Journal of Experimental Psychology:
Learning, Memory, and Cognition
1990, Vol. 16, No. 2, 163-178

0278-7393/90/$00.75

List-Strength Effect: I. Data and Discussion

Roger Ratcliff
Northwestern University

Steven E. Clark
University of California, Riverside

Richard M. Shiffrin
Indiana University

Extra items added to a list cause memory for the other items to decrease (the list-length effect). In one of the present studies we show that strengthening (or weakening) some items on a list harms (helps) free recall of the remaining list items. This is termed the *list-strength effect*. However, in seven recognition studies the list-strength effect was either absent or negative. This held whether strengthening was accomplished by extra study time or extra repetitions. The seven studies used various means to control rehearsal strategies, thereby providing evidence against the possibility that the findings were due to redistribution of rehearsal or effort from stronger to weaker items within a list. Current models appear unable to predict these results. We suggest that different retrieval operations underlie recall and recognition, as in the SAM model of Gillund and Shiffrin (1984), which can be made to fit the results with certain relatively minor modifications.

Forgetting is often studied in modern experimental psychology as the *list-length effect*: Items on a longer list are remembered less well than items on a shorter list. This effect is essentially universal; it holds in the paradigms of free recall, cued recall, and recognition, among others (see Gillund & Shiffrin, 1984) and is found with extreme reliability. Not surprisingly, all models of memory and forgetting have been designed with mechanisms that ensure prediction of the list-length effect, so it is of great theoretical interest to examine related phenomena. In this article we ask whether *strengthening* certain items (but not all), either by studying them longer or repeating them, reduces memory for the remaining items on a list. Essentially all current models predict that such a *list-strength effect* should occur. Figure 1 illustrates the effect. The list-length effect would be instantiated if Item 1 is better recognized in List 1 than List 2. The list-strength effect would be represented by better recognition of Item 1 in List 1 than in List 3.

Tulving and Hastie (1972, Experiment 1) showed that repeating some items on a list reduced free recall of the remaining, nonrepeated items when total number of different items was held constant. The list-strength effect has not been studied in cued recall and recognition paradigms and has not been studied in paradigms in which items are strengthened by longer study times rather than increased numbers of presentations. This article focuses on the list-strength effect in recognition paradigms utilizing free and cued recall in a few instances as control conditions.

This research was supported by National Science Foundation Grant BNS 8510361 and National Institute of Mental Health (NIMH) Grant 44640 to Roger Ratcliff, and NIMH Grant 12717 to Richard M. Shiffrin. The authors gratefully acknowledge the helpful reviews of the first version of this article by William K. Estes, Douglas Hintzman, and two anonymous reviewers.

Correspondence concerning this article should be addressed to Richard M. Shiffrin, Psychology Department, Indiana University, Bloomington, Indiana 47405.

The focus upon recognition derives from theoretical considerations. Most current models assume that the recognition decision is based upon a sum of activation across all the stored items. Such models include those of Gillund and Shiffrin (1984), Murdock (1982), Metcalfe Eich (1982), Hintzman (1986), Pike (1984), and Anderson (1973; Anderson, Silverstein, Ritz, & Jones, 1977). In models of this type, extra items decrease performance because they add "noise" to the summed activation. Recognition performance is often expressed as d':

$$d' = \frac{\bar{F}(I_T) - \bar{F}(I_X)}{\mathrm{Var}^{1/2}\,[F(I_X)]}. \quad (1)$$

In Equation 1, $\bar{F}(I_T)$ refers to the mean summed activation (familiarity) for a list item (target), and $\bar{F}(I_x)$ refers to the mean summed activation for a nonlist item (distractor). The denominator is the standard deviation of the activation produced by a distractor test. In almost all current models, adding items to a (long) list or strengthening some items in a (long) list, will not change the numerator because familiarity will increase equally whether targets or distractors are tested. However, the denominator will increase if extra items or stronger items add noise (variance). Extra items produce extra noise in the models because each list item adds another nonnegative term to the expression for the variance.

Not all the models to be considered are explicit with respect to the effects of strengthening items, although many reasonable assumptions imply that noise will be added; we shall consider this issue in detail in Part II (Shiffrin, Ratcliff, & Clark, 1990). In any event, the SAM model of Gillund and Shiffrin (1984) explicitly predicts noise to increase with item strength and (along with other models sharing this assumption) predicts a list-strength effect in recognition.

One might ask whether list-length effects and list-strength effects are predicted in free and cued recall for reasons similar to those applying in recognition tasks. The issue is more complex in recall settings. For example, the SAM model predicts both list-length and list-strength effects to hold in recall

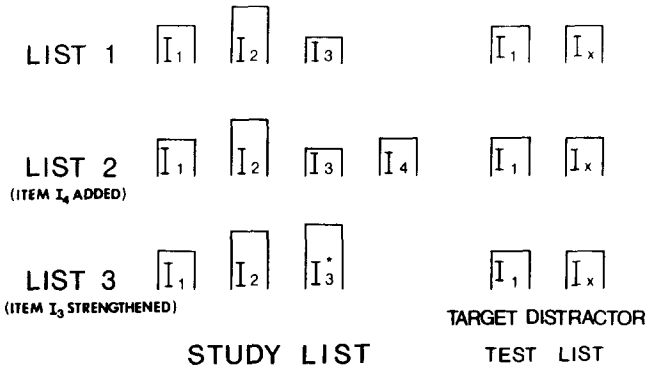

Figure 1. Simplified illustrations of the list-length effect (List 1 vs. List 2) and the list-strength effect (List 1 vs. List 3) in a recognition setting.

tasks, but for quite different reasons than the reasons applying in recognition tasks. In SAM, recall operates as a search based on sampling of memory images with replacement. Images are sampled in proportion to their strength. Thus stronger items on a mixed list tend to be sampled preferentially, and weaker items suffer as a result. Thus according to SAM, the variance of items is not very important in recall compared with the mean activation strengths. Most of the other models under discussion do not give explicit mechanisms for handling free recall. Cued recall in those models, however, operates in a manner similar in many respects to recognition: Cued recall is harmed to the degree that added items or stronger items introduce extra noise. For somewhat differing reasons, then, the various models all seem to predict list-strength effects in recall. We will discuss the details in the companion article (Shiffrin et al., 1990).

In summary, then, there are important theoretical reasons to examine the list-strength effect, especially in recognition paradigms. The models either make explicit predictions or will be constrained tightly by the results.

A test of the list-strength hypothesis necessarily involves presenting a study list containing items that are stored with at least two different levels of strength: Term these strengths S1 and S2, and call this list *mixed.* There are two natural control conditions to consider: In one list all items are of strength S1, and in the other all items are of strength S2; term these lists *pure.* Each of the three lists should contain the same total number of different items (not counting repetitions of the same item as different).

According to the list-strength hypothesis, memory for weak items in the pure-weak list should be better than for weak items in the mixed list because the mixed list is stronger on the average (perhaps producing higher variance of activation). Conversely, memory for strong items in the pure-strong list should be worse than for strong items in the mixed list because the mixed list is weaker on the average (perhaps producing lower variance of activation). The predictions can be summarized together by taking the ratio of strong item performance to weak item performance in the pure lists and comparing it with this ratio in the mixed list. Because strong items should be better remembered in the mixed list than in the pure-strong list and weak items worse remembered in the mixed list than in the pure weak list, the mixed strong-to-weak ratio should be larger than the pure strong-to-weak ratio. Note that both strong-to-weak ratios should be larger than 1.0, verifying the existence of a strength difference. The studies in this article use this design (or variants thereof) to assess the presence of a list-strength effect.

Because all of the studies to be discussed use at least one critical condition in which items of two different strengths occur in the same list, great care must be taken to control rehearsal strategies, coding strategies, and effort. For example, it would be possible for subjects to borrow some rehearsal time from the stronger items and give this rehearsal time to the weaker items. This would be a natural outcome of a "buffer" rehearsal strategy when items of different presentation times are mixed on a list (see Atkinson & Shiffrin, 1968). Such a redistribution of capacity, whether due to buffer mechanisms or otherwise, would tend to reduce the mixed list strong-to-weak ratio (the pure list ratio would be unaffected). In the studies in this article, we have used a variety of methods to reduce or eliminate such redistribution strategies. These methods will be discussed in the context of each study and then summarized in the General Discussion.

Experiment 1

Experiment 1 used single items and varied presentation time.

Method

Subjects. There were 5 paid subjects, each recruited from the student population at Northwestern University. Each subject took part in 6 to 9 sessions lasting about 45 min each, for a total of 37 subject-sessions in all.

Procedure. There were three conditions: Pure1, Pure2, and Mixed. *Pure1* involved presentation of 32 words presented for 1 s of study each. *Pure2* involved presentation of 32 words for 2 s each. *Mixed* involved presentation of 16 words presented for 1 s each and 16 words presented for 2 s each. The mixed-list items were presented in blocks as follows: One half the time, there were four 1-s presentations, then twelve 2-s presentations, then twelve 1-s presentations, then four 2-s presentations. The rest of the mixed lists used the same schedule with the times reversed.

Immediately following each list, an old-new recognition test for single words was carried out. The 32 list items and 32 new items were tested in random order. Subjects responded at their own pace, each test trial occurring 250 ms after the response.

Subjects were tested individually on a display terminal. Materials were presented and responses collected on a Radio Shack computer. Each subject received 14 lists each session, 4 of each mixed type and 3 of each pure type, in random order. All words in a session were unique (save for list items presented in the test phase).

