Memory & Cognition
1978, Vol. 6 (3), 274-282

Effect of temporal locus of a recitation attempt on learning and retention

EDWIN MARTIN, FREDERICK G. FLEMING, and PATRICK D. NALLY
University of Kansas, Lawrence, Kansas 66045

In three experiments, analyses of individual-subject data show that temporal point of interruption for a practice recitation of a serial list affected neither the ultimate amount of time needed to master that list nor the amount of time needed for remastery 24 h later. Learning and relearning were by the unpaced whole-presentation method, with scheduled test interruptions at different stages of original learning.

Tradition has it that for purposes of research, the learning procedure for verbal materials is a succession of study-test cycles. Some such cycles are batch defined in that the learner studies a list of material and then is tested on it, studies again, is tested again, and so on. Other such cycles are item defined in that on seeing or hearing one item, such as a stimulus in paired associate learning or the nth item in serial learning, the learner is tested by having to anticipate a certain other item, such as the corresponding response or the next item, followed immediately by a study phase wherein the learner is presented with the appropriate other item for confirmational or instructional purposes. Recurrences of these item cycles are generally separated by intervening cycles on other items.

Tradition aside, one might wonder at the practice of imposing so many test trials in such a regular and unrelenting manner. To so proceed is not simulative of memorization in nonlaboratory situations, nor is it necessary to keeping track of how well the laboratory learner is doing. More importantly, there arises the question of how test operations contribute to, or detract from, the learning process. The study and test phases of a cycle must certainly interact in complex ways, to the end that generalizations about learning beyond the several standard laboratory paradigms are of unknown value.

The literature on test effects in the area of verbal learning supports several conclusions. With respect to paired associate learning, it is clear that a good schedule of study and test events should include tests early in the sequence (LaPorte & Voss, 1974), but that multiple testing in the form of successive retesting is a waste of time (Bregman & Wiener, 1970; Izawa, 1967, 1970). A likely interpretation of this is that the results of a test are the basis for selective study of the more difficult pairs, with successive retests offering no further information. As for long-term retention, it seems that in a situation where there has been a sequence of study phases without any test phases, overnight memory for the pairs increases as the number of poststudy unreinforced tests goes from zero to one to five (Allen, Mahler, & Estes, 1969). This is probably due to what might be called practicing the retrieval process (LaPorte & Voss, 1975). With respect to free recall learning, study and test phases appear to be interchangeable in their effects on acquisition (Lachman & Laughery, 1968; Tulving, 1967, Experiment 2), provided there are not too many test phases in a row (Donaldson, 1971, Experiment 1), although perhaps this proviso is not cogent (Hudson, Solomon, & Davis, 1972). In any event, successive test phases produce a stereotypy of output order that is more marked than when study phases are intermixed with the test phases (Rosner, 1970). The effect of test-event density in the acquisition schedule on long-term retention is not clear (Hogan & Kintsch, 1971), with a good chance of there being none (Birnbaum & Eichner, 1971). When it comes to serial learning in its usual paced-anticipation form, inquiry into the matter of test effects offers an as yet unmet methodological challenge.

The different effects of tests in paired associate and free recall learning are traceable to the distinguishing peculiarities of the tasks themselves. In paired associate learning, the learner must map one set of items (responses) in a one-to-one fashion onto another set of items (stimuli), with no structure within either set for guidance. The result is the formation of numerous dyads that the learner tries to retain in memory as separate entities. Early tests serve to identify those dyads in need of intensive review, and poststudy tests serve as retrieval practice on those dyads that are at that point retrievable. Testing thus acts on isolated pieces, not on the whole. In the free recall learning task, the situation is quite different. Here the learner cannot treat with isolated pieces, since their number is generally greater than the memory span and since no part of a piece is given as a cue in testing, as the stimulus part of a dyad is given as a cue for the response part in

This research was supported by Grant BNS 75-15770 from the National Science Foundation. Reprints may be requested from Edwin Martin, Fraser Hall, Department of Psychology, University of Kansas, Lawrence, Kansas 66045.

paired associate learning. In order to cope with such a task, the learner must somehow string the pieces together so that one leads to another in memory. This may sound natural enough, but now we come to a peculiarity of the free recall learning task that can only be seen as bizarre: As each study phase comes up, the experimenter rescrambles the order of the items. This is not the place to inquire into how such an odd, unrepresentative procedure came to make its first appearance, or how it survived that appearance. Suffice it to say that we know with virtual certainty that even in the face of such unreasoning adversity, the learner persists in stringing the pieces together as best he can (Tulving, 1962). And what of test effects? The output in a test phase must be some combination of whatever strings of items are retrievable and whatever isolated items might come to mind. This output operation is similar to a study phase, in that the learner has the opportunity to link together the diverse strings he recalls and to attach isolated items to their neighbors. Thus, we rationalize the observed interchangeability of study and test phases in free recall learning. We see also that there should be a limit to this interchangeability, since sooner or later the learner must return to the list to pick up more items.

