Modeling word frequency effects in a continuous remember/know judgment paradigm

L. M. Reder[1], A. Nhouyvansivong[3], C. D. Schunn[1], M. Ayers[1], P. Angstadt[2], K. Hiraki[4]
[1]Department of Psychology & [2]Department of Computer Science
Carnegie Mellon University, Pittsburgh, PA 15213
[3] Educational Testing Service, Princeton, NJ
[4] Electrotechnical Laboratories, Tsukuba City, Japan
{reder, schunn, ayers, paige}@cmu.edu
anhouyvansivong@ets.org & hiraki@etl.go.jp

Abstract

Words of varying pre-experimental frequency were presented up to 10 times each. On each presentation, three responses were allowed-- *new, remember, and know*--the last for words that seem familiar, but give no conscious recollection of an earlier presentation. A novel pattern of results was predicted by the SAC memory model. SAC used the same parameter values used in fits to other tasks and provided good fits to the participants' remember and know responses.

Introduction

A dominant goal in cognitive science is to develop a theory of cognitive behavior within a unified framework that can explain a broad spectrum of human behavior without the necessity of postulating a new theory *denovo* each time a new task is to be explained. In this paper we describe a theory that has such a goal, to account for individual subject performance at a very detailed level of analysis, for a wide variety of cognitive phenomena. The model is restricted to trying to account for phenomena associated with *declarative memory* (cite Anderson, Squire). We have tested the SAC (*Source of Activation Confusion*) model of memory in a variety of research paradigms, and have been gratified by precise fits of theory to data without the need to postulate many new variables to fit a large quantity of data points in a qualitatively different task. We have recently reported our efforts to use this model to explain feeling of knowing and strategy selection decisions (Reder & Schunn, 1996; Schunn, Reder, Nhouyvansivong, Richards & Stroffolino, 1997). In this paper we describe a more recent test of the model's generality that extends the model to a new domain and tests the model's novel theoretical predictions against empirical data and the degree of fit to the behavioral data. The empirical test of novel predictions enables us to examine our theoretical constraints using both at the conceptual level and at the level of specific parameter values from previous model fits.

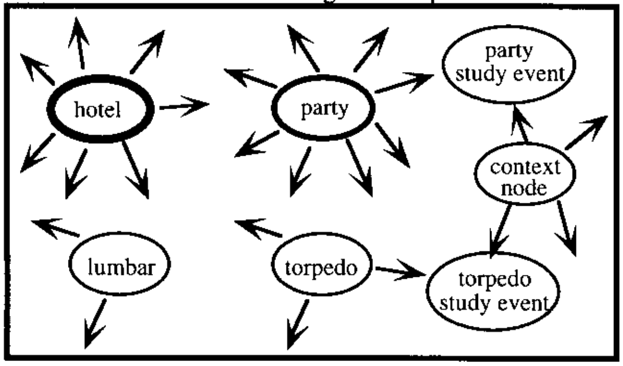
The domain that we have chosen to explore is called *The Mirror Effect* (Glanzer, Adams, Iverson & Kim, 1993), using the *Remember/Know Paradigm* (ref) as a magnifying glass that enables more fine grain predictions. The Mirror Effect refers to the phenomenon that two distinct classes of items (e.g., high and low frequency words) produce opposite orderings in liklihood to respond "old" in recognition tests, depending on whether the item had actually been studied. That is the "hits" (correct recognition judgments for presented items) are higher for low frequency words than high frequency words, while the "false alarms" (spurious recognition judgments for items not studied) are 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 that 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.

The Remember/Know Paradigm was first developed by Tulving (1985) to explore the recollective component of memory and has been a popular paradigm among researchers who subscribe to the view that there are two processes for recognition judgments (e.g., Mandler, 1980; Jacoby & Dallas, 1981) or that there exist multiple memory systems (e.g., Tulving, Schacter, Squire & McNaughton, refs). In this paradigm, participants study a list of words and are asked to make new/old judgments like in a standard recognition test of studied words. The difference is that after participants respond "old" to words that they believe had been presented on the list, they are asked to decide whether this "old" judgment is based on a recollective experience in which they can actually recall having seen the word presented on the list (in their "mind's eye") or whether they are basing this "old" decision on a sense of familiarity, i.e., the subject does not really remember seeing the word on the list, but just "knows" it must have been presented because it seems so familiar.