Materials. The 1,024 words for each hourly session were chosen randomly from the Toronto Pool of 1,040 words (Murdock, 1970). Each subsequent session used a new sampling of words from the same pool.

Table 1
Results From Experiment 1

Presentation time per item	Mixed list Hit rate	Mixed list F/A rate	Mixed list d'	Pure list Hit rate	Pure list F/A rate	Pure list d'	Mixed/pure ratio of ratios
2 s	.705	.227	1.30	.740	.202	1.48	
1 s	.646	.227	1.12	.646	.228	1.12	
Strong/weak ratio			1.16			1.32	.88

Note. F/A = false alarm.

Results and Discussion

Table 1 gives the results for hit rates, false alarm rates, and d' for each condition.[1] Statistical analyses are based on d' calculated for individual subject-sessions for given conditions.[2] The results show strong items to be better recognized than weak items, $F(1, 36) = 59.5$, $p < .001$, as suggested by strong to weak ratios of 1.16 and 1.32 for mixed and pure lists respectively; pure lists to be better than mixed lists, $F(1, 36) = 8.95$, $p < .01$; and a significantly negative list-strength effect assessed by the interaction term, $F(1, 36) = 9.60$, $p < .01$, and illustrated by a mixed/pure ratio of ratios of 0.88. Thus, the presentation rate manipulation had its desired effect, but there was no evidence for the predicted effect in the ratio of ratios. In fact, the ratio was significantly the opposite of what was predicted. In case that rehearsal borrowing could have occurred, we analyzed hit rates for the middle four positions of each block of 12 items of the same presentation time (9–12 and 21–24). These are positions for which rehearsal borrowing would come from adjacent items with the same presentation time. Results showed the same effects with the mixed/pure ratio of ratios even smaller, 0.80. In addition, serial position functions for input positions were examined, and there was little evidence for rehearsal borrowing around the switch points for the different presentation rates. Generally across serial position, items presented for shorter times were recognized less than items presented for longer times though the serial position functions were quite noisy. In addition, performance decreased with test position (e.g., Ratcliff & Murdock, 1976; Shulman, 1974).

None of these findings suggest a reason why the strong/weak ratio should have been higher in the pure condition. If rehearsal had been borrowed from the strong items in the mixed list and given to the weak items, then such borrowing should have been evident near the boundaries of the blocks of similar items, but not near the center of the blocks. Nonetheless, the list-strength effect is negative even when items near the center of such blocks are the only ones scored. It is possible that some sort of redistribution of effort occurs nonetheless, but in a manner not restricted to nearby positions. Rehearsal of distant items has been observed (e.g., Brodie, 1975; Brodie & Murdock, 1977), but most rehearsals, in fact, come from nearby positions. Nonetheless, it is possible that the grouping of items by presentation time encourages distant rehearsals or the transfer of coding effort between blocks of different presentation times taken as wholes. Thus additional and converging evidence is needed.

Experiment 2

Experiment 2 used pairs of words and varied presentation time. The items in this study were pairs so that rehearsal would tend to be allocated within item rather than between items.

Method

Subjects. There were 6 paid subjects, each recruited from the student population at Northwestern University. Each subject took part in 6 to 10 sessions lasting about 45 min each, for a total of 50 sessions in all.

Procedure. Experiment 2 was similar to Experiment 1 except items were studied in pairs instead of single items as in Experiment 1. There were 16 pairs of words studied, each pair studied for 2 or 6 s. The *Pure2* list used 2 s per pair throughout, and the *Pure6* list used 6 s per pair throughout. The *Mixed* lists used blocks of pairs at different times: either 2 pairs at 2 s, 6 pairs at 6 s, 6 pairs at 2 s, and 2 pairs at 6 s, or the reverse ordering. There were 14 lists per session followed by recognition tests (4 each of the mixed lists, and 3 each of the pure lists), and there were 4 lists followed by cued recall. Instructions to the subjects encouraged them to learn the pairs together for the cued recall tests. (A programming error prevented the recall results from being analyzed, but this error was rectified in Experiment 3.) Immediately following each study list, the subject pressed a key to begin the test phase. The 32 studied words and 32 new words were presented one at a time in random order for old/new recognition testing as in Experiment 1. Each response was followed by a 250-ms blank interval before the next test item was presented.

Results and Discussion

Table 2 gives the results for hit rates, false alarm rates, and d' for each condition. The results show strong items to be

[1] To eliminate spurious data, in Experiments 1–4, responses longer than 2,500 ms and less than 200 ms were discarded, as was the first response in each test sequence.

[2] Standard errors for the means underlying our statistical measures may be determined as follows. Let X be the mean value of a statistic under consideration. Then X/t or $X/(F^{1/2})$ give the standard error of the mean.

Table 2
Results From Experiment 2

Presentation time per item	Mixed list Hit rate	Mixed list F/A rate	Mixed list d'	Pure list Hit rate	Pure list F/A rate	Pure list d'	Mixed/pure ratio of ratios
6 s	.706	.166	1.52	.700	.159	1.53	
2 s	.620	.166	1.27	.665	.165	1.40	
Strong/weak ratio			1.20			1.09	1.10

Note. F/A = false alarm.

better recognized than weak items, $F(1, 49) = 26.0$, $p < .001$, as illustrated by strong-to-weak ratios greater than 1.0: 1.20 and 1.09 for mixed and pure lists, respectively; pure lists to be better than mixed lists, $F(1, 49) = 8.8$, $p < .05$, and a nonsignificant list-strength effect assessed by the interaction term, $F(1, 49) = 0.54$, $p > .05$, and illustrated by a mixed/pure ratio of ratios of 1.10. In case that rehearsal borrowing could have occurred, we analyzed the hit rates for the middle four pairs in the blocks of six pairs in the mixed list conditions and found a ratio of ratios of 1.07 (*ns*). Also, inspection of the serial position curves showed little evidence of rehearsal borrowing across boundaries (from long to short presentation times per item).

Experiment 3

Experiment 3 used pairs of items, varied presentation time, and cued recall, as well as recognition testing.

Method

Subjects. There were 22 subjects, students at Northwestern University who participated for one session each and who received credit in an introductory undergraduate psychology course.

Procedure. Experiment 3 was similar to Experiment 2, with list presentation and timing the same. However, in this experiment, each of the four kinds of study lists was used once for cued recall for each subject so that the mixed/pure ratio could be examined for cued recall. Instructions to the subjects were aimed at reducing the possibility of rehearsal sharing: Subjects were instructed to form interacting images of the concepts of the study pair and were instructed to rehearse only the pair presented in order to maximize recall performance.

Results and Discussion

Table 3 gives the results for hit rates, false alarm rates, and d' for each condition for recognition. The results show strong items to be better recognized than weak items, $F(1, 21) = 21.7$, $p < .01$, as illustrated by strong-to-weak ratios greater than 1.0: 1.14 and 1.23 for mixed and pure lists, respectively; mixed and pure lists did not differ significantly, $F(1, 21) = .004$, $p > .05$; and there was a nonsignificant list-strength effect as assessed by the interaction term, $F(1, 21) = 0.27$, $p > .05$, and illustrated by a mixed/pure ratio of ratios of 0.93. An analysis of the middle four study pairs in each block of six produced a ratio of ratios of .93 (*ns*), providing no evidence of rehearsal borrowing (nor did inspection of serial position functions).

Cued recall data are shown in Table 4. Strong items are recalled better than weak items, $F(1, 21) = 41.7$, $p < .001$, illustrated by strong-to-weak ratios greater than 1.0: 1.86 and 1.74 for mixed and pure lists, respectively; pure and mixed lists did not differ, $F(1, 21) = 0.64$, $p > .05$; and the list-strength effect was nonsignificant when assessed by the interaction term, $F(1, 21) = .005$, $p > .05$, illustrated by a mixed/pure ratio of ratios of 1.07. This result suggests that, at most, a small list-strength effect could have been present in cued recall.

The results of the first three studies argue fairly strongly against the view that there is a borrowing of effort or rehearsals from nearby serial positions. It remains a possibility that redistribution of coding, rehearsal, or effort occurs between the slow and fast blocks of serial positions, taken as a whole. The likelihood of this hypothesis is lowest for Experiment 3: Such redistribution would have had to occur despite explicit instructions to limit rehearsal to the currently presented pair of items and despite interactive imagery instructions for each pair.

The first three experiments used multiple lists within a session, each of differing "strength" composition, and distractors were always new to the entire session. Under these circumstances, subjects may not have restricted the focus of retrieval to the current list alone, but rather to the entire session (despite the fact that it would be more efficient to focus on the recent list, according to most theories). If so, the strength variations within the current list may have been rather unimportant in comparison with the many variations in strength occurring across the whole session: In effect, all testing may have been of items from a "mixed" list. In Experiment 4 we test this hypothesis by presenting distractors for test of a given list that had been presented and tested in earlier lists. This procedure should encourage subjects to focus retrieval upon the most recently presented list only (assuming it is possible to do so).

Experiment 4

Experiment 4 used pairs of items, varied presentation time, and used distractors from earlier lists.[3] The failure to obtain

[3] Doug Hintzman (personal communication, September, 1988) suggested the reasoning leading to Experiment 4. The authors thank Gail McKoon for implementing this study.

