Journal of Experimental Psychology:
Learning, Memory, and Cognition
1995, Vol. 21, No. 2, 267-287

0278-7393/95/$3.00

Effects of Category Length and Strength on Familiarity in Recognition

Richard M. Shiffrin, David E. Huber, and Kim Marinelli
Indiana University

In most recognition models a decision is based on a global measure often termed familiarity. However, a response criterion is free to vary across lists varying in length and strength, making familiarity changes immeasurable. We presented a single list with a mixture of exemplars from many categories, so that the criterion would be unlikely to vary with length or strength of the category of the test item. False alarms rose with category length but not category strength, suggesting that familiarity does not change much with changes in strength of other items but grows when additional items are studied. The results were well fit by an extension of the search of associative memory (SAM) model presented by R. M. Shiffrin, R. Ratcliff, and S. E. Clark (1990).

The present article explores recognition memory. Previous research (e.g., Ratcliff, Clark, & Shiffrin, 1990) has focused on sensitivity of recognition, usually measured as d'. In the studies reported here, we examined the values of familiarity, measured by the probabilities of giving *old* responses to singly presented test items,[1] that most models assume underlie the participants' recognition judgments. We did so by varying category length and category strength rather than the more commonly manipulated list length and list strength.

Most models of old–new recognition memory utilize concepts loosely borrowed from the theory of signal detection (e.g., Banks, 1970). It is assumed that the *old–new* decision is based on a single numerical value, variously termed *familiarity* (adopted in the present article for convenience), *match, activation,* or the like. In different models, this value arises from different underlying processes, such as a match of two vectors (e.g., Murdock, 1982), a sum of retrieval strengths (e.g., Gillund & Shiffrin, 1984), or a sum of activations (e.g., Hintzman, 1988), but the source of the familiarity value need not be considered for the time being. It is assumed that the value of familiarity (F) when an item is tested has a distribution with a higher mean for a target (an item from the studied list) than for a distractor (an item not from the studied list). When accuracy is the response measure of interest, the participant is assumed to choose a criterion (C). An old response is given when F is greater than C, and a new response is given otherwise. The participant presumably chooses the value of C to be suitable for the kind of item tested, the kind of list studied, the kind of experimental context, and the payoffs of the experiment.

The situation is illustrated in Figure 1a: A list of a given length (number of different words) and strength (number of repetitions of each word) has been studied. The distribution on the right gives the familiarity values when a target is tested, and the distribution on the left gives the familiarity values when a distractor is tested. Suppose the participant places a criterion where the two distributions cross, at the familiarity value labeled X. The hit rate, denoted P(H), is the probability of responding old to a target; it equals the area to the right of the criterion under the right-hand distribution. (The hit rate equals 1 minus the miss rate.) The false-alarm rate, denoted P(F), is the probability of responding old to a distractor; it equals the area to the right of the criterion under the left-hand distribution. (The false-alarm rate equals 1 minus the correct rejection rate.) P(H) and P(F) comprise the data usually available from experiments. It is typical to assume, at least as an approximation, that the distributions are normal and have equal variance (as depicted in the figure). One can then use the hit rate and the false-alarm rate to calculate a measure of sensitivity, d', and a measure of relative criterion placement. The measure d' is the distance between the means of the distributions, divided by the common standard deviation. Although criterion placement is often given as β (i.e., the ratio of the two ordinates at the criterion value), we find it facilitates exposition to use standard deviation units (called C by Snodgrass & Corwin, 1988) for referring to the placement of the criterion in terms of the distance of the criterion from the point at which the distributions cross.

A typical study involves the variation of some variable such as list length (the number of different words in a list), item strength (the number of study repetitions of the tested word), or list strength (the total number of repetitions of all studied items). Of interest is the way in which d' varies across such manipulations. However, Figure 1 shows that there are a number of different ways that familiarity can change and still predict the same pattern of d' changes. Suppose, for example, that a longer list is used than that giving rise to the distributions of Figure 1a. Figures 1b, 1c, and 1d all show possible

Richard M. Shiffrin, David E. Huber, and Kim Marinelli, Department of Psychology, Indiana University.

This research was supported by National Institute of Mental Health Grant MH12717.

Correspondence concerning this article should be addressed to Richard M. Shiffrin, Department of Psychology, Indiana University, Bloomington, Indiana 47405. Electronic mail may be sent via Internet to shiffrin@indiana.edu.

[1] Recognition paradigms sometimes require participants to choose which of two presented items is the old one (e.g., Glanzer & Adams, 1985). Such tests tend to provide information about the relative oldness of two items and therefore are a somewhat more indirect measure of familiarity than a direct judgment of oldness. Thus, in these experiments, we used single-item tests and individual judgments of old versus new, leaving forced-choice paradigms for future research.

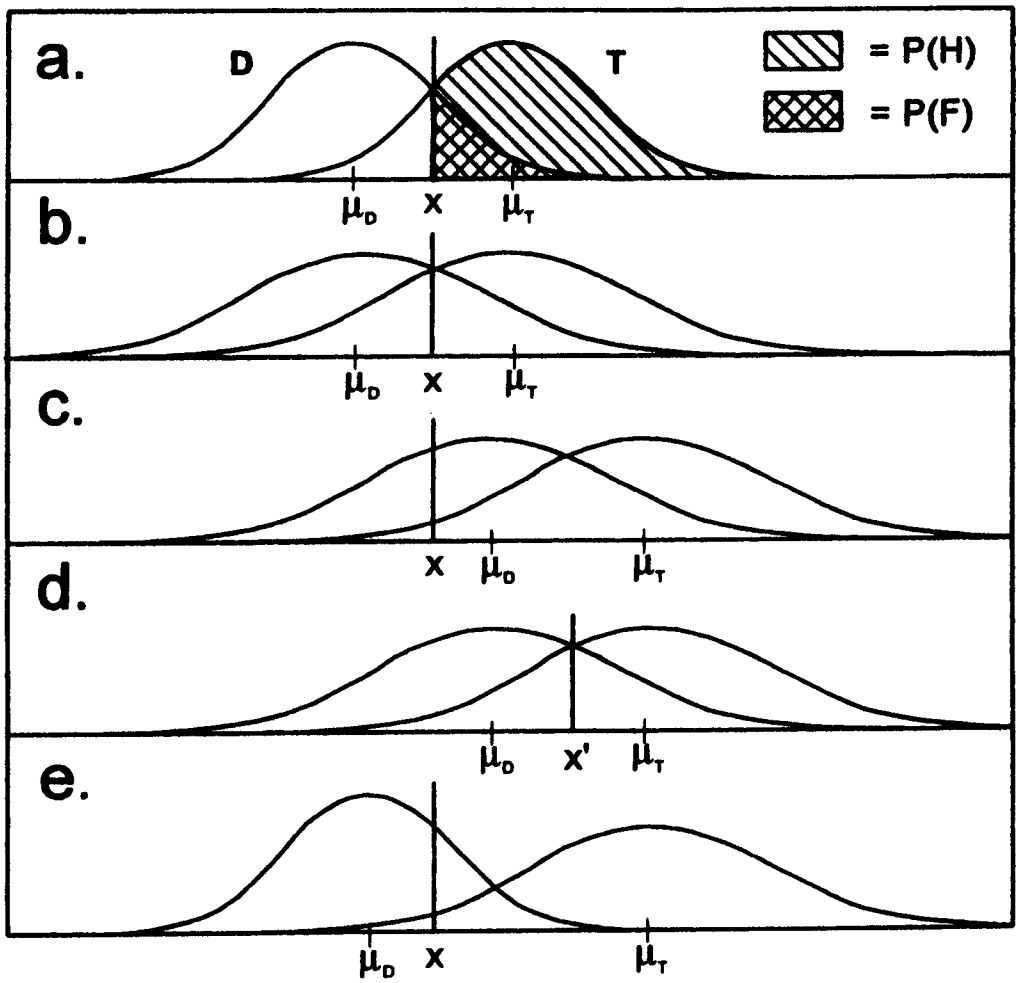

Figure 1. Examples of distributions of familiarity used to make a recognition decision. T represents the distribution for target tests, and D distractor tests, with means μ_T and μ_D, respectively. P(H) stands for hit rate; P(F) stands for false-alarm rate. The criterion above which an old response is given is labeled X and X' in different sections. Section a contains examples of distributions for a short list or category. Sections b, c, and d contain examples of distributions for a longer list or category under different models: Section b shows that extra length increases the variance, but not the means or criterion; section c shows that extra length increases both the means and variances, but not the criterion; section d shows that the distributions caused by increased length are as in section c, but that the criterion has increased accordingly. In section e, an increase in word strength, for a list or category of the same length as in section a, is depicted as leaving the distractor distribution unchanged but as increasing the mean and variance of the target distribution; the criterion is shown at the same point as in section a (appropriate for the category experiments in the present article).

distributions of familiarity for such a longer list. Figure 1b shows the case in which the target and distractor distributions have the same means as in Figure 1a, but have higher variance, producing a lower value of d'. Figure 1c shows the case in which both the means and variances of the target and distractor distributions rise relative to Figure 1a, producing a lower value of d'. Although the familiarities represented in Figure 1b and 1c are quite different, they produce the same value of d'.

It appears that one should be able to determine the changes in familiarity by reference to the values of P(H) and P(F). However, Figures 1a, 1b, and 1c are misleading because the criterion is shown at the same placement in all three panels. Because Figures 1b and 1c represent a longer list than Figure 1a, and because this difference would be readily apparent to a participant, the participant would be free to adjust the criterion to a new position. This is shown in Figure 1d: The distributions are those of Figure 1c, but the criterion has been moved upward so that the P(H) and P(F) values match those in Figure 1b. Thus, the hit and false-alarm rates can be used to determine the changes in familiarity only when the criterion can be assumed to remain constant across conditions. Finally, calculation of relative criterion placement (or β or any similar measure) does not allow determination of the changes in familiarity: Figure 1b and 1d have the same d' and the same

relative criterion placement, but different familiarity distributions.

These examples serve to illustrate our general point: When conditions are varied between lists, the participant is free to adjust the criterion, making it impossible to assess how familiarity changes between conditions. Because many of the extant models make explicit predictions about the absolute values of familiarity and make explicit predictions concerning how these values should change with experimental manipulations, we decided to carry out a study in which it would be plausible to assume that the criterion is not moved between conditions of interest. The basic idea involved presenting a single very long list of words for study (several hundred items), followed by an old–new recognition test. Embedded in the list were many categories of words. The words in each category were presented in a widely spaced fashion, such that participants were largely unaware of the categorical structure of the list and were seldom aware of the existence of any particular category. This allowed us to manipulate the variables we were exploring across categories within a given list.

For this paradigm to work, it is essential that there is preferential access to the stored members in the category of the test word (even when the participant is unaware of the length or strength of that category). If strength of activation of traces is based on similarity of the test word to the stored word, as incorporated in most models, then this requirement will be satisfied. The validity of this assumption can be tested by varying the lengths of categories and examining false alarms to distractors from each category; false alarms will rise with length only if there is preferential access to the category of the test item.

We manipulated two of the most common variables used to study recognition memory: length and strength. In the present study, we instantiated these variables as category length (the number of different words studied from a given category) and category strength (the number of presentations of the words from a given category). We hoped that the participant would adopt a criterion that was on the average the same for items tested from categories of differing length and strength. Although we have no independent test, the internal consistency of the results can provide some verification of this hypothesis.

Previous studies (e.g., Gillund & Shiffrin, 1984; Ratcliff et al., 1990) have varied list length and list strength, making it likely that criteria vary with these variables and making the absolute values of hit rates and false-alarm rates difficult to interpret. Therefore, we discuss only the implications of extant findings concerning the variations of sensitivity (i.e., d') with length and strength. The lowering of sensitivity as list length increases is a virtually universal finding. Whether a d' decrease with category length in the present study was to be expected depends on the model being examined. Many models predict d' to decrease with length because the extra items increase the target variance as much as the distractor variance, without changing the difference between the means. In an experiment in which length is varied across categories within one list, however, the variance increases only as a result of the small number of extra items in a given category, an increase that might be masked by the variance contributed by the many items from the other categories.