The reason that we were drawn to investigating the mirror effect, varying word frequency, and the Remember/Know paradigm is that our theory of memory makes novel predictions that integrate across these two paradigms, specifically addressing what happens to the Mirror Effect for words of different frequency when the recognition task requires participants to discriminate among old responses, i.e., make Remember/Know distinctions. The reason that we claim that these predictions are novel is that there are already some experiments in the literature that do not support our theoretical predictions; however, we felt sufficiently confident about the validity of our model[i] that we believed further tests of this prediction were justified. The prediction that we make is that there will be more "remember" judgments for low frequency words that are old, but that there will be more "know" responses for high frequency words than low frequency words, regardless of whether or not the word had actually been studied.[ii]

Theories are not difficult to generate that explain how a person correctly identifies that a word was studied or explain how a person correctly rejects a lure as not studied. Of more theoretical interest is to explain, without making additional assumptions, why people will incorrectly accept some not presented items as studied (false alarms) and why they will fail to recognize some items that were studied. Like other dual process models of memory, we assume that there are two ways to make a recognition judgment, one based on familiarity of the word and one based on retrieval of the encoding event. SAC postulates a node to represent the concept of the studied word and another node to represent the encoded memory episode. Figure 1 illustrates the memory representation that we assume. The familiarity of the word/concept node is affected by recent study but it is also affected by how frequently it has been seen in previous contexts (i.e., pre-experimental word frequency will affect this feeling of familiarity). The postulation that previous exposure will affect tendency to respond "old" motivated our interest in experiments that use pre-experimental word frequency as a factor. In our view, an accurate recognition judgment is based on the retrieval of the study event node, while responses based on the word node are error prone.

Figure 1: An example semantic network representation of nodes involved in a word recognition experiment.



Note in Figure 1 that higher frequency words not only have a higher starting strength or familiarity from more previous exposures, but that they also have more pre-experimental associations from all the contexts in which the word has been seen. According to SAC[iii] the amount of activation that can spread from a concept node to the event node must be divided among all links that *fan* out from the node. If we assume that all links have equal strength, then the amount of activation that can reach the event node is much less in the case of high frequency words than low frequency words because the former has much more competition for the activation than the latter. This type of theory would explain the Mirror Effect as follows: high frequency words have a higher strength/familiarity of the word node and hence there will be many "old" responses based on the word node, regardless of whether or not the word had actually been studied in the experiment. This will cause more false alarms for high frequency words than low frequency words, a common result. On the other hand, the greater *fan* out of high frequency words means that it is more difficult to send activation from the word node to the node that encoded the episodic study event, leading to fewer correct recognitions of high frequency words (fewer hits), the mirror result of the greater false alarms.

Until recently it seemed difficult to test the plausibility of this type of theoretical account; however, the advent of the Remember/Know paradigm has made it possible to bring evidence to bear on our predictions. Words whose concept nodes have greater strength should elicit more "know" judgments, i.e., higher frequency words should elicit more "know" responses regardless of presentation on the study list. This result has not been found in prior research (that we were able to find). As explained above, SAC also predicts more "remember" responses for low frequency words than for high frequency words, provided that the word had actually been studied.

In order to provide a rigorous test of our theory, we modified the traditional experimental paradigm to enable more precise or richer predictions. Specifically, we crossed pre-experimental frequency with experimental word frequency because our model also allows us to predict how much strengthening and forgetting there should be as a function of number of presentations of a word and the delay since it was last seen. So we opted to use a continuous recognition paradigm in which participants were required to make a remember /know/new judgment each time a word was presented. Each subject received a unique sequence of words and the same sequence was given to the computer simulation that predicted an specific individual's judgments for each word, each time the word was presented: new vs. remember vs. know. This enabled us to compare the observed proportion of each type of judgment for each word on each appearance of a given word with the predicted proportion generated by the computer simulation of SAC. More details of the model and how the precise model fits were generated are described after the experimental results are reported.

