonomic Bulletin & Review, 8,* 408–438. doi:10.3758/BF03196178

Smith, S. M., Glenberg, A., & Bjork, R. A. (1978). Environmental context and human memory. *Memory & Cognition, 6,* 342–353. doi:10.3758/BF03197465

Tulving, E. (1972). Episodic and semantic memory. In E. Tulving & W. Donaldson (Eds.), *Organization of memory* (pp. 381–402). New York, NY: Academic Press.

Tulving, E. (1993). What is episodic memory? *Current Directions in Psychological Science, 2,* 67–70.

Tulving, E., & Thompson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review, 80,* 352–373. doi:10.1037/h0020071

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. doi:10.1016/j.cogpsych.2003.08.001

Wagenmakers, E.-J. M., Zeelenberg, R., Huber, D., Raaijmakers, J. G. W., Shiffrin, R. M., & Schooler, L. J. (2003). REMI and ROUSE: Quantitative models for long-term and short-term priming in perceptual identification. In J. Bowers & C. J. Marsolek (Eds.), *Rethinkig implicit memory* (pp. 105–123). Oxford, England: Oxford University Press.

Wagenmakers, E.-J. M., Zeelenberg, R., & Raaijmakers, J. G. W. (2000). Testing the counter model for perceptual identification: Effects of repetition priming and word frequency. *Psychonomic Bulletin & Review, 7,* 662–667. doi:10.3758/BF03213004

Wixted, J. T., & Mickes, L. (2010). A continuous dual-process model of remember/know judgments. *Psychological Review, 117,* 1025–1054. doi:10.1037/a0020874

Zacks, J. M., & Tversky, B. (2001). Event structure in perception and conception. *Psychological Bulletin, 127,* 3–21. doi:10.1037/0033-2909.127.1.3

(*Appendices follow*)

Appendix A

Analysis and Statistics

Experiment 1

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, $t(7) = -2.30$, $p = .055$.

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, $t(7) = 1.96$, $p = .097$.

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

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 of this analysis of group means showed a (non-significant) negative effect of frequency for response time ($\beta = -.002$, $r^2 = .054$, $p = .23$), and a significant positive effect of frequency on accuracy ($\beta = .004$, $r^2 = .147$, $p < .05$).

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, $t(5) = 0.51$, $p = .63$.

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, $t(6) = -0.387$, $p = .71$, 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 ($r^2 = .053$, $p = .24$).

(*Appendices follow*)

Six of the seven subjects were tested again following a 6-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, $t(5) = 0.605$, $p = .57$. Analysis of d' after delay found no significant effect in the contrast analysis, $t(5) = -0.989$, $p = .37$, 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 versus immediate test. This analysis showed that there was no significant difference in the magnitude of the hit rate effect, $t(5) = 0.30$, $p = .77$, 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, $t(13) = -0.27$, $p = .79$, but there was a significant change in the false alarm rate effect, $t(13) = -2.24$, $p = .04$.

Appendix B

Parameter Descriptions and Values

Tables B1 and B2 give the value and a description, respectively, of each parameter used in the SARKAE (Storing and Retrieving Knowledge and Events) simulations described in this article. Where the same value is used for the three conditions, that value was assumed not to vary. 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.

Table B1
Parameter Values

	Lexical decision			Episodic recognition			2AFC
Parameter	Experiment 1	Experiment 2	Experiment 2 with delay	Experiment 1	Experiment 2	Experiment 2 with delay	Experiment 1
s_p	.571	.667	.667	.571	.667	.667	.571
s_k	.143	.167	.167	.143	.167	.167	.143
s_t	.143	.167	.167	.143	.167	.167	.143
u_c	.1	.1	.1	.1	.1	.1	.1
u_k	.5	.5	.5	.5	.5	.5	
u_e	.75	.75	.75	.85	.75	.55	
c	.8	.8	.8	.8	.8	.8	
R	.30/.60	.25/.75	.25/.75	.6	.35	.6	
α	.0008	.0008	.0008	.1	.1	.1	100
timestep	.1	.1	.1				
MRT	.59	.78	.80				
nf	.60	.35	.35				
s_i				.7	.7	.7	
s_l				.25	.25	.25	
M_c				15	15	15	
N_c				20	20	20	
m							.0075
b							.70

Note. 2AFC = two-alternative forced-choice; *MRT* = mean residual time.

(*Appendices follow*)

Table B2
Parameter Descriptions

Parameter	Description
s_p	Probability of storing target item features into knowledge trace
s_k	Probability of storing existing knowledge features of item into knowledge trace
s_t	Probability of storing previous target item features into knowledge trace
u_c	Rate of storage for context features into knowledge trace
u_k	Rate of storage for non-context features into knowledge trace
u_e	Rate of storage for all features into event trace/percept
c	Probability of copying a feature value correctly from non-knowledge sources
R	Odds criterion for a response (R_o = old criterion; R_N = new criterion in lexical decision)
α	Parameter that governs copy probability when contacting a knowledge trace
timestep	Number of milliseconds assumed to pass with the completion of one evidence gathering/comparison time-step in the lexical decision model
MRT	Mean residual time—non-decision time in the response times for lexical decision
nf	Number of new features in novel (unstudied) pseudo-lexical decision test items
s_i	Probability of storing from target item during construction of event trace
s_l	Probability of storing from item's knowledge trace during construction of event trace
M_c	Number of context features of an extra-list trace that must match in order for the trace to be activated and included in the episodic memory decision calculations
N_c	Number of times context change process is run between training sessions
m	Slope of the linear function used to determine high-level feature extraction in two-alternative forced-choice
b	Intercept of the linear function used to determine high-level feature extraction in two-alternative forced-choice