Table 3
Results From Experiment 3

Presentation time per pair	Mixed list Hit rate	Mixed list F/A rate	Mixed list d'	Pure list Hit rate	Pure list F/A rate	Pure list d'	Mixed/pure ratio of ratios
6 s	.750	.121	1.82	.745	.111	1.88	
2 s	.655	.121	1.60	.677	.143	1.53	
Strong/weak ratio			1.14			1.23	.93

Note. F/A = false alarm.

a list-strength effect in Experiments 1 through 3 could have been due to a failure to focus retrieval on the just presented list. If so, the inclusion of prior list distractors should force subjects to focus their retrieval, and a list-strength effect should emerge.

Method

Subjects. There were 15 subjects in the control condition and 28 subjects in the experimental condition, each paid and recruited from the student population at Northwestern University. Each took part in one session lasting about an hour.

Procedure. Experiment 4 was similar to Experiment 3 but had two between-subjects test conditions. In the control condition, the distractors for a given subject were all new. In the experimental condition, one half the distractors were new, and one half "old": One half of the old distractors were randomly chosen from the items that had been studied and tested on the previous list (termed *positive* distractors), and one half of the old distractors were randomly chosen from the items that had been tested but not studied on the previous list (termed *negative* distractors).

Pairs of words were studied, and the subjects were instructed to learn the pairs both for recognition testing and an occasional cued recall test. In fact, cued recall was not a concern of this study; although one list was followed by a cued recall test to validate the instructions, little data were obtained, and the results were not scored.

Each subject received 16 lists. One list of randomly chosen type was a starting warm-up list that did not use any repeated distractors. List 7 was tested for cued recall but not scored, and Lists 1 and 8 were tested for recognition but not scored because there were no distractors from the previous list that could be used for repeated negatives. Except for Lists 1 and 7, the other 14 consisted of three pure-strong lists of 16 pairs, three pure-weak lists of 16 pairs, four mixed lists in the pattern of two weak pairs, six strong pairs, six weak pairs, two strong pairs, and four mixed lists in the pattern of two strong pairs, six weak pairs, six strong pairs, and two weak pairs. These lists were presented in random order. Weak pairs were studied for 1 s, and strong pairs for 5 s.

Table 4
Results From Cued Recall in Experiment 3

Presentation time per pair	Probability of recall: Mixed list	Probability of recall: Pure list	Mixed/pure ratio of ratios
6 s	.458	.428	
2 s	.246	.246	
Strong/weak ratio	1.86	1.74	1.07

The test lists began with two old words randomly taken from the first two and last two pairs studied, and two new items, randomly ordered. The rest of the test list consisted of the remaining 30 studied words and 32 distractors, in random order. For control subjects, the distractors were always new to the session. For experimental subjects, 16 distractors were new to the session, 8 were positive words from the previous list, and 8 were negative words from the previous list. Four of the positive distractors were taken from pairs 3 to 8 of the previous list, and four from Pairs 9 to 14, and half of these had been studied in left positions of a pair, and half studied in right positions.

Results and Discussion

The primary data were d' calculated from a given hit rate and a false alarm rate based on the new distractors only. Table 5 gives the hit and false alarm rates, d', and d' ratios for the various conditions. The control subjects exhibited better performance on strong than weak items, $F(1, 14) = 64.4$, $p < .01$, illustrated by a d' ratio of 1.92 for the pure lists and 1.48 for the mixed lists. The pure and mixed lists did not differ, $F(1, 14) = 0.01$, $p > .05$. The pure lists had better strong items and worse weak items than did the mixed list, producing a significantly negative list-strength effect as assessed by the interaction term, $F(1, 14) = 7.20$, $p < .05$.

The experimental condition showed better strong than weak performance, $F(1, 27) = 49.7$, $p < .01$, illustrated by strong-to-weak d' ratios of 1.79 for pure lists and 1.44 for mixed lists. Pure and mixed lists did not differ, $F(1, 27) = 0.84$, $p > .05$. As in the control condition, the pure-strong items were better and the pure-weak items worse than the corresponding items in the mixed lists, but the interaction term was not significant, $F(1, 27) = 1.99$, $p > .05$; the mixed/pure ratio of ratios was 0.80. The list-strength effects did not differ significantly between the two groups, $F(1, 27) = .84$, $p > .05$.

In the experimental conditions, the false alarm rate was higher for old distractors than new (.45 old vs. .29) and only slightly different for positive and negative old distractors (.46 vs. .43).

Perhaps some experimental subjects may have been able to focus retrieval on the most recent list, improving performance and producing a positive list-strength effect, while others may have been unable to do so, leading to poorer performance and less of a list-strength effect. The data were therefore divided on the basis of the ability to discriminate studied items from old distractors into the 18 best performers (average hit rate minus old distractor false alarm rate = .293) and the 10 "worst" performers (average hit rate minus old distractor

Table 5
Results From Experiment 4

Presentation time per pair	Mixed list Hit rate	Mixed list F/A rate	Mixed list d'	Pure list Hit rate	Pure list F/A rate	Pure list d'	Mixed/pure ratio of ratios
Control subjects—no repeated negatives							
5 s	.804	.308	1.36	.827	.272	1.55	
1 s	.662	.308	.92	.659	.345	.81	
Strong/weak ratio			1.48			1.92	.77
Experimental condition—repeated negatives							
5 s	.717	.278	1.17	.748	.255	1.33	
1 s	.588	.278	.81	.625	.338	.74	
Strong/weak ratio			1.44			1.79	.80

Note. F/A = false alarm.

false alarm rate = .082). No evidence for the emergence of a positive list-strength effect among the "better" subjects was found.

Finally, an analysis of hit rates only from Positions 4 through 7 and 10 through 13 produced a ratio of ratios of .75 for the control group and .79 for the experimental group; this was not markedly different from the values calculated from all the data and provided no evidence for rehearsal borrowing from nearby positions.

In summary, Experiment 4 provides evidence against the hypothesis that the failure to obtain a list-strength effect is due to a failure to focus retrieval upon the most recent list. Under conditions that should have required such focusing, a positive list-strength effect failed to emerge.

Let us now summarize the results of Experiments 1–4, in which item strength variations were produced by manipulations of presentation time. Little evidence was found for a list-strength effect in recognition: The ratio of ratios, which should have been greater than 1.0 if a list-strength effect were present, was .88, 1.10, .93, .71, and .80 in the four studies (.80, 1.07, .93, .75, and .79 if only the center positions of a block were examined.) Strong items and weak items can also be considered separately. For strong items a list-strength effect would be seen as a higher level of performance in mixed lists: In fact, the mixed and pure levels were equal in Experiment 2 and went in the wrong direction in Experiments 1 and 4. For weak items, a list-strength effect would be seen as a lower level of performance in mixed lists: Although this pattern was observed in Experiment 2, the levels were equal in Experiment 1 and went in the wrong direction in Experiments 3 and 4. Experiment 3 also examined cued recall: The ratio of ratios was 1.07 in this case, nonsignificant, suggesting that at most a weak list-strength effect could have been present. In none of the studies was there evidence for a strategy of rehearsal borrowing.

Experiment 5

Experiment 5 was designed to extend the results of the first four studies by altering strength in a different fashion: Would a list-strength effect appear in recognition when strength is manipulated by varying the number of presentations of an item? The case of spaced repetitions proves to be of particular importance for models assuming composite storage of information (see Shiffrin et al., 1990).

Method

Subjects. The subjects were 110 Indiana University students taking part in a session of about 50 min to satisfy part of an introductory psychology course requirement.

Procedure. Experiment 5 varied strength by varying both presentation time and number of presentations. Lists of 24 distinct word pairs were studied in all conditions, with instructions to rehearse together the members of each pair, rather than words in different pairs. Examples of some coding techniques were given.

In lists that varied number of presentations, each word pair was presented two or four times, at spaced intervals, each pair presented for 1.25 s. Three lists were used in this condition: *Pure 2P* = all 24 pairs presented twice; *Pure 4P* = all 24 pairs presented four times; *Mixed-P* = 12 pairs presented twice and 12 pairs presented four times. In these three list types, the lags between repetitions were equated for the mixed and pure conditions and for the two and four repetition item. Table 6 illustrates the method by which this equality was attained. The Pure-4P list began with 6 filler pairs that were repeated in the following block of six presentations and then repeated again in two blocks of six presentations at the end of the list. The experimental pairs occurred in eight blocks of nine presentations in the central portion of the list: A block of 9 pairs (A) was followed by a block of 9 other pairs (B), these two blocks repeating in this fashion four times. Within each block of six or nine items, order of pairs was randomly permuted. On the average, then, there were 17 intervening pairs between successive presentations of an experimental pair, with a range of 9–25 intervening pairs. These experimental lags were matched in the Pure-2P condition by starting the list with a block of 6 filler pairs and repeating this block at the end of the list. In the central portion of the list, there was an A block of 9 experimental pairs, followed by a B block of 9 other pairs. These two blocks were then repeated (four blocks in all). These lags were matched in the mixed condition by constructing two blocks of 9 different pairs, each consisting of six 4P pairs (denoted A and C) and three 2P pairs (denoted B and D). These two 9-pair blocks were repeated in alternating fashion twice each (for the first four blocks of 9 pairs). Then

Table 6
Presentation Conditions for Experiment 5

Study-list organization		
Pure-4P list	Pure-2P list	Mixed list (4P + 2P)
6 fillers (F)	6 fillers (F)	6 pairs (A) + 3 pairs (B)
6 fillers (F)	9 pairs (A)	6 pairs (C) + 3 pairs (D)
9 pairs (A)	9 pairs (B)	6 pairs (A) + 3 pairs (B)
9 pairs (B)	9 pairs (A)	6 pairs (C) + 3 pairs (D)
9 pairs (A)	9 pairs (B)	6 pairs (A) + 3 pairs (E)
9 pairs (B)	6 fillers (F)	6 pairs (C) + 3 pairs (F)
9 pairs (A)		6 pairs (A) + 3 pairs (E)
9 pairs (B)		6 pairs (C) + 3 pairs (F)
9 pairs (A)		
9 pairs (B)		
6 fillers (F)		
6 fillers (F)		
Study list summary		
24 pairs	24 pairs	24 pairs
48 words	48 words	48 words
96 presentations	48 presentations	72 presentations
Test conditions		
36 experimental words (A and B) + 36 new words	36 experimental words (A and B) + 36 new words	48 old words + 48 new words

Note. Letters in parentheses identify groups of pairs.

the three 2P pairs in each 9-pair block (B and D) were replaced by 3 new 2P pairs (denoted E and F), and four more blocks were presented, each of the different 9-pair blocks presented twice, in alternating fashion.