How sensitivity changes with list strength has been examined several times in recent years (e.g., Murnane & Shiffrin, 1991a, 1991b; Ratcliff, Clark, & Shiffrin, 1990; Shiffrin, Ratcliff, & Clark, 1990). It has been found that increasing the strength of some list items (by means of extra study time or increased numbers of spaced repetitions) does not reduce d' for other items (and may even slightly improve d' performance). These effects are of interest because most models do not predict such effects. Shiffrin et al.'s (1990) search of associative memory (SAM) model, which is a slight modification of Gillund and Shiffrin's (1984) SAM model, does predict such findings. This model may be extended to the present category study by adding the assumption that there is preferential activation of the stored items that are in the category of the test item (perhaps due to a larger value of associative similarity between a test item and stored items in its category)—that is, these items are activated to a higher degree than noncategory items. Thus, each category acts as a mini-list, with all the items from other categories contributing noise. This model predicts no change in d' with category strength.

In the present research, of course, our main interest was not in d', but in the way that the hit and false-alarm rates varied with category length and category strength. Assuming that the criterion did not change across length and strength conditions, the false-alarm rate was expected to be particularly diagnostic. Shiffrin et al.'s (1990) SAM model extended to categories predicts that the false-alarm rate will remain approximately constant as the strength of items in a category increases and that the false-alarm rate will rise as the length of a category increases. Most alternative models predict a different qualitative pattern, in which strength and length have the same effect on the false-alarm rate. A more detailed exposition of the predictions of various models is deferred to the Results and Discussion section.

The categories we used were of two very different types. One type was *semantic:* A prototype word was selected to be long and of relatively low natural language frequency (e.g., *butterfly*). In the experiment, we presented exemplars semantically related to the prototype for study. We did not present the prototype, in order to reduce the possibility of the participant noticing the categorical nature of the list. The category exemplars also tended to be long, low-frequency words. The other category was *orthographic–phonemic:* A short, relatively high-frequency, monosyllabic word was selected as the prototype. The exemplars all had the same vowel and the same vowel sound as the prototype, but they differed in either the first consonant or the last consonant. These categories are not among the standardized categories often used in memory studies, but we chose them because some pilot testing suggested they would not be noticed by participants during the study phase, but would nevertheless affect performance. We used these two different types of categories to assess the generality of the findings. We reasoned that if the results were similar for these two types of categories, it would be difficult to argue that the pattern was due to an idiosyncratic choice of items.

In all, we carried out four experiments in this category paradigm. The two experiments carried out first were designed in part to answer some additional questions not entirely germane to the present issues, and as a result they are not

ideally suited to make our main points. In the present article, we have labeled these studies Experiments 3 and 4, and a summary of their relevant results is presented at the end of the article in Appendix D. In Experiments 3 and 4, frequency judgments were required during study. In Experiment 3, only category strength, and not category length, was varied, and both immediate and delayed tests were included. In Experiment 4, both category length and strength were varied, and at test, frequency judgments were required first, and recognition judgments were required second. The essential point to keep in mind is that the results of these studies (i.e., those reported in Appendix D) replicate the results to be reported in the body of this article in all important respects, both qualitatively and quantitatively.

The two experiments we report now (Experiments 1 and 2) are identical in all respects save one. In each, as a word was presented for study, participants gave a rating of its pleasantness. This was done primarily to combat lapses of attention and to increase the probability that multiply presented words were stored more strongly. Following study, there was a long series of single-word tests. For each, the participant gave a judgment on a 6-point scale of his or her confidence that the word had been studied. Test words included studied exemplars and various types of nonstudied words, namely, nonstudied category prototypes, nonstudied category exemplars, and nonstudied words that were not members of any of the studied categories. In both studies, category lengths were 2, 6, and 9 words. In Experiment 1, strength was varied for the categories of length 6. In Experiment 2, strength was varied for the categories of length 2. In both studies, strength was manipulated by varying the number of spaced repetitions of a given exemplar.

The following illustrative example helps clarify the terminology used in our article: Suppose Category 1 has the members *A* and *B*; Category 2 has *C*, *D*, and *D*; Category 3 has *E*, *E*, *F*, and *F*; and Category 4 has *G*, *H*, *I*, and *J*. The letters stand for the presented words in each category. Category length refers to the number of different words in a category that are studied, regardless of the number of repetitions of a given word. Thus, Categories 1, 2, and 3 have a category length of 2, and Category 4 has a category length of 4. Word strength refers to the number of presentations of a word. Thus, in Category 2, Word C has a word strength of 1 and word D has a word strength of 2. Category strength refers to the average word strength of a category. Thus, the category strengths of Categories 1 and 3 are 1 and 2, respectively. It is often of considerable theoretical importance to hold word strength constant while varying category strength. In our example, this occurs for Words A and B versus C in Categories 1 and 2, and it occurs for Words E and F versus D in Categories 3 and 2.

Method

Experiment 1

Participants. The participants were 47 Indiana University students taking part in a 30-min session to satisfy part of an introductory psychology course requirement.

Apparatus. Presentation of words and collection of data were carried out for each participant on IBM-compatible personal computer systems.

Materials. There were 15 semantic categories, each consisting of a prototype and 11 exemplars. For these categories the prototypes and exemplars tended to be relatively long words with relatively low natural language frequency (Kuçera & Francis, 1967; Nusbaum, Pisoni, & Davis, 1984). The exemplars were chosen on the basis of semantic relatedness to the target, where relatedness was defined informally on the basis of intuition. A certain amount of pruning of both categories and exemplars took place after some pilot testing, in which the aim was to use categories that would not be noticed during study but would nonetheless affect performance. For example, one semantic category comprised the prototype *butterfly* and the exemplars *nectar, cocoon, monarch, flutter, metamorphosis, dragonfly, flitting, wings, camouflage, fragile,* and *caterpillar.*

There were 10 orthographic–phonemic categories, each consisting of a prototype and 11 exemplars. The exemplars were chosen on the basis of orthographic and phonemic relatedness to the prototype. The prototypes and exemplars were either all three-letter or all four-letter monosyllabic words generally of high natural language frequency. All the exemplars of an orthographic–phonemic category shared the same vowel sound with the prototype and also shared exactly one of the consonant clusters. An example of an orthographic–phonemic category is that formed for the prototype *sip;* it contained exemplars that differed from the prototype only in the first or last letter, but not both: *tip, lip, hip, dip, nip, pip* and *rip* (first letter differs), and *sin, sit, sis,* and *six* (last letter differs).

In addition to these prototypes and exemplars, two other classes of words were used. The first was termed *extra semantic items* and comprised 30 words similar in structure to the semantic category items but not obviously related to them or to each other. To control for primacy, recency, and short-term memory effects, 24 of these words were used as follows: Each session began with 10 of these words; the study list ended with 10 more of these words; and the test list began with 4 more of these words. The second class was termed *extra orthographic–phonemic items* and comprised 20 words similar in structure (i.e., monosyllabic, high frequency, three or four letters) to the orthographic–phonemic category exemplars but were based on different vowel sounds than any of these categories. The category prototypes, exemplars, and extra items are listed in Appendixes A and B.

Procedure. Participants viewed a single study list of 255 successive word presentations, followed by a recognition test. In the study list, each word appeared on a computer screen for 3 s. Within that time, we asked participants to enter a rating of the pleasantness of the presented word on a 6-point scale. We informed participants at the start of the session that some words would be repeated during study and that there would be a final recognition memory test.

The recognition test consisted of 149 successively presented words. The participants gave a rating on a 6-point scale (1–6) of their confidence that the word had been studied, with the neutral point between 3 and 4. Once a participant's responses had been entered, we presented the next test word immediately.

During study, exemplars (but no prototypes) of each of the 25 categories were presented, with the category members and repetitions of a given word randomly spaced between the 10 primacy and 10 recency buffer words. Five categories were assigned to each of five conditions. There were three length conditions: two, six, or nine exemplars studied once each. There were two strength conditions. One was termed *strong,* in which six exemplars were studied three times each. The other was a mixed-strength condition, termed *mixed,* in which three exemplars were presented once and three were presented three times. Three semantic categories and two orthographic–phonemic categories were assigned to each of these five conditions. For each participant, the exemplars studied from each category, the

Table 1
Probability of Responding Old for Category Type and Length–Strength Condition

Test items	Length–strength conditions					
	Length 0	Pure 1, length 2	Pure 1, length 6	Pure 1, length 9	Mixed, length 6	Pure 3, length 6
			Experiment 1			
Semantic						
Distractors	.074	.085	.121	.195	.124	.149
Prototypes		.206	.220	.284	.270	.248
Strength 1 targets		.794	.801	.773	.787	
Strength 3 targets					.957	.968
Orthographic–phonemic						
Distractors	.069	.096	.213	.207	.213	.229
Prototypes		.234	.191	.309	.309	.372
Strength 1 targets		.766	.750	.777	.824	
Strength 3 targets					.915	.952
			Experiment 2			
					Mixed, length 2	Pure 3, length 2
Semantic						
Distractors	.046	.108	.163	.196	.105	.105
Prototypes		.163	.333	.405	.176	.176
Strength 1 targets		.833	.853	.876	.817	
Strength 3 targets					.974	.974
Orthographic–phonemic						
Distractors	.069	.147	.206	.225	.152	.152
Prototypes		.176	.333	.382	.167	.216
Strength 1 targets		.838	.873	.828	.804	
Strength 3 targets					.931	.971

order of study, and the assignment of conditions to categories were separately randomized.

After the 4 buffer items starting the test list, words on the test list were presented in an order randomized for each participant. Two studied exemplars (targets), the nonstudied prototype, and two nonstudied exemplars (distractors) were tested from each category other than the mixed categories. For the mixed condition, two words of each presentation frequency were tested, in addition to the nonstudied prototype and the two distractors. In addition to these items, 10 nonstudied extra items were tested, at randomly chosen test positions: 6 of these were the remaining extra semantic items not serving as buffer items, and the other 4 were chosen randomly from the 20 extra orthographic–phonemic items. These distractors may be thought of as arising from categories of length zero.

Experiment 2

Experiment 2 was identical in all respects to Experiment 1 save the following: The strong condition consisted of two exemplars presented three times each (instead of six exemplars presented three times each, as in Experiment 1), and the mixed condition consisted of one word presented once and one word presented three times. As a result, there were only 155 words in the study list and 139 words in the test list. The test list was unchanged, except that for the mixed categories, only one word of each presentation frequency was available for test. This study involved 51 Indiana University students who participated to satisfy an introductory psychology course requirement.

Results and Discussion

The primary analyses were carried out on the hit probabilities, defined as confidence ratings 4 through 6, and the false-alarm probabilities, defined as confidence ratings 1 through 3. When we carried out sensitivity analyses, the hit and false-alarm probabilities were converted to d' for each participant and condition. If either P(H) or P(F) was zero, it was replaced by a value equal to the inverse of $2n$, and if either was 1, it was replaced by 1 minus the inverse of $2n$.

In many articles, the sensitivity results would be given and discussed first. In the present article, our primary concern is with the changes in the hit and false-alarm rates that occur with variations in category length and category strength, so we begin with these findings. (We note that in every figure in the body of the article, and in the Appendixes, error bars around p(*old*) values are used to indicate the standard error of the mean.)