Appendix C

Probability Estimation Through Simulation

At the core of the SARKAE (Storing and Retrieving Knowledge and Events) model is the equation giving the likelihood ratio for a match of probe to trace: Equation 1. In this equation, there are two ratios expressing the feature likelihood ratios for matches and mismatches. There are four terms needed to calculate these ratios: P(m|s), P(m|d), P(nm|s), and P(nm|d). In the retrieving effectively from memory (REM) model (Shiffrin & Steyvers, 1997), these ratios could be written in simple form in terms of the model's parameters. Unfortunately, the complex construction of the traces, percepts, and test probes used in SARKAE does not allow us to write a closed form expression for these ratios in terms of the model parameters. As an alternative, the ratios were estimated for a given set of parameter values through a simulation process: We used a given set of parameter values to produce memory traces and target and foil probes for a large number of simulation runs. From these, we counted average numbers of matches (m) and mismatches (nm) for the two relevant cases: when an episodic trace or a percept is being compared to its own trace (s), and when it is being compared to a different trace (d). The resultant proportions were then used as "empirical" estimates of the four probabilities in Equation 1. This process must be carried out for each set of parameter values. We verified the accuracy of this procedure by testing it on predictions for the REM model, because that model is simple enough that explicit expressions were available for the same terms. The two approaches matched closely, as shown for one set of REM parameter values in Figure C1.

(*Appendices follow*)

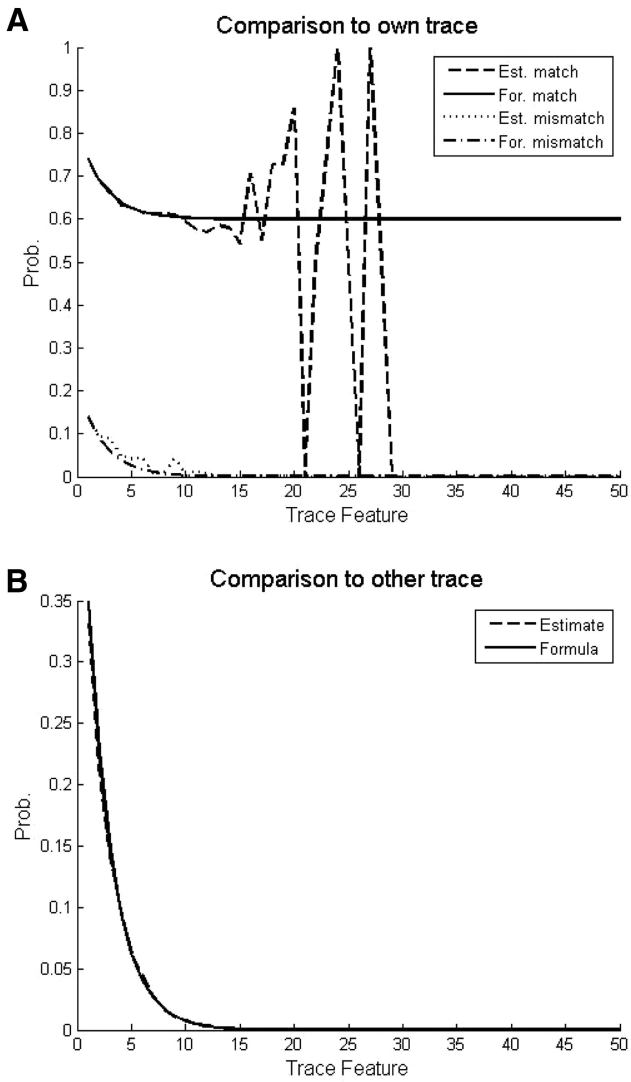

Figure C1. Retrieving effectively from memory (REM) model (Shiffrin & Steyvers, 1997): Estimated (Est.) probabilities (Prob.) compared to probabilities calculated by formula (For.) when an item is compared to its own trace (Panel A) or some other trace (Panel B). Note that the estimated probabilities in Panel A differ from the formula probabilities only for large feature values, due to the extremely low occurrence of these very low frequency values.