At test in the Mixed-P condition, all 48 studied words were mixed with 48 new words, for successive, single-item old–new recognition judgments. In the Pure-4P and Pure-2P conditions only the 36 experimental words were tested, along with 36 new words. Table 7 gives the average lags between first study of a pair and test and last study of a pair and test, for each condition and each number of presentations.

In lists that varied presentation time, each pair was presented once, either for 2.5 s or 5.0 s. In the *Pure-5T* condition all 24 pairs were presented in random order for 5.0 s. In the *Pure 2.5T* condition, all 24 pairs were presented in random order for 2.5 s. In the *Mixed-T* condition 12 pairs were presented for 5.0 s, and 12 for 2.5 s, in randomly intermixed fashion. All three list types were followed by 48 single-word tests of list words and 48 single-word tests of new words, intermixed. The study–test lags, therefore, in terms of number of items, were equal in the three conditions. Response time was not carefully controlled (see below), but the average times from study until the end of the list were 60 s in the Pure-5T, 30 s in the Pure-2.5T, and 45 s for both long and short items in the Mixed-T condition. Table 7 summarizes the study–test lags.

Each subject was given these six list types in a session. Subjects were tested in groups ranging from 1 to 6 in size. The subjects in a given group received the same order of list types because their testing had to be synchronous: Each subject received the next test item when all subjects had responded or when 8 s had expired. These times per trial averaged about 3.7 s. Otherwise, each subject received an independent random sample of words for study and test. Order of list types was randomized between groups.

Order of mixing of test words was random, subject to the constraint that each member of a studied pair was tested in a different half of the test list. Intervening between each study list and test was a 30-s distraction task consisting of the mental addition of a sequence of presented digits and the typing of the sum at the end of the sequence.

Materials. All words were high frequency (50+ occurrences per million) taken from the Kučera and Francis (1967) and Thorndike-Lorge (1944) norms. The presentation of words and digits to a terminal screen and the collection of responses was controlled by a DEC-PDP/11 computer.

Results and Discussion[4]

The mean hit and false alarm rates for each subject were used to construct d' values for each subject for each condition and each item type. The averages of these d' values across subjects are given in Table 8.

The strong-to-weak ratios were significantly greater than one in both the case of repetitions, $t(109) = 4.70$, $p < .01$, and presentation time, $t(109) = 2.52$, $p < .05$; however, the mixed list alone did not reach significance, $t(109) = 1.48$. For presentation time, the list-strength effect was in the wrong direction, just failing to reach the .05 significance level, $t(109) = 1.94$. For number of presentations the list-strength effect was not present, $t(109) = .663$, $p > .05$, nor was it present when considering the early or late 2P items separately. It should be noted that two 1.25-s spaced presentations produced performance about equal to one 2.5-s presentation, but four 1.25-s spaced presentations were much superior to one 5-s presentation. This may be likened to a spaced presentation advantage (Crowder, 1976, chap. 9) and suggests that gains due to extra study time become increasingly less strong than gains due to extra spaced presentations.

The 2P items that occurred early in the mixed list and late in the mixed list differed significantly with a slight advantage for late items, $t(109) = 2.21$, $p < .05$, suggesting the need to examine serial position functions. These functions were examined in detail for all the conditions in the experiment but are not shown because the results can be summarized easily: For any given condition, whether serial study position is scored by the first or last presentation position (for the presentation number conditions), the serial position functions are essentially flat. These flat functions do not rule out the possibility that the late-occurring 2P items are better because they have smaller study–test lags, but they do reduce the likelihood of this explanation. An alternative explanation for the advantage of late 2P items needs to be considered: Note that the late 2P items are relatively novel because they occur in a surrounding context of familiar and already studied 4P items. This may lead the subject to borrow some rehearsal time from the late 4P items and devote it to late 2P items. Such a process could explain why the list-strength effect is not found in this condition. Of course, this explanation would not apply for items varying in presentation time, but other types of rehearsal trading could apply in the case of mixed presentation times. We therefore decided to study further the possibility of rehearsal trading in Experiment 6: In some conditions in Experiment 6, repetitions are blocked rather than mixed.

[4] To simplify the tables for Experiments 5, 6, and 7, only the d's are given. The hit and false alarm results are available from Richard M. Shiffrin upon request.

Table 7
Study–Test Lags for Experiment 5

List condition	First presentation of pair: Av. no. pairs to list end	Av. no. single word tests	Total	Av. time[a] (s)	Last presentation of pair: Av. no. pairs to list end	Av. no. single word tests	Total	Av. time[a] (s)
			Repetition conditions					
Pure 4P	75	36	111	257	21	36	57	190
Pure 2P	33	36	69	204	15	36	51	182
Mixed 4P	63	48	111	287	9	48	57	219
Mixed 2P	45	48	93	264	27	48	75	242
Early study	63	48	111	287	45	48	93	264
Late study	27	48	75	242	9	48	57	219
			Presentation time conditions					
Pure 5T	12	48	60	268				
Pure 2.5T	12	48	60	238				
Mixed 5T	12	48	60	253				
Mixed 2.5T	12	48	60	253				

Note. Av. no. = average number.
[a] Study–test average lags in seconds are column 1 (or 4) times 1.25 plus column 2 (or 5) times about 3.7 plus 30.

In summary, then, Experiment 5 tends to replicate the results of Experiments 1 to 4: The list-strength effect in recognition is absent (or slightly negative). Could rehearsal borrowing in the mixed lists have produced the results? The use of pairs and the instructions mandating intrapair rehearsal should have helped reduce this possibility. Thus for the presentation time conditions, rehearsal borrowing seems unlikely. Turning to repetitions, we doubt that repetitions per se led to rehearsal borrowing: Informal questioning of subjects revealed that differences in repetition frequency were not very evident to the subjects. Nevertheless, the relative novelty of late-occurring 2P items could possibly have caused rehearsal borrowing, and Experiment 6 included conditions to further test the list-strength effect in the case of repetitions.

Table 8
Experiment 5 d′ Values

List condition	Condition: Mixed	Pure	Mixed/pure ratio of ratios
	Presentation time varied		
5T	1.56	1.54	
2.5T	1.48	1.31	
Strong/weak ratio	1.05	1.18	.89
	Number of repetitions varied		
4P	1.73	1.75	
2P	1.38	1.45	
Early 2P	1.28		
Late 2P	1.48		
Strong/weak ratio	1.25	1.21	1.03
Ratio mixed 2P early	1.35		1.11
Ratio mixed 2P late	1.17		0.97

Experiment 6

This study further tests the list-strength effect in recognition when repetitions are varied. A minor rationale for the study involved an attempt to remove the possible confound noted in Experiment 5. More important are conditions introduced to contrast, in the same study, the list-strength effect in recognition, cued recall, and free recall. The failure to obtain a list-strength effect in recognition is more meaningful if an effect is obtained in recall in similar conditions. In addition, conditions are included to examine the list-length effect: The failure to obtain a *list-strength* effect in recognition will have greater import if a *list-length* effect is obtained in recognition in similar conditions.

Method

Subjects. The subjects were 84 Indiana University students, undergraduate and graduate, paid for their participation.

