Memory & Cognition
1991, 19 (2), 119-130

Word repetitions in sentence recognition

KEVIN MURNANE and RICHARD M. SHIFFRIN
Indiana University, Bloomington, Indiana

When some items on a list are strengthened by extra study time or repetitions, recognition of other, unrelated, list items is not harmed (Ratcliff, Clark, & Shiffrin, 1990). Shiffrin, Ratcliff, and Clark (1990) accounted for this list-strength finding with a model assuming that different items are stored separately in memory, but that repetitions are accumulated together into a single stronger memory trace. Repeating words in the context of different sentences might cause separate storage of the repetitions of a given word, because either word or sentence traces are stored separately. Separate storage would, in effect, convert a list-strength manipulation into a list-length manipulation and thereby induce a positive list-strength effect. In Experiment 1, this result was produced for single-word recognition and for two types of sentence recognition. In Experiment 2, both words and sentences were repeated together, which should have caused repetitions to be stored in a single, stronger, trace. As expected, the list-strength effect was eliminated. A sentence trace model was fit to the data, supporting the account of Shiffrin et al. (1990) and supporting an account of word and sentence recognition in which activation is summed for representations of all list items. The results from the two studies are inconsistent with most current models of memory (as shown by the theoretical analyses of Shiffrin et al., 1990) and pose an additional challenge for theory.

Ratcliff, Clark, and Shiffrin (1990) studied the list-strength effect: If the strength of storage of some items on a list is increased, will memory for other list items be harmed? Conversely, will strength decreases improve memory for other items? If so, a positive list-strength effect is said to have occurred. The data showed no list-strength effect or a slightly negative list-strength effect when memory was assessed by single-item, yes/no recognition tests. This was true when strength was varied by changing presentation time or by changing the number of presentations of an item, even in spaced fashion. On the other hand, a large positive list-strength effect was found in free recall and, at most, a small positive list-strength effect was found in cued recall.

Shiffrin, Ratcliff, and Clark (1990) showed that such results are inconsistent with many current models of memory. Examining a number of current models, they could not find variants of composite memory models, including certain types of connectionist models, that can deal with the recognition results, especially when strength is varied via spaced repetitions. Models positing storage of separate traces in a way that does not produce mutual degradation of the traces could in theory handle the findings. However, one such model, MINERVA 2 (Hintzman, 1986), had difficulty predicting both the recall and recognition results. The pattern of data was consistent with certain new variants of the SAM model (e.g., Gillund & Shiffrin, 1984).

The SAM model and its successful variants will be discussed later. At this point, it is sufficient to mention certain general hypotheses that Shiffrin et al. (1990) incorporated in their proposed model: (1) storage of different items is in separate memory traces (without mutual degradations among traces), rather than in a composite memory trace (with mutual degradations among traces); (2) repetitions of an item within a list are accumulated into a single, stronger, memory trace (at least for the conditions examined by Ratcliff et al., 1990); (3) the variance of activation of each trace is roughly constant, regardless of the strength of the stored item (when the cue and the item encoded in the trace are not the same and had not been rehearsed together); (4) recognition and recall operate via different retrieval processes, the recognition decision based on the summed activations of all traces, and recall involving separate access to separate traces (possibly via a search-and-sampling process). Hypothesis 1 is needed so that degradation by repeated items will not be inevitable; Hypothesis 2 allows repetitions to act differently from the way in which new presentations act; Hypothesis 3 allows performance to be independent of strength of storage of other items; Hypothesis 4 allows different list-strength results to occur in recall and recognition. The first three of these hypotheses will be examined further in the present paper.

To set the stage, consider three lists schematically represented as follows: List 1, ABCD; List 2, ABCDCD; List 3, ABCDABCD. We are interested in recognition performance for a test of A in Lists 1 and 2 and D in Lists 2 and 3. That is, what effect will spaced repetitions of some other items on a list have upon recognition of a nonrepeating item (A in Lists 1 and 2) or upon a repeated item (D in Lists 2 and 3)?

This research was supported by NIMH Grant 12717 and AFOSR Grant 870089 to the second author. Send reprint requests to Richard M. Shiffrin, Psychology Department, Indiana University, Bloomington, IN 47405.

Recognition sensitivity is usually defined theoretically by

$$d' = \frac{E[F \mid T] - E[F \mid D]}{\mathrm{Var}[F \mid D]^{1/2}}, \qquad (1)$$

in which E refers to expectation, Var to variance, F to some measure on which a decision is based (often termed *familiarity*), T to a target (an old item from the study list), and D to a distractor (an unstudied item). This definition of d' assumes that distributions of familiarity for targets and distractors are normally distributed and of equal variance. Even though the models of interest do not necessarily predict equal variances, and even though the data show slightly higher variances for targets (Ratcliff & McKoon, in press), it facilitates discussion to adopt this definition.

There are basically two classes of extant models. In one class, the repetitions directly degrade the representation(s) of the tested item, causing the numerator in Equation 1 to decrease; usually, in such models, the denominator will increase due to increased noise caused by repetitions. Both factors lead to a decrease in performance consistent with a positive list-strength effect. Models in this first class include the marking theory of Glanzer and Bowles (1976; see also Bowles & Glanzer, 1983) and certain types of connectionist models (e.g., Ackley, Hinton, & Sejnowski, 1985; Kosko, 1987; see the discussions in Ratcliff et al., 1990, and Shiffrin et al., 1990). In the second class of models, the repetitions affect mean familiarity equally for targets and distractors so that the numerator of Equation 1 is unchanged. However, in these models repetitions increase the variance, which leads to the prediction of a positive list-strength effect. Examples of models of the second class include Eich (1982, 1985), Anderson (1973), Gillund and Shiffrin (1984), Hintzman (1988), Pike (1984), and Murdock (1982).

Among the assumptions Shiffrin et al. (1990) needed to eliminate the predicted positive list-strength effect due to spaced repetitions was the hypothesis that all repetitions of a given item (word) are accumulated into a single memory trace (of course, this trace could be quite complex, containing frequency information among other things). What is interesting about the trace accumulation hypothesis is the following corollary: If repetitions of an item could be forced to occupy different memory traces, then a positive list-strength effect would be predicted. One can argue that encoding repetitions in different traces in effect turns a list-strength manipulation into a list-length manipulation, with repetitions harming memory just as do extra presentations. Whatever the explanation, this idea provides the motivation for the present studies: In one experiment, we will induce separate storage of repetitions, thereby producing a positive list-strength effect in recognition. In a second, otherwise similar, experiment, we will induce accumulated storage of repetitions, thereby eliminating the list-strength effect.

Our approach to forcing separate storage of repeated items is fairly simple. Suppose a given word is repeated in quite different, distinct contexts so that the encodings it is given are specific to each context and largely independent of each other. It then seems reasonable that different memory traces of that word might be stored (the roles of distinctiveness and independence of encodings in producing different traces would be interesting to explore but are not the focus of the present research). It is fairly likely that different traces could be produced in this manner if one took the paradigm to the extreme, as could happen if two encodings produced different, unrelated meanings of a word. For example, suppose a subject received the following two sentences at a widely spaced interval: (1) "Two rubies and a diamond made the necklace valuable." (2) "The speedy shortstop circled the diamond for an inside-the-park home run." One would expect "diamond" to be stored in two distinct memory traces. In our research, we intentionally shied away from manipulations this extreme (though, of course, any two different contexts tend to produce different semantic encodings for a word). Instead, we repeated words in the context of different five-word sentences.