Method

Participants. The participants were 28 CMU undergraduates taking part in the experiment for course credit.

Materials and Procedure.

This experiment employed a continuous recognition paradigm (e.g., Shepard & Teghtsoonian, 1961). This design does not have the separate study and test phases typically found in memory experiments. Instead, the words are continuously presented for judgment, and the participants have to keep track constantly of which words have been presented and which words have not.

Within this paradigm, we manipulated two factors, pre-experimental word frequency and experimental presentation frequency. The first factor had two levels, using 192 low frequency and 192 high frequency words selected from the MRC psycholinguistic database (Coltheart, 1981). Low and high frequency words had Kucera and Francis normative mean frequency counts of 1.6 and 142, respectively, which were comparable to those used by Gardiner and Java (1990).

The second factor, presentation frequency involved randomly assigning words from each frequency category to be presented either 10, 5, 3 or 1 times, with N's of 8, 4, 4, and 160 respectively. This produced a total of 272 trials. The presentation order of the words was random.

The stimuli were presented to the participants on the computer over a single 25 minute session. The participants were asked to read each word silently and make one of three responses: "new" if they thought that the word had not been presented previously in the experiment; "R" if they recognized the word as having been presented earlier in the experiment and *had conscious recollection* of reading it earlier; or "K" if they recognized the word from earlier in the experiment but *did not have conscious recollection* of reading it earlier. Note that this differs from most remember/know experiments where participants first made new/old judgments before making remember/know judgments for "old" responses. We used this procedure in order to get the participants' first impressions. They were told to make the judgment as quickly as possibly without sacrificing accuracy.

To help participants understand the difference between the R and K responses, they were given real-world examples taken from Gardiner and Java (1990). In addition to the examples of remember and know experiences, it was stressed that the difference in the responses was not of memory strength, but rather of two different states of memory. Knowing did not necessarily entail a poorer memory. After the examples were presented to them, the participants were required to give two additional examples of their own to establish that they had understood distinction.

Results and Discussion

Six participants were dropped from the analyses: two due to procedural errors, and 4 misunderstood the distinction between R and K responses. The level of significance for this experiment was set to p < .05, unless otherwise noted.

Table 1 shows the mean probabilities of R and K responses for each presentation number for both low and high frequency words. The overall recognition was computed as the sum of R and K responses. Note that presentation one entails the lure trials for which the correct response was "new". Thus, these probabilities represent the false alarm rates. Presentations 2 – 10 then constitute the *old* trials. The overall hit rates were computed as the mean of the probabilities from presentation 2 – 10.

Table 1: Proportion of R and K responses as a function of word frequency and presentation number.

	1	2	3	4	5	6	7	8	9	10	(2-10)
Low											
R	.01	.49	.69	.79	.83	.89	.87	.91	.92	.92	.81
K	.04	.38	.28	.19	.16	.11	.13	.09	.07	.09	.17
High											
R	.03	.38	.52	.67	.73	.77	.79	.86	.84	.80	.71
K	.13	.44	.39	.30	.24	.23	.19	.12	.15	.16	.25