Procedure. Four pure lists and three mixed lists were used at study, as indicated in Table 9. For each of these seven study conditions, memory was tested by each of recognition, cued recall, and free recall. Thus the experiment consisted of 21 lists in all. Each pair in all conditions was presented for 1.25 s. The Pure-4P(16) condition consisted of 16 pairs repeated four times each. These 16 pairs were alternated in two blocks of eight each, with order randomized each time, so that at least eight items intervened between repetitions of a given pair. The Pure-1P(16) list consisted of 16 pairs presented once each in random order. The Pure-4P(10) consisted of 10 pairs presented four times each (in blocks of five so that minimum lag between repetitions was five). The Pure-1P(40) consisted of 40 pairs presented once each. The Mixed-4P/1P(16) condition used eight pairs presented four times each (in blocks of four, so that the minimum lag between repetitions was four) at the start of the list, followed by eight pairs presented one time each at the end of the list. The Mixed-1P/4P(16) condition simply reversed the order of the 4P and 1P items from the preceding condition. Finally, the Mixed-(16) condition used a mixture

Table 9
Conditions for Experiment 6

List	Total pairs	No. strong pairs	No. reps.	No. pres.	No. weak pairs	No. reps.	No. pres.	Total pres.
Pure-4P(16)	16	16	4	64	0	0	0	64
Pure-1P(16)	16	0	0	0	16	1	16	16
Pure-4P(10)	10	10	4	40	0	0	0	40
Pure-1P(40)	40	0	0	0	40	1	40	40
Mixed-4P/1P(16)	16	8	4	32	8	1	8	40
Mixed-1P/4P(16)	16	8	4	32	8	1	8	40
Mixed(16)	16	8	4	32	8	1	8	40

Note. The mixed conditions vary in the order of presentation of strong and weak pairs. Mixed-4P/1P(16) blocks had strong pairs first. Mixed 1P/4P(16) blocks had weak pairs first. Mixed(16) randomly intermixed the pairs. Reps. = repetitions; pres. = presentations.

of eight pairs presented four times and eight presented once: In effect, a Mixed 4P/1P or 1P/4P list was constructed, and then, without changing the order of the 4P or 1P items considered separately, the two types of items were randomly intermixed.

It should be noted that the Pure-4P(10) list used less than the 16 unique pairs in our main conditions but matched them in the total number of presentations and total study time, allowing us to examine these whole-list factors. Note also that the Pure-1P(40) list used more than the 16 unique pairs in our main conditions but matched them in the total number of presentations, allowing us to examine the *list-length* effect independently of total study time.

Each study list was followed by an arithmetic distractor task of 30-s duration (as in Experiment 5) prior to test.

The test procedures were as follows:

1. *Recognition.* For the lists with 16 unique pairs, one word from each pair was tested, along with an equal number of new words, in random order, for successive, single word, old–new recognition judgments. Thus 32 words were tested. For the Pure-4P(10) lists, 20 words were tested, one from each of the 10 unique pairs, and 10 new words, in random order. For the Pure-1P(40) list, one word was tested from each of 20 randomly selected pairs from the 40 presented, along with 20 new words.

2. *Free recall.* Subjects were given 4 min to recall in any order as many different words from the preceding list as they could.

3. *Cued recall.* On each trial either the left- or right-hand member of a study pair was presented as a cue. Subjects tried to report the other member of that pair.

The experiment required three sessions, each lasting about an hour, on consecutive days. Practice lists to illustrate cued recall, free recall, and the recognition procedures, followed by 5 of the 21 experimental lists, were presented in the first session. Eight lists were presented in each of the last two sessions.

Subjects were tested in groups of one to four. The order of conditions was the same for all subjects in a group, but the words were randomized within the group. The order of conditions and the words were randomized between groups.

The same equipment and word pool was used as in Experiment 5.

Results and Discussion

The top part of Table 10 gives the recognition results in terms of d'. For each subject, the hit rate and false alarm rate were used to produce a d' value for each item type in each condition; these d' values were then averaged across subjects. The results were straightforward: The 4P items were much better recognized than 1P items, $t(83) = 7.68$, 8.83, 10.24, 9.02, and 11.50 for the pure, three mixed, and pure cases, reading from left to right in the top section of Table 10; $p < .001$ in all cases. The 4P items in the lists with 16 unique pairs did not differ from each other. The 4P items in the Pure-4P(10) list were just significantly higher in lists with 40 total presentations than were the other 4P items, $t(83) = 1.97$, $p < .06$, and just higher than in the Pure-4P(16) list, $t(83) = 1.96$, $p < .06$. The latter represents a standard list-length effect. The former cannot be interpreted without a model delineating exactly the changes in strength with presentation time and the variances associated with different item types. The 1P items in the lists with 16 unique items did not differ from each other significantly, but the 1P items in the Pure-1P(40) list were significantly lower than the others, $t(83) = 4.95$, $p < .01$; again demonstrating a *list-length* effect. No *list-strength* effect was seen, either considering the mixed lists separately or aggregating them together (overall, in aggregate, the ratio of ratios was 1.01). Thus, list-strength effects do not occur even when the differing repetitions are blocked, eliminating the "novelty" effect in Experiment 5.

Cued recall results are tabulated in the middle portion of Table 10 in terms of the probability of correct response. 4P items were much better recalled than 1P items (t values ranged from 9 to 14; $p < .001$ in all cases). The 4P items in lists with 16 unique items did not differ significantly from each other, but the aggregate of these was significantly lower than the 4P items in the Pure-4P(10) list, $t(83) = 4.26$, $p < .01$. The 1P items did not differ significantly from each other in the lists with 16 unique items, but their aggregate was just significantly higher than the 1P items in the Pure-1P(40) list, $t(83) = 2.34$, $p < .05$, thus demonstrating a *list-length* effect. For lists with 16 unique items, the ratio of 4P to 1P items was significantly higher for the mixed than the pure cases only for the Mixed-4P/1P(16) list, when conditions were compared individually, although the ratios were higher in each case. The aggregate of the mixed cases was significantly higher than the pure ratio, $t(83) = 2.47$, $p < .05$. Thus, for cued recall the nonsignificant positive list-strength effect noted in Experiment 3 is significant in the present case but is quite small in both instances.

The probabilities of free recall for the various conditions are given in the bottom portion of Table 10. 4P items were much better recalled than 1P items (t values ranged from 11 to 16; $p < .001$ in all cases). For 4P items, the Pure-4P(16) condition was just significantly poorer than the others, $t(83) = 2.01$, $p < .05$, which differed nonsignificantly. For 1P items

Table 10
Experiment 6 Results

Measure	Pure-4P(16)	Pure-1P(16)	Mixed-4P/1P	Mixed-1P/4P	Mixed-(16)	Mixed average	Pure-4P(10)	Pure-1P(40)
				Recognition d′				
Strength								
4P	2.32		2.23	2.43	2.49	2.38	2.62	
1P		1.48	1.40	1.44	1.63	1.49		1.09
				Performance ratios				
Strength ratio								
4P/1P	1.58[a]		1.59	1.68	1.53	1.60	2.40[c]	
Mixed/pure			1.01[b]	1.06[b]	0.97[b]	1.01[b]		
				Probability of cued recall				
Strength								
4P	0.365		0.374	0.368	0.435	0.390	0.483	
1P		0.163	0.110	0.130	0.156	0.130		0.117
				Performance ratios				
Strength ratio								
4P/1P	2.24[a]		3.39	2.84	2.78	3.00	4.14[c]	
Mixed/pure			1.51[b]	1.27[b]	1.24[b]	1.34[b]		
				Probability of free recall				
Strength								
4P	0.296		0.374	0.404	0.376	0.385	0.363	
1P		0.126	0.118	0.054	0.099	0.090		0.081
				Performance ratios				
Strength ratio								
4P/1P	2.34[a]		3.16	7.54	3.80	4.26	4.50[c]	
Mixed/pure			1.35[b]	3.22[b]	1.62[b]	1.82[b]		

[a] Pure-4P(16)/Pure-1P(16). [b] Ratio of ratios (list-strength effect). [c] Pure-4P(10)/Pure-1P(40).

the Mixed-1P/4P condition was lower than the others, $t(83) = 2.20$, $p < .05$, which did not differ from each other. The ratio of 4P to 1P was much lower in the pure lists than in any of the mixed lists, or the aggregate of the mixed lists ($p < .01$ in all cases).

Excluding the Pure-1P(40) condition, the free recall results are consistent with Tulving and Hastie (1972, Experiment 1). However, the fact that on the average the 1P items in the mixed conditions were recalled better (but not significantly so) than the 1P items in the Pure-1P(40) condition is slightly at variance with the results of Tulving and Hastie (1972) and Hastie (1975), which showed poorer recall of items mixed with stronger items when total list study time was equated. We shall return to this point later.

The list-length effect is seen in this study for all three test paradigms. For once-presented items the Pure-1P(16) was superior to the Pure-1P(40) condition: 1.48 > 1.09 in recognition; .163 > .117 in cued recall; .126 > .081 in free recall. For four-times presented items, the Pure-4P(10) was superior to the Pure-4P(16) condition: 2.62 > 2.32 in recognition; .483 > .365 in cued recall; .363 > .296 in free recall. Although the list-length effect is found regularly in recognition, the cause is under debate. For example, Shulman (1974) has suggested that the effect is due to a combination of three effects: proactive interference during study, proactive interference at test, and forgetting due to delay between end of study and start of test. Such possibilities suggest examination of input and output serial position data from our recognition conditions. These functions may also shed light on the failure to obtain a list-strength effect in recognition.

