Journal of Experimental Psychology:
Learning, Memory, and Cognition
2000, Vol. 26, No. 2, 294–320

0278-7393/00/$5.00 DOI: 10.1037//0278-7393.26.2.294

A Mechanistic Account of the Mirror Effect for Word Frequency: A Computational Model of Remember–Know Judgments in a Continuous Recognition Paradigm

Lynne M. Reder, Adisack Nhouyvanisvong, Christian D. Schunn, Michael S. Ayers, Paige Angstadt, and Kazuo Hiraki
Carnegie Mellon University

A theoretical account of the mirror effect for word frequency and of dissociations in the pattern of responding Remember vs. Know (R vs. K) for low- and high-frequency words was tested both empirically and computationally by comparing predicted with observed data theory in 3 experiments. The SAC (Source of Activation Confusion) theory of memory makes the novel prediction of more K responses for high- than for low-frequency words, for both old and new items. Two experiments used a continuous presentation and judgment paradigm that presented words up to 10 times. The computer simulation closely modeled the pattern of results, fitting new Know and Remember patterns of responding at each level of experimental presentation and for both levels of word frequency for each participant. Experiment 3 required list discrimination after each R response (Group 1) or after an R or K response (Group 2). List accuracy was better following R responses. All experiments were modeled using the same parameter values.

Theories that explain how a person correctly identifies that an item was studied before (a recognition judgment) are not difficult to generate. Likewise, it is not theoretically challenging to explain how a person correctly rejects a lure that has not been studied. Of more theoretical interest is to explain, without making additional assumptions, why people incorrectly accept some not-presented items as studied (false alarms) and why they fail to recognize some items that were studied. Broadly construed, there are two classes of explanations for how people remember whether an item has been seen before. One class involves a measurement of familiarity on a unitary dimension, and the other class of explanations assumes two bases (e.g., recollection vs. familiarity) or systems (episodic vs. semantic) for making this type of determination. A challenge to either type of explanation is how to account for the word frequency effect, sometimes called the mirror effect.

The mirror effect (Glanzer & Adams, 1985, 1990; Glanzer, Adams, Iverson, & Kim, 1993) refers to the phenomenon that two distinct classes of items, such as high- and low-frequency words, produce opposite orderings in likelihood to respond "Old" in recognition tests, depending on whether the item had actually been studied. That is, the hit rate (correct recognition judgments for presented items) is higher for low-frequency words than high-frequency words, and the false alarm rate (spurious recognition judgments for items not studied) is higher for high-frequency words than low-frequency words. When these results are plotted as two functions, one for hits and one for false alarms, with frequency on the abscissa, they are mirror images, hence the name. One reason this effect has interested memory theorists is that, to the extent that psychology aspires to provide mechanistic explanations of phenomena, this pattern of data offers a clear set of constraints that any theoretical account must satisfy. Several other factors have been shown to produce mirror image performance on hits and false alarms (see Stretch & Wixted, 1998). In this article we focus on word frequency, but in the General Discussion we briefly discuss how SAC (Source of Activation Confusion) can account for other mirror effects.

Theoretical accounts of the mirror effect sometimes assume a unitary measure of familiarity for all words, but with different types of items having different distributions of preexperimental familiarity or pools of distinctive features (e.g., Glanzer et al., 1993). These different types of items are

Lynne M. Reder, Adisack Nhouyvanisvong, Christian D. Schunn, Michael S. Ayers, and Kazuo Hiraki, Department of Psychology, Carnegie Mellon University; Paige Angstadt, Department of Computer Science, Carnegie Mellon University.

Christian D. Schunn is now at the Department of Psychology, George Mason University; Michael S. Ayers is now at the Board of Education, Pittsburgh Public Schools; Paige Angstadt is now at Facset Research Systems, Greenwich, Connecticut; Kazuo Hiraki has returned to Electrotechnical Laboratory, Tsukuba, Japan.

Portions of this work were presented at the annual meeting of the Psychonomics Society, November 1996, and of the Cognitive Science Society, August 1997. Experiment 1 was reported as part of Adisack Nhouyvanisvong's master's thesis. Preparation of this article was supported by Grant 1R01 MH52808-01 from the National Institute of Mental Health, and in part by Grant N00014-95-1-0223 from the Office of Naval Research and by Air Force Office of Scientific Research Grant F49620-97-1-0054.

Correspondence concerning this article should be addressed to Lynne M. Reder, Department of Psychology, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213. Electronic mail may be sent to reder@cmu.edu.

differentially affected by the same experimental manipulations (e.g., Hilford, Glanzer, & Kim, 1997; Kim & Glanzer, 1993). Another approach that has been used to account for people's judgments about different types of items involves postulating multiple decision criteria (e.g., Hirshman, 1995).

Another challenge for memory theories is to account for the Remember–Know phenomena. Remember versus Know judgments refer to participants' classification of Old responses into those for which they can recollect a particular experience associated with the item (evoking the Remember response), and those for which the decision was based simply on a strong feeling of familiarity, thereby inferring that the item must have been seen recently (evoking the Know response). Interest in the Remember–Know paradigm (Tulving, 1985) has been especially strong since dissociations were reported for R versus K judgments with various manipulations. Some manipulations have produced effects on Remember responses but not on Know responses, whereas others have produced effects on Know responses leaving Remember responses unaffected, suggesting that Remember responses are a good measure of explicit memory and that Know responses are a good measure of implicit memory (e.g., Gardiner, 1988; Gardiner & Java, 1990; Gardiner & Parkin, 1990; Rajaram, 1993). One of the dissociations found with Remember–Know judgments involves the manipulation of word frequency. High-frequency words and low-frequency words yield different patterns of R and K responding (Gardiner & Java, 1990; Strack & Forster, 1995).

Recent work with the Remember–Know paradigm has been taken by some as support for the validity of separate memory systems or at least two qualitatively different processes (e.g., Rajaram, 1993). Others hold that these judgments merely reflect different levels of certainty on a continuous familiarity scale and that signal detection analyses support this view (e.g., Donaldson, 1996; Hirshman & Master, 1997).

There exist formal models for the mirror effect of word frequency (e.g., Hilford et al., 1997; Hintzman, 1994; Hirshman, 1995; Kim & Glanzer, 1993; Maddox & Estes, 1997; McClelland & Chappell, 1998; Shiffrin & Steyvers, 1997) and formal accounts for a dual-process signal detection model for recognition (e.g., Yonelinas, 1997). There also exist accounts for Remember–Know judgments (e.g., Gardiner & Java, 1990; Knowlton & Squire, 1995; Tulving, 1985), some of them formal (e.g., Donaldson, 1996; Hirshman & Master, 1997).

To date, despite all of the research and theories concerned with recognition memory effects for word frequency and Remember–Know judgments for word frequency, there have been no mechanistic accounts that simultaneously predict the recognition mirror effect and Remember–Know judgment pattern for words of high and low normative frequency. Most of the works cited above provide mathematical models that explain one or more types of mirror effects in memory (e.g., Kim & Glanzer, 1993; Shiffrin & Steyvers, 1997). Some theorists provide formal, quantitative accounts of the mirror effect for words of different classes (e.g., Hirshman & Arndt, 1997) and can separately provide a quantitative account for Remember–Know judgments (Hirshman & Henzler, 1998). Still others offer a mechanistic account for the mirror effect for word frequency and other mirror effects (e.g., McClelland & Chappell, 1998).

This article provides a mechanistic model of memory that both explains the word frequency mirror effect (why it occurs) and, using the same simple assumptions, accounts for the dissociation in Remember–Know judgments for low- and high-frequency words. The theory on which this model is based has been used to account for other memory phenomena and requires few additional assumptions in order to accommodate these results. The theory, named SAC, makes predictions that some would view as counterintuitive and that are at odds with claims in the literature of null effects. An important test of a theory is to compare precise predictions with data. A still stronger test of a theory is to make a prediction that runs counter to conventional wisdom and findings and to determine whether or not there remains support for the theory. Given that the theoretical predictions are clear and the contradictory findings do not seem definitive, it seems worthwhile to create another, stronger test that will either replicate and extend prior results or provide converging evidence for the SAC theory.

The goals of this article are several. One goal is to offer an account of the word frequency effect (WFE) without positing different decision criteria or differential shifts in familiarity for words of these two classes. Both preexperimental (or normative) frequency and experimental frequency are varied in order to examine the independent effects of both sources of familiarity on recognition judgments.

A second goal is to account for Remember and Know patterns of responding as a function of preexperimental and experimental manipulations of frequency without resorting to ad hoc assumptions. The theoretical position is motivated by our previous empirical and theoretical efforts (Reder & Schunn, 1996; Schunn, Reder, Nhouyvanisvong, Richards, & Stroffolino, 1997). This theoretical position shares some assumptions with other views in the literature but also makes predictions that differ from findings reported in the literature. In the three experiments we describe and in an informal meta-analysis of the literature, we find considerable support for predictions that contradict reported findings. The theory we propose is implemented in a computational model that is fit to individual participants' data on a trial-by-trial basis, making a large number of point predictions and using few degrees of freedom.

Overview of Model and Theory

An important premise of SAC that is shared by many others (e.g., Jacoby, 1991; Jacoby & Dallas, 1981; Mandler, 1980; Yonelinas, 1997) is that there are two ways to decide that a word has been seen before. One way is to note that the word seems familiar and to infer that it must have been studied because it seems so familiar. The second way is to actually retrieve the encoding event in which the word was studied, that is, use recollection. The SAC model of memory (e.g., Ayers & Reder, 1998; Reder et al., 1997; Reder & Schunn, 1996; Schunn et al., 1997) assumes one node

(concept) to represent the actual word and another node to represent the encoded memory event for that particular word. The word node has associated to it lexical information such as its phonemic and orthographic information, semantic information such as related concepts and its component features, and contextual information such as previous and current encoding events. The encoding-event node represents the knowledge that the word was encoded in the current experiment.[1] In the model, memory strength is represented by a node's level of activation; the greater the activation, the greater the strength of the memory representation. The base (or resting) level of activation of a node is determined by prior history of exposures; the more often seen (and the more recently seen), the higher the base-level activation. Likewise the strength of an association among concepts is a function of the frequency and recency of exposure. Formalizations of these relationships are given after Experiment 1. Figure 1 provides a schematic illustration of the memory representation that we assume for encoding words in a memory experiment, omitting for simplicity such aspects as componential features and semantic and lexical associations. In this figure, two words of high normative frequency are represented, namely, *grass* and *apple*, the former already presented in the experimental context (Figure 2 illustrates how apple increases in experimental strength). Also represented are two low-frequency words, *stoic* and *caveat*. The higher resting level of activation for words of higher normative frequency is denoted with thicker ovals for their conceptual (or word) nodes.

Using this dual-process model of word recognition, the question becomes what affects the familiarity-based judgment and what affects the judgment that is based on retrieval of the encoding event. In our view, the familiarity of the word–concept node is affected by whether or not the word has been recently seen and how frequently it has been seen. This means that preexperimental (normative) word frequency affects familiarity as would recent exposure to the word. Because familiarity can arise from causes other than an exposure during the experiment, an accurate recognition judgment is based on the retrieval of the study-event node (i.e., a true recollection). In other words, responses based on the word node (i.e., familiarity-based responses) are error prone. This view can explain why there are more false alarms for high-frequency words than low-frequency words: High-frequency words are more familiar (they have more prior exposures) and hence more likely to seem old when a response is made on the word node.

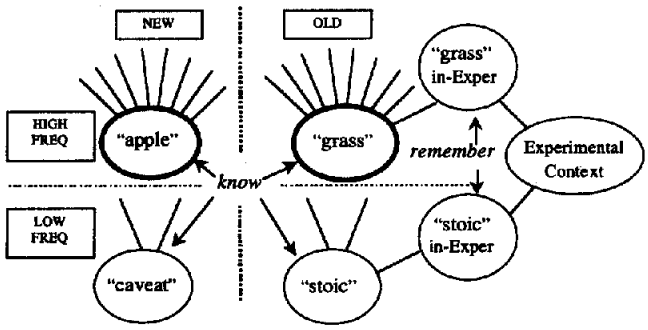

Figure 1. Schematic illustration of memory representation of words of different frequencies tested in a recognition memory experiment. FREQ = frequency; Exper = experiment.

The mirror effect for recognition of words of different frequencies refers to high-frequency words producing not only more false recognitions but also fewer correct recognitions than low-frequency words. Our explanation for more correct recognitions for low-frequency words involves the ease of retrieving the encoding event. This retrieval depends on the amount of activation that spreads from the word node when it is activated during the judgment. A high-frequency word has more contexts associated with it, so more concepts will share the activation that spreads from the word node.[2] The larger number of prior contextual associations for the high-frequency words is denoted in Figure 1 by a greater number of links emanating from the concept node. More associated contexts, called a greater *fan* (e.g., Anderson, 1974; Reder & Anderson, 1980) emanating from these nodes, means that it is more difficult to get sufficient activation to any particular associated context and therefore more difficult to access the relevant event node. (Formal specifications of the model are presented with the discussion of the computer simulation of the experiment after Experiment 1.)

The advent of the Remember–Know paradigm enables us to make strong tests of our theory. According to SAC, Remember judgments, assuming that they really derive from a recollection, should be based on activation of the encoding-event node and not the word node. In contrast, Know responses should be based on activation of the word node rather than the event node (consult Figure 1). This means that not only can we try to account for the mirror effect for word frequency, also we can try to account for the pattern of Remember and Know responses that vary with word frequency. SAC predicts more Remember responses for low-frequency words because low-frequency words have less fan, that is, fewer prior contextual associations competing with the current contextual association. With fewer competing associations, more activation reaches the encoding-event node, making the low-frequency word's encoding event more accessible and more likely to elicit a Remember response. This prediction has already been confirmed in the literature.

SAC also predicts more Know responses for high-frequency words for two reasons. First, for words that were actually presented (the Old words), high-frequency words are less likely to elicit a Remember response for the reason

[1] This idea bears similarity to other proposals (e.g., Anderson & Bower, 1973), but those models did not examine phenomena such as how recognition changes with multiple presentations of an item, how activation is converted into recollection versus familiarity judgments, and so on.

[2] The greater number of contextual associations should be distinguished from Glanzer and Bowles's (1976) postulation of a greater number of meanings for high-frequency words. Although there may be more meanings associated with high-frequency words, that is not part of our theoretical account. The greater fan that we refer to is *contextual* fan.