The use of differing sentence contexts could produce separate storage for either of two reasons. First, the unit of storage could be the individual word; different words would be stored separately, and repetitions of a given word in different sentences could be stored separately due to different and perhaps independent coding. Second, the unit of storage could be the sentence rather than the individual word (as argued by Shiffrin, Murnane, Gronlund, & Roth, 1989). In this case, the different sentences would very likely be stored separately (at least if their word overlap or similarity were not too high). Whether words or sentences are stored separately, the result could be viewed as a list-length effect: memory would contain more traces for the lists with more repetitions. If sentences were the units of storage, there would not only be more traces when repetitions were used, but these traces would all be different. Although either model would give rise to a positive list-strength effect, other predictions might differ, so it could be important to know what the units of storage are. The studies contain certain conditions that provide evidence concerning the appropriate storage units.

In our studies, we used what Ratcliff et al. (1990) termed the *mixed-pure paradigm*. All lists contain the same number of different, unique words. Three types of list were presented to the subject. The *pure-weak* lists contained only words that were presented once each. The *pure-strong* lists contained only words that were presented three times each. Half of the mixed lists were made up of words presented once each, and half were words presented three times each. In the first study, words were rearranged into different sentences so that word repetitions took place in the context of different sentences. In fact, no two sentences shared more than one word. It is in this case that repetitions are predicted to be stored separately, and hence a positive list-strength effect is predicted to appear.

The logic of the mixed-pure paradigm is straightforward. Assume, for the sake of argument, that other stronger items harm performance. Then strong items (those given three repetitions) in pure-strong lists have stronger list competitors than do strong items on mixed lists and thus should exhibit poorer performance. Weak items (those only appearing once) in pure-weak lists have weaker competitors than do weak items on mixed lists and thus should exhibit better performance. (Of course, these effects could occur because there are *more* traces in memory rather than *stronger* traces in memory; we use the term *list strength* nonetheless because the items repeated are themselves better remembered.) These two predictions (or their combination) will be of primary interest in the present experiments.

In addition to manipulating the form of repetitions, we also manipulated type of recognition test. In the previous studies, all recognition tests were successive *old-new* judgments of singly presented test words. We utilized these single-item tests in the present experiments, as well as two other types of test, both involving *old-new* judgments of whole sentences. In what we term the *one-new* condition, targets were studied sentences and distractors were (different) studied sentences with one word replaced by a new word that had not appeared in any studied sentence. In what is termed the *intact-rearranged* condition, targets were again studied sentences and distractors were sentences made of one word from each of five different studied sentences. Thus, all five words in the distractor sentences had appeared before, but no word had previously appeared in the same sentence with any of the other words. In both of these sentence test conditions, distractors were semantically coherent sentences and could not be discriminated from targets on this basis. These sentence test conditions were included to test certain predictions of global activation models of recognition and also to provide evidence concerning the hypothesis that sentences may have been the units of storage.

EXPERIMENT 1

In Experiment 1, words were repeated but sentences were not. In fact, no presented sentence shared more than one word with any other sentence.

Method

Procedure. There were three types of study lists: pure-weak (PW), pure-strong (PS), and mixed (M; strong words from the mixed list are denoted MS, and weak words MW). Each study list utilized 50 different words. The words were arranged in sentences that were of the form "The adjective noun verbed the adjective noun" (e.g., "The alert boy found the magic sword."). In PW lists, the 50 words were presented as 10 five-word sentences with no words being repeated. In PS lists, each word was repeated three times in three different sentences; the 50 different words were presented as 30 sentences in such a way that no sentence shared more than one word with any other. These 30 sentences actually consisted of two subgroups of 15, whose words did not overlap, presented in randomly intermixed order. The mixed lists consisted of one group of 15 sentences with repeated words, just as in one half of the PS lists, and one group of five sentences with no repeated words, just as in one half of the PW lists. These 20 sentences were presented in randomly intermixed order. Thus, the PW conditions used 50 words, each presented once, in the form of 10 sentences. The PS conditions used 50 words, each presented three times, in the form of 30 relatively unique sentences. The mixed conditions in effect consisted of one half of each of the pure lists: 25 once-presented words in five sentences and 25 three-times-presented words in 15 sentences, for a total of 50 words in 20 sentences.

Stimulus materials consisted of sentences grouped in sets of 12, one such set for each study-list-test-list combination. Each word was assigned a particular sentence position (e.g., first adjective, verb, etc.). The sets were constructed so that any word from any of the five sentence positions could be combined with any other word from another sentence position. This pair could then be combined with any third word from a third sentence position and so on. The result was a set of 12 sentences that could be rescrambled in any order and still make semantic sense, subject to the constraint that an individual word could only appear in one sentence position. The 12 sentences in each set were experimentally manipulated in two subsets of 6 sentences each. From each of these subsets, five sentences appeared as study sentences and the words in the sixth were used to construct distractors. The construction of both study and distractor sentences within each subset was randomized for each subject. To construct sentences for strong study lists in which words were repeated three times, the five study sentences were rearranged in such a way that no two words ever appeared together in the same sentence more than once. Overall, there were nine sets of 12 sentences; assignment of sentence set to the nine experimental conditions (three list types × three test types) was randomized over experimental sessions.

For each study condition, there were three test conditions for a given subject. Each study-list-test-type combination employed a different set of words. In single-item recognition, one word from each sentence position (five in all) was chosen from the 25 words (five sentences) in each subgroup presented at study and used as a target. All words from the distractor sentence for each subgroup were used as distractors. Since there were two subgroups per list, there were 10 targets and 10 distractors tested in random order for a given study list. Each test word was presented successively for *old-new* judgments that had to be made in a maximum of 5 sec.

In one-new recognition, target sentences in the PW condition were five studied sentences, randomly chosen. The distractors were the other five studied sentences with one word in each sentence replaced by a new word (one in each sentence position). In the PS condition, targets were 10 randomly chosen study sentences and distractors were 10 more study sentences, each with a word replaced by a new word (two in each study position). In the mixed condition, the five weak-item sentences were converted randomly into two targets and three distractors or the reverse. The strong-item sentences were converted randomly into five targets and five distractors.

For intact-rearranged recognition, target sentences were 10 studied sentences; distractor sentences were constructed by using one word (in correct sentence position) from each of the five studied sentences in a subgroup. Five weak distractors were made from once-presented words and five strong distractors were made from three-times-presented words. In repeated-word cases, the distractors and targets shared the property that no two sentences (study or test) shared more than one word.

The subjects received nine study-test blocks in one session in random order. A 10-sec visual warning signal appeared before each block. Each study sentence then appeared for 8 sec. After study, a 10-sec visual signal indicated the nature of the recognition test to follow. The subjects had a maximum of 5 sec to respond to each recognition test, all of which were presented in random order. There was a 20-sec break between blocks. Responses were made on two keys of a keyboard, designated *old* and *new*; the mapping of hand to response was varied across subjects.

Each session began with instructions and 15 practice trials (five study-test trials for each of the three test types) using words not appearing elsewhere in the session.

Apparatus. Presentation and randomization of stimuli, timing, and collection of responses were controlled by a DEC PDP-11/34 computer. Stimuli were presented on a CRT screen.

Subjects. Ninety-eight subjects at Indiana University participated in partial fulfillment of course requirements. We discarded the data of one subject who failed to follow instructions. The subjects were run in groups of 3 to 6 in separately controlled booths. Sessions took about 1 h to complete.

Materials. The requirements to choose words that could be rearranged into semantically coherent sentences defeated attempts to utilize a carefully selected and circumscribed set of words from standard sources. The words utilized tended to be relatively familiar and fairly short.

Table 1
p(H), p(FA), d', d'', and $\sigma(d'')$ Values for Experiment 1

Condition	p(H)	p(FA)	d'	d''	$\sigma(d'')$
		Single-Item Test			
PW	0.726	0.216	1.38	1.53	0.08
MW	0.655	0.259	1.04	1.21	0.10
PS	0.816	0.170	1.85	2.01	0.09
MS	0.835	0.149	2.01	2.15	0.09
		Intact-Rearranged Test			
PW	0.664	0.194	1.29	1.46	0.10
MW	0.620	0.196	1.16	1.28	0.11
PS	0.588	0.326	0.67	0.73	0.08
MS	0.627	0.339	0.74	0.84	0.10
		One-New Test			
PW	0.684	0.194	1.34	1.50	0.11
MW	0.536	0.190	0.97	1.16	0.13
PS	0.607	0.206	1.09	1.22	0.08
MS	0.669	0.196	1.29	1.45	0.10

Note—PW = pure-weak; MW = mixed-weak; PS = pure-strong; MS = mixed-strong; d' and d'' are two measures of performance (see text); σ = standard deviation; p(H) = probability of hit; p(FA) = probability of false alarm.