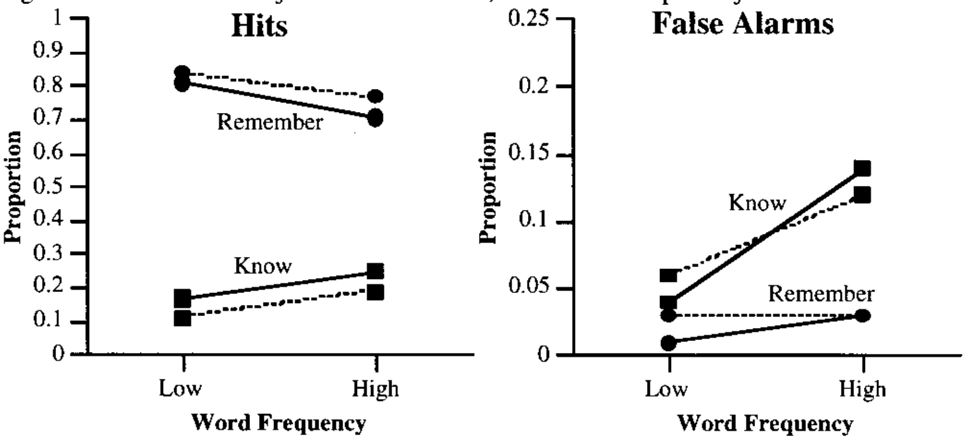
A separate repeated measures ANOVA was carried out for the hit and false alarm rates. For the hit rates, the ANOVA revealed a main effect of word frequency, F(1,21) = 7.40, MSe = 0.024 such that low frequency words were recognized more than high frequency words. Discriminability (*d'*) scores also show this difference. Low frequency words were better discriminated than high frequency words, *d'* of 4.17 and 3.02, respectively. There was also a main effect of presentation number, F(1,21) = 25.20, MSe = 0.049. This is evident in the increase in hit rates from presentation 2 to presentation 3. After presentation 3, participants appear to be at ceiling. The main effect of response was also significant, F(1,21) = 49.26, MSe = 60.45, such that there were more R responses than K responses. The interaction of word frequency by response type was significant, F(1,21) = 31.86, MSe = 1.70 as was the interaction between presentation number and response type, F(8,168) =29.8, MSe = 1.40. The word frequency by presentation number interaction was marginally significant, F(8,168) = 1.90, MSe = 0.002, p < .10. The three-way interaction was not significant, F(8,168) = 1.20, MSe = 0.021.

Of most interest to us is the word frequency by response type interaction. Figure 2 shows this interaction. Note that as predicted and consistent with the previous findings, R responses were greater for low frequency words than for high frequency words, t(21) = 4.47. However as our model predicts, this pattern is reversed for K responses. There were more K response for high frequency than for low frequency words, t(21) = 3.52.

The ANOVA conducted on the false alarms revealed a main effect of word frequency, F(1,21) = 30.26, MSe = 0.087, such that participants made more false alarms to high frequency words than to low frequency words. 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, as was the word frequency by response type interaction, F(1,21) = 16.62, MSe = 0.027. This last interaction is shown in the right panel of Figure 2. As predicted by SAC, a contrast between low frequency and high K false alarms revealed a reliable difference, t(21) = 8.02. The contrast between R responses for low frequency and high frequency words was also significant, t(21) = 2.26.

These findings are in agreement with SAC's predictions: The hit rates were greater for low frequency words than for high frequency words. The reverse pattern was found for the false alarm rates. Of more interest to us was the dissociation between the R and K responses due to word frequency. The proportion of R responses were greater for low frequency words than for high frequency words. This result, predicted by our model, is consistent with what other researches have found. The pattern of results for K responses was in the opposite direction. As predicted by our model, there were more K responses for high frequency words than for low frequency words. This result has not been found by previous research.

Figure 2. Data in filled objects and solid lines, simulation in open objects and dotted lines.



Simulation of Experiment Data

In this section, we present a simulation of the data from our experiment as a test of SAC's precise predictions. The computer simulation was given as input the same words presented to each participant. Since the presentation order of the words was randomly determined for each particpant, a separate simulation was conducted for each participant. This precise yolking of the simulation to participants was important because on a given trial the expected activation level for a word would vary depending on the exact sequence of trials. That is, for each participant on a given trial, the number of links, the current activation, and strength of the presented word would be different from any other participant's values. The simulation outputs a probability of responding R and K on each trial. We will now step through the process by which that probability is determined.