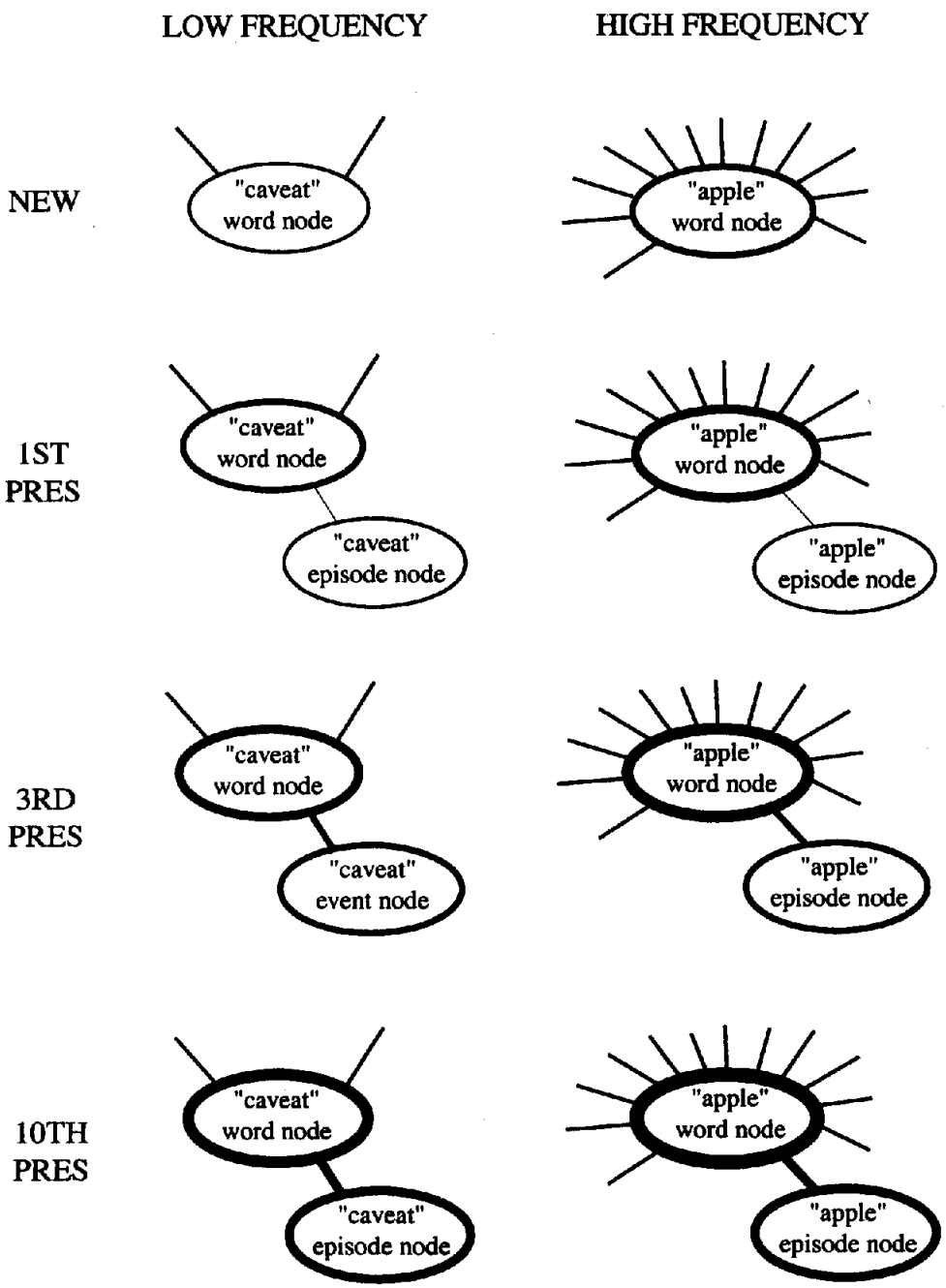

Figure 2. Experiments 1 and 2. Schematic illustration of memory representation for low- and high-frequency words over progressive numbers of presentations (PRES). Know responses are based on activation level of word nodes. Remember responses are based on activation level of event nodes.

given previously. Given that it is more difficult for high-frequency words to elicit a Remember response, those items should instead produce more Know responses.[3] That is, if a person is going to respond "Old" and chooses not to say "Remember," the only response left is "Know." Second, SAC predicts more Know responses for high-frequency foils (words that do not have an episodic trace of the experimental presentation) because the base-level activation of the word node is higher for high-frequency words than low-frequency words. The higher base-level activation of the high-frequency word node means that they are more likely to pass

[3] This, of course, assumes that the total percentage of Old responses is equal for low- and high-frequency words, which is not the case—there are fewer hits for high-frequency words than low-frequency words; however, we claim that the lower rate of Old responding for high- as compared with low-frequency words is due to the lower rate of recollection, that is, Remember responding.

over threshold, giving more erroneous Know responses to high-frequency foils.[4]

The theoretical prediction of WFEs for Know judgments seemed intuitive to us, and thus we were surprised to find that previous experimenters did not report such differences in Know responses for different levels of word frequency for either Old words or New words (e.g., Gardiner & Java, 1990; Strack & Forster, 1995). For example, Gardiner and Java (1990) wrote "word frequency influenced R responses but had no discernible effect upon K responses [and thus] the word frequency effect ... occurs only for recognition accompanied by recollective experience [and is not] attributable to increased familiarity" (p. 25). Given that this familiarity-based effect is a key prediction of the SAC model, we wanted to conduct a meta-analysis to determine whether the result may have been present in previous studies but too small to be detected in any single experiment.

We found five published articles that examined the mirror effect for word frequency and collected Remember–Know judgments. Some of those studies contained multiple experiments. Table 1 summarizes the results of those experiments and another from a talk given by Gardiner (1998). Gardiner, Richardson-Klavehn, and Ramponi (1997) asked participants to distinguish "Know" responses from "guesses." We combine those responses into "Know" responses because the authors suggested that under standard instructional conditions, guesses are included in Know judgments. A Wilcoxon matched-pairs signed-ranks test revealed a significant difference between Know responses to high- and low-frequency words, $p < .01$. Another Wilcoxon matched-pairs signed-ranks test, even excluding the "guesses" from Gardiner et al. (1997), also revealed a reliable difference, $p < .05$. Note that the above nonparametric tests did not include the data from Huron et al. (1995) because those authors did not report the response proportions for false alarms. A formal meta-analysis was not possible because in order to calculate a measure of effect size such as Cohen's *d* (1992), both the actual numbers in each condition and estimates of error in measurement of those numbers are needed. Few of the experiments in Table 1 provided both of these pieces of information.

Given the trends present in the literature and the significant Wilcoxon signed-ranks tests for more Know responses for high-frequency words, we thought it might be possible to find a stronger effect with a stronger manipulation. It seemed important to establish whether the *claims* of the literature or the model would be supported. We decided to test our prediction in several ways. In addition to a stronger manipulation of the standard study-test paradigm (Experiment 3), we also wanted a stronger test of our theory. Instead of simply comparing WFEs on Remember–Know judgments, we crossed preexperimental word frequency with experimental word frequency by varying orthogonally the number of exposures to the words in the experiment with preexperimental word frequency.

The top row of Figure 2 illustrates a situation in which the high-frequency word "apple" begins with more associated contexts (demonstrated by having more lines emanating from the word node) than the low-frequency word "caveat." Because neither word has yet been seen in the experiment, none of the links are associated to an event node for the current experiment. The word node for "apple" is thicker than the word node for "caveat," denoting a higher base level of activation. According to SAC, the greater base-level activation and greater fan of high-frequency words results in worse recognition memory performance—the former causing more false alarms, and the latter resulting in fewer accurate retrievals of the encoding event.

In Experiments 1 and 2 we cross preexperimental frequency with experimental frequency, making the predictions more interesting and complex. The second row of Figure 2 adds contextual associations to the low- and high-frequency examples. The strength of the connection between the word node and the context node is the same for both the high-frequency and low-frequency word, and the strength of the event node is also the same. The amount of activation that can spread from a concept node to an associated event node depends on the strength of that connection and the number and strength of all the competing links that fan out from the concept node. We predict that Remember judgments are less likely for high-frequency words because less activation will spread to a high-frequency word's associated study-event node because of its greater fan.[5] In other words, when a word is presented for the second time, more activation will arrive at the event node (to be added to the current base strength) if the word is of low frequency than of high frequency, because low-frequency words have less competition from other associations.

The probability of giving a Remember response depends on the activation level of the encoding event. The event node gets stronger from multiple presentations, as do both the link from the word node to the event node and the word node itself. This means that with each new presentation, the probability of eliciting a Remember response increases for both low-frequency words and high-frequency words. The greater fan associated with high-frequency words becomes less important as the strength of the association to the study-event node increases and as the study-event node itself gains strength. The bottom half of Figure 2 illustrates how this memory representation for multiple presentations of a high- and low-frequency word evolves with repeated presen-

[4] SAC does predict that the WFE on know judgments should be slightly larger for hits than for false alarms. For hits and false alarms there is an equal-sized direct effect on the base level of activation of the word nodes. However, for hits, low-frequency words are more likely to be remembered, and this reduces the proportion of Know responses indirectly. Because the size of the indirect effect is related to the proportional size of the WFE on Remember hits, the influence of the indirect effect will be relatively small and thus hard to detect empirically. Moreover, these predictions hold true for only midrange Remember–Know response rates. When response rates approach floor or ceiling, the relative size of the WFE on Know judgments for hits versus false alarms could be larger or smaller. For this reason, we will not make much of the presence or absence of the Word Frequency × Hits–False Alarms interaction on Know judgments.

[5] Our theory shares this assumption with other activation-based theories, most notably ACT-R (e.g., Anderson & Lebiere, 1998).

Table 1
Proportion of Know Judgments to High-Frequency and Low-Frequency Words in Published Experiments

Data set	Know hits: High-frequency words	Know hits: Low-frequency words	Direction of difference
Bowler et al. (1998)	.28	.22	+.06
Gardiner & Java (1990) Exp. 1	.16[a]	.17[a]	−.01
Gardiner et al. (1997) 50/50 target/lure	.34[b]	.32[b]	+.02
	.22[c]	.23[c]	−.01
Gardiner et al. (1997) 30/70 target/lure	.33[b]	.31[b]	+.02
	.21[c]	.22[c]	−.01
Huron et al. (1995)	.34[a]	.33[a]	+.01
Kinoshita (1995) Exp. 1	.15[a]	.15[a]	.00
Kinoshita (1995) Exp. 2	.37[a]	.33[a]	+.04
Strack & Forster (1995) Exp. 1	.35	.30	+.05
	Know false alarms		
	High-frequency words	Low-frequency words	Direction of difference
Bowler et al. (1998)	.10	.06	+.04
Gardiner & Java (1990) Exp. 1	.08[a]	.07[a]	+.01
Gardiner et al. (1997) 50/50 target/lure	.27[b]	.19[a]	+.08
	.09[c]	.09[a]	.00
Gardiner et al. (1997) 30/70 target/lure	.21[b]	.23[b]	−.02
	.11[c]	.12[c]	−.01
Huron et al. (1995)	NR	NR	NR
Kinoshita (1995) Exp. 1	.08[a]	.08[a]	.00
Kinoshita (1995) Exp. 2	.22[a]	.17[a]	+.05
Strack & Forster (1995) Exp. 1	.11	.10	+.01

Note. Exp. = experiment. NR indicates proportion was not reported.
[a]Exact means were not reported in these studies. Means in table are estimated from bar graphs. [b]These numbers combine Know judgments and "guesses." Those authors suggest that under standard instructional conditions, guesses are included in Know judgments. [c]These numbers include only Know judgments (i.e., exclude "guesses").

tations. Each time the word is presented, its word node, the corresponding study event, and the link between them are all strengthened. Increases in strength are moderated by delay—with the absence of repetition, activation decays over time, according to Equation 1 presented after Experiment 1.

Rather than relying on normative word frequency measures (e.g., Kučera & Francis, 1967), some recent investigations into the mirror effect have used artificial materials (e.g., pronounceable nonwords) and experimentally varied recency and frequency of participants' exposure to the materials prior to a recognition test result (e.g., Chalmers & Humphreys, 1998; Maddox & Estes, 1997). We, too, varied frequency and recency of participants' exposure to our materials; however, we also crossed experimental frequency with normative word frequency, and we asked participants to make Remember–Know judgments to items they claimed to recognize. One goal of this line of research was to test the SAC model's ability to account for individual participants' changes in recognition memory and Remember–Know judgments at a fine-grained, exposure-by-exposure level of detail. To accomplish this, rather than simply varying the number of exposures to the items and then testing each item once at the end of the study session, we used a continuous recognition procedure, much like Shepard and Teghtsoonian (1961), in which participants are required to judge whether the presented item is being shown for the first time or has been seen before in this set. In this way, we get multiple judgments on a word as it is building up experimental frequency. With this paradigm, we have a rich data set that allows us to keep track of variables such as time since a word was last seen and the number of times it was seen. We can then use these variables to predict the probability of responding Old, and the probability of responding Remember versus Know. This rich data set provides a rigorous test of our computational model.

In summary, our experimental paradigm crosses preexperimental frequency with experimental word frequency, allowing us to examine how judgments (New vs. Remember vs. Know) vary as a function of preexperimental history of exposure and how these judgments shift over time as a function of experimental exposure and delay.

Experiment 1

Method

Participants. Twenty-two undergraduates enrolled in psychology courses at Carnegie Mellon University participated in the experiment as partial fulfillment of a course requirement. All participants were native English speakers.

Design and materials. This experiment used a continuous recognition paradigm (e.g., Shepard & Teghtsoonian, 1961). This design is distinctive because it did not include the separate study and test phases that typify most memory experiments. Instead, the words were continuously presented for judgment. Consequently, participants had to constantly keep track of which words had been previously presented and which words were presented for the first time.

Within this paradigm, we manipulated two factors. One factor was normative frequency. Words were selected from the Medical Research Council (MRC) psycholinguistic database described by Coltheart (1981). We selected our words to have frequencies comparable to those used by Gardiner and Java (1990). Low- and high-frequency words had Kučera and Francis (1967) normative mean frequency counts of 1.6 and 142, respectively. A total of 192 low-frequency and 192 high-frequency words were selected. All 384 words were between 5 to 10 letters in length.

The second factor was experimental presentation frequency. The words were randomly selected to be presented either 1, 3, 5, or 10 times. From the pool of low-frequency words, eight words were randomly selected (without replacement) to be presented 10 times, four to be presented 5 times, and four to be presented 3 times. Another 80 low-frequency words were randomly selected to be shown only once. The same procedure was used for assignment of high-frequency words to conditions. This resulted in 384 trials, half of which were "New" trials (the first presentation trial of a word) and half "Old" trials (the subsequent presentation trials of a word). Five lists of 96 low-frequency and 96 high-frequency words were constructed. Participants were assigned one of these five lists of words, but the order of presentation of the 384 trials was randomly determined for each participant.

Procedure. The participants were tested individually in a single session that lasted about 25 min. All stimuli were presented on a Mac IIci with a black and white monitor. The words were presented individually in the middle of the computer screen. Participants were asked to read each word silently. After reading each word, they were asked to make one of three judgments: New, Remember, and Know. Participants indicated their responses by pressing labeled keys on the keyboard's number pad. They were asked to press the key labeled "New" when they thought that the word had not been presented previously in the experiment. Participants were told that the words could start repeating at any time, so they should always be prepared to respond Remember or Know. They were asked to press the key labeled "R" for Remember when they recognized the word as having been presented earlier in the experiment and had conscious recollection of reading it earlier. If they believed the word was seen earlier in the experiment but did not have conscious recollection of reading it earlier, they were told to press the key labeled "K" for Know. The number keys 1, 2, and 3 were labeled New, R, and K, respectively. Note that this procedure differs from most Remember–Know experiments in which participants first made a New versus Old judgment before proceeding to categorize the Old judgments into either an R or a K response.

To help participants understand the difference between the Remember and the Know responses, they were given the same examples used by Gardiner and his colleagues (Gardiner, 1988; Gardiner & Java, 1990). They were told that if the experimenter asked them what movie they saw last, they were to likely remember the name of the movie, when they saw it, and with whom they saw it. Thus, this means that they had conscious awareness of the movie experience and would warrant an R response. On the other hand, if the experimenter asked them what their name was, they would typically respond in a Know sense. They would not be conscious of when and where they learned their names, yet they would definitely know their own names. In other words, they typically would not consciously recollect any event or thing associated with their names. This would then typically warrant a K response. In addition, it was stressed to participants that the difference in the responses was not of memory strength. They were told that there are two different states of memory. Knowing did not necessarily indicate a poorer memory. This was illustrated by the "knowing your name" example. After the two examples were presented to them, the participants were required to give two additional examples of their own in order to establish that they had understood the two types of responses.

Participants were also told to make judgments as quickly as possible while remaining accurate. After they made a judgment, the next trial would begin after an intertrial interval of 1,500 ms. This process continued until all 384 trials were completed. Participants were given scheduled breaks at 60-trial intervals, the duration of which was determined by the participant.

Results and Discussion

There are several useful ways to analyze these data. The most straightforward way is to (a) analyze each event as an Old versus New item and (b) examine the effects of preexperimental or normative word frequency on tendency to give Remember versus Know responses for hits (previously presented items) and false alarms (not previously presented items). This analysis collapses all judgments from the second presentation or later into a single category of Old items. Another plausible way to analyze the data is to examine how tendency to respond R or K changes with each successive presentation for the two levels of normative (preexperimental) word frequency. In all analyses, across all three experiments, and in discussing other data in the General Discussion, we adopt a significance level of $p < .001$ and report the p value only in the few cases when the statistic is less reliable.

Figure 3 displays the proportions of Remember and Know responses for hits, that is, items correctly identified as Old, and for false alarms, that is, items incorrectly identified as Old. These are displayed as a function of normative word frequency. The proportion of responses R versus K is denoted directly on the functions. Note that these hits collapse over all judgments on words that had been presented at least once before, whereas false alarms refer to spurious R and K responses for words presented for the first time. Also note that some words were presented only three times, but other words were presented as many as 10 times. We conducted separate two-way ANOVAs (analyses of variance) for the two responses Remember and Know using preexperimental word frequency and whether the word had been previously studied (hits vs. false alarms) as within-participant factors. In addition we calculated one-way ANOVAs for the hits and false alarms separately.