Results and Discussion

Trials in which a subject did not respond within 5 sec were counted as 0.5 correct response and 0.5 incorrect response. Over conditions, the percentage of such trials ranged from 0.000 to 0.046, with the highest values for targets in the one-new condition and distractors in the intact-rearranged condition. We give the average hit and false-alarm rates for each condition in the tables, as well as d' scores calculated from these group averages. However, to carry out within-subject statistical comparisons, we calculated a d'-like measure for each subject for each condition, on the basis of that subject's hit and false-alarm rates. Since these rates were based on a maximum of 10 observations per condition for each subject, a number of cases occurred with probability of *old* given target or *old* given distractor equal to either 0.0 or 1.0. To analyze scores in terms of a d'-like measure, scores of 0.0 were converted to 0.05 and scores of 1.0 were converted to 0.95 (the pattern of results was not different when other methods of analysis were used). The scores that result from averaging individual subjects' performance scores calculated from these adjusted hit and false-alarm rates are referred to in this paper as d'' and are also given in the tables. The percentage of subjects with hit rates of 1.0 ranged over conditions from 0.000 to 0.459, with all but 1 subject in the range 0.000 to 0.224; the percentage of subjects with false-alarm rates of 0.0 ranged over conditions from 0.031 to 0.541, with all but 3 subjects in the range 0.153 to 0.347. The d'' scores across conditions were combined by contrasts into a single number for each subject and assessed by t tests (standard errors of the mean may be obtained using the t value).

In Table 1 are given the hit and false-alarm rates for each condition averaged across subjects, the d' calculated from these rates, the d'' values calculated by averaging d'' values for each subject, and the standard error of the mean for the d'' values. Obviously, performance was well above chance levels (d' or d'' = 0.0) in all conditions. Even the least significant result, the intact-rearranged test for strong items on mixed lists gave $t(96) = 8.07$, $p < .0001$.

The fact that the intact-rearranged results are well above chance at once rules out one simple model for recognition in that condition. If subjects probe memory with just single words and then combine the results for the words in the test sentence, discrimination between intact and rearranged sentences would not be possible, because each word by itself is equally familiar in both cases.

A major concern of this study is the comparison among conditions when there are single-item tests. Strong items were much better than weak items for both pure and mixed lists [for pure lists, $t(96) = 4.79, p < .001$; for mixed lists, $t(96) = 7.85, p < .001$]. To assess the presence of a list-strength effect, we calculated the sum of pure-weak minus mixed-weak plus mixed-strong minus pure-strong), because stronger other items contribute to the second term in each case and hence should make the second terms smaller. (In Ratcliff et al., 1990, a ratio of d' ratios was calculated to assess the presence of a list-strength effect rather than a sum of differences. Although the ratio of ratios simplifies the theoretical derivations, ratios are too sensitive to deviations to use for individual subjects when the number of observations per subject condition is as small as it is in the present study.) The list-strength results, given as differences and sums of differences, are given for convenience in Table 2 for all conditions. For single-item tests, the sum was significantly positive [$t(96) = 2.982, p < .005$]; taken separately, pure-weak/mixed-weak was significantly positive [$t(96) = 2.558, p < .01$], but mixed- strong/pure-strong, while positive, did not reach significance [$t(96) = 1.318, p > .05$].

Taking into account the studies reported in Ratcliff et al. (1990) and unpublished data of Caulton and Shiffrin (1988), to be described shortly, this is the first positive list-strength effect for recognition testing we have seen. Our ability to produce it tends to validate the empirical procedure employed in this experiment and the similar procedures employed in the studies of Ratcliff et al. (1990). The method used to produce the effect was based

Table 2
List-Strength Effects for Experiments 1 and 2

List-Strength Differences	Tests Single Item	 Intact-Rearranged	 One-New
	d'', Experiment 1		
PW−MW	0.322	0.180	0.341
MS−PS	0.145	0.110	0.234
(PW−MW)+(MS−PS)	0.467	0.290	0.576
	d'', Experiment 2		
PW−MW	0.086	−0.009	0.007
MS−PS	0.032	−0.132	0.250
(PW−MW)+(MS−PS)	0.118	−0.141	0.258

Note—PW = pure-weak; MW = mixed-weak; PS = pure-strong; MS = mixed-strong; d'' = measure of performance.

on the theory put forward by Shiffrin et al. (1990), so the finding also lends support to that theory.

The intact-rearranged conditions showed a substantially different pattern of results. Because the terms *strong* and *weak* are no longer appropriate, we use the terms *repeated-word* sentences (strong) and *nonrepeated-word* sentences (weak). First, it should be noted in Table 1 that repeated-word sentences are inferior to the nonrepeated-word sentences for both pure [$t(96) = 6.06, p < .001$] and mixed [$t(96) = 3.45, p < .001$] lists. Assuming that sentences are the units of storage and also the probe units for intact-rearranged tests, PS lists are three times as long as PW lists, and repeated-word sentences are worse because of the list-length effect. On mixed lists, repeated-word test sentences are similar to three times as many studied sentences as nonrepeated-word test sentences, making MS sentences worse than MW sentences. A detailed analysis of these results will follow presentation of the models.

Whatever the main effect of repeating words on sentence recognition, the predictions for the list-strength effect seem clear. Having more other sentences and more repetitions of other words should by hypothesis reduce performance. Thus, PW should be superior to MW and MS should be superior to PS. Although the data exhibited such trends, the differences in Table 2 did not reach significance [$t(96) = 1.51, p > .05$; $t(96) = 0.83, p > .10$]. In combination, the trend toward a positive list-strength effect also failed to reach significance [$t(96) = 1.58, p > .05$].

The one-new conditions exhibit yet another pattern of results. Overall, the results for repeated-word sentences and nonrepeated-word sentences did not differ. However, for pure conditions, nonrepeated-word sentences were superior to repeated-word sentences [$t(96) = 2.47, p < .02$]. In the mixed condition, repeated-word sentences were superior to nonrepeated-word sentences [$t(96) = 2.08, p < .05$]. Interpretation of these results must await the context of the models to be presented later. Despite the new relationship of repeated-word to nonrepeated-word performance seen in this condition, a positive list-strength effect occurs: PS items are worse than MS [$t(96) = 2.28, p < .025$], and PW items are better than MW [$t(96) = 2.41, p < .01$]. Overall, the list-strength effect is significant [$t(96) = 3.32, p < .005$].

Before turning to models of these results, we must ask whether the findings of positive list-strength effects in Experiment 1 are due to the use of sentences or to the nature of strengthening items (i.e., the rearrangement of repeated words into new groups). If the rearrangements are the key, as hypothesized, then the use of sentences in which strengthening occurs only by repeating entire sentences should result in the elimination of the list-strength effect. This idea is the basis for Experiment 2.

EXPERIMENT 2

Method

The method for Experiment 2 was identical to that of Experiment 1 in all respects, with the exception that repetitions of words on strong lists or the strong words on mixed lists were accomplished by repeating an entire sentence. One change also had to be made in testing to keep Experiment 2 as similar as possible to Experiment 1. In one-new testing of strong sentences, all presented sentences were tested as targets and the same sentences were tested as distractors, with a randomly chosen word replaced by a new word. The subjects were 95 Indiana University undergraduates, who participated for partial fulfillment of course requirements and who had not participated in Experiment 1.