In neither experiment did length or strength interact with category type (analyses of variance [ANOVAs] were carried out separately on the two experiments; the 12 interaction terms representing combinations of category lengths and word strengths and category strengths with category types for targets and distractors for Experiments 1 and 2 all had p values greater than .14), so the following results are summed across category type. (For reference, Table 1 gives the breakdown of the findings by category type.) The following statistical results are all given as t tests based on within-subject contrasts; one-tailed tests were used when the direction of an effect was an a priori prediction of the SAM model (any two-tailed tests are indicated).

Figure 2 gives the category-length results from Experiments 1 and 2 (zero length refers to tested extra semantic and extra

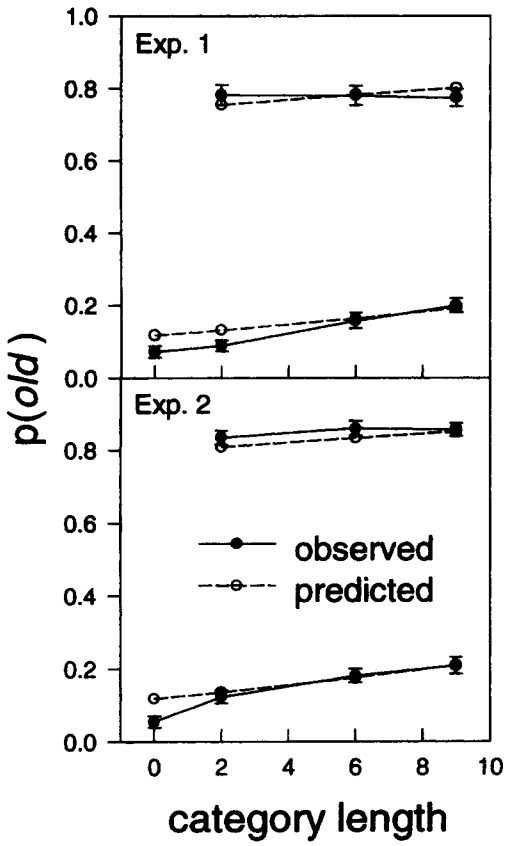

Figure 2. Hit and false-alarm rates as a function of category length for Experiment 1 (top section) and Experiment 2 (bottom section). Predictions of the search of associative memory model are the dashed lines. *p(old)* stands for the probability of responding old, and the error bars around each value represent the standard error of the mean.

orthographic–phonemic items). False alarms rose significantly with category length, $t(46) = 7.46, p < .001$, for Experiment 1 and $t(50) = 7.51, p < .001$, for Experiment 2. Hits exhibited an unanticipated slight downward trend in Experiment 1, but not significantly so; $t(46) = 0.37$, $p = .15$, and no trend in Experiment 2, $t(50) = 0.87, p = .20$. For reference, note that the trend was slightly upward in Experiment 4; this is reported in Appendix D.

Figure 3 gives the category-strength results for Experiments 1 and 2. False alarms did not vary with category strength, for fixed category length, $t(46) = 0.92, p = .36$ (two-tailed), for Experiment 1 and $t(50) = 0.00$, $p = 1.0$ (two-tailed), for Experiment 2. This is evidenced by the lowest horizontal line in the figures. There was a main effect of word strength for targets: Items presented three times had higher hit rates than did items presented once, $t(46) = 8.61$, $p < .001$, for Experiment 1 and $t(50) = 8.11, p < .001$, for Experiment 2. This is evidenced by the difference in level between the two upper lines in the figures. There was no significant effect of category strength for targets: The strength of the other items in the category had no effect on the hit rate when the strength of the target item was fixed, $t(46) = 1.64, p = .11$ (two-tailed), for Experiment 1 and $t(50) = -0.28$, $p = .78$ (two-tailed), for Experiment 2. This is evidenced by the horizontal nature of each of the two upper lines in the figures.

The contrast between the length and strength effects on false alarms is illustrated in Figure 4, which gives false-alarm rates as a function of the total number of presentations of all items in a category. False alarms clearly rose with length and also clearly did not rise with strength, at two different levels of performance, corresponding to short and long categories.

The various results reported above may be analyzed in a fine-grained fashion by examining in more detail the confidence ratings. One can analyze the entire distribution of ratings or, equivalently, analyze separately hits and false alarms defined by different cut points along the confidence scale. We looked at the results in both ways; because they are entirely consistent with the patterns reported above, they are not presented or discussed further. These claims may be

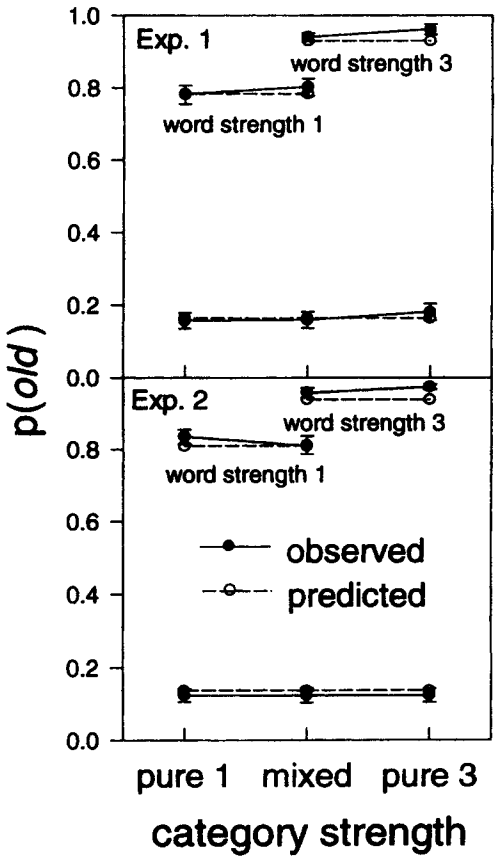

Figure 3. False-alarm rates as a function of category strength, and hit rates as a function of category strength for two levels of word strength for Experiment 1 (in the top section; category length is 6) and for Experiment 2 (in the bottom section; category length is 2). Predictions of the search of associative memory model are represented by the dashed lines. p(*old*) stands for the probability of responding old, and the error bars around each value represent the standard error of the mean. Category strength is indicated by whether the test items come from pure 1 categories (every exemplar presented once), pure 3 categories (every exemplar presented three times), or mixed categories (half the exemplars presented once and half the exemplars presented three times).

assessed by perusing the confidence rating breakdown, summed over category types, given in Appendix C.

The sensitivity results were analyzed statistically by calculating d' for all conditions within each participant's data and then by averaging across participants. However, Figures 5 and 6 graph d' calculated from the group hit and false-alarm rates; this was done because the model fit to the data predicts only group data (the d' patterns were the same in both methods, but all the d' values were higher when we used the average of individual participant d's).

There were no reliable interactions of length and strength with category type. ANOVAs were carried out on the d' values calculated per participant on the basis of truncated scores for each study. The six interaction terms (two studies by length and word strength and category strength) had p values all exceeding .12, so the results were summed across category type.

Figure 5 gives d' as a function of category length for Experiments 1 and 2: For Experiment 1, d' decreased significantly with increases in category length, $t(46) = 4.30, p < .001$; for Experiment 2, the trend downward was smaller but still significant, $t(50) = 2.20, p < .05$. Such results are consistent with an extensive amount of literature (e.g., Gillund & Shiffrin, 1984; Ratcliff et al., 1990) showing list-length effects. Although the size of the observed category-length effects was not enormous, larger effects would have been unlikely given that the length manipulation was carried out on a category embedded in a much larger list, without obvious demarcation. This argument is formalized in terms of variance contributions when the models are presented and discussed in the *Empirical Summary* section of the present article.

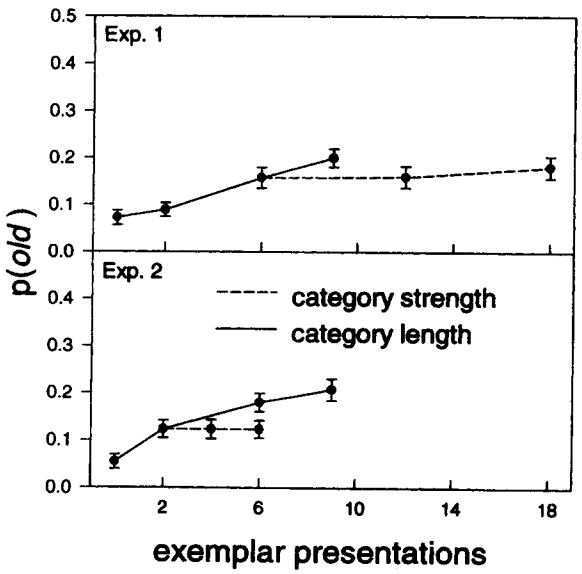

Figure 4. False-alarm rates as a function of category length and category strength, both given as the total number of exemplar presentations per category, for Experiment 1 (top section) and Experiment 2 (bottom section). p(*old*) stands for the probability of responding old, and the error bars around each value represent the standard error of the mean.

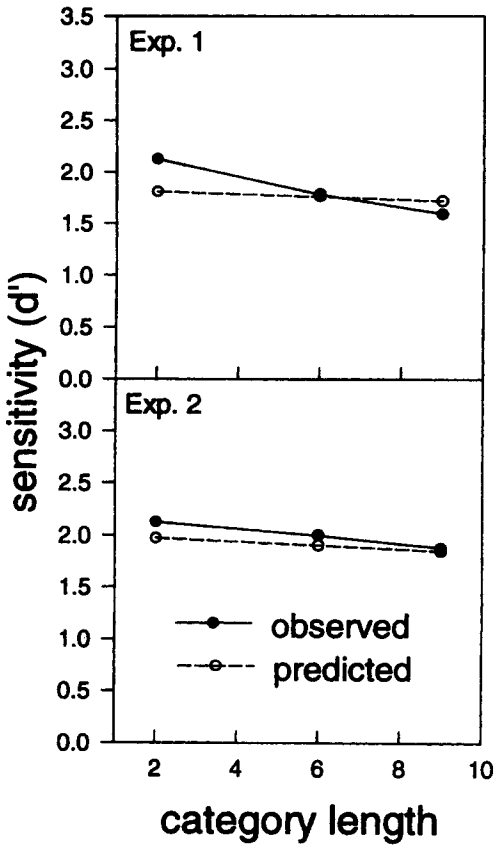

Figure 5. Sensitivity (d') as a function of category length for Experiment 1 (top section) and Experiment 2 (bottom section). Predictions of the search of associative memory model are represented by the dashed lines. The observed (and predicted) d' values were calculated from the group hit rate (and the group false-alarm rate) for a given condition.

Figure 6 gives d' as a function of category strength, when category length is held constant, for Experiment 1 (in which length was 6) and Experiment 2 (in which length was 2). Consistent with all previous findings, there was a main effect of word strength: d' increased with strength, $t(46) = 7.44, p < .001$, for Experiment 1 and $t(50) = 7.93, p < .001$, for Experiment 2. There was no effect of category strength (i.e., the strength of other items in the category) for a fixed value of target strength, $t(46) = 0.59, p = .56$ (two-tailed), for Experiment 1 and $t(50) = 0.03, p = .98$ (two-tailed), for Experiment 2. These category-strength results are consistent with quite a few recent articles on the list-strength effect (e.g., Murnane & Shiffrin, 1991a, 1991b; Ratcliff et al., 1990). It may be noted that for word-strength 3, category strength exhibited a trend consistent with a negative category-strength effect, although not significantly so. Several negative list-strength effects, a few significant, have also been noted in the articles just referenced.