Not surprisingly, there were significantly more Remember responses for Old words ($M = 0.76$) than New words ($M = 0.02$), $F(1, 21) = 390.6$, $MSE = 0.03$. There was also a significant main effect of word frequency, $F(1, 21) = 10.0$, $MSE = 0.003$, $p < .01$, such that low-frequency words ($M = 0.41$) were "remembered" more than high-frequency words ($M = 0.37$). As found previously, there were significantly more Remember responses for low-frequency words

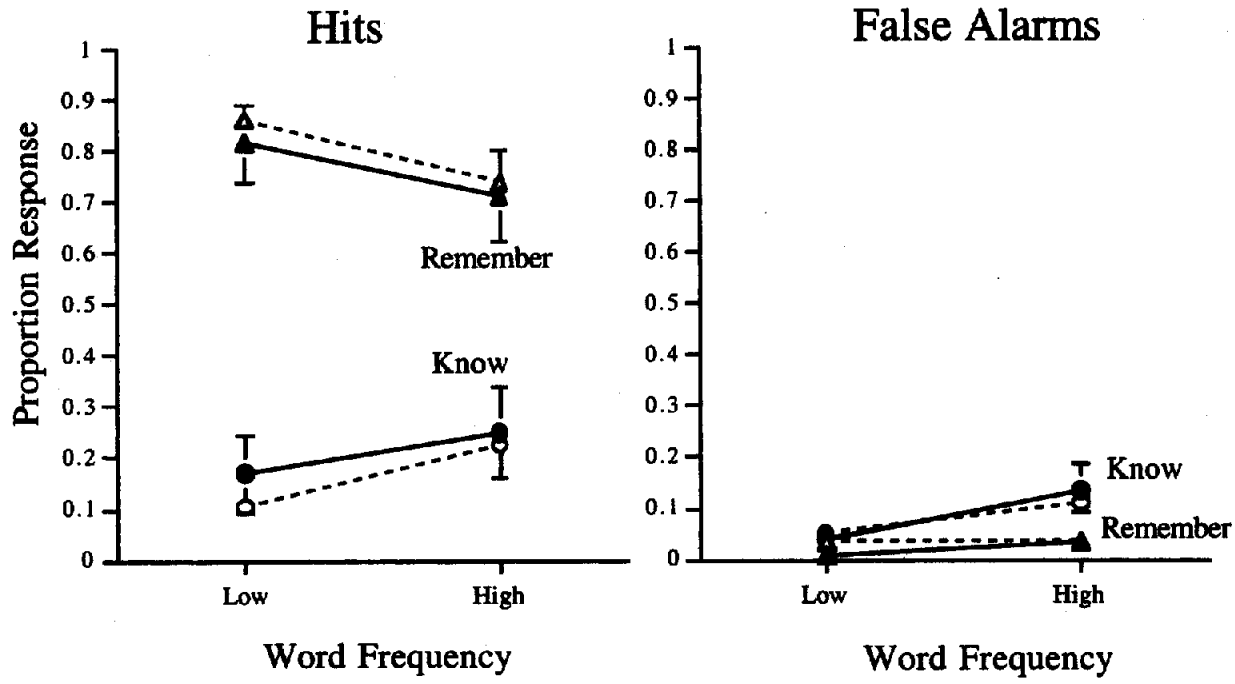

Figure 3. Experiment 1. Proportion Remember and Know for hits (left) and false alarms (right) as a function of word frequency. Triangles represent Remember responses; circles represent Know responses. Closed symbols with solid lines represent the actual data. Open symbols with dashed lines represent the model predictions. The error bars represent 95% confidence intervals.

($M = 0.81$) that had been presented before (hits) than for high-frequency words ($M = 0.71$), $F(1, 21) = 36.4$, $MSE = 0.003$. The pattern was different for new words (false alarms); there were few false Remember responses overall, and there were even fewer for low-frequency words ($M = 0.007$) than for high frequency ($M = 0.034$). This word frequency difference (in the opposite direction) was marginally reliable, $F(1, 21) = 6.0$, $MSE = 0.001$, $p < .05$; however, the interaction of word frequency with Old–New status was highly reliable, $F(1, 21) = 65.8$, $MSE = 0.001$, such that there were more Remember responses for low-frequency words that were Old but no effect for Remember responses for high-frequency words when the items were New.

The analyses concerning Know responses are of special interest. As expected, there were also more Know responses for Old words (hits; $M = 0.21$) than words not previously presented ($M = 0.09$), $F(1, 21) = 7.1$, $MSE = 0.045$, $p < .05$. As predicted, but in contrast to claims in the literature, we also found more Know responses for high-frequency words ($M = 0.19$) than low frequency ($M = 0.10$), $F(1, 21) = 47.4$, $MSE = 0.004$. This effect held for both hits (high $M = 0.25$, low $M = 0.17$), $F(1, 21) = 24.1$, $MSE = 0.003$, and for false alarms (high $M = 0.14$, low $M = 0.04$), $F(1, 21) = 33.9$, $MSE = 0.003$, and the interaction of Old–New status with word frequency was not reliable, $F < 1.2$.

For the most direct comparison between this experiment and previous studies, we compared the proportion Know responses given to each item's second presentation. For items' second presentation, high-frequency words received a greater proportion Know judgments ($M = 0.44$) than did low-frequency words ($M = 0.38$). This difference was marginally significant, $F(1, 21) = 4.00$, $MSE = 0.009$, $p < .06$. Given that this theoretical prediction differed from previous conclusions in the literature, we attempted to replicate it using more traditional Remember–Know procedures. These efforts are reported in Experiments 2 and 3.

When collapsing over all Old presentations, we found the pattern that we had predicted; however, it is useful to ask how this pattern changes over multiple presentations. Specifically, would we see the greater number of Know responses for high-frequency words on the second presentation (the first correct Old response)?[6] How does the pattern of Remember versus Know change over the course of multiple presentations? Does the preexperimental word frequency variable wash out such that there is ultimately no difference between high- and low-frequency words? Figure 4 presents the mean proportion of R and K responses for words as a function of preexperimental frequency and as a function of experimental frequency. The left panel displays the R and K responses for low-frequency words, and the right panel presents these functions for high-frequency words. Note that the first presentation of a word is considered a new trial. That

[6] All statistics were also calculated with just the second presentation for the old responses, that is, first versus second presentation, dropping all later trials. The significance pattern was identical.

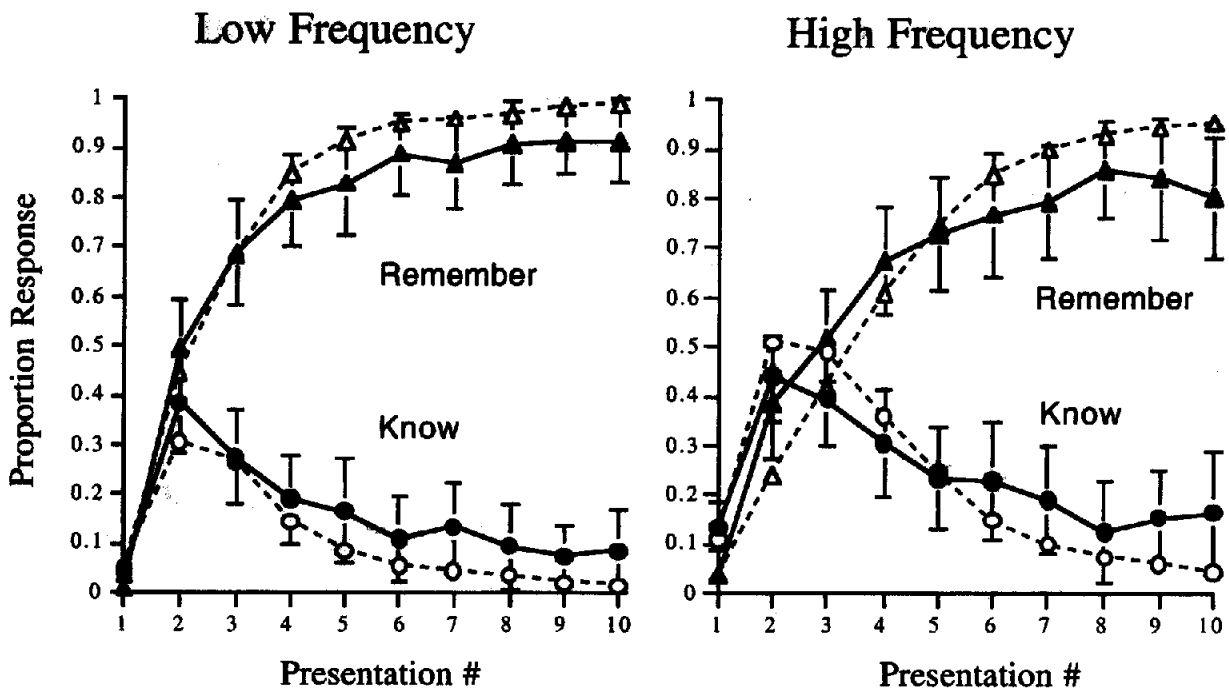

Figure 4. Experiment 1. Actual (closed) and predicted (open) proportion of Remember and Know responses for low and high frequency as a function of presentation number. The error bars represent 95% confidence intervals.

is, these are the lure trials under a typical memory experiment. Thus, these probabilities represent the false alarm rates. Presentations 2–10 constitute the Old trials.

Following a precedent set by Gardiner and Java (1990) and Gardiner (1988), we also conducted analyses in which we treated response type, R versus K, as an independent variable, on the grounds that it may be regarded as an instructional manipulation. We recognize that this assumption is somewhat questionable because of the dependence between the two responses; however, as Gardiner and Java noted, "it does have the advantage that interactions involving response type can be directly evaluated" (1990, p. 25). For these analyses, we analyzed hits and false alarms separately and used a repeated measures ANOVA with preexperimental word frequency (high vs. low) and response type (R vs. K) as factors for the false alarms, and these same factors plus presentation number as factors for the hits.

For hits (words presented two times or more), a 2 (low- vs. high-frequency words) × 9 (Presentations 2–10) × 2 (R vs. K response) repeated measures ANOVA also indicated the significant effect of word frequency, $F(1, 21) = 7.40$, $MSE = 0.003$, $p < .01$, such that low-frequency words ($M = 0.49$) were recognized more often than high-frequency words ($M = 0.48$). The interaction of word frequency with response type was significant, $F(1, 21) = 31.86$, $MSE = 0.053$, $p < .01$, such that the pattern for Remember and Know is significantly different for high- and low-frequency old words: Low-frequency words ($M = 0.81$) elicit more Remember responses than do high-frequency words ($M = 0.71$), but not more Know responses for low-frequency words ($M = 0.17$) than high-frequency words ($M = 0.25$). The interaction between presentation number and response type was also reliable, $F(8, 168) = 29.8$, $MSE = 0.047$, indicating that for both levels of word frequency, responses tend to move toward more and more Remember responses with increasing numbers of presentations.

For foils (first presentation of a word that should not be judged as Old), a 2 (low- vs. high-frequency words) × 2 (R vs. K responses) repeated measures ANOVA also revealed a main effect of word frequency, $F(1, 21) = 30.26$, $MSE = 0.003$, such that participants made more false alarms to high-frequency words ($M = 0.09$) than to low-frequency words ($M = 0.02$).[7] That is, there were more R and K false alarms to high- than to low-frequency words. The main effect of response type was also significant, $F(1, 21) = 19.65$, $MSE = 0.095$, such that there were more Know than Remember responses to new words, means of 0.08 and 0.02, respectively. The Word Frequency × Response Type interaction was significant, $F(1, 21) = 16.62$, $MSE = 0.002$. This reflects the fact that the WFE on false alarms is largely carried by the Know responses, that is, there are many more K responses for high frequency than low frequency, 0.14 and 0.04, respectively, while the word frequency difference is negligible for the R responses, 0.03 and 0.01, respectively.

We were pleased with our ability to find the qualitative

[7] Discriminability (d') scores also showed this difference: Low-frequency words were better discriminated than high-frequency words, d' of 4.17 and 3.02, respectively.

pattern predicted by SAC but were concerned that other researchers had not found that pattern. One possible explanation for the previous failures to find a reliable difference in Know responses for high versus low frequency might have been due to floor effects. Specifically, Gardiner and Java's (1990) participants had a 24-hr delay after a single presentation of a word, whereas our participants had only a few minutes delay between the first and second presentation of a word. Our participants had higher R and K responses for targets, which is not surprising given the delay difference. Strack and Forster's (1995) participants had a very large list with only 1 s of exposure to each word before any testing and then a delay of 45 min between study and test. Their participants' recognition accuracy was even lower than Gardiner and Java's participants, so it is not surprising that they too failed to find a reliable difference.[8]

The SAC Simulation

Quantitative–computational test of the theory. As a stronger test of our theory, we implemented the model in a computer simulation. There were two primary reasons for doing this. First, although some of the predictions seem straightforward from our conceptualization, others are less obvious without actually running the simulation. For example, it seems obvious that the model should produce more Know responses for high-frequency words for foils (false alarms) than for low-frequency words. Because of the higher base level of activation for high-frequency words, they should be more likely to spuriously exceed the threshold to respond Know. In other words, high-frequency words seem more familiar because they are more familiar. By contrast, the predictions concerning Remember and Know responses for previously presented words (targets) are more subtle. We have emphasized that there are more links (associations) emanating from a high-frequency word because of more prior exposures, and therefore the relative strength of any one link is much weaker for high-frequency words. However, because the resting level of activation of the base node is also higher for high-frequency words, there should be more activation to spread from a base node to the episodic node. The question is, would the higher resting level of activation from the high-frequency word node counteract its greater fan? Given that these two effects of frequency work in opposite directions, the qualitative prediction will depend on quantitative details. Therefore, it is important to first demonstrate that we can get the qualitative fits using principled parameter values.

The second goal of simulating the data was to see whether we could not only qualitatively fit the data, but also quantitatively fit the data without changing the parameter values from previous modeling enterprises. Our goal is to account for the pattern of results at a fine grain size without estimating many new parameters. The details of the modeling enterprise are described below.

Simulation overview. In this section, we present a simulation of participation in Experiment 1 using the SAC model of performance. Although this model is for Remember–Know judgments for words of various preexperimental and experimental frequencies, most of the assumptions and parameters are taken directly from previous efforts in our laboratory to model an arithmetic feeling of knowing experiment that also used continuous judgments (on whether to retrieve or calculate) to repeated problems (Reder & Schunn, 1996; Schunn et al., 1997).

The computer simulation received as input the exact order of words presented to a given participant and then predicted that specific participant's probability of saying New, Know, or Remember on each trial. Because each participant received one of the five lists of words in a different random order, the forgetting and strengthening history for a given word would vary depending on the specific order. Therefore, a separate simulation was executed for each specific order. This precise yoking of the simulation to participants was essential because on a given trial the expected activation level for a word varied depending on the exact sequence of trials: On a given trial, the strength of the links, the current activation of the word node, and the current activation of the event node might differ from any other participant's values. The model fits are described below.

Simulation details. The simulation estimates a probability of responding R and K for each trial based on current activation values. These values are affected by a number of variables. First, initial strength of the words is affected by the words' preexperimental history of exposure, which we estimate based on word frequency norms. It is also affected by exposures during the course of the experiment, which we can calculate more exactly. The base-level strength increases and decreases according to a power function,

$$B = B_w + c_N \sum t_i^{-d_N}, \quad (1)$$

in which B_w is the preexisting word base-level activation (determined from word frequency norms, and set to zero for episode nodes), c_N and d_N are constants, and t_i is the time since the ith presentation. This function captures both power-law decay of memories with time, and power-law learning of memories with practice.[9]

The base or resting level of activation of a node should be distinguished from its current activation values. The current level of a node will be higher than its baseline whenever it receives stimulation from the environment, that is, when the concept is mentioned or perceived, or when the concept receives activation from other nodes.[10] Whereas baseline strength decays according to a power function (i.e., first quickly and then slowly), current activation decays rapidly and exponentially toward the base level. Let A represent the current level of activation and B represent the base level of

[8] Gardiner and Java's (1990) Know response rate was under 20%, and Strack and Forster's (1995) was under 10%. In contrast, our Know response for the comparable second presentation was about 40%.

[9] See Anderson and Schooler (1991) for a discussion of the evidence for this function in learning and retention phenomena.

[10] It is this high level of activation that enables focus of attention on the concept.

activation. Then, the decrease in current activation will be

$$\Delta A = -\rho(A - B) \quad (2)$$

such that, after each trial, the current activation will decrease for every node by the proportion ρ times that node's current distance from its base-level activation.