At the beginning of the simulation, each participant's simulation is identical: the context node and the nodes for all of the words to be presented in the experiment are assumed to already exist, and the nodes for the study events are assumed not to exist (i.e., these study events are novel). The initial base-line strength of the word nodes are determined by their respective Kucera and Francis frequency counts, using a power-law learning function (i.e., raising the word-frequency to an exponent). The pre-existing experimental context base-level strength and fan are set to a constant amount, the amount being irrelevant to the simulations of the recognition process. When a word is seen for the first time in the experiment, a study event node is created for that word, as are the links from the word and context nodes. The initial base-line strength of the study event node and of links to it are determined by our standard learning and decay parameters (presented below).

Increases and decreases in each node's base-line strength change according to a power function:

$$B = c \Sigma t_i^{-d} \qquad (1)$$

where B is the base level activation, c and d are constants, and t_i is the time since the i^{th} presentation. This function captures both power law decay of memories with time, and power law learning of memories with practice.

In addition to the base or resting level of activation of a node, there is also the *current activation* level of a node. The current level of a node will be higher than its base-line whenever it receives stimulation from the environment. This *current* activation decays exponentially towards the base level. Let A represent the current level of activation and B represent the base level of activation. Then, the decrease in *current* activation will be:

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

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. In our simulations, ρ is set to 0.8. Thus, current activation drops quite rapidly, and only has noticeable effects on the trial on which it became activated, and perhaps the trial immediately thereafter.

Activation spreads between nodes via links. For example, links connect nodes representing the words to nodes representing the study event. These links will vary in strength depending on how often the word has been seen in that context. Strength of links also depends on the delay between exposures. Specifically, link strength is determined by a power function given by:

$$S_{s,r} = \Sigma t_i^{-d_L} \qquad (3)$$

where $S_{s,r}$ is the strength of the link from the node s to node r, t_i is the time since the i^{th} co-exposure, and d_L is the decay constant for links.

On each trial, all nodes representing the study event are activated by a constant amount. We assume that a basic perceptual process activates these nodes. For example, when the word *torpedo* is presented for the third time, the *torpedo* word node and the context node are activated (see Figure 1). Activation then spreads along the links from the word and context nodes to all connected nodes (e.g., the node representing the study event).

The amount of activation that is sent depends on the activation level of the source (sending) node and on the strength of the link from the source node to the receiving node, relative to the strength of all other links emanating from the same source node. The change in activation of some node r is computed by summing the spread of activation from all source nodes s directly connected to node r according to the equation:

$$\Delta A_r = \Sigma(A_s * S_{s,r} / \Sigma S_{s,i}) \qquad (4)$$

where 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. Equation 4 is very similar to one used by Anderson (1993) to account for data in fan effect paradigms (e.g., Anderson, 1974).

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. Note that the activation of the study event node is not just affected by the amount of activation it receives from the word node. It also gets stronger (has a higher base level strength) each time it is studied. So activation at the event node on a trial will be stronger when it has been presented many times, both because the event node itself is strengthened and because the word node is sending more activation -- the link from the word node is strengthened and even the word node has a somewhat increased base strength.

We assume that Remember decision involves a fixed activation threshold with normally distributed noise. Thus, 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. Specifically, this probability is computed by the formula:

$$P(R) = N[\ (A_E - T_E)/s_E\ ] \qquad (5)$$

where A_E is the activation of the event node, T_E is the participant's threshold for the study event node activation, s_E is the standard deviation of the study event node noise 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. Consequently, the probability of responding K is a calculated by the following formula:

$$P(K) = \{\ 1 - N[\ (A_E - T_E)/s_E\ ]\ \} * N[\ (A_W - T_W)/s_W\ ]\ . \qquad (6)$$

In essence, the probability of responding K is one minus the probability of the study event node passing over threshold times the probability of the word node being above its threshold. The probability of responding "new" is one minus the probability that the activation does not pass over the Know threshold.

After each trial, all the links strengths, node strengths and node activations are updated using Equations 1, 2, and 3. At this point, if a word is presented for the first time, then a new study event node would be created as well as the links connecting the new node to the word and context nodes. The nodes in the network are updated in this fashion regardless of whether the subject responds "new", R, or K.