Appendix D

Calculation of Probe-Trace Matching

Cox and Shiffrin (2012) suggested a much improved method for calculating the similarity/match between a probe and memory trace, and we give here a slightly generalized variant of that method. For the special and limited conditions of the studies in this article, the better method was not needed, but a general theory needs to have a well-tuned and coherent calculation. The improved calculation takes into account the evidence against a match when a feature (a feature class, like "color," not a feature value, like "red") is present in probe or trace but not present in the other. In addition, the method allows appropriate similarity calculation regardless of the size of the trace and size of the probe (size referring to the number of features with values), an essential property

(*Appendices follow*)

because the probe grows over time and the trace size varies with factors like study time. The exposition is clearer if we assume the following equation applies to the matching of a probe to an event trace (the extension to knowledge trace is described later). This equation applies to the current status of a probe, either at a time t after the presentation of the test item, or as limited by perceptual properties of test presentation. To produce the given equation, we assume that there is at most one value per feature in the trace and in the probe.

In principle, each event or knowledge trace can have a different number of features. However, in a given situation the limited capacity of short-term memory and the limited study time will mean that only a sample of these features will be involved in storage of an event (or added to a developing knowledge trace). Similar limitations will mean that only a sample of potential features will be available for incorporating in a dynamically developing memory probe. These limiting factors are represented (implicitly) in the parameters as defined below. We assume that the parameters are the same for each trace that is matched to the probe, and for each probe.

Also, in principle, the number of feature values per feature, and the base rates of those feature values, should vary. In the present applications, we do not know the features or the feature values, so we assume for simplicity that there is the same number of values per feature, with equal base rates. It is not unreasonable to assume that the system has "knowledge" of the number of feature values and their base rates, for each given feature. If so, the equations given would be easy to modify accordingly, although the expressions would be rather long to write out.

$n(T)$ = number of features in a trace that have no corresponding feature in the probe.

$n(P)$ = number of features in the probe that have no corresponding feature in the trace.

$n(M)$ = number of matching values for a common feature in both probe and trace, summed across all features.

$n(Q)$ = number of mismatching values for a common feature in both probe and trace, summed across all features.

$K(s)$ = number of features that are in short-term memory during event study, and potentially available for storage. $K(s)$ is a parameter to be estimated, but we assume it is fixed across conditions and items. If a very long study occurs, and new features are added to short-term memory as time passes, then it is unlikely but possible that the number of feature values stored in an event trace will exceed $K(s)$. When this happens, a random sample of $K(s)$ features is selected and used for comparison.

$K(p)$ = number of features potentially available for encoding into the test probe. $K(p)$ is a parameter that is estimated but fixed across conditions and items. We usually set $K(p) = K(s)$.

α = probability that a given feature will be present in both the $K(s)$ and $K(p)$ features when the trace and test item match. This parameter will be high but less than 1.0 due to fluctuation in the process of selecting features to join short-term memory.

γ = probability that a given feature will be present in both the $K(s)$ and $K(p)$ features when the trace and test item do not match. This parameter represents feature overlap between different items. The longer the delay and the greater the context change between study and test, the higher the true degree of fluctuation should be when probe and trace match. When probe and trace mismatch, the greater the similarity between the test item and the trace, the greater should be the true degree of feature overlap. It is an open question whether the system or participant can adjust these values for different experimental conditions.

$J + 1$ = number of values per feature, for every feature, assumed to have equal base rates.

c = probability of copying a studied feature value correctly, given a value is stored, assumed to be the same value for all features and values.

u = probability that a given feature value had been stored in the trace being compared = $[n(T) + n(M) + n(Q)]/K(s)$.

v = probability that a given feature value had been encoded in the current probe = $[n(P) + n(M) + n(Q)]/K(p)$.

There are five free parameters: α, γ, c, K, J:

$$\lambda\left[\frac{Match}{Mismatch}|n(T), n(P), n(M), N(Q)\right] = \left[\frac{\alpha(1-c)}{\gamma}\right]^{n(Q)}\left[\frac{\alpha(1+Jc)}{\gamma}\right]^{n(M)}\left[\frac{1-\alpha u}{1-\gamma u}\right]^{n(P)}\left[\frac{1-\alpha v}{1-\gamma v}\right]^{n(T)}$$

The larger the first and last two exponents, the less is the evidence of matching; the larger the second exponent the more is the evidence for matching. The fewer the number of feature values stored in a trace, the smaller will be the value of u; small u moves the third term toward 1.0 reducing its effect upon the likelihood ratio. The fewer the number of feature values in the probe, the smaller will the value of v; small v moves the fourth term toward 1.0 reducing its effect upon the likelihood ratio. Thus, features present in the probe or trace and not present in the other will reduce the likelihood ratio more to the extent that both are rich in feature values.

This equation can also be used to calculate similarity/likelihood ratio for a knowledge trace, under the following assumptions. A knowledge trace will often have more than $K(s)$ features; when this is the case, assume a sample of $K(s)$ features is made in proportion to the summed number of values stored in that feature. In addition, a knowledge trace will often have many values per feature; when this is the case, choose a single value for that feature in proportion to the number of values stored.

Finally, it should be emphasized that this equation is the method for calculating the probe trace match (activation, likelihood ratio). It says almost nothing about the factors that determine which features and values are selected for short-term memory, for storage in the event trace and knowledge trace, and for incorporation into the developing probe of memory. That selection will be a mixture of automatic and attentive processes and will reflect factors such as context, strategies, perceptual factors, and goals.

Received October 5, 2011
Revision received December 19, 2012
Accepted January 4, 2013 ■