Each time a word is presented its activation spreads to associated concepts (nodes) via links. For example, a word node will be connected to episodic nodes (see Figure 1) that represent the various contexts in which a word has been seen as well as to other word nodes that are semantically and experientially related. The amount of activation that spreads down any one link depends on the number of links emanating from a node and their relative strengths. The amount of activation any particular node r receives can be represented as

$$\Delta A_r = \sum (A_s * S_{s,r} / \sum S_{s,i}), \quad (3)$$

in which ΔA_r is the change in activation of the receiving node r, A_s is the activation of each source node s, $S_{s,r}$ is strength of the link between nodes s and r, and $\Sigma S_{s,i}$ is sum of the strengths of all links emanating from node s. The effect of the ratio $S_{s,r}/\Sigma S_{s,i}$ is to limit the total spread from node s to all connected nodes to be equal to the node s's current activation A_s.

Associative links vary in strength depending on how often the two concepts have been thought of at the same time and on the delay between exposures. Specifically, we assume a power function given by

$$S_{s,r} = c_L \sum t_i^{-d_L} \quad (4)$$

in which $S_{s,r}$ is the strength of the link from node s to node r, t_i is the time since the ith coexposure, and d_L is the decay constant for links.

All of these equations were used in simulating the Remember–Know judgments for each individual participant's exact experimental history. We assume that at the start of the experiment, the representation of memory for the simulation (for each participant) is identical regardless of the experimental stimuli to be seen. That is, representations of all of the words to be presented in the experiment are assumed to already exist in memory. Similarly, we assume that the experimental context node already exists.[11] However, the nodes for the study event are assumed not to exist (i.e., these study events are novel). The initial baseline strength of the word nodes is determined by their respective Kučera and Francis (1967) frequency counts. Specifically, each word node's initial baseline strength (B_w) was computed by raising each word's Kučera and Francis (1967) frequency count to an exponent of 0.4. Similarly, the fan (number of associations to each word) off each word was computed using an exponent of 0.7. The base-level strength and fan for the experimental context node are set to a constant amount that does not affect the simulations of the recognition process.

At the start of each trial, the word node and the context node for the presented word are activated by a constant amount. When a word is "seen" for the first time by the simulation, the judgment based on activation values of nodes can be derived from the activation of only the word node because a study event node has yet to be created. Following the first presentation of an item, a study-event node is created for that word because one was not there before, and links from the word and context nodes to the study-event node are built (see Figure 1). The initial base-level strength of the study-event node and of the links is simply determined by the equations determining power-law growth and decay. We assume that a basic perceptual process activates these nodes. For example, when the word "caveat" is presented for the third time, the caveat word node and the context node are given a boost in temporary activation that spreads to the study-event node. The amount that is spread is a function of that specific participant's word presentation history (the strength of the link depends on number of prior presentations and the delay between presentations as detailed above).

Once the activation has spread across these links, the activation of the study-event node and the word node can be used to make the R and K judgments. We assume that this decision follows a normally distributed function of activation. Rather than producing a binary decision, the simulation produces a probability of choosing R or K based on the activation values. This means that if the activation value of the study-event node is high, the probability of responding R is very high; conversely, when the activation is very low, the probability of responding R is very low. This probability of responding R is calculated by assuming a normal distribution of activation values with fixed variance and activation threshold for responding R. This probability is computed by the formula

$$P(\mathrm{R}) = N[(A_E - T_E)/\sigma_E] \quad (5)$$

in which A_E is the activation of the event node, T_E is the participant's threshold for the study-event node distribution, σ_E is the standard deviation of the study-event node distribution, and $N[x]$ is the area under the normal curve to the left of x for a normal curve with mean = 0 and standard deviation = 1. Recall that we assume an interdependence between R and K judgments.[12] Consequently, the probability of responding K is a calculated by the following formula:

$$P(\mathrm{K}) = [1 - N[(A_E - T_E)/\sigma_E] * N[(A_W - T_W)/\sigma_W]]. \quad (6)$$

[11] Just as we ignore details concerning the componential analysis of word nodes such as semantic and lexical features, so too we ignore the component features of the experimental context node. Of course we believe they exist, but for simplicity, we finesse that aspect of the representation.

[12] Note, however, that this interdependence is only partial and one directional (Remember proportions affect Know proportions but not vice versa) and therefore does not imply 100% symmetry of effects. It implies only that differences in Remember proportions will have at least a small effect on Know proportions.

In essence, the probability of responding K is the product of one minus the probability of the study-event node and the probability of the word node being above their respective thresholds. Aspects of this equation are reminiscent of the assumptions put forward by Yonelinas (1997). The correspondence between the two views can be made more direct by summarizing our theory as follows: P(Remember) = r, and r decreases with contextual associations (word frequency). P(Know) = $(1 - r) * k$, and k increases with base-level strength (which increases with frequency of exposure, i.e., word frequency).

After each trial, the activation levels of all the nodes are updated using Equations 1 and 2. The strength of each link is also updated using the same kind of power-law function used to determine changes in base-level activation (Equation 4). Specifically, all the links connecting the word and context nodes to the study-event nodes are strengthened, whereas all other links in the network are weakened.[13] The nodes in the network are updated in this fashion regardless of whether the participant responds New, R, or K.

This simulation involves 12 parameters that are listed in Table 2. The first two parameters, discussed previously, convert Kučera and Francis's (1967) frequency counts to a preexisting baseline strength and a preexisting fan. Two of the other parameters are related to the initialization and decay of current activation. First, the input-activation parameter, set to 40, determines the current activation setting of the word and event nodes when the word is presented. Second, the fast-decay parameter, ρ, is the exponential decay constant at which the current activation of all nodes decays. For simplicity, the unit of decay is trials rather than time. This value was set to 0.8, the same value used for the simulation of the feeling-of-knowing phenomenon described in Reder and Schunn (1996) and Schunn et al. (1997).

The parameters necessary for changing base activations, c_N and d_N in Equation 1, were set to 25 and 0.175, respectively. Thus, the initial strength value of a study-event node after its creation was 25, and decayed with time and grew with repeated presentations from there. As with fast decay, we used trials as the unit rather than time for simplicity. The two parameters used in the computation of link strength, c_L and d_L from Equation 3, were set to 25 and 0.12, respectively. Thus, the new link that was created in connection to the study-event node was initially set to 25. Both decay constants, d_N and d_L, were the same values used for the feeling-of-knowing experiments.

To convert these activation values to probabilities of responding R or K, four other parameters are necessary. Recall that we assumed this decision follows a normally distributed function of activation. Correspondingly, there are two parameters used to determine the shape of this normal function: the threshold that is the center of this distribution and the standard deviation. Thus, for both R and K judgments, there are the respective threshold and standard deviation parameters. We used a single value for the standard deviation parameter for each word node and event node for all simulations, $\sigma_W = 8$ and $\sigma_E = 40$.

However, in contrast to the single standard deviation for a type of node, we assume that participants vary in their thresholds for responding R and K. That is, some participants are conservative and have high thresholds. Others, however, might be more liberal and have lower thresholds. The R decision threshold (T_E) and K decision threshold (T_W) values reflect the participant's overall base rate of responding R and K, respectively. The best fitting participant R thresholds ranged from 36 to 270, with a mean threshold of 91.7 (SD = 49.8). The best fitting K thresholds ranged from 46 to 124, with a mean of 59.5 (SD = 16.0). Although the participants might have differed on other dimensions as well, there were no other obvious differences, so for parsimony's sake, the other eight parameters were held constant across participants. In sum, there are 12 parameters used in the present simulation. Ten of the parameters were held constant for all simulations. Table 2 presents a summary of these parameters. Table 3 provides the six equations underlying the SAC model.

Model fits. To compare SAC's predictions to participants' actual R and K responses, we regressed the model's predicted R and K probabilities to the participants' actual R and K probabilities for each condition. We present Pearson's r^2 between predicted and actual values for the overall recognition rates (i.e., sum of R and K) as well as for each response type separately. The fit of the model to the data was defined as the sum of the squared error between the model's predicted R rate for each participant in each condition and each participant's actual R rate in each condition plus the sum of squared error between the models' predicted K rate and the participant's actual K rate. Correspondingly, the quality of the fits will be described in terms of RMSD (root mean squared deviation). We did not search the full, exhaustive combinatorial space of possible parameters. Instead, we used the same parameters from Schunn et al. (1997) when possible (three parameters), selected reasonable ballpark values for some of the new parameters (five parameters), and iteratively tried a range of values for the other two new general and two participant specific parameters (consult Table 2). For this last type of parameter, we selected the value on each parameter producing the lowest sum squared error.

Figure 3 displays the predicted as well as the observed proportion of Remember and Know responses for hits and false alarms as a function of word frequency. Note that consistent with the empirical data, the predicted R judgments are higher for low-frequency than for high-frequency words; whereas for K judgments, the model again correctly predicts more K judgments for high-frequency than for

[13] We assume that all increases in strength occur according to a power law; however, we had to approximate prior history of strengthening for words of varying preexperimental frequency. We varied their base strengths by converting normative frequency values into an initial base strength. Ideally each increment would vary depending on this base strength; however, because we cannot actually represent all prior occurrences of these words, accretions to the words were not varied as a function of word frequency. This was not a problem for the event nodes of high- and low-frequency words because they all start from no prior history (no strength). It was also not a problem for differences in fan.

Table 2
SAC Model Parameter Descriptions and Values

Parameter name	Function	Value
Preword strength	Converts Kučera and Francis frequency to preexisting baseline activation	0.4[b]
Preword fan	Converts Kučera and Francis frequency to starting preexisting fan	0.7[b]
Input activation	Input current activation for component nodes	40[b]
ρ	Exponential decay constant for current activation	0.8[a]
c_N	Power-law growth constant for base-level activation	25[b]
d_N	Power-law decay constant for base-level activation	0.175[a]
c_L	Power-law growth constant for link strength	25[b]
d_L	Power-law decay constant for link strength	0.12[a]
T_E	Study-event node decision threshold	36–270[d]
E	Study-event node decision standard deviation	40[c]
T_W	Word node decision threshold	46–124[d]
W	Word node decision standard deviation	8[c]

Note. SAC = Source of Activation Confusion.
[a]Parameter value taken from Schunn et al. (1997). [b]New parameter, but not selected to optimize the fits to data. [c]New parameter value, one value selected to optimize the fits to data for all experiments. [d]New parameter value, a different value for each participant.

low-frequency words. A stronger test of the model is whether it can fit the learning trends and shifts from New to Know to Remember for words of different presentation frequencies. Also, displayed in Figure 4 are the predicted proportion of Remember and Know responses for hits and false alarms as a function of word frequency and number of experimental presentations. The predicted curves represented by the open symbols are averages of separate simulations that are run to fit each individual participant with the unique presentation order for a given participant. The aggregate simulation appears very close to the aggregate data.[14] This makes clear that not only can the model account for the basic pattern of types of recognition judgments that accrue for words of different preexperimental frequency, also it can account for how these judgments change with increasing experimental exposure.

The SAC model produced a good fit to the data, producing a Pearson's r^2 of 0.86 (440 data points) for the overall recognition rate. In other words, the SAC model accounted for a large percent of the variance of the participant's R and K judgments even at the individual participant level, with only two free parameters per individual and two estimated parameters[15] over all participants. The fits of the model to each type of response were also very good. The overall RMSD was 0.077 for Figure 4. For the R judgment probabilities, a fit of the SAC model's predicted probabilities to the participants' actual R judgment probabilities produced a Pearson's r^2 of 0.80. For the fit of the K responses, the Pearson's r^2 was 0.61. Both of these fits were made comparing 440 observed to predicted data points (880 overall), so we were pleased with the quality of fit given how few parameters were estimated specifically for this experiment. Note that most of the predicted data points fall within the 95% CI error bands of the empirical data points.

Summary of Experiment 1 Results

It was gratifying that our theoretical predictions were confirmed at both a qualitative and a quantitative level. This was especially true because the greater proportion of Know responses for both New and Old high-frequency words as compared with low-frequency words had not been found in the earlier published studies. Although this particular pattern had not been reported earlier, other findings in the literature seem consistent with our theoretical interpretation. For example, Kinoshita (1995) found evidence that WFEs for recognition memory and repetition priming were differentially affected by attention, and she concluded the loci of these effects are different. We concur: Repetition priming effects are caused by the different base rates for low- and high-frequency words (Erickson & Reder, 1998); the recognition advantage for low-frequency words is due largely to the lower fan from the word node but is also helped by the word node's lower resting level of activation (causing fewer false alarms). Hockley (1994) found that the mirror effect for word frequency held only for item recognition but not for pair recognition. That too is consistent with our view in that base-level activation of a word node is irrelevant because the participant is not judging the familiarity of a word, but rather whether two words were presented together or swapped.

Given the novelty of our results and given that our

[14] Although the model appears to consistently overpredict Remember responses, this does not occur in Experiment 2 nor consistently for individual fits. Examples of the individual participant plots (observed against predicted) are presented later in the article.

[15] Although there were more than two parameters that were new to the these simulations (relative to the Schunn et al., 1997, simulations), only two were systematically varied to get a better fit (these were the sigma parameters). The other parameters either were chosen arbitrarily (e.g., the growth parameters) or were calculated on the back of an envelope to be roughly consistent with other parameters in the model (e.g., the preword-strength and preword-fan parameters for converting from the word frequency norms were selected to be consistent with the d_N and d_L parameters).

Table 3
SAC Model Equations

Equation	Description
(1) $B = B_w + c_N \Sigma t_i^{-d_N}$	Base-level activation as a function of delay and repetitions
(2) $\Delta A = -\rho (A - B)$	Change in current activation from one trial to the next
(3) $\Delta A_r = \Sigma(A_s S_{s,r} / \Sigma S_{s,i})$	Change in receiver's current strength due to activation spread
(4) $S_{s,r} = c_L \Sigma t_i^{-d_L}$	Link strength as a function of delay and repetitions
(5) $P(R) = P$(Event)	Probability of responding R as a function of current activation, in which P(Event) $= N[(A_E - T_E)/\sigma_E]$
(6) P(K) $= [1 - P$(Event)$]$ * P(word)	Probability of responding K as a function of current activation, in which P(Word) $= N[(A_W - T_W)/\sigma_W]$

Note. R = remember; K = know. SAC = Source of Activation Confusion.

procedure differs from the existing paradigm in two ways, we felt it was premature to conclude that our theory was supported. First, we used a three-alternative force choice (3-AFC) procedure rather than a series of two binary decisions. That is, previous research in this area has required that participants first respond Old or New, and after making that decision, they are required to respond Remember versus Know if they responded Old. It was suggested that our results might have been different from those in the literature because of this change in procedure.[16] It therefore seemed important to verify in Experiment 2 that our results would replicate with the more standard Remember–Know procedure.

Experiment 1 also differed from the standard paradigm in that we used a continuous recognition judgment rather than a study phase followed by a testing phase. Although we can argue that the second time an item is presented is effectively the testing phase, the differences in procedure, delays, and so on are undeniable. Experiment 3 tests the generality of our conclusions using the more standard study test procedure (while concurrently testing other issues). In Experiment 2 we chose again to use the continuous recognition procedure because of the power it affords in terms of sophisticated modeling predictions.

Experiment 2

The primary goal of this experiment was to test the generality of our results. Specifically we wished to ensure that our predicted but novel results would hold when the Remember–Know judgments were done more conventionally, that is, after making an Old–New judgment. In addition we wanted the opportunity to model another data set to test whether our computational model would do as good a job of fitting the data without needing to estimate any parameters (e.g., the scaling factor for word frequency into base strength) except the individual participant parameters.

Method

Participants. Fourteen Carnegie Mellon University undergraduates enrolled in psychology courses participated in the experiment in partial fulfillment of a course requirement. All participants were native English speakers.

Design and materials. The design and materials were identical to those for Experiment 1, except that instead of assigning the five lists sequentially to participants, lists were assigned randomly to participants, with the constraint that each list had to be used N times before any list was used $N + 1$ times.

Procedure. The procedure was identical to that of Experiment 1, with the exception that Experiment 2 did not ask participants to select a single response among three alternatives (New, Remember, and Know); instead, participants made possibly two successive binary judgments on each trial. The first judgment, made on all trials, was an Old–New judgment. If an item was judged Old, participants then judged whether they "remembered" or "knew" that the word had been presented before. The instructions given to participants followed as closely as possible those used by Knowlton and Squire (1995). Participants were instructed to respond Remember if they recognized the word and had conscious recollection of reading it earlier, and to respond Know if they recognized that it appeared earlier but had no conscious recollection of reading it. Following Knowlton and Squire's procedure, we further explained Remember and Know as follows:

> Often, when we remember some event or thing, we consciously recollect and become aware of aspects of our previous experience with it. At other times, we simply "know" that something has occurred before. We are not able to consciously recollect anything about its occurrence or what we experienced at that time.
>
> An "R" response would be appropriate in a circumstance such as when one remembers a recent television program and is able to recollect specific details about the experience, such as when and with whom it was viewed.
>
> A "K" response would be appropriate in a circumstance such as when one has the experience of recognizing a person, but is unable to recollect any specific details at all about the person, such as the person's name.