Although category type did not interact with category length and category strength, there were differences in both studies between the category types in the probabilities of responding old, and in performance. In summary, sensitivity (d') was better for the semantic than for the orthographic–phonemic

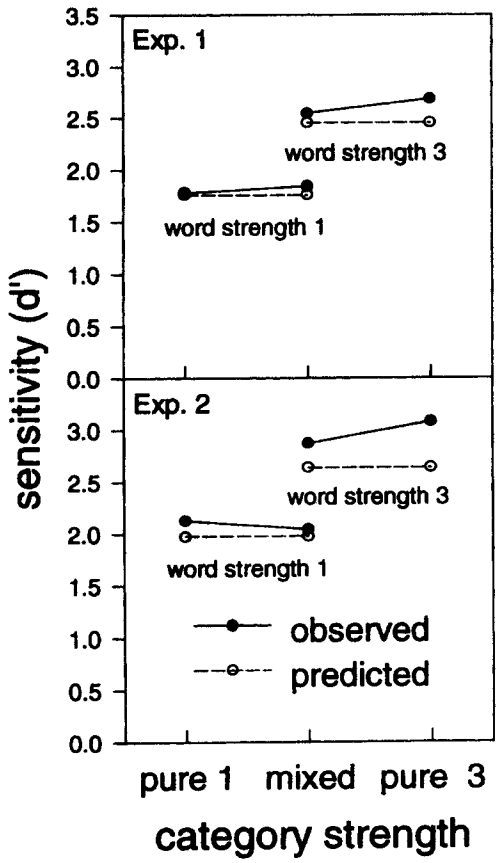

Figure 6. Sensitivity (d') as a function of category strength for two levels of word strength for Experiment 1 (in the top section; category length is 6) and for Experiment 2 (in the bottom section; category length is 2). Predictions of the search of associative memory are represented by the dashed lines. The observed (and predicted) d' values are calculated from the group hit rate and the group false-alarm rate for a given condition. Category strength is indicated by whether test items come from pure 1 categories (every exemplar presented once), pure 3 categories (every exemplar presented three times), or mixed categories (half the exemplars presented once and half presented three times).

categories, $t(46) = 5.01$, $p < .001$, for Experiment 1 and $t(50) = 2.93$, $p < .01$, for Experiment 2. This is of course consistent with the literature, because recognition performance is well known to be better for low (natural language) frequency words (e.g., Gregg, 1976). This effect was due largely to a higher false-alarm rate for orthographic–phonemic distractors than for semantic distractors, $t(46) = 4.09$, $p < .001$, for Experiment 1 and $t(50) = 2.46$, $p < .01$, for Experiment 2. There may be some differences here from the usual mirror effect (e.g., Glanzer & Adams, 1985), according to which the semantic targets ought to have had a higher hit rate than that for orthographic–phonemic targets. However, the differences between the two types of category words were evident enough that the participants could have chosen different criteria for words from the two types of categories, making it difficult to come to any firm conclusions. Such results may suggest that participants were attending to and coding semantic properties of items to a greater degree than they were attending to and coding orthographic–phonemic properties of items, or the results may simply reflect a different similarity structure for the two types of categories.

Regardless of the overall differences between the two types of categories, the fact that the category type did not interact with the strength and length manipulations led us to collapse across category type for statistical analyses, discussion, and in the drawing of conclusions. Even more important, the fact that the length and strength pattern was the same for these two different categories suggests a generality to the findings that transcends the particular category choices we made.

False alarms to prototypes can also be examined as a function of length and strength, although the data are less stable because of smaller numbers of observations. The results are summarized in Figure 7. For both experiments, prototype false alarms rose with category length, $t(46) = 1.94$, $p < .05$, for Experiment 1 and $t(50) = 6.24$, $p < .001$, for Experiment 2. The factors responsible for this rise might well be those responsible for the rise for regular category distractors.

For Experiment 1, prototype false alarms rose slightly with category strength, $t(46) = 2.72$, $p < .01$ (two-tailed). For Experiment 2, prototype false alarms did not change significantly with category strength, $t(50) = 0.70$, $p = .49$ (two-tailed). Although the data are somewhat noisy, it still may be asked why false alarms to prototypes showed a greater tendency to rise with category strength than with other types of category distractors. One answer is related to the hypothesis of implicit associative response (IAR), which was proposed for recognition memory by Underwood (1965). During the study of a list, when an exemplar from a category is encountered,

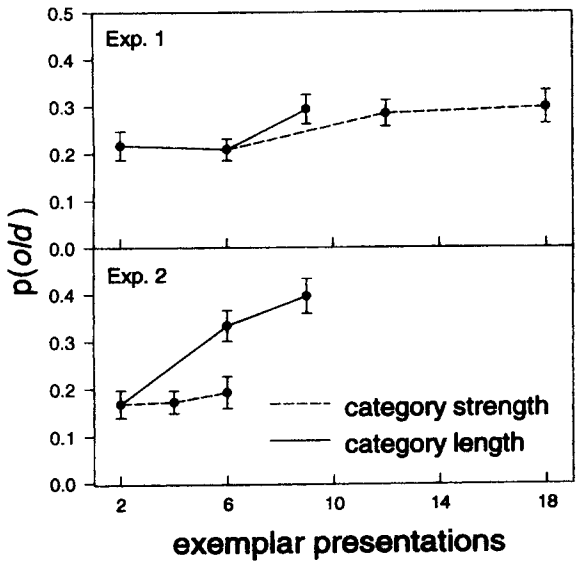

Figure 7. False-alarm rates for prototype tests as a function of category length and category strength, both given as the total number of exemplar presentations per category, for Experiment 1 (top section) and Experiment 2 (bottom section). p(*old*) stands for the probability of responding old, and the error bars around each value represent the standard error of the mean.

certain unstudied words may be generated spontaneously during the encoding process. It seems natural that the chances of this happening will be higher for the category prototype than for some other unstudied category exemplar. Furthermore, the probability of prototype generation during study would likely rise with additional exemplar repetitions (see Leicht, 1968). Once generated during study, a prototype would tend to be judged old on a later test, thereby producing the (somewhat noisy) observed increase with increases in category strength.

In the remainder of this section, we consider data concerning the position of a word during the study list, the position during the test list, and the study–test lag. Certain models make predictions concerning the way in which activation should change across positions. For example, the theory of distributed associative memory (TODAM) model (Murdock, 1982) predicts that the activation caused by a test word should be a geometric function of its recency of presentation.

Table 2 gives the effect of recency in session; it gives the hit probability for fifths of the study list, for the length conditions. In Experiments 1 and 2 there was a small effect of recency: Words closer to the end of the list had marginally higher hit rates: In Experiment 1, $t(46) = 1.01$, $p = .16$, and in Experiment 2, $t(50) = 2.28$, $p < .05$. Whatever effect there was, it could have been due to position in the session or the position in the category (our pool of data did not have sufficient power to separate these possibilities). Some possible explanations of the recency effect include better storage of information as the session or category presentation proceeded or a drift of context over the session so that the test context better matched recently stored context.

Table 3 gives the breakdown of the probability of responding old, along with d', for different types of test items as a function of test position, partitioned by fifths of the test list. The effects were quite small: The only trends reaching significance were the slight decrease in the probability of responding old for targets in Experiment 2, $t(50) = 5.88$, $p < .001$, and the corresponding decrease in d', $t(50) = 3.88$, $p < .001$ (two-tailed). All the other p values were greater than .30 (two-tailed). The d' finding may suggest some degree of retrieval inhibition (see Bjork, 1989).

Table 4 gives the effect of the lag between study and test for the length-condition targets, with lags partitioned by fifths of the range. There was a significant tendency for the probability of a hit to decrease with lag, $t(46) = 1.82$, $p < .05$, for Experiment 1 and $t(50) = 3.45$, $p < .001$, for Experiment 2. This analysis combined the effects of study and test position and therefore lends itself to a number of different interpretations.

Table 2
Probability of Responding Old for Position Within Study List

Study list position	p(old)
Experiment 1	
10–57	.775
58–104	.763
105–151	.770
152–198	.790
199–245	.801
Experiment 2	
10–37	.803
38–64	.841
65–91	.833
92–118	.889
119–145	.889

Note. Study list position is broken into quintiles for length-condition targets. p(old) = probability of responding old.

Table 3
Probability of Responding Old and d' for Position Within Test List

Test list position	p(old)			d'
	Distractors	Prototypes	Targets	
Experiment 1				
4–33	.184	.272	.851	2.086
34–62	.150	.248	.841	2.152
63–91	.118	.268	.848	2.347
92–120	.189	.202	.833	2.038
121–149	.145	.328	.830	2.098
Experiment 2				
4–31	.155	.220	.939	2.510
32–58	.149	.296	.894	2.311
59–85	.154	.254	.885	2.320
86–112	.136	.231	.849	2.238
113–139	.165	.255	.837	2.076

Note. Test list position is broken into quintiles for all conditions.

Empirical Summary

The primary results of our studies are those showing the effects of length and strength.

Probabilities of responding old. Hits and false alarms in response to category exemplars did not change appreciably with category strength. The number of hits rose markedly with

Table 4
Probability of Responding Old for Lag Between Position Within Study List and Position Within Test List

Lag between study and test	p(old)
Experiment 1	
14–90	.848
91–166	.791
167–242	.747
243–318	.796
319–394	.765
Experiment 2	
14–68	.940
69–122	.895
123–176	.859
177–230	.776
231–284	.783

Note. Lag is broken into quintiles for length-condition targets. p(old) = probability of responding old.

increases in word strength. The number of prototype false alarms rose slightly with category strength in both experiments but rose significantly only in Experiment 1.

The number of false alarms (to both exemplars and prototypes) rose as category length increased. The number of hits remained relatively constant as category length increased.

Sensitivity measured by d'. Sensitivity did not vary with category strength (what trends there were indicated higher d' with larger category strength, which is a negative category strength effect), but d' rose markedly with word strength. Sensitivity decreased slightly with category length.

We ascribe the length effects to a pooling of trace activations, on the basis of similarity.[2] Longer categories have more images similar to the test item. In this explanation, increased total activation arises from the greater number of within-category traces, not from an increase in the activation of the trace of the test item itself. The length effects seemed to occur even though the participants tended to be unaware of the categorical structure of the list. We judged this to be the case on the basis of a debriefing of about 25% of the participants at session's end. When asked first if they noticed anything special about the list presented, almost no one volunteered anything concerning categories (1 or 2 participants mentioned a single associated group of studied items). When told that there were categories of items on the list, and asked to name any they could remember, a few more participants named a single category. It is interesting that these failures to report much awareness of categories occurred after the test portion of the experiment, a test that included all the category prototypes.

The pattern of length and strength results, excluding prototypes, is consistent with the pattern of familiarity distributions illustrated in Figure 1, if the following assignments are made. Assume that a short category of singly presented items results in familiarity distributions and the criterion depicted in Figure 1a. An increase in length is then assumed to produce an increase in mean (and somewhat in variance), but no change in absolute placement of the criterion, as illustrated in Figure 1c. An increase in category strength, for a fixed target strength, is assumed to produce no change in the distributions or criterion, so Figure 1a represents the situation. An increase in word strength is assumed to leave the distractor distribution and criterion unchanged, but increase the mean (and somewhat the variance) of the target distribution, as illustrated in Figure 1e.

We next demonstrate that these patterns and assumptions are consistent with the predictions of the differentiation version of the SAM model extended to the category setting. We do so by fitting this model to the data. For the purposes of model fitting, we ignore the differences between categories and category types and simply fit the model to the combined data. After fitting and discussing the SAM model, we discuss the implications for other models.