The present simulation just described involves ten parameters. The values for each of these parameters are listed in Table 2. The ρ, d_N, and d_L parameters were the same parameter values used in a simulation of feeling-of-knowing phenomena (Reder & Schunn, 1996; Schunn et al., 1977). Because of differences in design and stimuli used in the experiments, the KN, KL, c_L, and σ_E parameters are new parameters not found in the previous simulations. For parsimony, c_L and c_N were given the same value.

However, in contrast to all the other values, which were held constant across participants, 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. While 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.

Table 2: SAC model parameters descriptions and values.

Parameter	Function	Value
K_N	convert K-F frequency to word node strength	0.3
K_L	convert K-F frequency to word fan	0.7
ρ	decay constant for current activation	0.8
c_N	node power-law growth constant	25
d_N	node power-law decay constant	0.175
c_L	link power-law growth constant	25
d_L	link power-law decay constant	0.12
T_E	Study event node decision threshold	36-308
σ_E	Study event node decision standard deviation	40
T_W	Word node decision threshold	46-80
σ_W	Word node decision standard deviation	8

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 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. The full, exhaustive combinatorial space of possible parameters was not searched. Instead, we used the same parameters from our earlier model fits (Reder & Schunn, 1996; Schunn et al, 1997) when possible, and iteratively tried a range of values for each of the new parameters. We selected the value on each parameter producing the lowest sum squared error.

The best fitting participant R thresholds ranged from 36 to 308, with a mean threshold of 97.7 (SD=56.3). The best fitting K thresholds ranged from 46-80, with a mean of 57.4 (SD=9.6). Using these values, the SAC model fit the data quite well, producing an r^2 of 0.98 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.

The fits of each type of response was also very good. For the R judgment probabilities, a fit of the SAC model's predicted probabilities to the participants' actual R judgment probabilities, produced an r^2 of 0.95. For the fit of the K responses, the r^2 was 0.86. Again, even after breaking down the recognition judgment into the R and K components, SAC accounted for a good portion of the variance. Figure 3 plots the empirical R probabilities on the left and model simulation on the right. Similarly, Figure 4 plots the K probabilities. Note that, consistent with the empirical data, R judgments are consistently 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 low frequency words.

Figure 3: The mean proportion of Remember responses by word frequency and presentation number. A) Data B) Simulation.

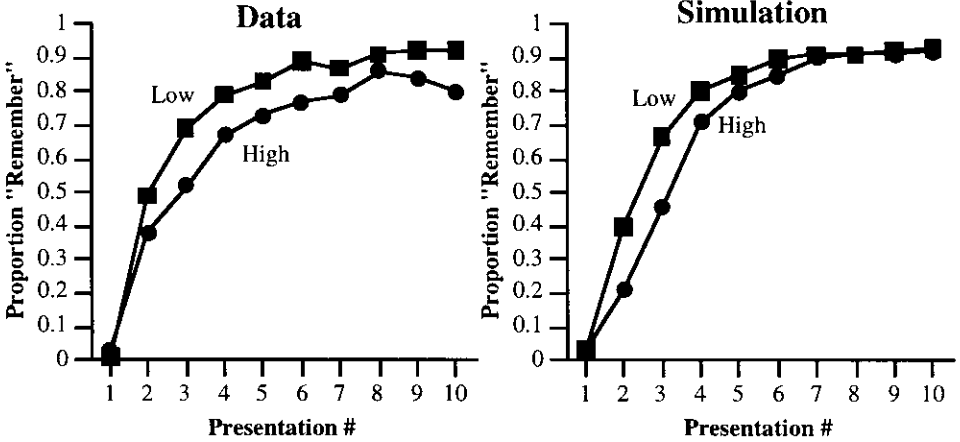
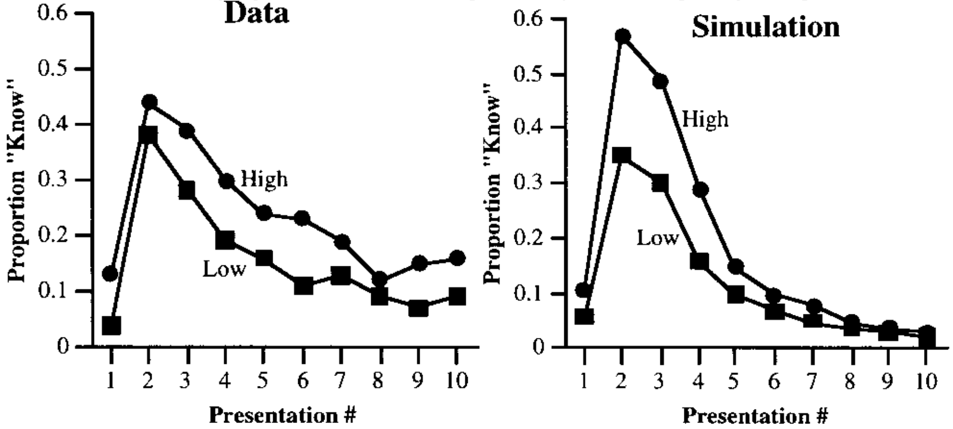