To establish that participants understood the task, we asked them to give one example of their own for each type of judgment. If participants were unable to generate examples or if the experimenter felt that the examples did not clearly demonstrate understanding of the judgments, the experimenter clarified the instructions sufficiently for the participant to generate adequate examples. The

[16] We thank Barbara Knowlton for this suggestion.

experiment never started until the participant had generated appropriate examples for each type of judgment. As in Experiment 1, participants were informed that both speed and accuracy were important.

Each trial proceeded as follows. First, a word appeared in the center of the screen, along with the word NEW below and to the left and the word OLD below and to the right of the stimulus word. To make the Old–New judgment participants pressed with either middle finger the "D" key or the "K" key, which were labeled "NEW" and "OLD," respectively. If the participant judged the item to be NEW, the screen went blank for 1 s, followed by the presentation of the next item. Immediately following an OLD judgment, the NEW and OLD prompts disappeared and the letters R and K appeared on the screen below and more centrally than the NEW and OLD positions. Thus, the layout of the keyboard positions matched that of the prompts on the screen, and participants could keep their fingers poised over the appropriate keys for the entire list of words. To make the Remember–Know judgment, participants pressed with either index finger either the "C" or the "M" keys, which were labeled "R" and "K" respectively. As in Experiment 1, this process continued until all 384 trials were completed. All stimuli were presented in lowercase black letters on white background, using a Mac IIci running PsyScope (Cohen, MacWhinney, Flatt, & Provost, 1993).

Results

We again analyzed the data as responses to Old versus New items, looking at the effects of normative word frequency on tendency to give Remember versus Know responses for hits (collapsing all presentations from the second presentation or later into a single category) and false alarms (not previously presented items). Figure 5 displays the data for Experiment 2 in an analogous fashion to Figure 3 for Experiment 1, that is, the proportion of Remember and Know responses for hits and for false alarms, as a function of (preexperimental) word frequency. As before, there were significantly more Remember responses for Old items ($M = 0.84$) than New items ($M = 0.01$), $F(1, 13) = 1{,}189.8$, $MSE = 0.008$, and significantly more for low-frequency words ($M = 0.45$) than high-frequency words ($M = 0.40$), $F(1, 13) = 22.1$, $MSE = 0.001$. Moreover the interaction between Old–New status and word frequency was highly significant, $F(1, 13) = 63.6$, $MSE = 0.054$, such that the greater proportion of Remember responses for low-frequency words occurred exclusively for Old items. The means for Remember responses for Old low- and high-frequency words were 0.89 and 0.78, respectively, whereas the means for Remember responses for New low- and high-frequency words were 0.01 and 0.02, respectively.

The pattern for Know judgments is also very similar to that for Experiment 1. There are significantly more Know judgments for Old words ($M = 0.13$) than New words ($M = 0.05$), $F(1, 13) = 18.13$, $MSE = 0.005$. Of special interest is our replication of significantly more Know judgments for high-frequency words ($M = 0.12$) than low-frequency words ($M = 0.06$), $F(1, 13) = 20$, $MSE = 0.002$. This effect was found for both hits and false alarms. There was a marginally significant interaction such that the tendency to find more Know responses for high-frequency words was greater for Old words than New words, $F(1, 13) = 8.472$, $MSE = 0.003$, $p < .05$. The means for Know responses for Old low- and high-frequency words were 0.08

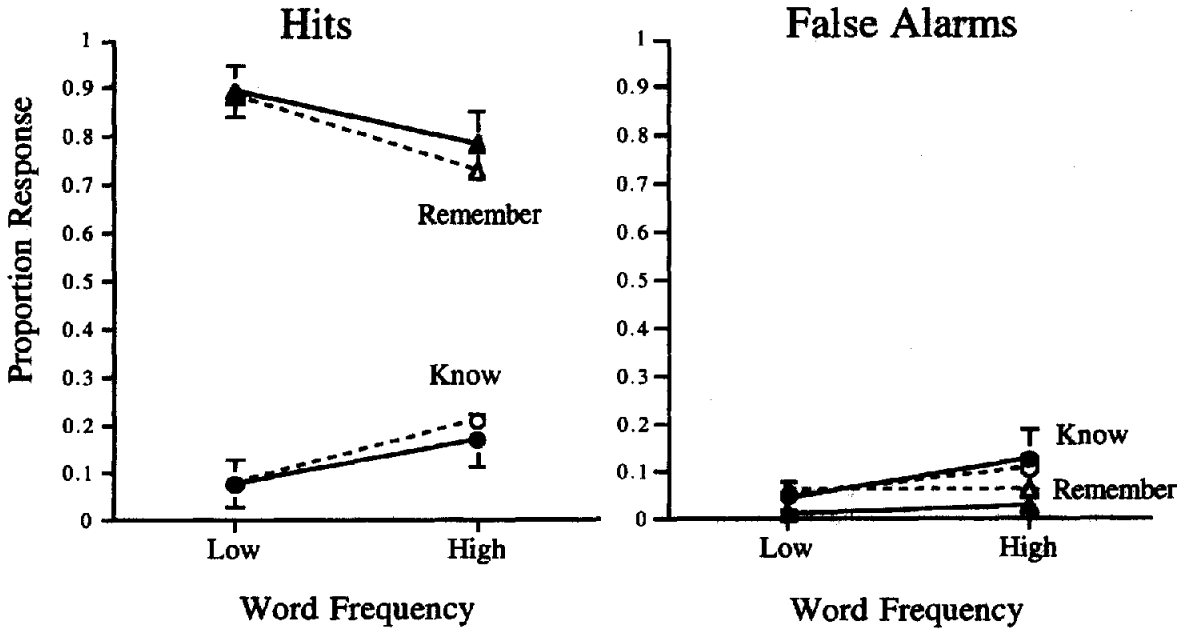

Figure 5. Experiment 2. Proportion Remember and Know for hits (left) and false alarms (right) as a function of word frequency. Triangles represent Remember responses; circles represent Know responses. Closed symbols with solid lines represent the actual data. Open symbols with dashed lines represent the model predictions. The error bars represent 95% confidence intervals.

and 0.18, respectively, whereas the means for Know responses for New low- and high-frequency words were 0.04 and 0.05, respectively. As discussed in Footnote 4, SAC does predict a slightly larger WFE for hits than for false alarms. Again, for the most direct comparison between this experiment and previous studies, we analyzed separately the proportion Know responses given to items' second presentation. Replicating the result from Experiment 1, for items' second presentation, high-frequency words received a greater proportion Know judgments ($M = 0.36$) than did low-frequency words ($M = 0.16$), $F(1, 13) = 23.97$, $MSE = 0.012$.

Figure 6 likewise displays the data in an analogous fashion to Figure 4, that is, with the proportion Remember and Know judgments plotted as a function of number of experimental presentations for low-frequency words in the left panel and for high-frequency words in the right panel. Given the claims of previous researchers such as Gardiner and his associates (e.g., Gardiner & Java, 1990) that there is no WFE for Know judgments, one might wonder whether our finding of more Know judgments for high-frequency than low-frequency words might be an artifact of the negative dependency between Remember and Know rates and the limited response range into which most of our data fall (i.e., our high hit rate). This hypothesis is not sufficient, however, on close examination of the data. First, we find the same pattern of more Know judgments to high-frequency words for false alarms as we do for hits, and restricted range is clearly not an issue. Second, the WFE is greater at lower levels of absolute memory performance than at higher levels for Experiment 2. On an item's second presentation the mean proportions of Know judgments were 0.36 and 0.26 for high-frequency words and low-frequency words, respectively, yielding a difference of 0.10. By the 10th presentation, the corresponding proportions were 0.02 and 0.00 for high-frequency and low-frequency words, respectively, yielding a WFE of only 0.02. It appears that, if anything, high absolute levels of memory performance tend to diminish, not increase, the difference in the proportion Know judgments for high-frequency versus low-frequency words.

We also analyzed the data separately for hits and false alarms so that we could compare Remember and Know responses directly.[17] An ANOVA performed on hits (i.e., participant said Old and it was the second or later presentation) indicated significantly more Remember ($M = 0.84$) responses than Know responses ($M = 0.12$), $F(1, 13) = 206.7$, $MSE = 0.035$, and a significant interaction between tendency to say Remember versus Know and preexperimental frequency, such that there were more Remember responses for low-frequency words ($M = 0.89$) than high-frequency words ($M = 0.78$) but more Know responses for high-frequency words ($M = 0.17$) than low-frequency words ($M = 0.07$), $F(1, 13) = 44.1$, $MSE = 0.003$. For false alarms, that is, those trials on which participants responded Old to New items, there were significantly more Know responses ($M = 0.08$) than Remember responses ($M = 0.01$), $F(1, 13) = 10.2$, $MSE = 0.006$. There was also a significant interaction between word frequency and type of response, $F(1, 13) = 11.4$, $MSE = 0.001$. This reflects the fact that there was a large difference in percentage of Know false alarms for high- ($M = 0.12$) versus low- ($M = 0.04$) frequency words but there was little difference in the percentage of Remember false alarms for high-frequency ($M = 0.01$) and low-frequency ($M = 0.02$) words.

As one would expect there was a significant effect of number of presentations on tendency to respond Old, $F(9, 117) = 292.6$, $MSE = 0.007$, and on tendency to respond Remember, $F(9, 117) = 127.9$, $MSE = 0.019$, or Know, $F(9, 117) = 8.79$, $MSE = 0.018$, such that the tendency to respond Old and give the Remember response grew with increased presentations; however, the Know responses initially increased and then decreased. All three dependent measures also showed significant interactions of number of presentations with preexperimental word frequency. That is, the difference in WFEs tended to decrease with increasing experimental frequency or as one approached the ceiling or floor of responding, $F(9, 117) = 6.52, 11.29, 3.3$, $MSE = 0.001, 0.006, 0.010$, for Old, Remember, and Know proportions, respectively.

The pattern of data is strikingly similar to that for Experiment 1, giving support to the view that the 3-AFC (New, Remember, Know) mode of responding does not produce qualitatively different results from the more traditional two-pass paradigm in which participants first respond Old, and then discriminate between Remember and Know responses.

It was especially important to replicate the novel result of significantly more Know responses for high-frequency words; however, an even stronger test of the theory is to see whether we can fit a new set of empirical data while holding constant the parameter values used for Experiment 1.

Simulation Model

To simulate the second experiment, the only parameters estimated were the two individual participant parameters representing an individual's threshold to respond Remember or Know for a given amount of activation. The eight nonindividual parameters from Experiment 1 were kept the same. As before, we made quantitative predictions for 40 data points per participant, using only two free parameters.

Figure 5 also plots the predicted proportions of Remember and Know judgments for hits and false alarms. As in Experiment 1, the model predictions and the empirical findings support the double dissociation between word frequency and tendency to give Remember versus Know responses: more Remember responses for low-frequency words and more Know responses for high-frequency words, the latter dissociation being a novel prediction and finding. Likewise, Figure 6 plots the predicted as well as the observed Remember and Know responses as a function of

[17] Because Remember and Know responses are not truly independent, we also calculated separate F statistics for Remember and Know (above). These statistics are reported so that they can be compared with those of Gardiner and Java (1990).

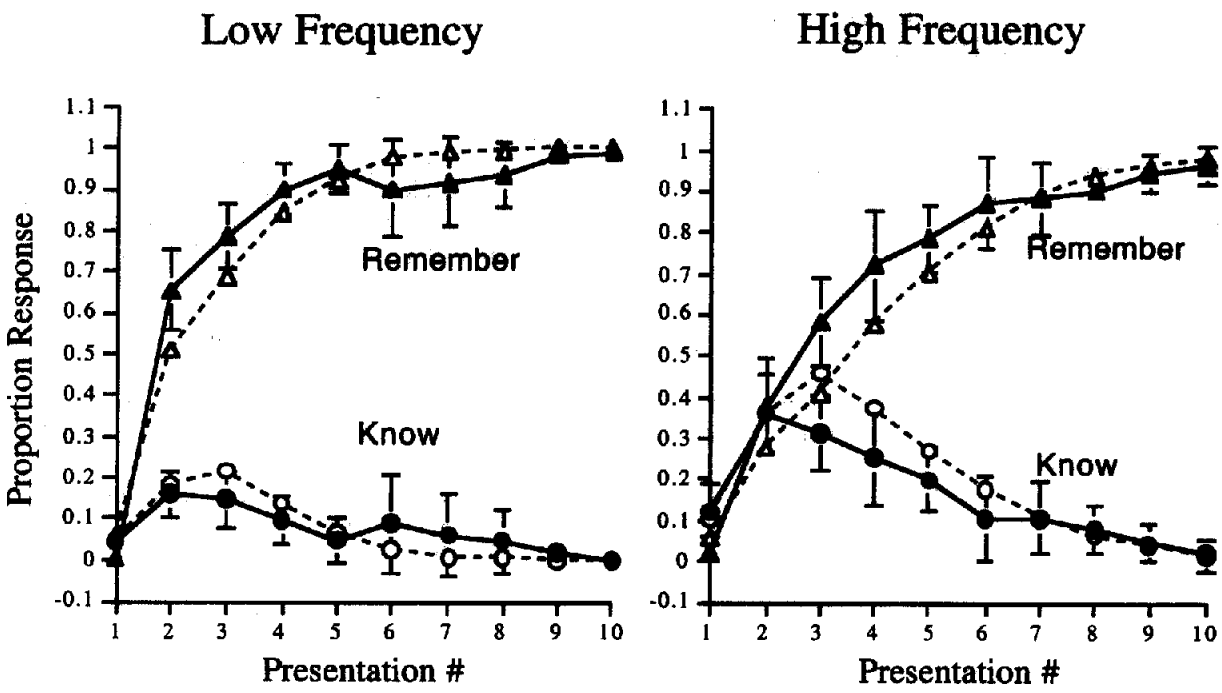

Figure 6. Experiment 2. Actual (closed) and predicted (open) proportion of Remember and Know responses for low and high frequency as a function of presentation number. The error bars represent 95% confidence intervals.

number of presentations in the experiment (on the abscissa) and as a function of preexperimental word frequency (low frequency in the left panel, high frequency in the right panel).

Overall, the SAC model produced very good fits to the data. The range of study-node threshold parameters was 50–188 with a mean threshold of 67.8 (SD = 36.3). The range of event-node threshold parameters was 52–78, with a mean threshold of 64.5 (SD = 7.8). Using these values, SAC fit the data well, producing a Pearson's r^2 of 0.88 for the overall recognition rate (280 data points). The fits for each type of response were also very good: For the R judgment probabilities, a fit of SAC's predicted probabilities to the participants' actual R judgment proportions produced a Pearson's r^2 of 0.83. For the fit of the K responses, the Pearson's r^2 was 0.52. The RMSD for Figure 6 overall was 0.068. Especially impressive is that we fit 560 (280 × 2) data points while holding constant all parameters from Experiment 1, estimating only the two individual participant thresholds (i.e., fitting 40 data points per participant with only 2 free parameters).

Fitting response patterns across participants varied considerably, presumably because they had different thresholds to respond Remember versus Know. Does SAC actually fit the various individual patterns well by just letting these two parameters vary for each participant, or was it luck that we obtained a close fit of the average predictions to the average empirical data? To show that the SAC model can predict proportion of responses for each individual participant, a few interesting participants' data were chosen to show the individual participant fits. The top and bottom halves of Figure 7 show the data and model predictions by type of response and number of times presented for two participants in Experiment 2. For each participant the left panel shows the data and model predictions for low-frequency words. The right panel shows the data and simulation predictions for high-frequency words. These two participants' data show somewhat different patterns (e.g., when the Know and Remember responses diverge for high-frequency words), yet the simulation predicts the proportion of responses quite well for each, by adjusting only two parameters per participant.

Another test of the model is predicting data for participants whose data patterns differ considerably from the normal trend. We went back to Experiment 1 and selected some nonrepresentative participants. One participant's data from the first experiment (Participant 12) exhibited responses with a K judgment much more often than with an R judgment. Participant 13 also had a higher threshold to give an R response than many participants (or the aggregate participant) but did not seem to have as strict a threshold as Participant 12. As shown in Figure 8, the simulation can also predict both participants' pattern of data quite well by just varying the two individual parameters.