The SAM Model

The applicable version of SAM is particularly simple. It is essentially the same model introduced by Shiffrin et al. (1990) and differs only in a differentiation assumption from earlier versions of SAM (e.g., Gillund & Shiffrin, 1984). It is assumed that each different word is stored in a different image in long-term memory but that repetitions of a word are stored in a single (stronger) image. At test, memory is probed with context and word cues. Each image is activated, to an extent determined by the match to the test probe, and then the image activations are summed. The participant gives a confidence rating based on the sum. We fit the full set of confidence rating data, so we needed a criterion for each cut point in the confidence scale (five parameters): A confidence rating i is given when the summed activation (familiarity) is greater than criterion $i - 1$ and less than criterion i. To apply the model, one need know only the mean and variance of the activation of each image in response to a given probe. We assumed independence for different images. For a long list like ours, the law of large numbers assures us that the sum will be approximately normal, with a mean equaling the sum of the individual means and a variance equal to the sum of the variances. To generate predictions, then, we needed to know only the means and variances of the individual activations; these were generated according to the following assumptions:

1. A target activates its own n-times studied image with mean strength S_n, $n = 1$ or 3.
2. A test word other than the prototype activates the image of any word in its category, other than its own image, with mean strength S_i.
3. A test word, whether distractor or target, activates the image of any word not in its category with mean strength S_o.
4. The variance of activation of any image is α times the square of the mean strength for that image.

The differentiation assumptions that enabled Shiffrin et al.'s (1990) version of SAM to predict list-strength findings are implicit in Assumptions 2 and 3: Parameters S_i and S_o do not change with strength of the image being activated. The idea, in short, is that the context cue tends to cause more activation of a stronger image but the item cue tends to cause less activation of a stronger image (because it differs from that image). The two tendencies cancel, leaving activation unaffected by strength. In this instantiation of SAM, as in previous applications, we assumed that exact cancellation occurs. There is nothing in the conceptualization of SAM to demand exact cancellation, and it would be quite sensible to introduce another parameter controlling the degree of cancellation. Nonetheless, for simplicity, and because it allows fairly accurate predictions of all the extant data, we formulated the equations so as to incorporate implicitly the assumption of exact cancellation.

Equations 1 and 2 give the means of the familiarity distributions for tests of targets and distractors, and Equations 3 and 4 give the corresponding variances. In the equations, L is the category size (excluding repetitions) for the category of the test item (we assumed $L = 0$ for tested extra semantic and extra

[2] The existence of length effects could conceivably be attributed to priming: The first presentation of each word could cause activation of images of associated (category) words (as in the IAR hypothesis proposed by Underwood, 1965). In this explanation, the increased probability of responding old would be due to the increased activation of the trace of the tested word itself. However, the priming explanation is not very likely for several reasons, chiefly because long-term associative priming is quite weak (as compared with identity priming) and usually cannot be found (e.g., Joordens & Besner, 1992; Roediger & Challis, 1992).

orthographic–phonemic items), and K is the total number of different items in the study list (excluding repetitions). For Experiment 1, $K = 145$; for Experiment 2, $K = 105$. In the following, F (standing for familiarity) is used to denote the total summed activation.

$$\mathrm{E}[F \mid \text{target}, n] = (K - L)S_o + (L - 1)S_i + S_n. \quad (1)$$

$$\mathrm{E}[F \mid \text{distractor}] = (K - L)S_o + LS_i. \quad (2)$$

$$\mathrm{Var}[F \mid \text{target}, n] = \alpha[(K - L)S_o^2 + (L - 1)S_i^2 + S_n^2. \quad (3)$$

$$\mathrm{Var}[F \mid \text{distractor}] = \alpha[(K - L)S_o^2 + LS_i^2]. \quad (4)$$

It would be easy to extend this model to predict tests of prototypes by adding a parameter, S_p, representing mean strength of prototype, to images of exemplars from its own category: S_p would replace S_i in Equations 2 and 4. However, Equations 2 and 4 vary only with length and not strength, so such a model would fail to predict the observed increase in p(old) for prototypes with category strength. We think the observed increase may have been due to an increasing chance of thinking of the prototype during study as repetitions increased, which is a hypothesis with some support (e.g., Underwood, 1965). However, we decided not to fit a model augmented by this hypothesis to the prototype data, because we would have no other independent validation of the augmented model.

The parameters were chosen so as to minimize the chi-square measure of discrepancy between the observed and predicted probabilities of giving confidence rating i for all the probabilities in both experiments, excluding prototype tests. The parameters were the same for both experiments, except for the 5 criteria. We allowed the criteria to vary between studies because the different study-list lengths ought to have affected the criterion choices. The 10 criteria parameters are denoted $C(i, 1)$ and $C(i, 2)$. The best fitting parameter values are given in Table 5. Because the participants were aware of the length of the list at test, one might expect them to choose lower criteria for the shorter list; such a pattern is evident in Table 5. This model fit quite well: The predictions are those that are graphed in Figures 2, 3, 5, 6, and 8 and are presented in tabular form in Appendix C.

The match of predictions and data demonstrate the adequacy of the version of the SAM model in incorporating differentiation (e.g., Shiffrin et al., 1990) to predict hits, false alarms, and d'. The pattern of observations was predicted in advance under the assumption that the response criterion does not shift with length and strength, lending support to this assumption as well as to the model.[3]

Finally, we discuss the receiver operating characteristic (ROC) functions from our studies (our model was fit to these implicitly because we fit the full range of confidence ratings, but a more detailed look is useful). Using the group confidence rating data, we constructed for each condition in each experiment an ROC function, giving cumulative probability of confidence ratings up to a given cut point for targets on one axis, versus distractors on the other axis, for the various conditions. Plotted on normal–normal axes these functions will be linear if the underlying familiarity distributions are normal.

Table 5
Parameters for Fit of the Search of Associative Memory Model to Experiments 1 and 2

Variable	Value
Activation	
S_o	1.0 (fixed)
S_i	1.57
S_1	13.0
S_3	26.1
Variance multiplier	
α	.210
Criteria	
Experiment 1	
$C(1, 1)$	145.3
$C(2, 1)$	148.8
$C(3, 1)$	151.5
$C(4, 1)$	153.0
$C(5, 1)$	155.0
Experiment 2	
$C(1, 2)$	105.5
$C(2, 2)$	108.6
$C(3, 2)$	110.5
$C(4, 2)$	111.8
$C(5, 2)$	113.2

The slope of the function gives the ratio of distractor standard deviation to target standard deviation. Our slopes ranged from .39 to .62 across conditions and studies. The functions are given in Figure 8. In other recognition studies these slopes have also tended to be below 1.0, typically in the range .6 to .9 (e.g., Ratcliff, Sheu, & Gronlund, 1992). The figure also gives predictions from the model. Although the model had no difficulty producing low slopes, and at least roughly fitted the ROC functions, some caveats are in order. The fit, and the low slopes, came about because the parameters controlling the mean and variance of target strengths were estimated to be very high relative to the parameter controlling distractor strength (13 and 26.1 vs. 1.87). Despite the high mean

[3] Although the slight downward trend of hits with increases in length in Experiment 1 was not reliable, we were led to explore model variants that might predict a decrease. We augmented our model by assuming that the recognition decision is sometimes based on a recall of a particular stored image. For simplicity, we did this in a parameter-free fashion: We assumed that on presentation of a test word, a single sample is made from the list words stored in memory. The probability of sampling an image is simply its (mean) strength divided by the summed (mean) strengths of all the images of the list words, the strengths being determined by the probe cues. If the test word is the one sampled, an *old* recognition response is made (actually, the highest confidence rating is given). If not, the decision is based on the summed activation. Because the sampling probability decreases with category length, we hoped that the model could combine the processes of sampling (recall) and summed activation (global recognition) in such a way as to produce a decrease in hit rate with increases in category length. We discovered that this was indeed possible, and a best fit to the data from Experiment 1 alone indeed produced a predicted decrease of hits with category length. However, when this model was jointly fit to the data from both experiments with the same parameters (except for the criteria), the results were not significantly better overall than the fit shown in Figures 2, 3, 5, 6, and 8.

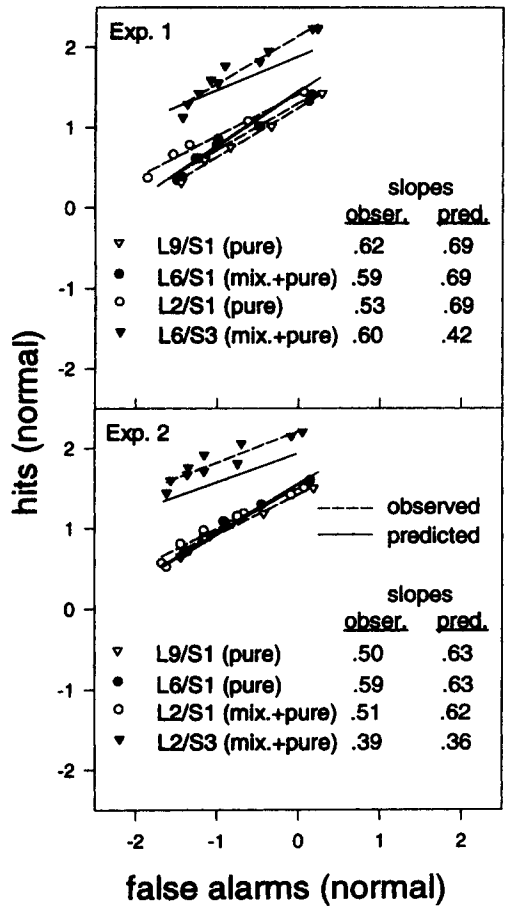

Figure 8. Receiver operating characteristic functions on normal–normal coordinates, with best linear fit to the observed points, and observed slope values. Predictions of the search of associative memory model are represented by the solid lines. Category-length (L) and word-strength (S) conditions are indicated, as well as the type of list (pure, mixed, or data collapsed across pure and mixed) for Experiment 1 (top) and Experiment 2 (bottom). obser. = observed; pred. = predicted; mix. = mixed.

strengths for targets, performance remained in the right range because the variances grew correspondingly large. Some researchers may regard these parameter estimates as conceptually implausible. In addition, the large target variances contributed to the failure of the model to predict as large an effect of category length on d' as was observed (see Figure 5).

In future research it may be worth exploring alternative models in which there is a recall component to recognition. Suppose, for example, that participants try to sample once from memory. If an image of the test word is sampled, the participant assesses the recovered episodic information and responds old or new on this basis, without using a feeling of familiarity of the recovered information. If an image of the test item is not sampled, the participant uses the familiarity value. Such a process might produce effects that, when fit with a familiarity-only model, appear to be due to much larger means and variances for targets than for distractors, even if the variances of the *familiarity* distributions are actually near equal for targets and distractors.

These alternatives notwithstanding, the current version of the SAM model captures most of the findings. In particular, it predicts the results that are the most important and diagnostic for discriminating theories—those illustrated in Figure 4: False alarms rose with category length and did not rise with category strength. If the criterion for a recognition decision does not change across these conditions, then we have good reason to conclude that the familiarity distributions shift uniformly upward with increases in category length but do not move with increases in category strength. This is exactly what is predicted by SAM. Strictly speaking, it would be possible to obtain similar predictions under other assumptions (in models other than SAM, for instance) if the variances of the familiarity distributions are assumed to change in ways not corresponding with the changes in means. However, the fact that the results reported hold for each cut point in the range of confidence ratings makes such esoteric possibilities unlikely.

How might we assess the fundamental assumption that the criterion does not change with category length and category strength? First, the results were predicted in advance, given the prior known results with lists that vary in strength and length, and given the added assumption that the criterion does not change with these variables in our category study. Second, the participants' reports indicated little knowledge of the type of categories that were on the study list, or even the existence of such categories, thus making it difficult to see how any process other than recognition itself could be used to assess the length or strength of the category to which the test item belonged. Third, if participants were adjusting criteria in accord with the list characteristics of the category of the test item, it is hard to see why a different result would have been obtained for length and strength. We would be the first to admit that these arguments are inconclusive, but as far as we know, no better design to assess this question currently exists.