The role of testing in learning and remembering might well be pursued by introducing additional variations in the scheduling of study and test phases. An attendant result, though, must be further refinement of learning paradigms we see as unnecessarily unnatural. The paired associate paradigm focuses on the simple one-link dyadic structure, but does so in a situation whose most obvious characteristic is massive interference among the many dyads that must be formed simultaneously. The free recall learning paradigm suffers from the curious rescrambling feature discussed earlier. Instead, we adopt a task wherein the learner has the entirety of the to-be-learned material before him to study in any way he chooses in preparation for a perfect serial recitation (Derks, 1974; Martin, Fleming, Hennrikus, & Erickson, 1977), an unpaced whole-presentation procedure. The learner can scan the list for the broad overview necessary to a comfortable choice of subjective segments, can return to difficult segments at will and can do whatever he might ordinarily do in memorizing a sequence of items. The question we address in this paper is: What is the effect of a premastery test (a recitation attempt) on ultimate learning difficulty and on 24-h retention, where the time of interruption for the test varies from early to late in the study period? We will answer this question as it applies to initial learning by noting how temporal locus of the recitation attempt affects the overall study time to mastery. As for overnight retention, we will measure the study time required to relearn the material and then relate that time to the point of interruption for the test during initial learning.

EXPERIMENT 1

Method

The subjects were nine graduate students who volunteered for paid participation. Five of them were assigned to Version A of the experiment. Their task was to memorize 11 lists of 12 words each, where a given list was composed of unrelated, common, one-syllable, four-letter words typed horizontally on a card. For the first three lists, the subject studied each in turn until he was ready for a perfect serial recitation, at which time he performed the recitation. The subject turned up a face-down card to begin a given study period, turned it face down to end it, and delivered his recitation into a microphone. The three study times were recorded and averaged, thus providing a personalized mean time for that subject. This much we call the preliminary stage.

The five subjects then went on to memorize eight further lists. The study periods for these lists were interrupted at one of four points en route to mastery. If the mean time from the preliminary stage is denoted t, then the time from the beginning of the study period for a given list to the interruption was either .1 t, .3 t, .4 t, or .5 t. These interruption times we refer to as experimental conditions. Two lists were assigned to each condition, so that each subject was tested twice in each interruption condition. If the schedule for the first lists was, say, .4, .1, .3, and .5 t, then the schedule for the second four lists was just the reverse, .5, .3, .1, and .4 t, thus balancing the interruption conditions with respect to earliness and lateness in the experimental session. Different subjects were given different schedules, and lists were newly assigned to conditions for each subject. After the interruption recitation attempt on a given list, during which the list card was turned face down, the subject resumed his study of that list until he was ready for a perfect recitation. The total study time is the critical dependent variable. Having mastered the final list, the subject was dismissed until the next day, with the information that the next day would involve more of the same.

On his return for the second session, the subject relearned all the lists, including the three preliminary lists, to the perfect-recitation criterion. No list was interrupted for a test. The list order for relearning was the same as for initial learning on the day before.

There were four subjects in Version B of the experiment. This version proceeded exactly as did Version A except that here the start-to-interruption times were .1, .3, .5, and .7 t. For one of the subjects, t was the smallest of his three preliminary-stage times instead of the average, a variation of no apparent consequence.

In both versions the subjects were exhorted "to use the minimum amount of time possible without making a mistake." The instructions emphasized, in several places, the shortest study time consistent with perfect performance.

Results

Version A involved start-to-interruption times of 10%, 30%, 40%, and 50% of the preliminary time t, with two lists assigned to each condition. The mean number of words recalled, out of 12, on these interruption tests were 3.8, 5.4, 8.7, and 7.4, respectively. An analysis of variance yielded F(3,12) = 5.839, MSe = 4.004, p = .011. In Version B the conditions were 10%, 30%, 50%, and 70%, for which mean recall was 5.2, 8.6, 10.1, and 11.0 words, respectively [F(3,9) = 13.565, MSe = 1.889, p = .001]. Thus the experimental conditions indeed