When the hit rates for the 4P items were scored at each of the study positions they occupied and then accumulated at each study position, the resultant functions all exhibited flat performance profiles and are therefore not shown. The study position performance functions for once-presented items are shown in Figure 2. Because of low numbers of observations, the d' values were based on average hit rates and average false alarm rates, the averages being taken over subjects and several adjacent input positions and then converted to d'. (This procedure produces lower average d' values than those given in the tables.) The Pure-1P(16) and Pure-1P(40) conditions exhibited primacy effects: better performance for items studied earlier (see Shulman, 1974). The Mixed-1P items showed recency, possibly because the 1P items late in the list were surrounded for the most part with familiar 4P items receiving their later repetitions. This effect may be related to the one seen in Experiment 5 for 2P items first occurring late in a list in a context of familiar 4P items. Nonetheless, although singly presented items occurring late in a mixed list may benefit from extra rehearsal, the list-strength effect is absent even

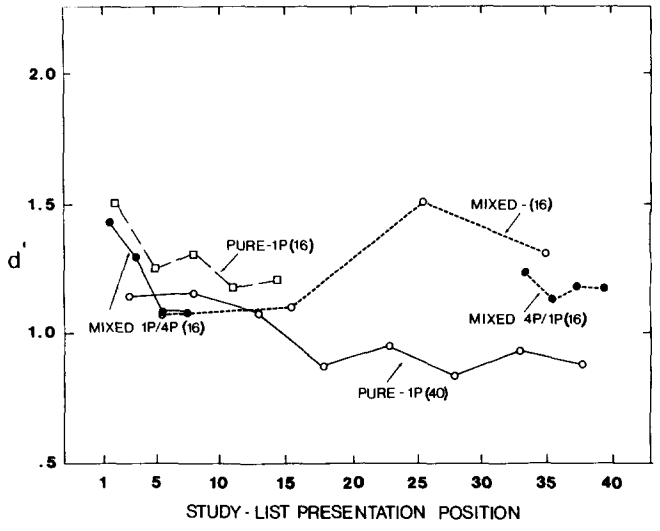

Figure 2. Experiment 6: d' as a function of serial presentation position for once presented items. (The d' measure is based on data summed over subjects and several adjacent positions.)

when the once presented items are blocked at the start or end of the study list, so that this factor cannot generally be used to explain the absence of a list-strength effect.

Similar to many earlier findings (Ratcliff & Murdock, 1976; Shulman, 1974) recognition d' tended to drop with test position. Figure 3 shows test position effects for the Pure-1P(16) and Pure-1P(40) conditions (along with the corresponding input functions). It is clear for these functions that the list length effect is not simply explicable in terms of input position or test position. To elaborate, consider two list-length comparisons. First, suppose that redistribution of rehearsal is not an important factor. Then the single items in the mixed

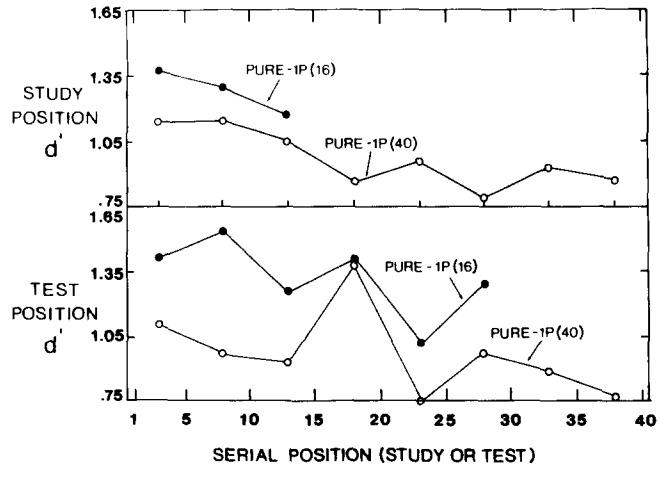

Figure 3. Experiment 6: d' results for the Pure-1P(16) and Pure-1P(40) conditions. (Top Panel: d' as a function of study position [see Figure 3]. Bottom panel: d' as a function of test position [summed over study position].)

conditions with 16 items can be compared with the Pure-1P(40) conditions. All these conditions have 40 presentations, the same total study time, and the same time until start of test. The Pure-1P(40) condition does have an average test position four greater than the other conditions, but the effect of this is easy to estimate from Figure 3. A regression function fit to the data reveals a decrease in d' of .0071 per each test position. Thus the four extra test positions contribute about .028 of the observed list-length difference of .39, or about 7%. In these cases, study position is not an important factor because blocking weak items early or late or mixing weak and strong items has no consistent effect on performance. Similar calculations show that only 18% of the list-length difference for 4P items between the Pure-4P(10) condition and the mixed 16 conditions can be attributed to test position effects.

If the possibility of redistribution of effort in mixed lists calls into question the above analyses, then the Pure-1P(16) and Pure-1P(40) conditions can be compared. In this case only .028 of the observed list-length differences is attributable to test position effects. This would be about 7% of the overall list-length difference. Furthermore, if Shulman (1974) were correct and proactive effects were to be taken into account, then only the first 16 study positions of these conditions should be considered (see Figure 3). In this case, 16% of the observed list-strength difference can be attributed to test position effects. Thus retroactive study position effects, as well as proactive effects, are likely to be an important factor.

In summary, this study demonstrates list-length effects in recognition even when list-strength effects are not found. The list-length effects are due (slightly) to proactive test position effects and, apparently, to both proactive and retroactive study position effects. Most current models could accommodate these list-length findings (though, interestingly, none could handle Shulman's condition that *only* proactive study position effects are operating).

The fact that a large list-strength effect occurs in free recall, at most a small one in cued recall, and none in recognition, poses problems for models that posit similar mechanisms to apply in all three paradigms. For such models, also, the different outcomes suggest that rehearsal borrowing is not the explanation for the failure to find a list-strength effect in recognition. If it were, then list-strength effects should have failed to occur in the recall tasks as well. However, this reasoning does not hold for models like SAM that posit quite different mechanisms to underlie recall and recognition. In SAM such a large list-strength effect is predicted for recall tasks that a list-strength effect could be seen even in the face of rehearsal borrowing or effort sharing. We therefore decided to carry out an incidental learning study to further test this possibility.

Experiment 7

Experiment 7 was designed to eliminate task demands that might encourage rehearsal borrowing or sharing of coding effort. Subjects performed an imagery task during list presentation. With no expectation of an impending memory test, subjects should have processed each item only as long as the imagery task required.

Method

Subjects. Twelve introductory psychology students at Indiana University participated in each of the four conditions, satisfying part of a course requirement.

The design deviated somewhat from the mixed–pure comparison in the previous experiments. Each item was presented once or three times. One list consisted of 100 repeated and 20 nonrepeated items. The other list consisted of 20 repeated and 100 nonrepeated items. If strong items harm recognition of weak items (i.e., if a list-strength effect holds), performance should be poorer for both repeated and nonrepeated items in the list containing mostly repeated (strong) items.

Words were repeated consecutively. For repeated words, different subjects were instructed either to (a) generate the same image three times, increasingly vividly or (b) generate three very different images for the repeated word. The logic behind this manipulation was as follows: Extra different items harm recognition (i.e., the list-length effect). If repetitions of a given item are made to have different memory representations, will these repetitions act like different items? To illustrate: If RABBIT is repeated on the list, the subject using the same-image instructions might image it each time as "Bugs Bunny," resulting in one stored representation; the subject under different-image instructions might image RABBIT as "Bugs Bunny," "rabbit stew," and "a stuffed toy animal," resulting in three representations. If so, a list-strength effect might be seen only in the different-image condition.

Procedure. For one group of subjects, 100 of the 120 items in the list were presented three times in immediate succession, and the remaining 20 were presented once. For another group, 100 were presented once and 20 were presented three times. The first condition is called the *Mostly-3* condition, and the second condition is called the *Mostly-1* condition. The two types of items were randomly intermixed on each list, a new randomization being used for each subject.

Subjects were told that the experiment was concerned with how people generate mental images. Subjects were instructed to generate an image for each item presented and to rate the vividness of their image on a 1–10 scale. Each item was presented for 2.5 s, after which subjects were signaled to enter their vividness rating. Subjects were allowed 1.75 s to enter their responses while the word was still on the screen. Including 0.25 s between items, total presentation time/word was 4.5 s. Half of the subjects in each group were told to generate different images for items that were repeated. Examples were provided to subjects in each group.

Following list presentation, the experimenter revealed to the subject the nature of the memory test. Old–new single item recognition was tested for the 120 list items plus 120 distractors, arranged in randomly intermixed order. Subjects were tested in groups of size 1 to 6, the next test item occurring when all subjects had responded or when 8 s had expired.

Materials. Each subject was presented with one list of 120 high-frequency words (50+ occurrences/million in Kučera and Francis (1968) norms), rated high in imageability in Toglia and Battig (1978) and Paivio, Yuille, and Madigan (1968) norms. Two hundred and forty words of this type were generated. The average imageability score (1–7 scale) was 4.64. Average frequency per million words was 139.30 (from Kučera & Francis, 1968).

Equipment. The same equipment was used as in Experiments 5 and 6.

Results and Discussion

The hit and false alarm rates for each subject were used to generate d' values. The average of these d's across subjects is shown in Table 11 for each condition. A 2 × 2 × 2 (Number of Presentations × Encoding Instruction × Proportion Repetitions) analysis of variance showed that (a) repeated items were recognized better than nonrepeated items, $F(1, 76) = 6.24$, $p < .0001$; (b) contrary to a list-strength effect, items in the Mostly-3P list were recognized slightly better than items in the Mostly-1P list, $F(1, 76) = 3.82$, $p < .07$; (c) the small difference between same and different encoding conditions, favoring different encoding operations, did not reach statistical significance, $F(1, 76) = 1.60$, $p > .20$; and (d) the interaction between encoding instruction and number of presentations was significant, $F(1, 76) = 4.03$, $p < .05$. No other interactions reached statistical significance.