Alternative Models

Consider now the predictions of some alternative models (for the old vs. new results). In previous articles (e.g., Murnane & Shiffrin, 1991a, 1991b; Ratcliff et al., 1990; Shiffrin et al., 1990; Shiffrin, Ratcliff, Murnane, & Nobel, 1993), models that fail to predict the way recognition performance depends on list length and list strength have been discussed. We tried, without success, to extend these models in the simplest and most natural fashion that would allow them to be applied to the category length and category strength findings from the present paradigm. We decided not to discuss these, however, because our extensions might not be optimal choices and because our interest is in models that can predict both the present results and the list length and list strength findings from previous studies.

As discussed by Shiffrin et al. (1990), a modification of the MINERVA model (e.g., Hintzman, 1988) is capable of handling the earlier set of recognition findings under certain special assumptions. In MINERVA, items are represented by vectors of features. At study, each item is stored as a vector in a separate trace. Each feature is stored independently with some probability, L. If a second presentation of an item results in replacement of the first trace with a new trace based on a higher value of L (or, equivalently, if the features in the first

stored trace that match the presented item are retained and the other features are replaced by correct features with some learning probability L'), then list-length and list-strength findings will be correctly predicted (approximately). This same model could be extended to the category paradigm by assuming that same-category traces have positive correlations. If so, the present pattern of old versus new results (Figure 5) would be predicted correctly, at least qualitatively. We regard the replacement assumption as conceptually implausible, and it certainly alters the multiple-trace assumption of MINERVA, but we mention this model for the sake of completeness.

Very recently, there have been reports of new models formulated to predict the list-length and list-strength findings. We discuss two of these. Murdock and Kahana (1993a) proposed a continuous memory version of the TODAM model, which they claimed predicted performance measures correctly. Shiffrin, Ratcliff, Murnane, and Nobel (1993) disagreed with this contention (see also the reply by Murdock & Kahana, 1993b). The present data bear on the critical issue. In the TODAM model, items are represented by vectors, and each presented item is added to a common memory vector. This vector is then multiplied by a forgetting factor α before the next item is added. At test, the inner product of the test vector with the memory vector (a measure of similarity) serves as the decision statistic and plays the role of activation or familiarity in SAM. Murdock and Kahana suggested that all past images, not just the current list, contribute to the decision statistic. If so, under reasonable side assumptions, it may be shown that neither list length nor list strength should alter variance of the decision statistic, and hence should not alter performance, for a test item at a fixed recency. To explain list-length effects, then, Murdock and Kahana argued that longer lists contain older serial positions and greater lags until test. Because the model predicts decreasing performance as lag until test increases, list-length effects are produced by averaging across serial positions.

It is not hard to apply this version of TODAM to the present study. It would be both necessary and plausible to assume that vectors of items in the same category are correlated more than vectors of items in different categories. It is critical to note that average serial position, average test position, and average study–test lag are the same in our study for items from categories of different lengths and strengths, so this factor plays no differential role in the predictions. Nonetheless, higher hit and false-alarm rates for longer categories are predicted as a result of the higher activation produced by additional within-category vectors being added to memory. Unfortunately, higher hit and false-alarm rates would also be predicted for additional repetitions of other within-category items, so this model would not handle the present findings. Other assumptions could conceivably be adopted that would eliminate the strength misprediction, but it is hard for us to imagine how this could be managed without also eliminating the predicted changes with length.

The second model we consider is an array-similarity model proposed by Estes (1994). The claim is again made that the basic pattern of list-length and list-strength effects is predicted by the model. In our view, this is debatable (because the model predicts positive list-length and list-strength effects, even though they would be of small magnitude for certain parameters). As with TODAM, the critical issue is clearer when the model is extended to the present study. The model is quite simple and easy to extend to the category paradigm. In essence, a test item has similarities to exemplars attached to an old response category and to exemplars attached to a new response category. For a distractor test, the sum of similarities, $A(N)$, to the new response category is simply a pre-experimental new bias, $s(N)$. For a distractor test, the sum of similarities, $A(O)$, to the old response category will equal the number, n, of within-category list items multiplied by a within-category similarity parameter, $s+$, plus the number, m, of out-of-category list items multiplied by the out-of-category similarity parameter, $s-$, plus a pre-experimental old bias, $s(O)$:

$$A(O) = ns + ms- + s(O). \quad (5)$$

The probability of responding old (and thereby giving a false alarm) is then

$$p(\text{old}) = A(O)/[A(O) + A(N)]. \quad (6)$$

For target tests, assume there are r repetitions of the target; then r of the within-category similarities in Equation 5 (i.e., $s+$) are replaced by the value 1.0.

It is clear from these equations that distractors from categories of equal total numbers of presentations, whether these presentations come from different items or repetitions, are predicted to have the same p (old), regardless of parameter choices. The critically relevant data is given in Figure 4: For six total presentations, length and strength give different values, $t(50) = 2.48, p < .01$.[4]

Conclusion

We gave participants a single long list with distributed exemplars from many categories. We did so in the hope that the participants would utilize a recognition criterion in the subsequent test that would not vary with the number of exemplars, or with the number of repetitions of exemplars, in the category of the test item. Assuming this to be the case, the hit and false-alarm rates give evidence concerning the changes in familiarity caused by changes in length and strength. The results showed that changes in the strength of category exemplars (other than the target itself) produced no reliable change in hits or false alarms, which suggests that the distributions of familiarity did not change. However, extra category items produced increases in both hits and false alarms, which suggests that the distributions of familiarity moved upward. Given the assumption that the criterion is independent of length and strength, the SAM model previously used to explain

[4] Some other recent models developed to predict the effects of list length and list strength in recognition include a network-similarity version of Estes's model (discussed very briefly by Estes, 1994) and the neural network models of Chappell and Humphreys (1994) and Dennis (1993). It is not clear to us how best to extend these to the present paradigm. In addition it is not obvious what the predictions would be for a given extension; predictions might well require extensive computer simulations.

list-length and list-strength results predicts the observed pattern of findings. In fact, a simple version of this model gave a good quantitative account of the results. This outcome gives some credence to the hypothesis of criterion independence.

Many recognition models assume that the recognition decision is based on the magnitude of some globally produced value, termed *familiarity* here for convenience. The present results suggest that constraints beyond those already imposed by performance measures need to be taken into account: Changes in category strength (and by extension, list strength) do not change familiarity, but increases in category length (and by extension, list length) do increase familiarity. The SAM model predicts such findings. Two recent models (namely, the continuous memory version of TODAM [Murdock & Kahana, 1993a] and the array-similarity model [Estes, 1994]) that have been claimed to correctly predict performance measures for list length and list strength failed to predict these category results, so the category results may be of considerable diagnostic utility.

Some caveats concerning these relatively simple results and conclusions need to be kept in mind. First, the assumption of criterion independence has not been verified directly. It is not clear to us how to carry out a direct test of this assumption, but future research using forced-choice paradigms may provide converging evidence. If the criterion does change with length and strength, it must covary with the changes in means and variances in a fairly intricate fashion to produce the observed data, but this possibility cannot be ruled out. Second, the model represented by the proposed version of SAM, in which all decisions are based on a single, normally distributed value of familiarity, may be too simple. However, we leave the exploration of more complex models to future research.

References

Banks, W. P. (1970). Signal-detection theory and human memory. *Psychological Bulletin, 74,* 81–99.

Bjork, R. A. (1989). Retrieval inhibition as an adaptive mechanism in human memory. In H. L. Roediger III & F. I. M. Craik (Eds.), *Varieties of memory and consciousness: Essays in honour of Endel Tulving* (pp. 309–330). Hillsdale, NJ: Erlbaum.

Chappell, M., & Humphreys, M. S. (1994). An auto-associative neural network for sparse representations: Analysis and application to models of recognition and cued recall. *Psychological Review, 101,* 103–128.

Dennis, S. J. (1993). *Integrating learning into models of human memory.* Unpublished doctoral dissertation, University of Queensland, Brisbane, Australia.

Estes, W. K. (1994). *Classification and cognition.* New York: Oxford University Press.

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.

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

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review, 93,* 411–428.

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple trace memory model. *Psychological Review, 95,* 528–551.

Hintzman, D. L., & Curran, T. (1994). Retrieval dynamics of recognition and frequency judgments: Evidence for separate processes of familiarity and recall. *Journal of Memory and Language, 33,* 1–18.

Hintzman, D. L., Curran, T., & Oppy, B. (1992). Effects of similarity and repetition on memory: Registration without learning? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 667–680.

Hintzman, D. L., & Ludlam, G. (1980). Differential forgetting of prototypes and old instances: Simulation by an exemplar-based classification model. *Memory & Cognition, 8,* 378–382.

Jones, C. M., & Heit, E. (1993). An evaluation of the total similarity principle: Effects of similarity on frequency judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 799–812.

Joordens, S., & Besner, D. (1992). Priming effects that span an intervening unrelated word: Implications for models of memory representation and retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 483–491.

Kuçera, H., & Francis, W. N. (1967). Information retrieval from long term memory: Category size and recognition time. *Journal of Verbal Learning and Verbal Behavior, 7,* 291–295.

Leicht, K. L. (1968). Recall and judged frequency of implicitly occurring words. *Journal of Verbal Learning and Verbal Behavior, 7,* 918–923.

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

Murdock, B. B., & Kahana, M. J. (1993a). Analysis of the list-strength effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 689–697.

Murdock, B. B., & Kahana, M. J. (1993b). List-strength and list-length effects: Reply to Shiffrin, Ratcliff, Murnane, and Nobel (1993). *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 1450–1453.

Murnane, K., & Shiffrin, R. M. (1991a). Interference and the representation of events in memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 855–874.

Murnane, K., & Shiffrin, R. M. (1991b). Word repetition in sentence recognition. *Memory & Cognition, 19,* 119–130.

Nusbaum, H. C., Pisoni, D. B., & Davis, C. K. (1984). *Sizing up the Hoosier mental lexicon: Measuring the familiarity of 20,000 words.* (Research on Speech Perception Progress Report No. 10). Bloomington: Indiana University.

Posner, M. I., & Keele, S. W. (1970). Retention of abstract ideas. *Journal of Experimental Psychology, 83,* 304–308.

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

Ratcliff, R., Sheu, C.-F., & Gronlund, S. D. (1992). Testing global memory models using ROC curves. *Psychological Review, 99,* 518–535.

Roediger, H. L. III, & Challis, B. H. (1992). Effects of identity repetition and conceptual repetition on free recall and word fragment completion. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 3–14.

Shiffrin, R. M., Huber, D., & Marinelli, K. (1993). *Effects of length and strength on familiarity in recognition* (Tech. Report No. 94). Bloomington: Indiana University, Cognitive Science Program.

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

Shiffrin, R. M., Ratcliff, R., Murnane, K., & Nobel, P. (1993). TODAM and the list-strength and list-length effects: Comment on Murdock and Kahana (1993a). *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 1445–1449.

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General, 117,* 34–50.

Underwood, B. J. (1965). False recognition produced by implicit verbal responses. *Journal of Experimental Psychology, 70,* 122–129.