Discussion

The fit of the simulation data to the behavioral data is impressive and provides strong support for the model. Nonetheless, it would be more convincing if there were converging evidence that the model accurately represents the mental structures that participants create and access in

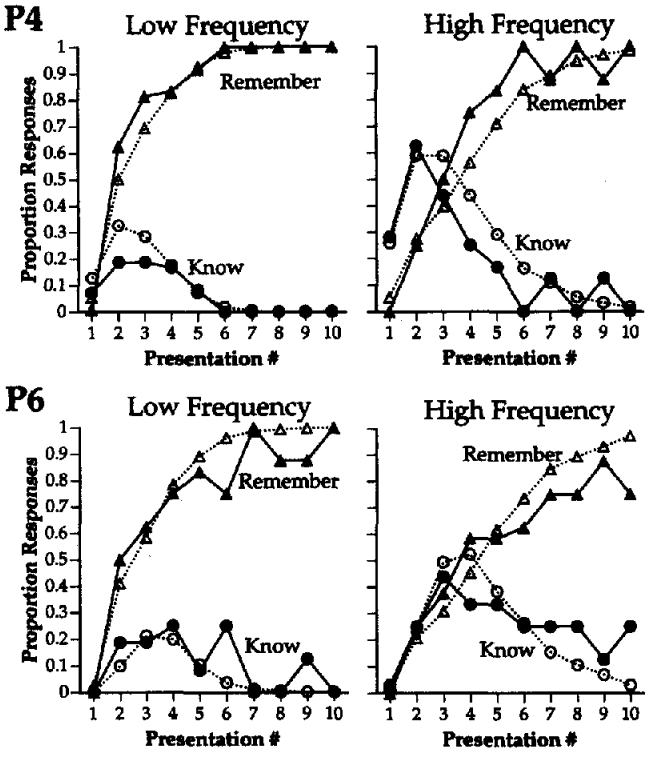

Figure 7. Experiment 2. For Participants (P) 4 and 6, actual (closed) and predicted (open) proportion of Remember and Know responses.

the Remember–Know task. In particular, how do we know that when a participant says Remember he or she is actually accessing the event node instead of the word node? As others have argued (e.g., Donaldson, 1996), it is possible that participants respond Remember when the activation or trace strength is stronger as opposed to actually accessing a different type of memory (the familiarity of the word rather than the episodic memory trace).

In order to provide converging evidence to support our contention that participants are indeed responding on the basis of different types of memories, we conducted another Remember–Know experiment that also varied word frequency, but additionally required list discrimination. By requiring participants to discriminate among lists when they respond Remember, we can determine how accurate they are at list discrimination when they respond Remember, how accurate they are at list discrimination when they respond Know, and whether the act of making list discriminations affects tendency to respond either way. We can also see whether we replicate the same pattern of responding as a function of word frequency when list discrimination is required, and finally see whether we replicate our results when we do not have a continuous recognition task.

Experiment 3

One motivation for this experiment was to test whether we would still get more Know judgments for high-frequency words in a more conventional paradigm. Unlike Experiments 1 and 2, this experiment did not involve a continuous recognition procedure. Instead participants studied each word only once and after all were studied, they made Old–New followed by Remember–Know judgments.

This study was slightly unusual in that the presentation format varied among the words: We varied the background color of the computer screen and the font in which the words were presented. All the words seen in a given color and font were blocked in presentation such that these variations in context (color and font) could be thought of as comprising different lists of words of different colored backgrounds and fonts. Participants studied words with four of these color–font context lists before proceeding to the recognition phase.

In the recognition phase, all words were presented on a neutral background in a generic font and the words from these four lists were presented in a random order, intermixed with an equal number of New words. In addition to making New–Old and Remember–Know judgments, participants were asked to identify the list (color) in which a word judged as Old was originally studied.

A between-participants variable manipulated whether participants made this list discrimination for all words that they judged as Old or only for those Old words for which they gave a Remember response. We were interested in seeing whether the R–K distinction was predictive of ability to recall the list context in which a word was seen. If the Remember response is actually tapping an episodic memory,

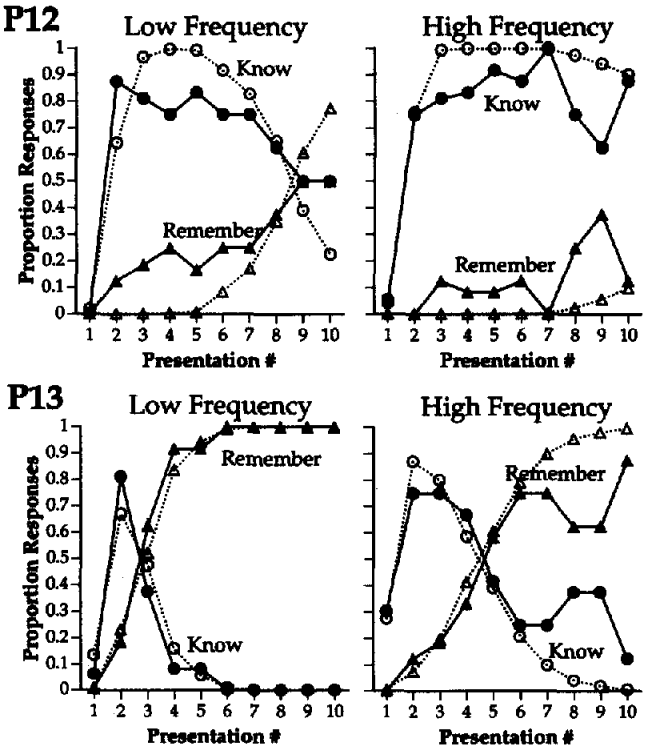

Figure 8. Experiment 1. For Participants (P) 12 and 13, actual (closed) and predicted (open) proportion of Remember and Know responses.

then it is likely to contain contextual information about the encoding event. If so, then participants should be more accurate at reporting the original background color–font information when they respond Remember than when the word only seems familiar and the participant responds Know. It is important to emphasize that participants were not told during encoding that they would be responsible for background color information or font information. Also, participants were not told that a Remember response meant that they had stored or could retrieve the contextual information that allows such a discrimination. Rather, we were interested in seeing the extent to which R responses predicted greater accuracy of report of contextual details.

In order to make the comparison, we needed a group that was asked to make the list discrimination regardless of whether a Remember response was given. On the other hand, we also worried that forcing participants to make such a discrimination even when they only felt that the word seemed familiar (a Know response) might make the R–K seem meaningless. We included both groups to evaluate the effect of forcing the list discrimination and to see how well participants did on list discrimination.

Method

Participants. Thirty-eight Carnegie Mellon University undergraduates participated in the experiment in partial fulfillment of a course requirement.

Design and materials. The experiment used a 2 (high-frequency words vs. low-frequency words) × 2 (Old words vs. New words) × 2 (list discrimination–Remember vs. list discrimination–Old) mixed design, with the first two factors varying within participants and the final factor varying between groups. Participants were randomly assigned to the List Discrimination–Old or List Discrimination–Remember group, with the constraint that each group had to have N participants before either group could have $N + 1$ participants. Each factor is further defined below.

Eighty high-frequency and 80 low-frequency words were randomly selected from the pools of 192 high-frequency and 192 low-frequency words used in Experiments 1 and 2. Separately for each participant, 40 low-frequency words and 40 high-frequency words were randomly drawn from this pool to be presented in the study phase. From these sets of 40, 10 high-frequency and 10 low-frequency words were randomly assigned to each of four study lists. Each of the four study lists was assigned a font and background color. The background colors and fonts used were green-Eglantine, blue-Harrington, red-Los Angeles, and orange-Durendal. The fonts were chosen to be rare, easily distinguishable from one another, and easy to read. Pairings of font and background color were the same for all participants. Order of presentation of the four study lists was determined randomly for each participant. The test list consisted of the entire pool of 160 words, also in a separate random order for each participant. All stimuli were presented on a 13-in. color display using a Mac IIci running PsyScope.

Procedure. The experiment consisted of three phases: study, filler task, and test. In the study phase, participants were shown each of the four study lists described above. We instructed participants to try to remember the words as best they could for a future memory test. They were given no further information regarding the nature of the eventual test. Each study list was preceded by a screen indicating the color of the list and that the participant should try to remember the words for a future memory test. The background color and font for the screen containing the instructions matched that of the succeeding study list. After reading these brief instructions, participants pressed the space bar to begin the presentation of a list. Each word was presented in the center of the screen in 24-point typeface for 2 s with a 200 ms ISI.

Following the final study list the participants were taken to a different lab room where they performed an unrelated spatial working memory task for approximately 20 min. After completing the filler task, the participants returned to the original room for the test phase of the experiment. All participants made one or two binary judgments on each trial. They first made an Old–New judgment. If an item was judged Old, participants then judged whether they "Remembered" or "Knew" that the word had been presented before. As in Experiment 2, the instructions given to participants followed as closely as possible those used by Knowlton and Squire (1995). All test words were presented in a generic font in black on a white background.

Each trial proceeded as follows. The procedure for the Old–New and Remember–Know judgments was identical to the procedure in Experiment 2. Following an Old decision, all participants made a Remember–Know judgment; however, unlike Experiment 2, participants in this experiment then might make another judgment. Participants in the List Discrimination–Old group made a list-identification judgment for every word they had called Old, regardless of whether they had given it a Remember judgment or a Know judgment. Participants in the List Discrimination–Remember group were asked to identify the study list only for items given Remember judgments.

For the list-identification judgment, the stimulus remained on the screen, and the R and K from the Remember–Know judgment disappeared. Displayed below the stimulus word were the words "red," "orange," "blue," and "green," in positions analogous to the numbers 2, 4, 6, and 8, respectively, on the number pad. These words appeared in their namesake colors and in the same fonts as their corresponding study lists. The number keys to be used for judgments were marked with stickers. We instructed participants to press the key on the number pad that corresponded to the color and font in which they had originally seen the word.

After the final judgment for a given trial, the stimulus disappeared and a prompt appeared for the participant to press the space bar for the presentation of the next item. When the participant pressed the space bar, the screen went blank for 200 ms before the next item was presented. This process continued until judgments were made for all 160 words, at which time the participants were fully debriefed as to the purpose of the experiment.

Results

The following analyses were intended to answer several questions. First, in this more conventional paradigm, would we replicate our novel prediction of more Know responses for high-frequency words as well as replicating the conventional finding of more Remember responses for low-frequency words? Second, given that Remember responses are supposed to tap a specific recollection, would participants display greater accuracy at retrieving contextual information about the study event when they make a Remember response than a Know response? Third, would the between-participant manipulation of requiring list discrimination for all Old judgments versus just after giving a Remember response affect the tendency to respond Old or Remember? We address this third question first.

Is tendency to respond Old affected by assignment to list discrimination condition? Figure 9 displays the proportion of Old responses for the two groups as a function of old (hits) versus new (false alarms). There was a main effect of condition on tendency to respond Old, $F(1, 35) = 8.98$, $MSE = 0.036$, such that participants in the List Discrimination–Old condition ($M = 0.35$) were less likely to respond Old than participants asked only to discriminate lists after making a Remember response ($M = 0.48$). This pattern suggests that the List Discrimination–Old participants found the list discrimination task difficult. Presumably they raised their threshold to respond Old to avoid having to constantly make list discriminations. It is also possible that the very process of making list discriminations for each word judged Old also caused output interference (e.g., Bjork, 1975) and thereby made participants less prone to respond Old and also less accurate at making the list discriminations. Consistent with this explanation the difference between the two groups was much larger when just examining the hits (correct Old judgments): The List Discrimination–Remember group's proportion of Old responses to Old words was 0.70, whereas the List Discrimination–Old group's hit rate was only 0.52, $F(1, 36) = 13.123$, $MSE = 0.05$.

Word frequency did not affect tendency to respond Old because the effects went in opposite directions for hits and false alarms: That is, there was a significant effect of word frequency for both hits and false alarms, $F(1, 36) = 26.4$, $MSE = 0.011$ and $F(1, 36) = 15.3$, $MSE = 0.013$, respectively, with more hits for low-frequency words ($M = 0.67$) than high-frequency words ($M = 0.55$) and more false alarms for high-frequency words ($M = 0.27$) than low-frequency words ($M = 0.17$). These effects did not interact with the assignment to group (List Discrimination–Old vs. List Discrimination–Remember). Moreover, in our view, this main effect of differential tendency to respond Old based on condition does not have implications for the other questions we wish to address.

Do we replicate our Remember–Know pattern using a more conventional paradigm? Figure 10 displays the proportion of Remember and Know responses as a function of word frequency for hits (on the left) and false alarms (on the right) for the List Discrimination–Old group. Figure 11 plots the same information for the List Discrimination–Remember group. The proportion of Remember responses varied as a function of word frequency, $F(1, 35) = 41.8$, $MSE = 0.006$, such that there were many more Remember responses for low-frequency words ($M = 0.19$) than high-frequency words ($M = 0.11$). For Remember responses, there was a significant interaction of hits versus false alarms with word frequency, $F(1, 35) = 73.13$, $MSE = 0.004$, such that there were a greater number of accurate Remember responses for low-frequency words ($M = 0.36$) than high-frequency words ($M = 0.19$) but a few more spurious Remember responses for high-frequency words ($M = 0.03$) than low-frequency words ($M = 0.02$). This pattern replicates previous findings and our results from Experiments 1 and 2.

Of special interest is the effect of word frequency on tendency to give a Know response. As in Experiments 1 and 2, we found significantly more Know responses for high-frequency words ($M = 0.30$) than low-frequency words ($M = 0.23$), $F(1, 35) = 10.1$, $MSE = 0.02$. This pattern held for hits ($M = 0.36$ vs. $M = 0.31$) and for false alarms ($M = 0.24$ vs. $M = 0.15$), and the interaction was not significant, $F(1, 35) = 2.6$, $MSE = 0.005$, $p > .1$. None of these effects interacted with assignment to treatment condition. In sum, we replicated our novel prediction using a more conventional paradigm.

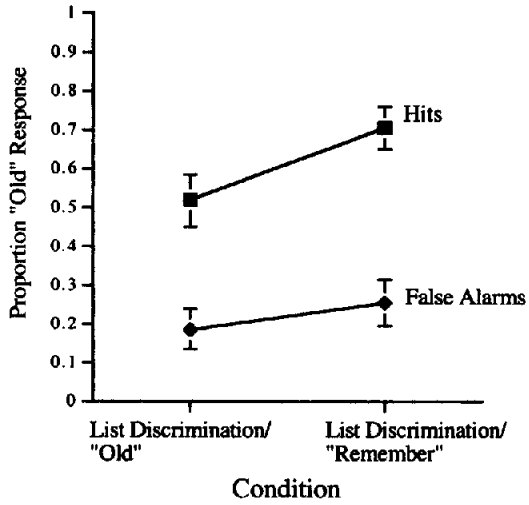

Figure 9. Experiment 3. Proportion Old response for hits and false alarms for the List Discrimination–Old and List Discrimination–Remember conditions. The error bars represent 95% confidence intervals.

Is list discrimination accuracy related to the Remember versus Know response? We conducted an ANOVA on Remember responses (hits and false alarms), using word frequency as a within-participant variable and assignment to group (List Discrimination–Old versus List Discrimination–Remember) as a between-participant variable.[18] Participants were marginally more accurate at list discrimination for Remember responses if they were not required to make these judgments regardless of type of Old response, $F(1, 32) = 3.52$, $p < .07$. Again, we speculate that this is because the requirement to make list discriminations when there is no memory trace to support it adds interference to the context-list nodes and makes subsequent discriminations more difficult.

For the group asked to make list discriminations regardless of R versus K response (the List Discrimination–Old group), we can ask whether they were more accurate at naming the list when they gave a Remember response. Participants were more accurate at the list discrimination in the Remember condition ($M = 0.39$) than in the Know conditions ($M = 0.29$) despite the fact that they were not told to base their Remember responses on an ability to select

[18] Because List Discrimination–Remember participants did not make list discrimination judgments for Know responses, we could make this comparison for only Remember responses.