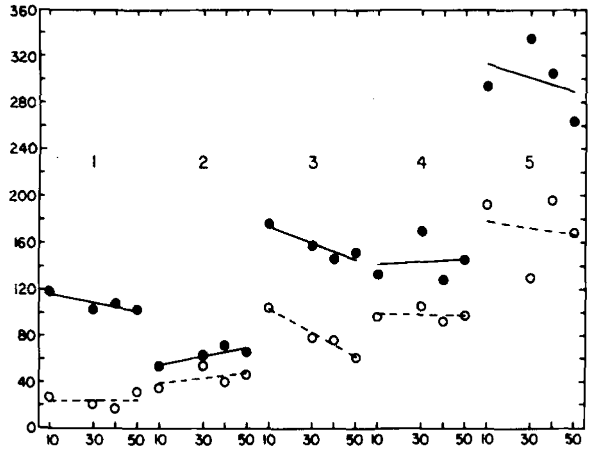

Figure 1. Study time in seconds as a function of percentage interruption time for individual subjects in first-day learning (filled circles) and second-day relearning (open circles).

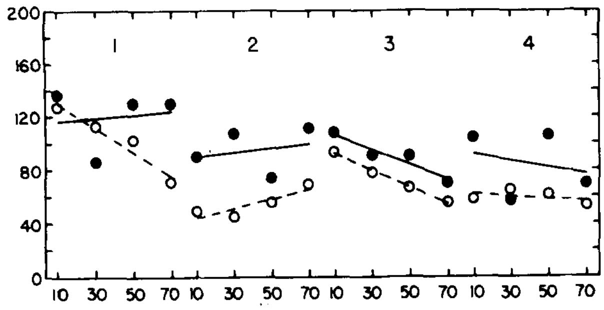

Figure 2. Study time in seconds as a function of percentage interruption time for individual subjects in first-day learning (filled circles) and second-day relearning (open circles).

imposed interruption recitation attempts at different stages of learning.

The total time to mastery of a given list is the sum of the pre- and postinterruption study times. With two lists per interruption condition, these total times were geometrically averaged, thus giving a single study time per condition for each subject.[1] These study times, in seconds, are plotted as filled circles for the five individuals in Version A in Figure 1 and for the four individuals in Version B in Figure 2. The open circles are the corresponding relearning times from the second session 24 h later.

The fitted straight lines in Figures 1 and 2 are least-squares regression lines. The slopes of these lines, and their standard errors, are listed in Table 1. No first-day slope differed significantly from 0, suggesting that the point of interruption for a recitation attempt did not affect overall learning time. Three of the second-day slopes differed significantly from 0: The 95% confidence intervals for Subjects A3, B1, and B3 excluded the possible value of 0.

The slopes in Table 1 are for individual subjects. They may be averaged as follows. We weight each slope in direct proportion to its reliability by dividing it by its variance. This procedure gives an unbiased estimate of a presumed underlying populational slope, an estimate with the further property of having the least variance relative to other possible schemes of weighting (Hald, 1952, p. 243-245). The result for the first-day learning times was $\bar{b} = -.281$, with standard error $s_{\bar{b}} = .081$, while for the second-day relearning times it was $\bar{b} = -.508$, with standard error $s_{\bar{b}} = .036$. The 95% confidence interval for the first-day mean slope was $-.474, .089$, which includes 0, while that for the second-day mean slope was $-.593, -.424$, which does not include 0.

Errors in recitation when the subject had presumably mastered (first day) or remastered (second day) the list were rare, averaging fewer than one error per subject for the entire experiment. The kinds of errors made were single omissions and single inversions.

Discussion

Apparently the temporal locus of a recitation attempt in the total study period does not affect the ultimate duration of that period. That the interruption occurred at different stages in learning is clear enough, as shown by the subjects' performance on those attempts. But no advantage in overall learning time is assignable to an earlier or later interruption. As for long-term retention, there seems to be a greater saving in relearning time the later in initial learning the interjected recitation attempt, significantly so for three of the nine subjects. The pooled data reinforce this conclusion.

The first-day study times argue for a no-effect null result. The second-day study times argue against such a result but are far from overwhelming when laid out subject by subject. In the next experiment, the learners were introduced to interruptions for tests in the preliminary stage, the notion being that they would thereby be better prepared for the experimental conditions, thus improving the picture of what is going on.

Table 1
Slopes and Standard Errors for Figures 1 and 2

	Day 1		Day 2	
Subject	b	s_b	b	s_b
		Version A		
1	–.369	.172	.011	.257
2	.366	.171	.209	.304
3	–.711	.237	–1.046*	.148
4	.094	.754	– .040	.224
5	–.626	1.149	– .286	1.259
		Version B		
1	.115	.635	– .905*	.165
2	.150	.453	.355	.142
3	–.570	.138	– .625*	.042
4	–.280	.648	– .080	.114

**95% confidence interval excludes 0.*

EXPERIMENT 2

Method

The preliminary stage of this experiment consisted of memorizing five lists in the following manner. Lists 1, 3, and 5 were studied in the usual way, to a subjective criterion of perfection, without interruption. If t_1 and t_3 are the resulting study times for Lists 1 and 3 for a given subject, then on Lists 2 and 4 that subject was interrupted for a recitation attempt at times $.5\ t_1$ and $.5\ t_3$, respectively. Subsequent to these recitation attempts, he resumed study until the criterion of perfection was reached. If an error was made in the final recitation of the list, the subject had to return to the list for further study until a perfect recitation was indeed given. The reason for this more elaborate preliminary stage is the idea that by introducing interruptions prior to the experiment itself, and requiring additional study in the event of an error, we might better sensitize the subject to his readiness to recite.