Appendix A

Semantic Categories

CATERPILLAR	FORTRESS	TRICKSTER	MADMAN
COCOON	STRONGHOLD	HYPNOTIST	INSANITY
MONARCH	MEDIEVAL	SORCERER	MANIAC
FLUTTER	CHATEAU	RABBIT	DEMENTED
METAMORPHOSIS	COURTYARD	CONJURE	PSYCHOTIC
DRAGONFLY	DUNGEON	VANISH	ASYLUM
WINGS	TOWERS	JUGGLING	DELIRIOUS
FLITTING	FEUDAL	ENCHANTED	RANTING
CAMOUFLAGE	THRONE	DECEPTION	HALLUCINATION
FRAGILE	MANSION	WIZARD	DERANGED
SLIGHT	VILLA	SPELLS	PSYCHOPATH
BUTTERFLY	**CASTLE**	**MAGICIAN**	**LUNATIC**
EMERALD	BETTOR	FOSSILS	WORKOUT
RUBIES	WAGER	EXTINCTION	EXERTION
RHINESTONES	BLUFF	REPTILES	JOGGING
CARAT	BOOKIE	SWAMPS	NUTRITION
BRILLIANCE	ROULETTE	GLACIERS	CALISTHENICS
PRECIOUS	CASINO	SKELETONS	AEROBICS
HARDNESS	POKER	BRONTOSAURUS	TONING
SPARKLE	STAKES	AMPHIBIANS	PHYSIQUE
FACET	JACKPOT	MAMMOTH	BICEPS
GLITTERING	LOTTERY	GEOLOGY	SWEATING
PRICELESS	BLACKJACK	ARTIFACTS	BARBELLS
DIAMOND	**GAMBLER**	**DINOSAUR**	**FITNESS**
SPACEMAN	MUMMIES	BURGLARY	PREMATURE
COSMONAUT	HIEROGLYPHICS	THEFT	DIAPERS
SHUTTLE	SPHINX	WALLET	TEETHING
SATELLITE	EGYPTIAN	HOLDUP	HIGHCHAIR
PROPULSION	PHARAOH	BOOTY	STROLLER
WEIGHTLESSNESS	TOMBS	MUGGING	CRADLE
GRAVITY	TRIANGULAR	STICKUP	RATTLE
ATMOSPHERES	CATACOMBS	STEALING	LULLABY
VOYAGER	EMBALMING	BANDIT	PACIFIER
ROCKET	UNDERWORLD	ASSAILANT	BABBLE
ORBITING	VAULT	EMBEZZLEMENT	STORK
ASTRONAUT	**PYRAMID**	**ROBBERY**	**INFANT**
CYCLONE	SPECTER	CLOWN	
TYPHOON	GHOUL	JOKER	
TWISTER	GOBLIN	HUMORIST	
FUNNEL	APPARITION	SLAPSTICK	
WHIRLWIND	GHOST	COMIC	
SIRENS	HAUNTING	LAMPOON	
SPINNING	SPOOKY	CUTUP	
WHIRLING	LAMENTATION	BUFFOON	
GUSTS	BECKON	MONOLOGUE	
SPIRAL	PARANORMAL	PUNSTER	
WINDSTORM	GLOOMY	IMPROVISATION	
TORNADO	**PHANTOM**	**COMEDIAN**	

Note. Prototypes appear in bold type.

(*Appendixes continue on next page*)

Appendix B

Orthographic–Phonemic Categories and Extra Words

MATE	**POP**	**DEAL**	*Extra semantic words*
LATE	HOP	HEAL	
DATE	TOP	MEAL	JARGON
FAZE	COT	PEAL	PAUPER
FADE	COD	REAL	APARTMENT
FACE	COG	SEAL	STATIONER
HATE	MOP	VEAL	OPOSSUM
GATE	SOP	TEAK	MONO XIDE
FAME	COB	TEAM	ANTIQUITY
FA KE	CON	TEAS	STOREROOM
RATE	BOP	TEAT	BAGEL
FATE	**COP**	**TEAL**	TRIBESMAN
			JASMINE
SIRE	**GUN**	**WEFT**	LINGUISTICS
TIRE	NUN	WEPT	BISON
DIRE	PUN	WENT	TRIPLICATE
MICE	BUS	WELT	CARPORT
MIKE	BUT	BEST	WARMHEARTED
MIME	BUG	TEST	THICKET
FIRE	RUN	PEST	PICCOLO
HIRE	FUN	VEST	DACHSHUND
MINE	BUM	LEST	UNFORMED
MILE	SUN	NEST	INFERNO
WIRE	BUD	REST	UNDERGROWTH
MIRE	**BUN**	**WEST**	CANVAS
			TORTILLA
BAT	LOON	*Extra orthographic–*	THESAURUS
MAT	SOON	*phonemic words*	SYNOPSIS
PAT	COON		HOUSECOAT
CAD	BOOM	RAW	PODIUM
CAM	BOOB	PULL	SABLE
CAP	BOOT	FUR	CONVENIENCE
SAT	NOON	NOW	
FAT	MOON	TOY	
CAB	BOOS	SAW	
HAT	GOON	WOOD	
RAT	TOON	BIRD	
CAT	**BOON**	LOUD	
		COIN	
TIP	MOLE	YAW	
LIP	HOLE	BOOK	
HIP	SOLE	PERK	
SIS	ROSE	BOUT	
SIN	ROPE	FOIL	
SIT	RODE	LAWN	
DIP	POLE	NULL	
RIP	DOLE	BUR	
SIX	ROBE	COW	
NIP	ROTE	JOY	
PIP	ROVE		
SIP	**ROLE**		

Note. Prototypes appear in bold type.

Appendix C

Full Probability Breakdown for Confidence Ratings

Confidence	Length 0 Obs.	Length 0 Pred.	Pure 1, length 2 Obs.	Pure 1, length 2 Pred.	Pure 1, length 6 Obs.	Pure 1, length 6 Pred.	Pure 1, length 9 Obs.	Pure 1, length 9 Pred.	Mixed, length 6 Obs.	Mixed, length 6 Pred.	Pure 3, length 6 Obs.	Pure 3, length 6 Pred.
Experiment 1												
Distractors												
1	.536	.519	.477	.492	.451	.439	.389	.400	.436	.439	.411	.439
2	.260	.234	.257	.239	.232	.244	.245	.246	.251	.244	.238	.244
3	.132	.130	.177	.137	.160	.152	.166	.162	.153	.152	.170	.152
4	.038	.043	.028	.047	.055	.055	.077	.061	.051	.055	.043	.055
5	.013	.039	.030	.044	.034	.054	.049	.062	.032	.054	.053	.054
6	.021	.035	.032	.041	.068	.056	.074	.069	.077	.056	.085	.056
Strength 1 targets												
1			.074	.071	.091	.060	.077	.053	.079	.060		
2			.066	.079	.064	.071	.079	.064	.077	.071		
3			.077	.094	.064	.087	.070	.081	.043	.087		
4			.038	.058	.051	.054	.051	.052	.072	.054		
5			.100	.092	.096	.088	.100	.085	.083	.088		
6			.645	.606	.634	.641	.623	.666	.647	.641		
Strength 3 targets												
1									.013	.025	.013	.025
2									.021	.020	.013	.020
3									.026	.024	.013	.024
4									.017	.015	.017	.015
5									.053	.026	.043	.026
6									.870	.889	.902	.889
Experiment 2												
									Mixed, length 2		Pure 3, length 2	
Distractors												
1	.592	.544	.471	.512	.441	.449	.425	.404	.535	.512	.480	.512
2	.257	.236	.276	.243	.233	.251	.239	.254	.237	.243	.278	.243
3	.096	.102	.129	.110	.145	.124	.127	.134	.104	.110	.118	.110
4	.031	.043	.049	.048	.045	.058	.082	.065	.037	.048	.035	.048
5	.008	.034	.027	.039	.055	.049	.051	.057	.033	.039	.029	.039
6	.016	.041	.047	.049	.080	.069	.075	.087	.053	.049	.059	.049
Strength 1 targets												
1			.067	.062	.055	.051	.067	.044	.078	.062		
2			.051	.068	.043	.059	.051	.053	.047	.068		
3			.047	.062	.041	.056	.025	.051	.063	.062		
4			.045	.046	.045	.043	.047	.040	.051	.046		
5			.075	.062	.059	.058	.069	.054	.063	.062		
6			.716	.701	.757	.734	.741	.757	.698	.701		
Strength 3 targets												
1									.016	.027	.014	.027
2									.020	.019	.006	.019
3									.008	.016	.008	.016
4									.004	.012	.012	.012
5									.027	.017	.016	.017
6									.925	.909	.945	.909

Note. Predicted probabilities are based on SAM parameters listed in Table 5. Obs. = observed; pred. = predicted.

(*Appendixes continue on next page*)

Appendix D

Experiments 3 and 4

We report in condensed form in this appendix two additional studies bearing on the results reported in the body of the article (and essentially replicating them). Experiment 3 varied category strength, but not category length. In addition to an immediate test, Experiment 3 included a group of participants tested after a delay of one week. Experiment 4 included both category-strength and category-length conditions and used only immediate tests. In both studies, participants gave frequency ratings to words during study (rather than the pleasantness ratings used in Experiments 1 and 2) and gave two successive ratings at test: (a) a frequency judgment, and (b) recognition confidence judgment. There were other minor differences between Experiments 3 and 4, and between both of these and Experiments 1 and 2, and these are mentioned below. Details not mentioned below may be assumed to be similar to those reported for Experiments 1 and 2.

Method

Experiment 3

Participants. Forty-nine participants composed the immediate test group, and 48 participants composed the group tested after a delay of one week. These Indiana University students participated to fulfill an introductory psychology course requirement.

Materials. Sixteen semantic categories were used (including a category with the prototype ***TAILOR*** that is not listed in Appendix A). Eight orthographic–phonemic categories were used (excluding the categories listed in Appendix B that had the prototypes ***WEST*** and ***TEAL***).

Table D1
Probability of Responding Old for Category Type and Strength Condition

Test items	Strength condition: Pure 1, length 6	Mixed, length 6	Pure 3, length 6
Experiment 3—immediate test			
Semantic			
Distractors	.170	.163	.187
Prototypes	.480	.546	.510
Strength 1 targets	.782	.786	
Strength 3 targets		.964	.958
Orthographic–phonemic			
Distractors	.337	.347	.316
Prototypes	.327	.520	.551
Strength 1 targets	.782	.724	
Strength 3 targets		.944	.944
Experiment 3—delay test			
Semantic			
Distractors	.337	.370	.351
Prototypes	.677	.797	.750
Strength 1 targets	.672	.768	
Strength 3 targets		.904	.899
Orthographic–phonemic			
Distractors	.465	.500	.531
Prototypes	.531	.542	.604
Strength 1 targets	.696	.688	
Strength 3 targets		.906	.891

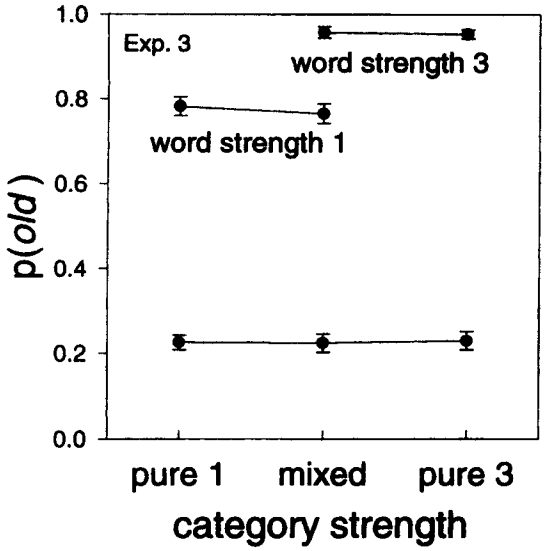

Figure D1. False-alarm rates as a function of category strength, and hit rates as a function of category strength, for two levels of word strength (1 and 3), for the immediate test condition in Experiment 3. p(old) stands for the probability of responding old, and the error bars around each value represent the standard error of the mean. Category strength is indicated by whether the test items come from pure 1 categories (every exemplar presented once), pure 3 categories (every exemplar presented three times), or mixed categories (two exemplars presented once, two exemplars presented two times, and two exemplars presented three times).

The exemplars making up the categories differed slightly from those listed in Appendixes A and B. In addition, for the orthographic–phonemic categories, some test exemplars were lower in similarity than those studied: They retained the vowel sound of the prototype but differed in both of the consonant clusters.