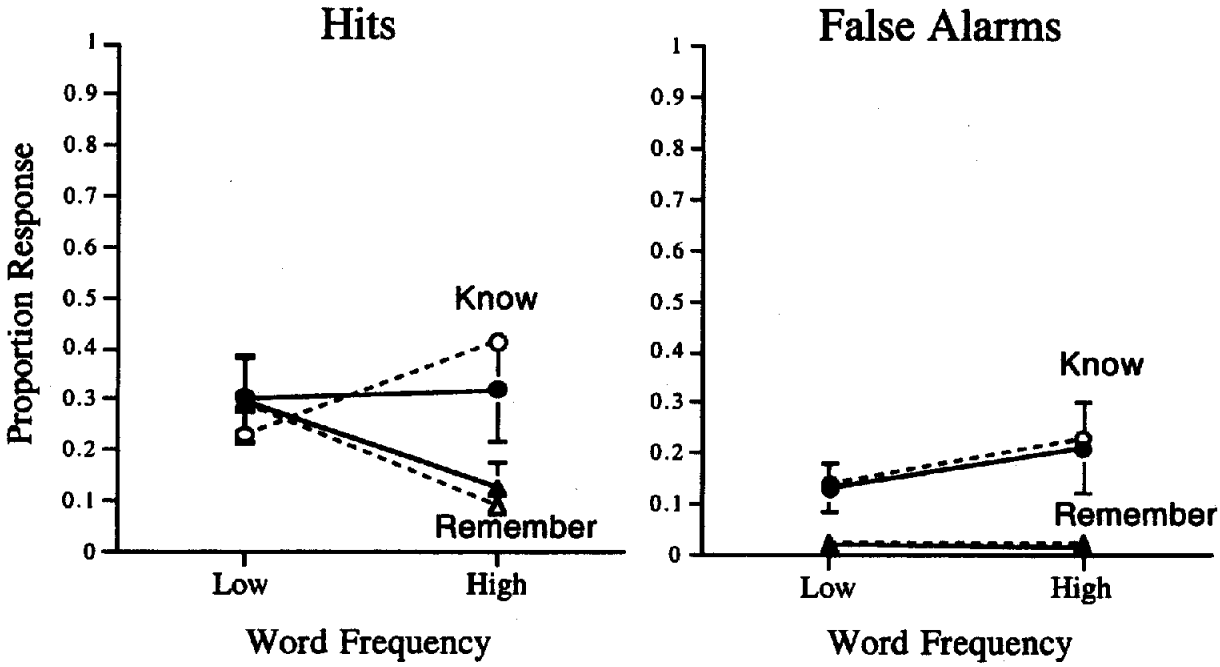

Figure 10. Experiment 3. Proportion Remember and Know for hits and false alarms as a function of word frequency for the List Discrimination–Old condition. Triangles represent Remember responses; circles represent Know responses. Closed symbols with solid lines represent the actual data. Open symbols with dashed lines represent the model predictions. The error bars represent 95% confidence intervals.

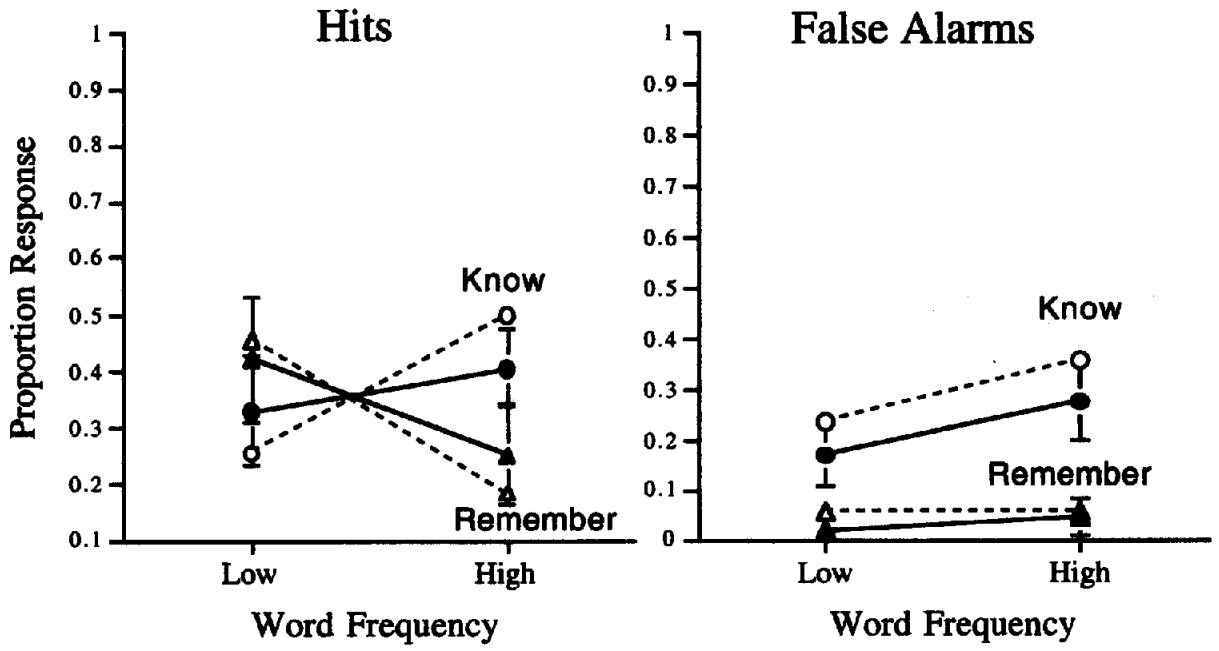

Figure 11. Experiment 3. Proportion Remember and Know for hits and false alarms as a function of word frequency for the List Discrimination–Remember condition. Triangles represent Remember responses; circles represent Know responses. Closed symbols with solid lines represent the actual data. Open symbols with dashed lines represent the model predictions. The error bars represent 95% confidence intervals.

the correct list. This effect is not reliable by a contrast; however, for words given a Remember response, the list discrimination accuracy is reliably different from chance $t(18) = 2.91$, $p < .01$, but is not for words given a Know response. We suspect that this effect is an underestimate of the true difference because of the interference caused by forcing list discrimination for words judged as just familiar.

Discussion

We take the aforementioned pattern of results as strong support for our theoretical position and predictions. Others who have examined the relationship between Remember–Know judgments and memory for source include Conway and Dewhurst (1995) and Mather, Henkel, and Johnson (1997). Both studies revealed that making source monitoring judgments reduced the proportion of R responses. Mather et al. concluded that more R responses are given for correctly attributed sources than for incorrectly attributed sources because of greater memory for perceptual detail (more accurate memory for source). Our data are consistent with this conclusion.

The explanation that the word frequency mirror effect is due to greater contextual confusions for high-frequency words (greater fan making it more difficult to access the event node) has also been tested by Guttentag and Carroll (1994). They found list discriminability to be worse for high-frequency words, consistent with our view that it is more difficult to access the event node for words with greater fan (i.e., high-frequency words). Interestingly, although they did not ask participants to make Remember versus Know judgments, they nonetheless concluded that greater familiarity for high-frequency words does not contribute to their poorer recognition performance. We believe both factors, contextual confusion and familiarity, are involved. However, as in Experiments 1 and 2, we again found that not only were participants more inclined to give Remember responses for low-frequency than high-frequency words, they also gave more Know responses for high-frequency than for low-frequency words, regardless of whether the word had been studied previously.

The replication of the WFE for Know judgments was important given the implication in the work of Guttentag and Carroll (1994), Gardiner and Java (1990), and Strack and Forster (1995). We have now demonstrated this result in three different experiments, using different experimental procedures with many levels of experimental frequency in Experiments 1 and 2. In those experiments, besides confirming our predictions, we showed how our computational model can deliver these novel qualitative predictions at a fine level of granularity.

Computational Model

For the model fits of Experiments 1 and 2, the parameter estimates were held constant, allowing only the individual participant thresholds for responding R versus K to vary. For Experiment 3, we also held the parameter values constant to those that have been used in earlier modeling efforts. However, the modeling enterprise is necessarily different in this case.

Unlike Experiments 1 and 2, this experiment does not require continuous recognition judgments in which activation values for a given judgment depend on the number of previous exposures and on the delay since it was last presented. In this case, the number of presentations is identical for all words, and we assume that the decay from study to test is the same for all words. The only things that varied were (a) the normative word frequency, (b) whether the word was studied or not, and (c) whether the participant made a list discrimination after each Old response or only after Remember responses. Because of this, there are few (only 4) data points per participant to fit, in contrast to the 40 data points in Experiments 1 and 2. In order to have a reasonable number of degrees of freedom (many more data points than free parameters), we chose to aggregate the data over participants and over words in a condition.[19]

By avoiding continuous recognition procedure, we were able to use a simple Microsoft Excel based model that implemented all the same equations and constants as before, but only one set of Remember and Know values were derived for all words of a given type. We chose to fit the observed aggregate data because there were few data points. Therefore, we did not estimate separate thresholds for each participant; we estimated only one Remember threshold and one Know threshold for all participants assigned to a condition. We fit participants from the two groups separately because it seemed clear that participants had a different bias to respond Old depending on treatment and presumably that also affected their tendency to respond R or K. Also, by fitting the two groups separately, we doubled the number of data points we could fit, but only added two extra parameters.

The predicted Remember–Know responses are also displayed in Figure 10 (for the List Discrimination–Old group) and Figure 11 (for the List Discrimination–Remember group). These theoretical points typically fall within the error bars of the empirical points. Indeed the overall r^2 is 0.93 with 16 data points and five free parameters. The r^2 between the observed and predicted data is 0.95 for the Remember judgments and 0.65 for the Know judgments. The overall RMSDs for Figures 10 and 11 were 0.063 and 0.046, respectively. Despite the very different paradigm (not multiple presentations, not continuous recognition), we were able to keep most of the parameters constant. Given the difference in the paradigms and given that we fit to average data (i.e., not participant by participant), it is impressive that we could keep almost all of the parameter values constant from previous modeling efforts. The only value we changed was the standard deviation for the word-node decision

[19] In Experiments 1 and 2, the aggregation was over words by activation values, which was determined by number of presentations and delay since last presentation. We did not aggregate over participants. In none of the experiments do we aggregate over words of different frequency status (high vs. low) or over presentation status (New vs. Old).

(Know judgment) from 8 to 20. This same value was used to estimate all conditions.

The two parameters that had been estimated for each participant were now estimated just once for the List Discrimination–Old participants and once for the List Discrimination–Remember participants. For List Discrimination–Old participants, the event threshold (for Remember judgments) was estimated at 83 and the word-node threshold (for Know judgments) at 63; for the List Discrimination–Remember participants, the corresponding thresholds were estimated at 65 and 55. It is understandable that the List Discrimination–Remember thresholds were lower, given that those participants responded Old more often.

General Discussion

We have presented three experiments that tested our account of the mirror effect for words of different frequencies and our account of how participants make Remember versus Know judgments and why there should be dissociations in those judgments as a function of word frequency. Experiments 1 and 2 used a continuous recognition procedure, which allowed us to examine the effects of preexperimental word frequency on Remember versus Know judgments. In addition, those experiments manipulated experimental word frequency in order to examine how R versus K responses would change as a function of experimental word frequency for words of varying normative frequency.

Our computational model of the mirror effect and Remember–Know produced excellent fits to the data both qualitatively and at a fine-grained level, fitting individual participant data trial by trial. It is worth noting that very few new parameter values were estimated in order to fit these data; most parameters were assigned default values established in earlier modeling enterprises. Differences in base-level strength and fan, representing word frequency, were determined simply by converting individual words' frequency ratings from Kucera and Francis (1967), obviating the need to postulate any type of metacognitive knowledge of a word's frequency class in order to account for the mirror effect.

We designed Experiment 3 to attempt to replicate our novel predictions using a somewhat more traditional paradigm. An additional goal was to see whether Remember responses were indeed associated with better memory for episodic details, as theorists assume. Participants were required to make list identifications after they responded Remember; furthermore, for half the participants, there was the additional requirement of making list discriminations even when the response was Know, rather than just for Remember, that is, for any Old response.

As expected, we again found significantly more Know responses for high-frequency words than low-frequency words for both hits and false alarms. We also replicated the established result of more Remember responses for low-frequency words. We were able to fit the experimental data using parameter values derived from fitting Experiments 1 and 2. Finally, despite considerable list interference introduced by forcing participants to select a list even when they gave a Know response, the data indicated that participants still possessed a significant ability to discriminate lists when a Remember response was given but not when a Know response was given. This result suggests that R responses are indeed associated with better memory for episodic details than are K responses.

Given that previous researchers have consistently claimed no reliable differences in the proportion of Know judgments to high-frequency versus low-frequency old words, which contradicts a key prediction of the SAC model, it was gratifying to confirm our predictions and provide close and detailed fits to those data. The results reviewed in Table 1 had suggested that other researchers had merely failed to notice the patterns SAC predicted because their effects were weak. Indeed, not included in Table 1 are data from another lab, for which 9 of 10 comparisons show more know judgments for high-frequency than low-frequency words. Those data are part of a manuscript in preparation (Chappell & Seth-Smith, 1999). Recently we have been made aware of a manuscript (Joordens & Hockley, 1999) that also reports multiple experiments finding more Know responses for high-frequency words than low-frequency words, for both hits and false alarms. Given all the published and unpublished data of which we are aware, we are confident that there is indeed an effect of word frequency on proportion Know judgments such that Know judgments are made more often to high-frequency words than to low-frequency words.

Comparisons With Other Models

Recently a number of mathematical models have been proposed that can produce the mirror effect of word frequency (e.g., Hirshman & Arndt, 1997; Murdock, 1998) including the Attention Likelihood Theory (ALT) of Glanzer and his colleagues (e.g., Glanzer & Adams, 1985, 1990; Glanzer et al., 1993; Kim & Glanzer, 1993; see also Murdock, 1998). ALT posits that old-item differences result from differential attention across conditions during study, resulting in differential marking or tagging of features in the low-frequency condition compared with the high-frequency condition and then rescaling the strength or familiarity to log likelihood ratios to determine the placement of the distributions of Old and New, high- and low-frequency words. According to Hintzman and his colleagues (e.g., Hintzman & Curran, 1997; Hintzman, Caulton, & Curran, 1994), there is a problem with attempts to explain the mirror effect by rescaling the familiarities of high- and low-frequency words to roughly equate bias for the two categories of items (e.g., Gillund & Shiffrin, 1984; Glanzer & Adams, 1990; Glanzer et al., 1993; Hintzman, 1994). Specifically, Hintzman, Caulton, and Curran (1994) ruled out a late, consciously controlled rescaling process, suggesting that any decision concerning word class must occur very rapidly and automatically.

In defense of ALT, it should be noted that it can explain the mirror effect even when the values of word frequency vary continuously. On the other hand, like many other models of the mirror effect it does not provide an account of Remember–Know judgments. It is possible that ALT could

be extended such that it could be applied successfully to our data. However, such an extension would involve many completely new details and would become a substantially different theory.

There do exist signal detection accounts of Remember–Know judgments (e.g., Donaldson, 1996; Hirshman & Henzler, 1998; Hirshman & Master, 1997; Inoue & Bellezza, 1998). Our account of Remember–Know shares some features with a strength account in terms of Know judgments being based on familiarity; however, we do not think that a unitary account such as signal detection can explain all the Remember–Know results. For example, in Experiment 3, our participants were more accurate in list discrimination when they responded R than when they responded K, suggesting that R corresponds to the presence of episodic information as well as a greater familiarity. In the model of Hirshman and Arndt (1997), for example, there exist different criteria for different word classes, but the mechanism that enables participants to know a priori this assignment so as to shift more for one word class than another is not totally specified. In other words, it is not that signal detection is inconsistent with our results, but rather it does not explain them because it does not say why there are word class effects on memory. One could say that the SAC model is an application of signal detection to a more detailed model of memory.

Donaldson (1996) argued that Remember and Know simply reflect two different thresholds or criteria on the same memories. If that were true, then the accuracy of Remember judgments should be the same as the accuracy of overall recognition or Old judgments (i.e., combining Remember and Know judgments). To examine whether this prediction holds for existing data, Donaldson conducted a meta-analysis of many data sets showing that, indeed, A' Remember and A' Recognition (or d' Remember and d' Recognition) are highly correlated (*r*s of .95 and .96).

Although these data can be taken as support for a double threshold model for Remember–Know, it is not unambiguous support. First, other models can also predict such a positive correlation. For example, in the SAC model, the activations of the word and event nodes are positively correlated, thereby also producing positive correlations between Remember accuracy and overall accuracy. Second, the correlation of accuracies itself is not the strongest test of the two-threshold model of Remember–Know. The two-threshold model makes the stronger prediction that the accuracies of Remember and overall Recognition should be identical (i.e., correlated with Slope 1 and Intercept 0). Yet, in Donaldson's data set, the slopes were significantly less than 1 and the intercepts significantly more than 0. For A', the slope was 0.80, different from 1, $t(38) = 5.3$, and the intercept was 0.15, different from 0, $t(38) = 4.6$. For d', the slope was 0.87, different from 1, $t(38) = 2.8$, and the intercept was 0.34, different from 0, $t(38) = 3.7$. Thus, the accuracies of Remember judgments and overall recognition, while correlated, are not identical. In other words, even in the data that Donaldson took as strong support for the two-threshold model, there is evidence against it.