If repeated sentences act as did repeated words in the studies of Ratcliff et al. (1990), then the repetitions might be accumulated into a single memory trace and list-strength effects might disappear.

Results and Discussion

As was the case in Experiment 1, performance was well above chance in all conditions. Table 3 gives average hit and false-alarm rates, d'' and d' for each condition, and the standard deviations of the mean for d''. Table 2 gives the list-strength comparisons.

For single-item tests, strong items were superior to weak items for both pure [$t(94) = 4.72, p < .001$] and

Table 3
$p(H)$, $p(FA)$, d', d'', and $\sigma(d'')$ Values for Experiment 2

Condition	p(H)	p(FA)	d'	d''	$\sigma(d'')$
		Single-Item Test			
PW	0.760	0.192	1.57	1.69	0.07
MW	0.693	0.164	1.47	1.61	0.10
PS	0.827	0.125	2.09	2.17	0.07
MS	0.874	0.164	2.11	2.20	0.08
		Intact-Rearranged Test			
PW	0.712	0.140	1.64	1.78	0.10
MW	0.693	0.120	1.66	1.79	0.10
PS	0.844	0.092	2.35	2.41	0.09
MS	0.854	0.135	2.15	2.27	0.11
		One-New Test			
PW	0.661	0.177	1.35	1.51	0.11
MW	0.637	0.166	1.32	1.50	0.15
PS	0.809	0.118	2.06	2.18	0.09
MS	0.867	0.101	2.39	2.43	0.09

Note—PW = pure-weak; MW = mixed-weak; PS = pure-strong; MS = mixed-strong; d' and d'' are two measures of performance (see text); σ = standard deviation; p(H) = probability of hit; p(FA) = probability of false alarm.

mixed [$t(94) = 5.54, p < .001$] conditions. The sum of differences indexing the list-strength effect (see Table 2) did not reach significance [$t(94) = 0.86, p > .10$], nor did the mixed-strong-pure-strong difference [$t(94) = 0.371$, $p > .10$] or the pure-weak-mixed-weak difference [$t(94) = 0.85$, $p > .10$]. These results replicate those of Ratcliff et al. (1990) using pairs of study words and single study words.

For intact-rearranged testing, strong items were superior to weak items for both pure [$t(94) = 6.08$, $p < .001$] and mixed [$t(94) = 4.62, p < .001$] conditions. The list-strength effect indices were all negative, though not significantly so [combined, $t(94) = -0.98$, $p > .10$].

For one-new testing, strong items were superior to weak items for both pure [$t(94) = 5.63, p < .001$] and weak [$t(94) = 6.12$, $p < .001$] conditions. The list-strength effect indices were positive, but not significantly so [combined, $t(94) = 1.49, p > .10$; for the strong conditions only, $t(94) = 2.62, p < .01$; for weak conditions, $t(94) = 0.047$, $p > .10$].

The results from Experiments 1 and 2 differed in many ways. Overall, performance was higher in Experiment 2 [$t(190) = 8.5$, $p < .001$]. For single-item tests, however, the difference was slight [d'' was higher by 0.19 in Experiment 2; $t(190) = 2.29, p < .05$]. The strong-item advantage was higher by 0.175 d'' in Experiment 1, but not significantly so [$t(190) = 1.15, p > .05$]. However, the list-strength effect combined index was significantly higher in Experiment 1 [d'' was 0.35 higher in Experiment 1; $t(190) = 2.80$, $p < .005$].

For intact-rearranged testing, performance was much superior in Experiment 2 [$t(190) = 10.18$, $p < .001$], especially for strong items [for strong items, $t(190) = 15.7, p < .001$; for weak items, $t(190) = 3.19, p < .01$]. The comparison of strong to weak went significantly in opposite directions in the two studies. The list-strength effect was 0.4 d'' higher in Experiment 1 [$t(190) = 1.84$, $p < .05$].

For one-new testing, Experiment 2 was superior overall [$t(190) = 6.97, p < .001$] and superior for strong items [$t(190) = 9.14$, $p < .001$], but not for weak items [$t(190) = 0.95$, $p > .10$]. The strong-item advantage was much higher overall in Experiment 2 [$t(190) = 6.4$, $p < .001$] for both pure tests alone and mixed tests alone. The list-strength effect was higher in Experiment 1, but not significantly so [$t(190) = 1.29$, $p > .05$].

In summary, a significant, positive list-strength effect was observed in Experiment 1 for the single-item and one-new conditions; the positive list-strength effect for the intact-rearranged condition did not quite reach significance. This result is consistent with the hypothesis that the use of sentence contexts would cause separate storage of words or sentences, thereby producing interference. The results from Experiment 2 showed two positive and one negative list-strength effect, none near significance overall, although the one-new effect for strong items only reached significance. These results are generally consistent with those of Ratcliff et al. (1990) and pose problems for models positing substantial interference among items during the process of storage. Significantly more positive list-strength effects were found in Experiment 1 than in Experiment 2 for single-item and intact-rearranged tests, but not quite so for one-new tests. These results suggest that the change of context for word repetitions, rather than the use of sentences, was crucial.

Although we interpret the differences between Experiments 1 and 2 in terms of separate versus single storage of repetitions, these differences could be due to different rehearsal effects in mixed lists in the two studies. If Experiment 1 induced separate storage of all sentences, then there may have been no inducement to shift rehearsal or effort from the sentences with repeated words to those without. However, in Experiment 2, the subjects may have borrowed rehearsal or effort from repeated sentences to give to nonrepeated sentences. In Ratcliff et al. (1990), little evidence for such borrowing was found when study lists contained single words or word pairs. However, evidence bearing on this issue would be desirable for Experiment 2.

To test this hypothesis, several additional analyses were carried out for Experiment 2. First, weak items in mixed lists were analyzed in terms of their study positions, by fifths of the list (i.e., every four items). If rehearsal is borrowed from repeated sentences and given to nonrepeated sentences, the borrowing ought to be at a maximum late in the study list when repeated items have begun to receive second and third repetitions. The data are shown in Table 4. There was, at most, weak evidence for improvement of weak items late in the lists [for single-item tests, $t(94) = 1.1, p > .10$; for intact-rearranged tests, $t(94) = 2.16, p < .05$; for one-new tests, $t(94) = -0.38, p > .10$].[1] Of course, the serial position trends seen for weak items, if real, could be due to quite a number of factors other than rehearsal borrowing.

Although the serial position analysis is hardly conclusive, we shall assume in the remainder of this article that redistribution is not an important factor in our experiments. At first glance, our results seem to exhibit a complex pattern of differences among the three test conditions and differences between the two studies. We predicted that a positive list-strength effect would be found in Experiment 1 and eliminated in Experiment 2. This pattern was found; although the list-length effect in Experiment 2 was positive in two of the three conditions, the effects

Table 4
Hit Rates in Successive Fifths of Study Positions for Experiment 2 Mixed Condition, Once-Presented Items

	Single Item	Intact-Rearranged	One-New
1st	0.58	0.67	0.61
2nd	0.68	0.66	0.60
3rd	0.75	0.72	0.62
4th	0.73	0.77	0.61
5th	0.74	0.77	0.80

in either direction were far from significant. However, interpretation of the pattern of results is greatly facilitated by models. We begin with the model proposed by Shiffrin et al. (1990), slightly modified to deal with the sentences used in the present experiments.