The crucial comparison is between lists consisting mainly of repeated items versus lists consisting mainly of nonrepeated items. The Mostly-3P list was indeed stronger, with d' ratios of 1.23 and 1.26 for the two instructional conditions, as shown in Table 11. Strong and weak items should be better recognized in the Mostly-1P list if a list-strength effect holds. However, recognition of both repeated and nonrepeated items was *worse* in the Mostly-1P list, contrary to the prediction. This is seen in Table 11 as 1P/3P ratios that are all less than 1.0. Individual contrasts showed that the two same-image conditions reached the .05 significance level. One could argue that the negative effect observed here is due to rehearsal borrowing from strong items to weak items that is especially pronounced in the Mostly-3P lists. Although this possibility should be considered, it is strange that such an effect should be seen so strongly in an incidental learning study that should have reduced rehearsal strategies to a minimum.

The rather small interaction between encoding instructions and number of presentations reflects the fact that the effect of encoding instructions was slightly greater for repeated than for nonrepeated items. Although the instructional effect was small, it makes sense that it would occur only for repeated items because the instructions were not relevant for once-presented items.

Although the ratio of performance between Mostly-1P and Mostly-3P appears slightly greater for the different-image instructions, this effect did not approach statistical significance ($p > .20$). In retrospect, our reasoning that different images would produce different representations, and hence a list-strength effect, seems suspect. The key question may be whether the different images are independent and whether our procedure of immediate repetitions works against independence. In fact, our instructions to form different images may require the subject to refer to the previous images to comply. Thus to obtain a list-strength effect, it may be necessary to present repetitions in spaced fashion, in rather different contexts, so that encoding is both different and independent.[5]

The serial position functions (at study) are shown in Figure 4, counting each presentation as a separate input and combining results over instructional condition. To generate this figure, hit rates and false alarm rates are averaged over both subjects and several adjacent input positions and then converted to d'. The items in the Mostly-3P lists exhibit no strong

[5] This prediction has been verified in our laboratory (Murnane & Shiffrin, 1989).

Table 11
d' Values and Ratios in Experiment 7

List	Mostly 1P	Mostly 3P	Mostly 1P/mostly 3P	Strong list/weak list
		Different images		
3P	2.83	3.07	0.92	
1P	2.34	2.57	0.91	
3P/1P	1.21	1.20		
List average	2.42	2.99		1.23
		Same images		
3P	2.47	2.81	0.88	
1P	2.13	2.56	0.83	
3P/1P	1.16	1.10		
List average	2.19	2.77		1.26

recency or primacy. In the Mostly-1P lists, the 3P items at the start show some advantage (primary); on the other hand, the 1P items exhibit some recency. Such a result might suggest a gradual shift of attention from 3P to 1P items during study of the Mostly-1P lists. Even if this hypothesis were true, it would not explain the finding of a negative list-strength effect because it is the entire 1P list that is exhibiting poorer performance than ought to be the case. Thus the study position functions do not suggest rehearsal redistribution as a basis for the failure to obtain a list-strength effect.

General Discussion

The relation between the effect of strength variation and the list-strength effect is summarized in Table 12. Throughout the table, a ratio greater than one indicates a positive list-strength effect. In free recall, such an effect is quite strong. In cued recall, the effect is smaller, but positive and just significant in one study. In recognition, there is some variability of results, but little evidence overall for a list-strength effect. Ten ratios were less than one, and four (just) greater than one. Nine were not significantly different from one, and five were significantly less than one. The mean ratio overall was 0.92. The results suggest that the list-strength effect (defined as a ratio of ratios) usually lies in a small range near 1.0, with occasionally significant excursions away from 1.0 (below 1.0 in our present studies). We have not been able to identify a consistent factor enabling prediction of the conditions that will be significantly negative.

From Table 12, the correlation between the average strong–weak ratio for a condition and the list-strength effect ratio for that condition is −0.15, in contrast to the predictions of current models (see Shiffrin et al., 1990). Even if list-strength effects were masked by redistribution of rehearsal or effort, one would expect to see a positive correlation were current models correct.

The results in Table 12 give a broad summary of the list-strength findings. A finer look at the results may also be useful. A list-strength effect is seen if weak items in mixed lists give worse performance than weak items in pure lists or if strong items in mixed lists give better performance than strong items in pure lists. These comparisons are given in Table 13, arranged so that a positive difference indicates a positive list-strength effect. In addition, insofar as possible, the comparisons in Table 13 are given for items in corresponding study positions (e.g., weak items in the first half of mixed and pure lists are compared, etc.). The results confirm those given in Table 12: For weak items, 11 of 18 cases were negative, with a mean difference of −.007; for strong items, 16 of 18 cases were negative, with a mean difference of −.078. Overall, 27 of 36 cases were negative, with a mean difference overall of −.043. (As suggested by the statistical results given in the text, and because of the small numbers of observations,

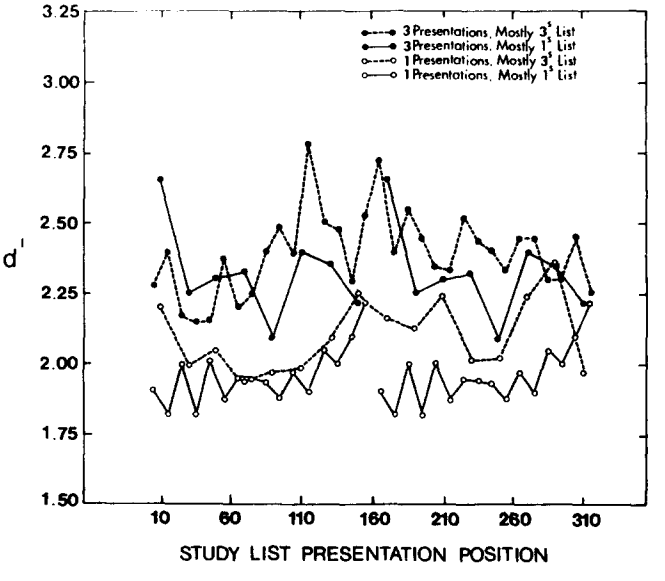

Figure 4. Experiment 7: *d'* as a function of serial presentation position (counting each presentation separately). (For convenience, the results from the Mostly 1S list are graphed twice—once corresponding to initial positions in the longer list and once corresponding to final positions in the longer list.)

Table 12
Relation of the Strength and List-Strength Effects

Experiment/condition	Mean strong/ weak ratio	List-strength ratio
Recognition		
Experiment 1: Time, single items, blocking	1.24	.88 (Mixed/pure)
Experiment 2: Time, pairs, blocking	1.15	1.10 (Mixed/pure)
Experiment 3: Time, pairs, blocking	1.24	.93 (Mixed/pure)
Experiment 4		
Time, pairs, blocking; control	1.70	.77 (Mixed/pure)
Time, pairs, blocking; experimental	1.62	.80 (Mixed/pure)
Experiment 5		
Time, pairs, mixing	1.12	.89 (Mixed/pure)
Reps, pairs, mixing, spacing	1.23	1.03 (Mixed/pure)
Experiment 6		
Reps, pairs, mixing, spacing	1.56	.97 (Mixed/pure)
Reps, pairs, blocking, spacing, strong first	1.59	1.01 (Mixed/pure)
Reps, pairs, blocking, spacing, weak first	1.59	1.01 (Mixed/pure)
Experiment 7: Reps, single items, mixing, no spacing		
Instruction—form same image		
Strong items (3P)	1.26	.92 (Mostly 1P/3P)
Weak items (1P)	1.26	.91 (Mostly 1P/3P)
Instruction—form different image		
Strong items (3P)	1.26	.88 (Mostly 1P/3P)
Weak items (1P)	1.23	.83 (Mostly 1P/3P)
Cued recall		
Experiment 6		
Reps, pairs, mixing, spacing	2.51	1.24 (Mixed/pure)
Reps, pairs, blocking, spacing, strong first	2.82	1.51 (Mixed/pure)
Reps, pairs, blocking, spacing, weak first	2.54	1.27 (Mixed/pure)
Experiment 3 (Time, pairs, blocking)	1.78	1.07 (Mixed/pure)
Free recall		
Experiment 6		
Reps, pairs, mixing, spacing	3.17	1.62 (Mixed/pure)
Reps, pairs, blocking, spacing, strong first	2.75	1.35 (Mixed/pure)
Reps, pairs, blocking, spacing, weak first	4.94	3.22 (Mixed/pure)

Note. Reps = repetitions.

essentially none of these differences deviated significantly from zero.)

The results may be compared with some earlier ones. Tulving and Hastie (1972) included one condition (part of Experiment 1) examining the list-strength effect in free recall: A items in a list of $A + B$ items were better recalled than A items in a list of $A + 2B$ items (where $2B$ refers to B items repeated twice). This finding was replicated in our Experiment 6. Hastie (1975) included a recognition test in a paradigm where total list time was held constant, but the relative numbers of once and twice presented items were varied (so that different lists had different numbers of unique items). If testing had occurred immediately after each list, our present results suggest that a list-length effect would have been seen, with performance determined by the total number of unique items (see our Experiment 6). However, recognition testing in Hastie (1975) was carried out at the end of session, after many lists of different types. The models introduced at the start of this article suggest that performance would be based on activation summed over all items in the session; if so, total number of unique items in the session, rather than in each list, would determine performance. Indeed Hastie (1975) found that recognition did not differ for lists of different compositions when subjects focused on items given the same number of repetitions.