Finally, although there exist other computational models that explain the mirror effect of word frequency that do not involve rescaling (e.g., McClelland & Chappell, 1998; Shiffrin & Steyvers, 1997), we know of no other formal model that simultaneously accounts for Remember versus Know judgments, for the mirror effect for normative word frequency, and for experimental variations in word frequency while maintaining a small set of parameters, importing the parameter values from previous modeling efforts in other domains whenever possible. Moreover, we fit individual participant profiles using only two parameters per participant.

Gardiner and Java (1990, 1991) have suggested that K judgments arise from a separate perceptual representational system and that R judgments arise from a declarative system (but see Richardson-Klavehn, Gardiner, & Java, 1996, for a more current view). Knowlton and Squire (1995) and Richardson-Klavehn et al. have argued that R and K judgments are based on declarative processes, K judgments reflecting semantic memory and R judgments reflecting semantic plus episodic memory (see Knowlton, 1998, for a review). SAC can be seen as formally instantiating the aforementioned interpretation by representing semantic information as word nodes and episodic information as event nodes. On the other hand, unlike Knowlton and Squire and others, we do not see the need to posit a separate implicit memory system. Elsewhere we provide evidence that implicit memory effects may be understood as effects operating on the word node (Erickson & Reder, 1998).

Accounting for Other Mirror Effects

The goal of this article was to account for the mirror effect of word frequency within the same framework that had also explained feeling-of-knowing results (e.g., Reder & Schunn, 1996; Schunn et al., 1997) and to test our controversial prediction that a mirror effect should also occur for Remember versus Know judgments as a function of word frequency. The strength of the model is its ability to account for these results using the same assumptions and even the same parameter values across experiments and tasks. Nonetheless, one might wonder whether SAC can account for the other mirror effects that are described in the literature, for example, the list-length and the list-strength mirror effects (Ratcliff, Clark, & Shiffrin, 1990). The list-length mirror effect refers to the result that recognition tests of longer lists produce fewer hits and more false alarms than shorter lists; the list-strength mirror effect refers to the result that items that are practiced more (get stronger) produce more hits and fewer false alarms than lists that are practiced less.

The list-length mirror effect is explained rather naturally using the assumptions of the SAC model. When more items are studied on a list, there will be greater fan out of the context node (not the episode node, but the node that represents the features associated with studying the items on that experimental list—refer to Figure 1). At test, activation spreads from both the concept node of the test item (that may or may not have been studied on the list) and the node that represents the experimental context. If sufficient activation arrives at an episode node from these two sources, then a recollection (and Remember) response will be given; however, if the item was not studied or the fan out of the concept

node or context node is too great to allow enough activation to accrue at the corresponding episode node, a recollection does not occur. In other words, SAC predicts fewer hits for items from long lists because less activation gets to the episode node. (It would also predict more Know judgments than Remember judgments.) Finally, SAC predicts more false alarms occur because there is a greater reliance on the concept node since it is more difficult to access the episode node. In other words, participants will lower their word-node thresholds for longer lists because episode nodes are not being sufficiently activated, producing too few Old responses. This predicts that the rise in false alarms with longer lists would be due to a rise in Know judgments. Recall that the only parameters allowed to vary by participant are the thresholds. The amount of activation that accrues is completely specified, while thresholds are assumed to vary with individual and are probably affected by the situation as well.

The account offered by SAC for the list-strength mirror effect is also a straightforward extension of the existing assumptions. Our own Experiments 1 and 2 demonstrate that there are more hits for items presented more times. As an item is repeatedly presented, its episode node gets stronger and the link from the concept node to the episode node also strengthens, making it easier to recollect and give a Remember response or give more valid hits. Because our experiments varied the strength within the same list, one cannot compare false alarms for few versus many presentations. This paradigm contrasts with those designed to examine list-strength mirror effects. Those experiments typically use separate study lists to vary the number of presentations. Each test follows a list of different strength so that the false alarms can be contrasted with the strength of the items on a given list. In our experiments, testing was continual, and although one can compare the hit rates for items as a function of the number of presentations, the false alarms cannot be assigned to items of different strength. The SAC explanation for false alarms in a list-strength experiment involves the same assumption of differential reliance on the concept node (shifting the threshold as a function of perceptions of ease of accessing the episode node) as given previously to explain the list-length mirror effects for false alarms.

Conclusion

Two key features of this model enabled us to account for the data from both paradigms (mirror effects for word frequency and Remember–Know judgments). First, people are not able to distinguish between activation values that come from recent exposure and activation that comes from a buildup of prior exposures. The name of the model, SAC, standing for Source of Activation Confusion, refers to our inherent inability to determine source of activation; source must be inferred. Frequently our attributions are correct because we can retrieve a contextual trace that allows us to infer why something seems reasonably familiar; however, even though we may attempt to compensate for differences in activation value (familiarity) due to preexperimental word frequency, these adjustments are insufficient for the most part. Participants are more inclined to spuriously accept high-frequency words as old because they have a higher base level of activation. In our model, these spurious Old judgments are based on misattributions of familiarity. They are reported as Know judgments because they are based on the activation level of the word node.

This explanation for more false alarms for high-frequency words, specifically that they have a higher base level of activation than lower frequency words, is consistent with results in the literature. For example, it has been found that participants are significantly more likely to false alarm to a word if the word is primed with a subliminal (i.e., unconscious) flash prior to its test presentation (e.g., Jacoby & Whitehouse, 1989). That result is consistent with the view that it is an elevation in base-level activation that gives rise to the spurious attribution that the word was studied earlier. Indeed, Rajaram (1993) found that this type of brief flash led to an increase of Know responses, but not Remember responses, adding further support for this view.

The second key feature of our model is that the number of contexts associated with the word node reflects the number of prior contexts in which a word has been seen. Therefore high-frequency words will tend to have many more contextual associations than low-frequency words. This difference in number of (contextual) associations is often referred to as the *fan effect* (e.g., Anderson, 1974; Lewis & Anderson, 1976; Reder & Ross, 1983; Reder & Wible, 1984). The amount of activation spread across any link is a function of that link's strength relative to the sum of the strength of all the competing associations. Therefore, the amount of activation sent down a given link from a word with greater fan will necessarily be less. Our study manipulated the number of exposures to the words for both high- and low- (normative) frequency items and the strength of the episodic (event) nodes. The strength of the connection between the word node and the event node should vary only with experimental exposures, not preexperimental exposures; however, the amount of activation that reaches the episodic node from the word node when it is presented for test will depend on the relative strength of the link, not the absolute strength. Because low-frequency words have fewer competing associations, more activation reaches their event nodes, making it more likely that the event node will pass over threshold. This means that it is easier to give a veridical Old response for low-frequency words and also means that there should be more Remember responses for low-frequency words. This is precisely what we found.

In summary, these three experiments and computational simulations provide the first formal, mechanistic account of how human memory encodes and retrieves episodic information in a way that simultaneously predicts word frequency patterns for Remember–Know judgments and the mirror effect for hits and false alarms. Given that the model also accounts for a variety of other phenomena, for example, feeling-of-knowing judgments in a continuous learning procedure, negative priming effects in a continuous selection and identification procedure, we feel confident that models

that share these architectural features are close to an accurate functional description of the operation of human memory.

References

Anderson, J. R. (1974). Retrieval of propositional information from long-term memory. *Cognitive Psychology, 6,* 451–474.

Anderson, J. R., & Bower, G. H. (1973). *Human associative memory.* Washington, DC: Holt, Rinehart & Winston.

Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought.* Mahwah, NJ: Erlbaum.

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science, 2,* 396–408.

Ayers, M. S., & Reder, L. M. (1998). A theoretical review of the misinformation effect: Predictions from an activation-based memory model. *Psychonomic Bulletin & Review, 5,* 1–21.

Bjork, R. A. (1975). Retrieval as a memory modifier. In R. Solso (Ed.), *Information processing and cognition: The Loyola Symposium* (pp. 123–144). Hillsdale, NJ: Erlbaum.

Bowler, D. M., Gardiner, J. M., & Grice, S. (1998). Episodic memory and remembering in high-functioning individuals with autism. *Journal of Cognitive Neuroscience, SS,* 49.

Chalmers, K. A., & Humphreys, M. S. (1998). Role of generalized and episode specific memories in the word frequency effect in recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 610–632.

Chappell, M., & Seth-Smith, M. (1999). *Know response rates reflect the word frequency mirror effect.* Manuscript in preparation.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155–159.

Cohen, J. D., MacWhinney, B., Flatt, M., & Provost, J. (1993). PsyScope: A new graphic interactive environment for designing psychology experiments. *Behavioral Research Methods, Instruments, and Computers, 25,* 257–271.

Coltheart, M. (1981). The MRC Psycholinguistic Database. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 33A,* 497–505.

Conway, M. A., & Dewhurst, S. A. (1995). Remembering, familiarity, and source monitoring. *Quarterly Journal of Experimental Psychology, 48A,* 125–140.

Donaldson, W. (1996). The role of decision processes in remembering and knowing. *Memory & Cognition, 24(4),* 523–533.

Erickson, M., & Reder, L. M. (1998). *More is better: The effects of multiple repetitions on implicit memory across long durations.* Unpublished manuscript.

Gardiner, J. M. (1988). Functional aspects of recollective experience. *Memory & Cognition, 16(4),* 309–313.

Gardiner, J. M. (1998, March). *Remembering and knowing, 1988–1998: Findings, theories, and problems.* Paper presented at the First Annual Tsukuba International Conference on Memory Consciousness and Memory, Tsukuba, Japan.

Gardiner, J. M., & Java, R. I. (1990). Recollective experience in word and nonword recognition. *Memory & Cognition, 18,* 23–30.

Gardiner, J. M., & Java, R. I. (1991). Forgetting in recognition memory with and without recollective experience. *Memory & Cognition, 19,* 617–623.

Gardiner, J. M., & Parkin, A. J. (1990). Attention and recollective experience in recognition memory. *Memory & Cognition, 18,* 579–583.

Gardiner, J. M., Richardson-Klavehn, A., & Ramponi, C. (1997). On reporting recollective experiences and "direct access to memory systems." *Psychological Science, 8,* 391–394.

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

Glanzer, M., & Adams, J. K. (1985). The mirror effect in recognition memory. *Memory & Cognition, 13,* 8–20.

Glanzer, M., & Adams, J. K. (1990). The mirror effect in recognition memory: Data and theory. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 16,* 5–16.

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

Glanzer, M., & Bowles, N. (1976). Analysis of the word-frequency effect in recognition memory. *Journal of Experimental Psychology: Human Learning & Memory, 2,* 21–31.

Guttentag, R. E., & Carroll, D. (1994). Identifying the basis for the word frequency effect in recognition memory. *Memory, 2,* 255–273.

Hilford, A., Glanzer, M., & Kim, K. (1997). Encoding, repetition, and the mirror effect in recognition memory: Symmetry in motion. *Memory & Cognition, 25,* 593–605.

Hintzman, D. L. (1994). On explaining the mirror effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 201–205.

Hintzman, D. L., Caulton, D. A., & Curran, T. (1994). Retrieval constraints and the mirror effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 275–289.

Hintzman, D. L., & Curran, T. (1997). Comparing retrieval dynamics in recognition memory and lexical decision. *Journal of Experimental Psychology: General, 126,* 228–247.

Hirshman, E. (1995). Decision processes in recognition memory: Criterion shifts and the list-strength paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 302–313.

Hirshman, E., & Arndt, J. (1997). Discriminating alternative conceptions of false recognition: The cases of word concreteness and word frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23,* 1306–1323.

Hirshman, E., & Henzler, A. (1998). The role of decision processes in conscious recollection. *Psychological Science, 9,* 61–65.

Hirshman, E., & Master, S. (1997). Modeling the conscious correlates of recognition memory: Reflections on the remember-know paradigm. *Memory and Cognition, 25,* 345–351.

Hockley, W. E. (1994). Reflections of the mirror effect for item and associative recognition. *Memory & Cognition, 22,* 713–722.

Huron, C., Danion, J., Giacomoni, F., Grangé, D., Robert, P., & Rizzo, L. (1995). Impairment of recognition memory with, but not without, conscious recollection in schizophrenia. *American Journal of Psychiatry, 152,* 1737–1742.

Inoue, C., & Bellezza, F. S. (1998). The detection model of recognition using know and remember judgments. *Memory & Cognition, 26,* 299–308.

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language, 30,* 513–541.

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

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

Joordens, S., & Hockley, W. (1999). *Recollection and familiarity through the looking glass: When old does not mirror new.* Manuscript submitted for publication.

Kim, K., & Glanzer, M. (1993). Speed versus accuracy instructions, study time, and the mirror effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 638–652.

Kinoshita, S. (1995). The word frequency effect in recognition

memory versus repetition priming. *Memory & Cognition, 23,* 569–580.

Knowlton, B. J. (1998). The relationship between remembering and knowing: A cognitive neuroscience perspective. *Acta Psychologica, 98,* 253–265.

Knowlton, B. J., & Squire, L. R. (1995). Remembering and knowing: Two different expressions of declarative memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 699–710.

Kučera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English.* Providence, RI: Brown University Press.

Lewis, C. H., & Anderson, J. R. (1976). Interference with real world knowledge. *Cognitive Psychology, 7,* 311–335.

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.

Mandler, G. (1980). Recognizing: The judgement of previous occurrence. *Psychological Review, 87,* 252–271.

Mather, M., Henkel, L. A., & Johnson, M. K. (1997). Evaluating characteristics of false memories: Remember/know judgments and memory characteristics questionnaire compared. *Memory & Cognition, 25,* 826–837.

McClelland, J. L., & Chappell, M. (1998). Familiarity breeds differentiation: A participantive-likelihood approach to the effects of experience in recognition memory. *Psychological Review, 105,* 724–760.

Murdock, B. B. (1998). The mirror effect and attention-likelihood theory: A reflective analysis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 524–534.

Rajaram, S. (1993). Remembering and knowing: Two means of access to the personal past. *Memory & Cognition, 21,* 89–102.

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

Reder, L. M., & Anderson, J. R. (1980). A partial resolution of the paradox of interference: The role of integrating knowledge. *Cognitive Psychology, 12,* 447–472.

Reder, L. M., Nhouyvanisvong, A., Schunn, C. D., Ayers, M. S., Angstadt, P., & Hiraki, K. (1997). Modeling the mirror effect in a continuous remember/know paradigm. *Proceedings of the Nineteenth Annual Meeting of the Cognitive Science Society* (pp. 644–649). Mahwah, NJ: Erlbaum.

Reder, L. M., & Ross, B. H. (1983). Integrated knowledge in different tasks: The role of retrieval strategy on fan effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 9,* 55–72.

Reder, L. M., & Schunn, C. D. (1996). Metacognition does not imply awareness: Strategy choice is governed by implicit learning and memory. In L. M. Reder (Ed.), *Implicit memory and metacognition* (pp. 45–77). Hillsdale, NJ: Erlbaum.

Reder, L. M., & Wible, C. (1984). Strategy use in question-answering: Memory strength and task constraints on fan effects. *Memory & Cognition, 12,* 411–419.

Richardson-Klavehn, A., Gardiner, J. M., & Java, R. I. (1996). Memory: Task dissociations, process dissociations and dissociations of consciousness. In G. Underwood (Ed.), *Implicit cognition* (pp. 85–158). Oxford, England: Oxford University Press.

Schunn, C. D., Reder, L. M., Nhouyvanisvong, A., Richards, D. R., & Stroffolino, P. J. (1997). To calculate or not to calculate: A source activation confusion model of problem familiarity's role in strategy selection. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23,* 3–29.

Shepard, R. N., & Teghtsoonian, M. (1961). Retention of information under conditions approaching a steady state. *Journal of Experimental Psychology, 62,* 302–309.

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

Strack, F., & Forster, J. (1995). Reporting recollective experiences: Direct access to memory systems? *Psychological Science, 6,* 352–358.

Stretch, V., & Wixted, J. T. (1998). On the difference between strength-based and frequency-based mirror effects in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 1379–1396.

Tulving, E. (1985). Memory and consciousness. *Canadian Psychologist, 26,* 1–22.

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

Received July 31, 1998
Revision received September 21, 1999
Accepted September 21, 1999 ■