A SAM MODEL FOR SENTENCE RECOGNITION

Units of Storage

The SAM model for recognition assumes that probe cues activate separately stored memory traces (termed *images*) and that the recognition decision is based on the value of the activation summed over all images. Thus, the first step requires determining the nature of the stored images; in the present instance, it is natural to ask whether the images are words or sentences (or both). Evidence bearing on this issue is obtainable from a comparison of the single-item and one-new performance levels. If single items are images, it is not hard to come to the intuitive judgment that one-new performance ought to be inferior to single-item performance. For example, if one treats each test word separately, the discrimination between target and distractor in one-new (ABCDE vs. ABCDX) differs from that for single-item tests (E vs. X) by the addition of four redundant words. These redundant words will add noise to the decision process and reduce performance. Alternatively, if all five items are used together to probe memory, the relative activations of the relevant image (E) will determine performance. Intuitively, the difference between activation of E by E alone as a cue and X alone as a cue should be greater than the difference between ABCDE and ABCDX: If overall activation is similar regardless of number of concatenated cues (e.g., if the individual cues are weighted inversely according to their total number), then the redundant words will directly reduce the difference. Alternatively, if activation is in proportion to total number of cues, then the overall list activation (and hence variability) will be higher when extra redundant cue words are used. We have analyzed a number of model variants within SAM, but in all cases where the images were assumed to be single words, the ratio of single-item d' to one-new d' was predicted to be greater than $\sqrt{5}$. In light of the virtually equal performance levels observed in our data, it seems clear that single-word images will not suffice within the SAM framework.

This line of reasoning suggests that, in cases where words in groups are likely to be stored as separate images rather than as group units, single-item testing would be superior to one-new. One example is seen in Clark and Shiffrin (1987). Triples of unrelated words were studied and a wide variety of testing conditions were used. A model assuming separate storage of word images was fit to the data with good success. The condition we term *one-new* in this report was termed *all-old* by Clark and Shiffrin. The single-item d' value was 1.05. The two-item one-new d' (AB vs. AX) was 0.90. The three-item one-new d' (ABC vs. ABX) was 0.83. In general, when stored images are single words, one would expect the advantage of single-item testing to grow as the number of redundant words added to the probe in one-new testing grows. This pattern is seen in the Clark and Shiffrin data, and should have occurred in the present data had all stored images been single words.[2]

A similar phenomenon is seen in a related testing paradigm termed *cued recognition*, which is similar to one-new testing except that the redundant items are all specified (and therefore may be thought of as "cues") and attention may be focused on only a single item. Although it would seem as if the cues could only help, in Gillund and Shiffrin (1984) and Clark and Shiffrin (1987), cued recognition was inferior to single-item testing. This would be expected if words were stored separately and if the redundant cue items were incorporated in the retrieval or decision process (and hence reduced the attention paid to the relevant test item). Such a result would not be expected if the group of presented items were stored as a unit. Clark and Shiffrin (1991) varied word frequency and presentation time in a similar paradigm; the results were consistent with the view that these factors controlled the likelihood of formation of higher order units.

The prediction that one-new recognition should be inferior to single-item recognition when words are stored separately may be true of models other than SAM, but perhaps not universally so. Bain and Humphreys (1988) discussed issues related to this in a paper arguing against higher order units. They based their conclusions mainly on studies using pairs of words. It may be that unitization of random word pairs is more difficult to produce than unitization of short sentences of the type used in the present article. Bearing on this issue is unpublished research by Caulton and Shiffrin (1988). In pilot work, they followed the logic of Experiment 1, but used random word pairs rather than sentences. In the experimental conditions, words were repeated but pairs were not. The results differed from those with sentences. The list-strength effect was actually slightly negative. It may be that the use of pairs did not produce an encoding context different enough to induce separate storage for repetitions.

Our present conclusion that sentences are stored as units coincides with that reached by Shiffrin et al. (1989). Using sentences similar to those of the present experiments and using recall paradigms and accuracy and reaction-time measures, they obtained evidence strongly favoring the view that sentences are stored as units in memory. On balance, the evidence suggests that single words are not the units in memory, and we shall assume so for the remainder of the article. Although various alternative models could be considered, including those positing *both* sentence and word units, we shall proceed with analysis of a model assuming the storage of sentence units only.

A SAM Model with Sentence Units

Assume that the units of storage (the memory images) are whole sentences, denoted I_i. These sentence units are complex patterns of information containing word information, among other things. Thus, the similarity of a test group of words to a given sentence image, and the concomitant activation caused by that test group, will depend on the number of words in common, among other factors. Assume that each sentence produces a distinct image, but that repetitions of a given sentence are accumulated into a single stronger image. When memory is probed for recognition with a context cue, C, and a test item, I_j, each memory image is activated and the sum of the activations (termed *familiarity*), F, is used for a decision (if F is greater than a criterion, an *old* response is given):

$$F(C,I_j) = \sum_{i=1}^{N} A(I_i|C,I_j). \tag{2}$$

We adopt a model of the type described by Shiffrin et al. (1990). Image activation is a function of the retrieval strength, S, for each cue separately, and the weights, ω_C and ω_I, given to the cues:

$$A(I_j|C,I_j) = [S(C,I_i)]^{\omega_c}[S(I_j,I_i)]^{\omega_I}. \tag{3}$$

For single-item testing, we assume context and the test item are the cues, given weights ω_C and ω_I. For intact-rearranged and one-new testing, we assume context and the sentence are the cues, given weights ω_C and ω_I (for now, we shall assume the weights are the same for single-item and sentence testing).

To derive predictions, a number of parameters are needed. Let $\{E[S(C,I_K)]\}^{\omega_c} = a(j)$, when the sentence encoded in image I_K has been repeated j times during study (j equals 1 or 3 in our studies). Let $\{E[S(I_l, I_K)]\}^{\omega_I} = e(i,j)$, when the sentence encoded in image I_K has been repeated j times during study and the test word I_l shares i words with that sentence (i equals 0 or 1; j equals 1 or 3 in our studies). Let $\{E[S(I_l,I_K)]\}^{\omega_I} = c(i,j)$, when the

Single item tests:

Experiment 1:

$$d'(pw) = \beta[e(1,1) - e(0,1)] / [10e^2(0,1)]^{1/2}$$
$$d'(mw) = \beta[e(1,1) - e(0,1)] / [20e^2(0,1)]^{1/2}$$
$$d'(ps) = 3\beta[e(1,1) - e(0,1)] / [30e^2(0,1)]^{1/2}$$
$$d'(ms) = 3\beta[e(1,1) - e(0,1)] / [20e^2(0,1)]^{1/2}$$

Experiment 2:

$$d'(pw) = \beta[e(1,1) - e(0,1)] / [10e^2(0,1)]^{1/2}$$
$$d'(mw) = \beta[e(1,1) - e(0,1)] / [5e^2(0,1) + 5\{a^2(3)/a^2(1)\}\, e^2(0,3)]^{1/2}$$
$$d'(ps) = \beta[e(1,3) - e(0,3)] / [10e^2(0,3)]^{1/2}$$
$$d'(ms) = \beta[e(1,3) - e(0,3)] / [5e^2(0,3) + 5\{a^2(1)/a^2(3)\}\, e^2(0,1)]^{1/2}$$

Intact-rearranged tests:

Experiment 1:

$$d'(pw) = \beta[c(5,1) + 4c(0,1) - 5c(1,1)] / [5c^2(1,1) + 5c^2(0,1)]^{1/2}$$
$$d'(mw) = \beta[c(5,1) + 4c(0,1) - 5c(1,1)] / [5c^2(1,1) + 15c^2(0,1)]^{1/2}$$
$$d'(ps) = \beta[c(5,1) + 4c(0,1) - 5c(1,1)] / [15c^2(1,1) + 15c^2(0,1)]^{1/2}$$
$$d'(ms) = \beta[c(5,1) + 4c(0,1) - 5c(1,1)] / [15c^2(1,1) + 5c^2(0,1)]^{1/2}$$