Tulving and Hastie (1972) and Hastie (1975) confounded list-length and list-strength manipulations in most of their studies: They used mixed lists of strong and weak items (because of repetition variation) but equated the total list time (or, equivalently, the total number of separate presentations, counting repetitions separately). In their studies there was a small but significant effect such that stronger items on such lists inhibited free recall of the weaker items. The result appeared related to rehearsal effects, at least in part, because the direction of the effect was determined by whether subjects were instructed to attend to and remember the frequency of the repeated items (Hastie, 1975). Apparently, instructions to do so induced subjects to use rehearsal that would ordinarily have gone to singly presented items for the purpose of rehearsing frequency information for the repeated items. This rehearsal was probably not necessary (e.g., Hasher & Zacks, 1979) and did not much improve memory for the repeated items but did harm memory for the nonrepeated items. The lesson for us is that rehearsal effects in mixed lists are impor-

Table 13
Performance (d') in Pure Versus Mixed Lists

Condition	Weak items			Strong items		
	Pure	Mixed	Diff	Mixed	Pure	Diff
E1, early	1.12	1.08	+.04	1.25	1.38	−.13
E1, late	1.12	1.15	−.03	1.32	1.53	−.21
E2, early	1.46	1.26	+.20	1.53	1.54	−.01
E2, late	1.35	1.30	+.05	1.50	1.71	−.21
E3, early	1.59	1.60	−.01	1.77	1.81	−.04
E3, late	1.57	1.63	−.06	2.01	2.03	−.02
E4, control, early	.83	.84	−.01	1.30	1.48	−.18
E4, control, late	.78	.98	−.20	1.36	1.55	−.19
E4, exp, early	.64	.60	+.04	1.12	1.24	−.12
E4, exp, late	.74	.80	−.06	1.09	1.31	−.22
E5, time, early	1.07	1.13	−.06	1.17	1.19	−.02
E5, time, late	1.04	1.12	−.08	1.23	1.32	−.09
E5, pres	1.45	1.38	+.07	1.73	1.75	−.02
E6, 4P/1P	1.19	1.19	±.00	2.23	2.32	−.09
E6, 1P/4P	1.37	1.23	+.14	2.43	2.32	+.11
E6, mixed late	1.19	1.31	−.12	2.49	2.32	+.17
E6, mixed, early	1.37	1.11	+.26			
	Mostly 1	Mostly 3	Diff	Mostly 1	Mostly 3	Diff
E7, late in M3	1.96	2.16	−.20	2.33	2.42	−.09
E7, early in M3	1.96	2.06	−.10	2.33	2.37	−.04

Note. E = experiment; diff = difference; exp = experimental condition; M3 = mostly 3; pres = repeated presentations condition.

tant and need to be controlled. In the Tulving and Hastie (1972) and Hastie (1975) studies, single words were used, and rehearsal was free to be allocated to strong and weak items at the subject's discretion.

Our studies therefore used a variety of means to try to prevent the transferal of rehearsal or coding effort from strong to weak items in mixed lists: Items of a given number of repetitions were presented together in a block so that rehearsal borrowing from strong to weak would tend to occur at the boundaries of the blocks; items were presented in pairs, with instructions to rehearse and code together only the two members of each pair; items that were repeated different numbers of times were presented at spaced intervals, in such a fashion that it was not obvious to the subjects that items were repeated different numbers of times; items were studied under incidental learning instructions (an imagery task), presumably reducing interitem rehearsal. Although a few conditions gave indications that rehearsal strategy and rehearsal redistribution might have been playing some role (e.g., the cases with late-studied weak items appearing in a context of familiar strong items), the size of such effects and their existence only in a few special conditions could not have explained the general absence of list-strength effects in those studies. Nonetheless, redistribution not directly observable in our data could have been playing a role. What kind of redistribution would this be? Our blocking manipulations provide fairly good evidence against redistribution of effort to nearby study positions, but more general redistribution effects are possible. In the absence of positive evidence for such effects, however, it seems appropriate to accept on a provisional basis the conclusion that the attempts to control redistribution have been successful. As a corollary, it is important to consider mechanisms other than redistribution to explain both the absence of list-strength effects and negative list-strength effects (a mechanism such as the *differentiation* hypothesis to be described below).

Whatever view one takes of the list-strength findings in recognition, one must explain the striking contrast to the results in free recall. Because type of test is not generally known until testing begins, differences in storage cannot be invoked as explanatory mechanisms, so that different retrieval mechanisms are surely implicated. What could these be? We suggest, following Gillund and Shiffrin (1984), that recall operates according to a search of memory involving pseudo-random access to individually stored memory images, whereas recognition operates by summing activation across all relevant memory images. (The details are given in Shiffrin et al., 1990.) Other possibilities need to be considered, however.

It is likely that the cues used to probe memory in recognition and free recall are quite different. In the SAM model of Gillund and Shiffrin, for example, free recall involves probing memory with either a list-context cue alone or a context cue combined with an item already recalled. Recognition involves probes with a context-cue plus the test item. Even when an item plus context cue is used in both free recall and recognition, the weighting of or attention given to the item cue might well be much higher in the case of recognition tests. To the extent that different cues, or cue weightings, are used in the two tasks, differences between the results can be expected. The critical question then remains: How do these different probe cues produce the results we have observed? One possible answer, that less specific context cuing in recognition would tend to cause activation of all prior lists from a session, was not supported by Experiment 4. In this study, distractors chosen from prior lists did not cause a list-strength effect to emerge. A more plausible hypothesis is that the item probe in recognition produces much more focusing on the target trace

and less on traces of other items in the same list than does the context probe presumably used in free recall. The problem here is that this hypothesis predicts a reduction in both list-strength and list-length effects because both depend on interference from other list items than the target (e.g., when recognition accumulates or sums across many items from the recent list). To resolve this conundrum, we propose that many list items are activated during recognition but that the interference (i.e., "noise" added by these items) is different for extra items than for stronger items. For our studies we assume that stronger items are represented by a single memory trace even when strength is increased through repetitions. We then assume that activation of such a stronger trace by an unrelated test item is subject to two opposing influences: Activation engendered by the context cue tends to rise with trace strength, but activation engendered by the item cue tends to decrease with trace strength. The latter process is one of *differentiation*. These opposing influences tend to cause the mean and variance of activation to remain close to constant as trace strength is varied, for cases where both a context and an item cue are used to probe memory. Thus small and possibly negative list-strength effects can be predicted for recognition and cued recall. One the other hand, when only the context cue is used, as occurs sometimes during free recall, then the mean and variance of activation rise with trace strength, and a list-strength effect is predicted. Explication of this approach and consideration of a variety of extant models are given in Shiffrin et al. (1990).

In summary, our results show a sizeable list-strength effect in free recall, at most a small list-strength effect in cued recall, and a missing or negative list-strength effect in recognition, though with some variability from study to study. The recognition findings are unlikely to be the result of redistribution of rehearsal or effort from strong to weak items in mixed strength lists, though this possibility cannot yet be ruled out completely. These results have important implications for theory, especially so if they are not due to redistribution, as shown in Shiffrin et al. (1990).

References

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

Anderson, J. A., Silverstein, J. W., Ritz, S. R., & Jones, R. S. (1977). Distinctive features, categorical perception, and probability learning: Some applications of a neural model. *Psychological Review, 84*, 413–451.

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K. W. Spence, & J. T. Spence (Eds.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 2, pp. 89–195). New York: Academic Press.

Brodie, D. A. (1975). Free recall measures of short-term store: Are rehearsal and order of recall data necessary? *Memory & Cognition, 3*, 653–662.

Brodie, D. A., & Murdock, B. B., Jr. (1977). Effect of presentation time on nominal and functional serial-position curves of free recall. *Journal of Verbal Learning and Verbal Behavior, 16*, 185–200.

Crowder, R. G. (1976). *Principles of learning and memory.* Hillsdale, NJ: Erlbaum.

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

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

Hasher, L., & Zacks, R. T. (1979). Automatic and effortful processes in memory. *Journal of Experimental Psychology: General, 108*, 356–388.

Hastie, R. (1975). Intralist repetition in free recall:Effects of frequency attribute recall instructions. *Journal of Experimental Psychology: Human Learning and Memory*, 104, 3–12.

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

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

Murdock, B. B., Jr. (1970). Short-term memory for associations. In D. E. Norman (Ed.), *Models of human memory* (pp. 285–306). New York: Academic Press.

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

Murnane, K., & Shiffrin, R. M. (1989). *Word repetitions in sentence recognition.* (Research Report 8, Indiana University Cognitive Science Program Research Report Series). Bloomington, IN: Indiana University.

Paivio, A. V., Yuille, J. C., & Madigan, S. A. (1968). Concreteness, Imagery, and meaningfulness values for 925 nouns. *Journal of Experimental Psychology Monograph, 76*(1, Pt. 2).

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

Ratcliff, R., & Murdock, B. B., Jr. (1976). Retrieval processes in recognition memory. *Psychological Review, 83*, 190–214.

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

Shulman, A. I. (1974). The declining course of recognition memory. *Memory & Cognition, 2*, 14–18.

Thorndike, P. W., & Lorge, I. (1944). *The teacher's word book of 30,000 words.* New York: Columbia University Press.

Toglia, M. P., & Battig, W. F. (1978). *Handbook of semantic word norms.* Hillsdale, NJ: Erlbaum.

Tulving, E., & Hastie, R. (1972). Inhibition effects of intralist repetition in free recall. *Journal of Experimental Psychology, 92*, 297–304.

Received August 19, 1988
Revision received July 18, 1989
Accepted July 24, 1989 ■