Table D2
Sensitivity (d') for Strength Condition

Test items	Strength condition: Pure 1, length 6	Mixed, length 6	Pure 3, length 6
Experiment 3—immediate test			
Semantic and orthographic–phonemic			
Strength 1 targets	1.700	1.695	
Strength 3 targets		2.557	2.543
Experiment 3—delay test			
Semantic and orthographic–phonemic			
Strength 1 targets	0.892	1.037	
Strength 3 targets		1.741	1.670

Note. *d'* was calculated per participant and then averaged.

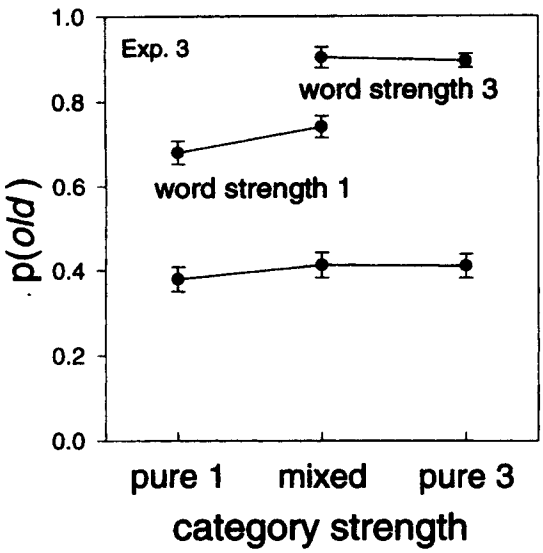

Figure D2. False-alarm rates as a function of category strength, and hit rates as a function of category strength, for two levels of word strength (1 and 3), for the delayed test condition in Experiment 3. p(old) stands for the probability of responding old, and the error bars around each value represent the standard error of the mean. Category strength is indicated by whether the test items come from pure 1 categories (every exemplar presented once), pure 3 categories (every exemplar presented three times), or mixed categories (two exemplars presented once, two exemplars presented two times, and two exemplars presented three times).

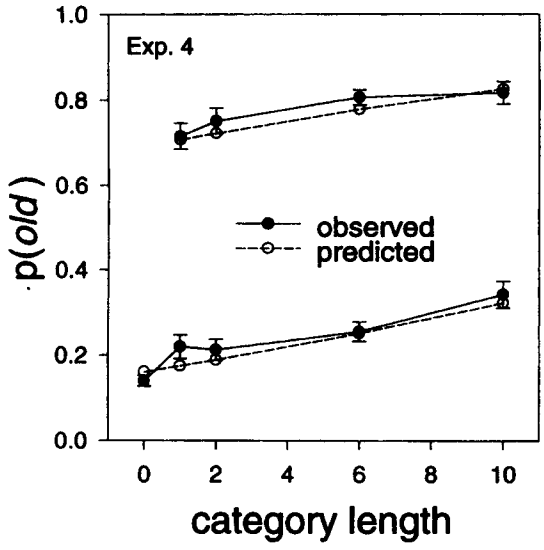

Figure D3. Hit and false-alarm rates as a function of category length for Experiment 4. Predictions of the search of associative memory model are represented by the dashed lines. p(old) stands for the probability of responding old, and the error bars around each value represent the standard error of the mean.

Procedure. Participants viewed a single study list of 288 successive word presentations, followed by a 30-s counting task, followed by a recognition test, either immediately or after one week. In the 3 s that each word was studied, we asked participants to enter a rating of the number of times the word had appeared within the experimental session.

The recognition test consisted of 296 successively presented words. Participants gave two ratings for each test word, taking as much time as necessary for each. The first judgment required a rating of the number of times a word had been presented in the study phase of the experiment, plus one. The second judgment was a confidence rating on a 1–6 scale of the likelihood that the test word had been seen at least once during the study phase of the experiment.

The category lengths were always six words. There were three pure strength conditions, in which the six exemplars studied were all presented either one, two, or three times. Mixed strength categories were created by presenting two words once, two words twice, and two words three times. Four semantic categories and two orthographic–phonemic categories were assigned to each of these four strength conditions.

Experiment 4

Participants. Ninety-four Indiana University students participated to fulfill an introductory psychology course requirement.

Materials. Forty categories of words were used in this experiment, including the 16 semantic categories and the 8 orthographic–phonemic categories used in Experiment 3. Sixteen additional semantic catego-

Table D3
Probability of Responding Old for Category Type and Length-Strength Condition in Experiment 4

	Length–strength condition						
Test items	Length 0	Pure 1, length 1	Pure 1, length 2	Pure 1, length 6	Pure 1, length 10	Mixed, length 6	Pure 3, length 6
Semantic							
Distractors	.140	.176	.161	.230	.340	.195	.206
Prototypes		.284	.314	.400	.504	.411	.456
Strength 1 targets		.676	.738	.786	.816	.803	
Strength 3 targets						.915	.941
Orthographic-phonemic							
Distractors		.298	.330	.309	.351	.330	.351
Prototypes		.245	.255	.378	.415	.394	.521
Strength 1 targets		.787	.777	.846	.819	.862	
Strength 3 targets						.936	.957

(*Appendixes continue on next page*)

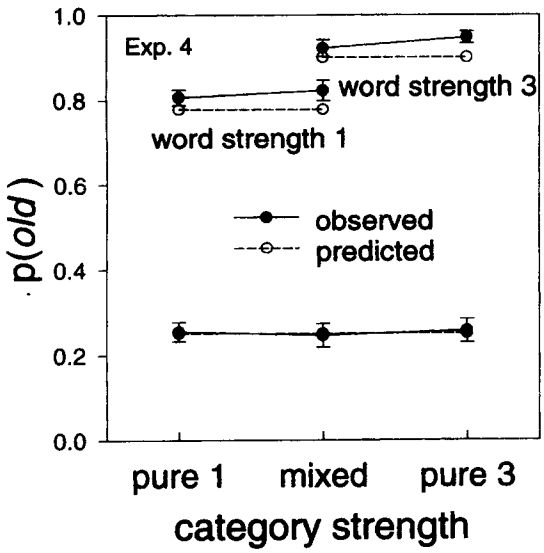

Figure D4. False-alarm rates as a function of category strength, and hit rates as a function of category strength for two levels of word strength (1 and 3), for Experiment 4. Predictions of the search of associative memory model are represented by the dashed lines. p(old) stands for the probability of responding old, and the error bars represent the standard error of the mean. Category strength is indicated by whether the test items come from pure 1 categories (every exemplar presented once), pure 3 categories (every exemplar presented three times), or mixed categories (two exemplars presented once, two exemplars presented two times, and two exemplars presented three times).

ries were created to fill out the design. These extra categories were generally similar to the old semantic categories but were not as carefully chosen, and the results from these were not analyzed.

Procedure. Participants viewed a single study list of 335 presentations, followed by a counting task and an immediate 170-item recognition test. The judgments given during study and at test were the same as in Experiment 3.

Five categories were assigned to each of eight different length-strength conditions. Of these five, two were old semantic categories, two were new semantic categories, and one was an orthographic–phonemic category. The eight conditions included the same four strength conditions used in Experiment 1: a mixed-strength and three pure-strength conditions. The four length conditions consisted of 1, 2, 6, or 10 words all shown once each.

Results and Discussion

Partly because the second judgment may have been biased by the first, the frequency rating representing zero prior presentations, and the summed confidence ratings representing a judgment of new, were almost entirely consistent. This fact makes it reasonable to report analyses of the confidence rating data, thereby making our analyses consistent with those used for Experiments 1 and 2. All statements we make concerning results are statistically significant at at least the .01 level of significance unless otherwise indicated.

Experiment 3. We start with the hit and false-alarm data corresponding to those given in the main body of the article. Table D1 gives the response rates for the immediate and delayed tests. Figures D1 and D2 give the false-alarm and hit rates as a function of strength for the immediate and delayed tests. The results of this experiment replicate those of Experiments 1 and 2: There was a large main effect of word strength, no effect of category strength for distractors or hits, and a small effect of category strength for prototypes (this last effect is not shown in the figures).

Table D2 gives the d' values (calculated per participant and then averaged) for the immediate and delayed conditions. There was a main effect of word strength but no effect of category strength, which is a replication of the results from Experiments 1 and 2.

Generally speaking, there is, of course, a good deal of forgetting from immediate to delayed test. The one exception is intriguing and occurs when one compares prototypes and distractors. For the orthographic–phonemic categories, there was some slight forgetting for prototypes vis-à-vis distractors, but the results were small because performance was so close to chance. However, for semantic categories, discrimination of prototypes from distractors did not change over the one-week delay.

Although category studies have shown less forgetting for prototypes than for targets (e.g., Posner & Keele, 1970), a finding of no loss of discrimination is unusual. It would not be easy to explain such a finding within the standard SAM framework, and the way to augment the model to handle such a finding is far from clear. The usual theoretical approach to interpreting such prototype-forgetting findings utilizes an assumption of a nonlinear relation between similarity and strength of activation. In one version, all similarities rise with delay and targets are already near ceiling, so prototypes gain relatively. In another version, similarity drops with delay; targets are on the steep portion of the function and drop more than prototypes (similar to MINERVA—see Hintzman, 1986; Hintzman & Ludlam, 1980). Such reasoning requires specific implementation in a model like SAM. We leave such issues for future research.

Experiment 4. Table D3 gives the probabilities of responding old as a function of the length and strength conditions. Figure D3 depicts the length results, and Figure D4 depicts the strength results. There was an effect of category length for distractors, a main effect of word strength, no effect of category strength for distractors, and a small effect of category strength for prototypes (the last effect is not shown in the figures); these effects replicate all earlier findings. In this study there was a small but significant increase in hits with category length; this differed slightly from the results of Experiments 1 and 2.

Table D4 gives the d' results for the length and strength conditions: d' was not systematically affected by length, rose with word strength, and did not change with category strength.

Table D4
Sensitivity (d') for Length–Strength Conditions in Experiment 4

	Length-strength conditions					
Test items	Pure 1, length 1	Pure 1, length 2	Pure 1, length 6	Pure 1, length 10	Mixed, length 6	Pure 3, length 6
Strength 1 targets	1.340	1.473	1.522	1.307	1.614	
Strength 3 targets					2.106	2.278

Note. Values are for semantic and orthographic-phonemic categories combined. d' was calculated from rates averaged across participants.

Table D5
Parameters for Fit of the Search of Associative Memory Model to Experiment 4

Variable	Value
Activation	
S_o	1.0
S_i	1.25
S_1	8.96
S_3	12.9
Variance multiplier	
α	.111
Criteria	
$C(1)$	217.8
$C(2)$	219.1
$C(3)$	219.8
$C(4)$	220.8
$C(5)$	221.9

The model described in the body of the article was fit to the confidence rating data from Experiment 4, and the collapsed predictions are those given in Figures D3 and D4. The best fitting parameter values are given in Table D5.

We note finally the following puzzle: When we fit the same model (with different criteria) to the frequency ratings, it did not fare well. Furthermore, ROC analyses of the frequency ratings revealed slopes greater than 1.0, in contrast to the confidence rating data from Experiments 1 through 4, all of which revealed slopes below 1.0. We suspect that frequency judgments for items judged to have been seen involve additional sources of information beyond the total sum of activation (see Shiffrin, Huber, & Marinelli, 1993, for additional information on our analyses, and Hintzman, Curran, & Oppy, 1992; Hintzman & Curran, 1994, and Jones & Heit, 1993, for additional data and discussion of these issues).

In summary, the results of Experiments 3 and 4 replicate those reported in the body of this article in all important respects. As before, the most diagnostic results are the rise of false alarms with category length in conjunction with the failure of category strength to have an effect. The fact that the same pattern of results was found in Experiments 3 and 4, which required frequency judgments during study, and in Experiments 1 and 2, which required pleasantness ratings during study, adds further generality and power to the results and conclusions.

Received August 11, 1993
Revision received April 14, 1994
Accepted May 2, 1994 ■