Experiment 2:

$$d'(pw) = \beta[c(5,1) + 4c(0,1) - 5c(1,1)] / [5c^2(1,1) + 5c^2(0,1)]^{1/2}$$
$$d'(mw) = \beta[c(5,1) + 4c(0,1) - 5c(1,1)] / [5c^2(1,1) + 5\{a^2(3)/a^2(1)\}c^2(0,3)]^{1/2}$$
$$d'(ps) = \beta[c(5,3) + 4c(0,3) - 5c(1,3)] / [5c^2(1,3) + 5c^2(0,3)]^{1/2}$$
$$d'(ms) = \beta[c(5,3) + 4c(0,3) - 5c(1,3)] / [5c^2(1,3) + 5\{a^2(1)/a^2(3)\}c^2(0,1)]^{1/2}$$

One-new tests:

Experiment 1:

$$d'(pw) = \beta[c(5,1) - c(4,1)] / [c^2(4,1) + 9c^2(0,1)]^{1/2}$$
$$d'(mw) = \beta[c(5,1) - c(4,1)] / [c^2(4,1) + 19c^2(0,1)]^{1/2}$$
$$d'(ps) = \beta[c(5,1) - c(4,1) + 2\{c(1,1) - c(0,1)\}] / [c^2(4,1) + 8c^2(1,1) + 21c^2(0,1)]^{1/2}$$
$$d'(ms) = \beta[c(5,1) - c(4,1) + 2\{c(1,1) - c(0,1)\}] / [c^2(4,1) + 8c^2(1,1) + 11c^2(0,1)]^{1/2}$$

Experiment 2:

$$d'(pw) = \beta[c(5,1) - c(4,1)] / [c^2(4,1) + 9c^2(0,1)]^{1/2}$$
$$d'(mw) = \beta[c(5,1) - c(4,1)] / [c^2(4,1) + 4c^2(0,1) + 5\{a^2(3)/a^2(1)\}c^2(0,3)]^{1/2}$$
$$d'(ps) = \beta[c(5,3) - c(4,3)] / [c^2(4,3) + 9c^2(0,3)]^{1/2}$$
$$d'(ms) = \beta[c(5,3) - c(4,3)] / [c^2(4,3) + 4c^2(0,3) + 5\{a^2(1)/a^2(3)]c^2(0,1)]^{1/2}$$

Figure 1. Predictions for the conditions of both experiments.

sentence encoded in image I_K has been repeated j times during study and the test sentence I_l shares i words with that sentence (i equals 0, 1, 4, or 5; j equals 1 or 3 in our studies).

The derivations cannot be carried out without making distributional assumptions about the retrieval strength values. Following Shiffrin et al. (1990), we assume that a strength value has a distribution whose standard deviation rises linearly with the mean. We can then calculate theoretical d' predictions for the various conditions in the two experiments. The method is illustrated in the Appendix, and the predictions are given in Figure 1.

Can we assume that the parameter values are constant across conditions and experiments? Although the pure-weak condition was identical in both experiments in all three conditions, performance in Experiment 2 was higher in all three pure-weak conditions. Thus, we decided to let parameters vary between the two studies. Within a given study, since subjects do not know the test condition in advance, only the weights assigned to cues might change between conditions; however, we decided to see how well the model would fare when parameters are held fixed across the conditions of a given study.

How well does this model do? If, as assumed by Gillund and Shiffrin (1984), $e(0,1)$ equals $e(0,3)$ and $c(0,1)$ equals $c(0,3)$, then it would not be possible to predict correctly the Experiment 2 results—a strongly positive list-strength effect would be predicted. Thus, we adopt Shiffrin et al.'s (1990) differentiation assumption, which suggests that $c(0,3)$ and $e(0,3)$ should be less than $c(0,1)$ and $e(0,1)$, respectively. The idea is that the activation of a *stronger* (different) image is *less*, presumably because the differences between text item and image become more salient as the image strength increases. For simplicity, let us assume that $a(3)/a(1) = c(0,1)/c(0,3) = e(0,1)/e(0,3)$. For each condition of Experiment 2, we then can substitute this result into the equations in Figure 1 and get $d'(\text{PW}) = d'(\text{MW}) < d'(\text{PS}) = d'(\text{MS})$. This prediction is quite close to the data.

It is not so easy to predict the results of Experiment 1. In particular, the single-item predictions for $d'(\text{MS})$, $d'(\text{PS})$, $d'(\text{PW})$, and $d'(\text{MW})$ must be in the ratios $(3/\sqrt{2})::(\sqrt{3})::(1)::(1/\sqrt{2})$. Although the ordering is predicted correctly, the ratios in the data are all smaller than the predictions. To take one example, the strong-to-weak ratio in the mixed condition is predicted to be 3.0, whereas the observed ratios are 1.78 for d'' and 1.93 for d'.

Table 5
Experiment 1 Pure-Strong Conditions: Probability of an *Old* Response for Successive Fifths of Study Positions

	One-New Condition		Intact-Rearranged Condition
	Hit Rate	False-Alarm Rate	Hit Rate
1st	0.71	0.31	0.59
2nd	0.67	0.24	0.62
3rd	0.70	0.28	0.61
4th	0.65	0.28	0.68
5th	0.62	0.35	0.70

The three-to-one prediction and the other predicted ratios are a direct consequence of the assumption that three-times-presented sentences are represented three times in memory. Even if words, rather than sentences, were the units of memory storage, separate representations for repetitions of a word would result in similar predicted ratios. However, if extra sentences or repeated words do not produce a linear increase in strength, then smaller ratios could be predicted. For example, if sentences in later study positions in lists with repeated items were stored less strongly, then such an outcome would occur.

Evidence concerning the possibility can be obtained from serial position functions. Relevant data are given in Table 5. For the pure-strong, intact-rearranged test conditions of Experiment 1, hit rates are given for successive fifths of the study list; for the pure-strong, one-new test conditions of Experiment 1, hit and false-alarm rates are given for successive fifths of the study list. The hit rates for intact-rearranged testing did not decrease with serial position. The hit rates for one-new testing decreased, but the false-alarm rates increased (both hits and false alarms should decrease if storage strength drops with serial position). These data provide little evidence in favor of a serial position explanation of the single-item discrepancies, especially when one notes that quite large serial position shifts would be needed to deal with the discrepancy between the predicted ratio of 3:1 and the observed ratios less than 2:1.

Although quite different models could be considered, the difficulties enunciated by Shiffrin et al. (1990) in finding *any* model capable of predicting the list-strength findings leads us to make a minimal change in the proposed model. This change involves context-sensitive coding (e.g., Bain & Humphreys, 1988; Clark & Shiffrin, 1987). Suppose a word's coding varies depending on the sentence in which it appears. Suppose further that such a coding induces a tendency for a similar semantic encoding to be given to that word when presented for test. Much evidence for this assumption is available in the literature on coding effects in implicit learning. For example, Jacoby and Witherspoon (1982) showed that the use of an orienting question to prime a particular meaning of a homophone later increased the probability of the corresponding spelling to be produced to an auditory presentation, even in the absence of any explicit memory for the prior occurrence (see also Richardson-Klavehn & Bjork, 1988; Schacter, 1987). Such results suggest that the semantic encoding of a test item (which forms part of the probe of episodic memory) is determined by prior encoding contexts.