Figure 4: The mean proportion of Know responses by word frequency and presentation number. A) Data B) Simulation.



The right hand panels of each figure show that the model gives very good quantitative fit to the data. The fit of the model to the data in Figure 2 shows that the model also accounts for the predicted the pattern of results for hits and false alarms rates. Note that the model not only predicts the dissociation in R judgments (which is consistent with SAC predictions and previous findings), it also predicts the reverse dissociation for the K responses. This reverse pattern for K judgment is a novel finding that is accounted for by SAC.

In the case of false alarms, the model accounts for the increase in K judgment due to word frequency, which is a novel finding that was predicted by SAC. For the R false alarm judgments, the theory predicts no effect of word frequency; however, there were slightly more R false alarms for high frequency than for low frequency in the behavioral data. Overall, the simulation from the SAC model produced very good qualitative fits.

General Discussion

We were pleased with our ability to confirm our predictions, especially given that they went against the existing literature. There are probably other theories that could make this prediction, e.g., by postulating different thresholds for saying remember. However, we consider these empirical and modeling results strong support for our theory for several reasons: (1) Our explanation comes from assumptions that have been tested and confirmed in very different experimental paradigms; (2) We made these predictions prior to conducting the experiments: (3) we fit our data at a very fine grain size, crossing pre-experimental exposure and experimental presentation frequency; (4) we account for a lot of data using few parameters, many of which were estimated for previous research in a different domain.

All this said, we think it unlikely that this theoretical account is exactly right--no theory is likely to stand the test of time without modification. Nevertheless, we think this example of detailed fitting of behavioral data with precise theoretical predictions is the way theorizing should be done.

Acknowledgements

This work was supported by a grant from the National Institute of Mental Health, XXXXand in part by a grant from the Office of Naval Research, XXXXto the first author.

References

References follow everything else. Use a first level section heading for the references. Use a hanging indent style, with the first line of the reference flush against the left margin and subsequent lines indented by 1/8 inch. Please use standard APA-style references, and list references alphabetically. List all authors in the references, but abbreviate citations with four or more authors in the text, listing the first author followed by *et al.* Examples of conference paper, book, journal article, thesis, and book chapter references are provided below, respectively.

Resnick, L.B. (1983). Towards a cognitive theory of instruction. In S.G Paris, G.M. Olson, & H.W. Stevenson (Eds.), *Learning and Motivating in the Classroom* (pp. 6--38). Hillsdale, NJ: Lawrence Erlbaum Associates.

[i] This is not to say that we believe that SAC as currently specified is exactly right, nor that similar models articulated by others could not account for the data. Rather, we believe that this is a good approximation to the truth and that is among the most precise models that make such specific predictions and has been more rigorously tested than any model of which we are aware.

[ii] . The study by Gardiner and Java (ref) and Strack and Forster (ref) did not find an effect of word frequency on know judgments for either old words or new words.

[iii] Our theory shares this assumption with other activation based theories, most notably ACT-R (ref).