These ideas have a direct application in the present experiments. When a word is presented in only one sentence, its coding at test will tend to match that of the single-sentence context, producing a high value of $e(1,1)$. However, when a word appears in three different sentences, its coding will take on three different forms; no one of these ought to be as highly primed at test as the single coding that occurs when only a single sentence has been used. The effect will be to reduce the value of $e(1,1)$ for

Table 6
d'' Values and Predictions

Condition	Experiment 1 Predicted	Experiment 1 Observed	Experiment 2 Predicted	Experiment 2 Observed
	Single-Item Test			
PW	1.61	1.53	1.65	1.69
PS	1.87	2.01	2.18	2.17
MW	1.14	1.21	1.65	1.61
MS	2.29	2.15	2.18	2.20
	Intact-Rearranged Test			
PW	1.45	1.46	1.78	1.78
PS	0.84	0.73	2.34	2.41
MW	1.16	1.28	1.78	1.79
MS	0.93	0.84	2.34	2.27
	One-New Test			
PW	1.47	1.50	1.50	1.51
PS	1.22	1.22	2.30	2.18
MW	1.26	1.16	1.50	1.50
MS	1.39	1.45	2.30	2.43

Note—PW = pure-weak; MW = mixed-weak; PS = pure-strong; MS = mixed-strong; d'' is a measure of performance (see text). Under the differentiation assumption described in the text, the predictions for pure and mixed cases in Experiment 2 are equal for a given strength; hence, there are just two free parameters per condition and these are chosen for convenience to be the expressions given in Figure 1 for the PW and PS conditions. The best-fitting values are those given in this table. For Experiment 1, there are six identifiable parameters, whose best-fitting values are $[e(1,1)/e(0,1)] = 6.09$; $f = 0.725$; $c(0,1) = 1.97$; $c(1,1) = 3.16$; $c(4,1) = 7.09$; $c(5,1) = 14.0$.

the two strong-item test conditions. This hypothesis is most simply captured by multiplying the value of this parameter by a fraction, *f*. The predictions for the strong single-item test conditions of Experiment 1 then become

$$d'(\text{ps}) = 3\beta[fe(1,1)-e(0,1)]/[30e^2(0,1)]^{1/2} \quad (4a)$$

and

$$d'(\text{ms}) = 3\beta[fe(1,1)-e(0,1)]/[20e^2(0,1)]^{1/2}. \quad (4b)$$

Otherwise, the equations given in Figure 1 remain unchanged. The predictions for this context-sensitive coding version of the model are given in Table 6, as are the estimates of the identifiable parameter values (six for each experiment, since group differences led us to estimate parameters separately for the two experiments). The fit is fairly good overall, though the predictions are still a bit discrepant for the single-item conditions of Experiment 1.[3] Admittedly, we estimated 12 parameters to fit 24 data points. Yet, the data exhibit a complex pattern of interactions that is captured by the model. Furthermore, any reasonable choices of parameter values will predict quite well the qualitative pattern of the data; the parameterization is needed to achieve quantitative accuracy. Most important, however, is the fact that any model at all can be found to predict this pattern of data. As Shiffrin et al. (1990) took pains to show, no then-current model could handle the list-strength results (equivalent to our present Experiment 2) and, for many model types, no variants could be found to do so. The results from the present Experiment 1 place additional constraints on models, so it should not be assumed that it will be easy to find alternative models capable of predicting the present data set, whatever the number of parameters employed.

Discussion of the Model

We have assumed that (1) different sentence images are stored for different sentences (even if a word is shared between sentences), (2) a stronger single-sentence image is stored for a repeated sentence, (3) context and sentence cues are used for the sentence test conditions, (4) context and a word cue are used for the single-word test condition, (5) differentiation produces roughly constant activation of images differing in strength when the test word or test sentence shares no word with the activated image, and (6) context-sensitive encoding determines activation by single-word probes when that word has been studied in different sentences.

The most important factor underlying the Experiment 1 predictions is the additivity of activations across all list images. The predictions are based simply on counts of the number of images that match the test item in *i* words. In effect, one makes such a count for a target item, subtracts a similar count for a distractor item, and divides by the square root of the count for a distractor item, in order to derive the d' prediction for a condition. The different list types and test types vary in the number of sentences stored and in the way that they are divided into classes of overlap with the test items. A similar process operates in Experiment 2, except that one must take into account the strength of the stored sentence image in addition to the aforementioned counts. The ability of the model to predict the complex patterns of data lends some credence to the basic assumption that recognition operates as a global sum of activations of representations of presented items (whether the summing occurs at retrieval, as in SAM and MINERVA 2 (Hintzman, 1988), or at storage, as in models such as those of Pike (1984), Murdock (1982), Eich (1982, 1985), and Anderson (1973), to name a few. However, note that the Experiment 2 results would pose difficulties for all the models positing summation at storage.

IMPLICATIONS FOR OTHER MODELS

Shiffrin et al. (1990) take up implications of the lack of a positive list-strength effect, as illustrated in the present Experiment 2, for a large number of current models. In particular, they could not easily find variants of composite storage models that could predict such findings. MINERVA 2 (Hintzman, 1988) could be altered to handle the recognition list-strength data taken alone, probably because it assumes separate storage, but it has difficulty with some ancillary findings, such as the list-length effect in recognition and list-strength effects in recall. If one were trying to predict the data of Experiment 1 only, ignoring the difficulties of the list-strength findings of Experiment 2, it is possible that models other than SAM could be applied successfully, as long as they assume recognition to be based on global activation processes. Although such models have not been applied to the present sentence recognition paradigms, they have the potential for success because they sum activations across presented items in a manner mathematically akin to SAM (albeit at storage rather than retrieval). Of the models other than SAM, only MINERVA 2 appears to have the potential to handle the findings of both of the present experiments, although a good deal of theoretical work would be needed to verify this speculation.

CONCLUDING REMARKS

In the present paper, we show that a list-strength effect in recognition may be produced or eliminated through manipulations of the context in which words are repeated. When words are embedded in sentences, repeating entire sentences produces the typical finding: a list-strength effect does not occur. When words are repeated in the context of different sentences (while holding constant the total number of words per list), a positive list-strength effect occurs. This result was predicted by the theory of Shiffrin et al. (1990) and helps to validate it.

The results from the single-word tests, together with the data from the two kinds of sentence recognition tests, provide evidence that the significant units of storage in this paradigm are sentences rather than words, at least if one is working within the SAM theoretical framework. The complex pattern of results from the two experiments were generally well predicted by a SAM model suggested by Shiffrin et al. (1990), modified to deal with sentence units. The Experiment 1 data, in particular, provided evidence in favor of models of recognition based upon summation of activation across all list items. The difficulties of the model in handling the quantitative details of the single-word data from Experiment 1 suggest a version of the model involving context-sensitive coding at storage and retrieval; this assumption considerably improved the fit of the model.

The present evidence is consistent with a SAM model in which sentences are the traces in memory, repetitions of a sentence are stored in the same memory trace, and different sentences are stored in separate traces. When such a model is augmented with a differentiation assumption and a context-sensitive encoding assumption, a good account of the findings can be obtained. Other than MINERVA 2, we are not presently aware of any models having the potential to predict the results of Experiments 1 and 2 and the related recognition findings of Ratcliff et al. (1990). Even if MINERVA 2 could be made to handle the present findings, the additional recall results reported by Ratcliff et al. (1990) would raise considerable difficulties. In any event, the results favor a model in which different items are stored separately, without mutual interference. Repetitions appear to be stored in a single, stronger, trace, unless markedly different storage contexts for the separate repetitions (such as different sentences in which the items are embedded) lead to storage in separate traces.

REFERENCES

Ackley, D. H., Hinton, G. E., & Sejnowski, J. J. (1985). A learning algorithm for Boltzmann machines. *Cognitive Science*, **9**, 147-169.

Anderson, J. A. (1973). A theory for the recognition of items from short memorized lists. *Psychological Review*, **80**, 417-438.

Bain, J. D., & Humphreys, M. S. (1988). The relational context effect: Cues, meanings, or configurations? In G. M. Davies & D. M. Thomson (Eds.), *Memory in context: Context in memory* (pp. 97-137). London: Wiley.

Bowles, N. L., & Glanzer, M. (1983). An analysis of interference in recognition memory. *Memory & Cognition*, **11**, 307-315.

Caulton, D., & Shiffrin, R. M. (1988). [Re-pairing of word pairs and the list-strength effect]. Unpublished raw data.

Clark, S. E., & Shiffrin, R. M. (1987). Recognition of multiple-item probes. *Memory & Cognition*, **15**, 367-378.

Clark, S. E., & Shiffrin, R. M. (1991). *Associations, retrieval capacity, and cued recognition*. Manuscript submitted for publication.

Eich, J. M. (1982). A composite holographic associative recall model. *Psychological Review*, **89**, 627-661.

Eich, J. M. (1985). Levels of processing, encoding specificity, elaboration, and CHARM. *Psychological Review*, **92**, 1-38.

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review*, **19**, 1-65.

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

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

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

Jacoby, L. L., & Witherspoon, D. (1982). Remembering without awareness. *Canadian Journal of Psychology*, **36**, 300-324.

Kosko, B. (1987). Adaptive bidirectional associative memories. *Applied Optics*, **26**, 4947-4960.

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

Pike, R. (1984). Comparison of convolution and matrix distributed memory systems for associative recall and recognition. *Psychological Review*, **91**, 281-294.

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

Ratcliff, R., & McKoon, G. (in press). Item recognition and theories of text processing. In W. E. Hockley & S. Lewandowsky (Eds.), *Relating theory and data: Essays on human memory in honor of Bennet B. Murdock*. Hillsdale, NJ: Erlbaum.

Richardson-Klavehn, A., & Bjork, R. A. (1988). Measures of Memory. *Annual Review of Psychology*, **39**, 475-543.

Schacter, D. L. (1987). Implicit memory: History and current status. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **12**, 432-444.

Shiffrin, R. M., Murnane, K., Gronlund, S. A., & Roth, M. (1989). On units of storage and retrieval. In C. Izawa (Ed.), *Current issues in cognitive processes: The Tulane Flowerree Symposium on Cognition* (pp. 25-68). Hillsdale, NJ: Erlbaum.

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

NOTES

[1] A related analysis considered triples of study items. When a weak item is preceded and followed by repeated items, borrowing should cause higher performance than when a weak item is preceded and followed by other weak items. Conversely, weak items surrounding a repeated item should cause reduced performance for the repeated item

when compared with repeated items surrounding a repeated item. The data showed no such trends, but the experimental design provided too few cases with an item bracketed by two weak items to come to any meaningful conclusions.

2. The disadvantage of one-new testing when words are the storage units may be lessened to a degree by an assumption of context-sensitive encoding and retrieval. The idea is that an item will better activate its own image if tested in the storage context of four redundant items than if tested alone. Indeed, Clark and Shiffrin (1987) needed to make this assumption to fit their data. Nonetheless, the resultant model still predicted an advantage of single-item testing over one-new testing. To produce equality of predictions for the case of five-word sentences would require too large context-sensitive effects to predict other aspects of the present data. However, context-sensitive encoding may well occur within a model, positing storage of sentence units.

3. One could consider models in which single-word images are stored in addition to sentence images. If there is just one, stronger, single-word image for repeated words, and if retrieval is a mixture of access to sentence and single-word images, then the single-word test results of Experiment 1 might be fit even more closely. Such a model is much too complex to take up in this article.

APPENDIX

Before turning to the derivations, a few preliminaries are needed. Following Shiffrin et al. (1990), assume for any strength of mean χ the following strength distribution:

$$p[S=g_i\chi] = p_i; \quad \sum_i p_i = 1; \quad \sum_i p_i g_i = 1; \quad g_i > 0.$$

Then

$$\mathrm{Var}^{1/2}[S] = \chi\left(\sum_i g_i p_i - 1\right)^{1/2},$$

so the standard deviation is linearly related to the mean. For derivations, we note that

$$E[S^{\omega}] = (E[S])^{\omega}\theta_{\omega}, \text{ where } \theta_{\omega} = \sum_i g_i^{\omega} P_i.$$

Then,

$$\begin{aligned} E[A_i|C,I_j] &= E[S(C,I_i)^{\omega_c}]E[S(I_j,I_i)^{\omega_I}] \\ &= E[S(C,I_i)]\}^{\omega_c}\theta_{\omega_c}\{E[S(I_j,I_i)]\}^{\omega_I}\theta_{\omega_I} \end{aligned}$$

$$\begin{aligned} &\mathrm{Var}[A_i|\mathrm{C},I_j] \\ &= \{E[S(C,I_i)]\}^{2\omega_c}\{E[S(I_j,I_i)]\}^{2\omega_I}[\theta_{2\omega_c}\theta_{2\omega_I}-\theta^2_{\omega_c}\theta^2_{\omega_I}]. \end{aligned}$$

The term $\beta = \theta_{\omega_c}\theta_{\omega_I}[\theta_{2\omega_c}\theta_{2\omega_I}-\theta^2_{\omega_c}\theta^2_{\omega_I}]^{-1/2}$ is common to all the d' calculations that follow. The context strength term cancels out most of the d' expressions, as given in Figure 1.

The derivations for the model are quite simple. The calculations of the three quantities in Equation 1 depend only on counting the number of things in memory that overlap with the test item to various degrees, since both the mean and variance of familiarity are simply a sum of activation values, one for each image in memory. Only a few examples are needed to make the procedure clear. Let $\theta_{\omega_c}\theta_{\omega_I} = g$, and $[\theta_{2\omega_c}\theta_{2\omega_I}-\theta^2_{\omega_c}\theta^2_{\omega_I}] = h$.

Mixed-Strong, Single-Item Tests (Experiment 1)

There are 5 weak sentence images and 15 strong sentence images in memory. The strong target word does not match any of the words in the weak images, but matches one word in each of three strong images. There is no match with the remaining 12 strong images. A distractor matches no word in any image. Thus,

$$\begin{aligned} E(F|T) &= \{3a(1)e(1,1)+17a(1)e(0,1)\}g; \\ E(F|D) &= \{20a(1)e(0,1)\}g; \\ \mathrm{Var}(F|D) &= \{20a^2(1)e^2(0,1)\}h. \end{aligned}$$

The difference between the first two terms divided by the square root of the third gives the result stated in the text.

Pure-Strong, Intact-Rearranged Tests (Experiment 1)

There are two groups of 15 strong sentences in memory. The target sentence matches 1 image in all 5 words, 10 in 1 word (each word appears in three sentences), and 19 in no words. A distractor sentence matches 15 sentences in 1 word and 15 in no words. Thus,

$$\begin{aligned} E(F|T) &= \{a(1)c(5,1)+10a(1)c(1,1)+19a(1)c(0,1)\}g; \\ E(F|D) &= \{15a(1)c(1,1)+15a(1)c(0,1)\}g; \\ Var(F|D) &= \{15a^2(1)c^2(1,1)+15a^2(1)c^2(0,1)\}h. \end{aligned}$$

Mixed-Weak, One-New Tests (Experiment 2)

There are five weak sentence images and five strong sentence images in memory. The weak target matches all five words in one weak image and no words in the remaining four weak and five strong images. A distractor matches four words from one weak image and no words in four weak images and five strong images. Thus,

$$\begin{aligned} E(F|T) &= \{a(1)c(5,1)+4a(1)c(0,1)+5a(3)c(0,3)\}g; \\ E(F|D) &= \{a(1)c(4,1)+4a(1)c(0,1)+5a(3)c(0,3)\}g; \\ Var(F|D) &= \{a^2(1)c^2(4,1)+4a^2(1)c^2(0,1)+5a^2(3)c^2(0,3)\}. \end{aligned}$$

The remaining derivations are carried out in this vein (for the one-new conditions, it is useful to note that the number of sentence images sharing no words with a distractor sentence are 9 [pure-weak], 19 [mixed-weak], 11 [mixed-strong], and 21 [pure-strong]).

It should also be noted that there is a correspondence between the present parameters and those of SAM in Gillund and Shiffrin (1984): $c(i,j)$ was termed c, $a(j)$ was termed a, and $c(0,j)$ and $e(o,j)$ were termed d.

The term β depends on the weights only and may be calculated. However, it is a common scaling factor in all the expressions and may be set equal to 1.0 without loss of generality.

A best fit using a least squares criterion was carried out, with the results given in Table 6.

(Manuscript received November 11, 1989;
revision accepted for publication October 26, 1990.)