The rest of the experiment proceeded as in Versions A and B of Experiment 1. The start-to-interruption times were .05, .20, .35, and .50 t, where t is the shortest study time from Lists 1, 3, and 5 in the preliminary stage for a given subject. There were two 12-word lists for each of the four interruption times (conditions), arranged as in Experiment 1. Any error in the final recitation of a list meant additional study of that list, the additional time being counted as part of the total study time to criterion. The subjects were six volunteers who answered a campus newspaper ad for paid participation.

Results

That the 5%, 20%, 35%, and 50% conditions indeed interrupted learning at progressively later stages was shown by the mean number of words recited: 4.2, 5.1, 6.4, and 8.1, respectively [F(3,15) = 6.803, MSe = 2.558, p = .004].

The geometric mean study times, in seconds, for individual subjects are plotted in Figure 3. As before, filled and open circles are for first-day learning and second-day relearning, respectively. The slopes of the least-squares straight lines shown in the figure are listed in Table 2, together with their standard errors. Only one of the 12 slopes had a 95% confidence interval that excluded 0, and that was the positive slope for Subject 5 from first-day learning. The weighted mean slope for first-day learning was $\bar{b} = .430$, $s_{\bar{b}} = 1.63$, with 95% confidence interval −.022, .882, which includes 0. For second-day relearning, $\bar{b} = .001$, $s_{\bar{b}} = .077$, with −.213, .214 as the 95% confidence interval, which also includes 0.

Table 2
Slopes and Standard Errors for Figure 3

	Day 1		Day 2	
Subject	s	s_b	s	s_b
1	.460	.428	.220	.628
2	−.007	.430	−.073	.271
3	−.440	.415	−.347	.349
4	.160	.590	.060	.086
5	.980*	.242	−.093	.605
6	−.647	.901	−.613	.371

**95% confidence interval excludes 0.*

Discussion

This experiment differed from Experiment 1 in several ways. First, it involved slightly different interruption times. Second, even though errors were rare in Experiment 1, here we required additional study should an error occur. Third, we elaborated the preliminary stage so as to better prepare the learner for the experimental conditions. The result of these variations is that the no-effect conclusion for first-day learning times from Experiment 1 is reaffirmed, and the statistically significant but observationally doubtful effect in second-day relearning from Experiment 1 does not reappear.

The next experiment was another attempt to evaluate the possibility that the temporal locus of a practice recitation is not a determiner either of overall task difficulty, as measured by total study time to criterion, or of long-term retention, as measured by relearning time the next day. Experiment 3 also addressed the further question of whether a practice recitation is better or worse than no practice recitation, its temporal locus aside: In the preceding experiments, the learner was always interrupted with a test; in the next experiment, there were uninterrupted study times

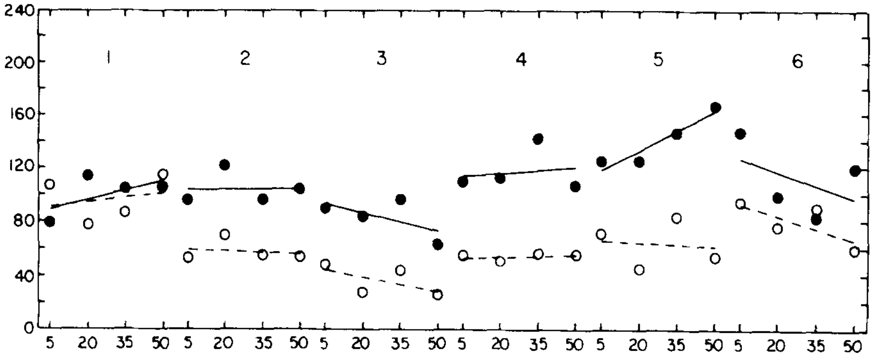

Figure 3. Study time in seconds as a function of percentage interruption time for individual subjects in first-day learning (filled circles) and second-day relearning (open circles).

against which interrupted study times may be compared. Finally, we decided to let each learner interrupt himself when he felt he had spent a certain percentage of the total time he thought he would need for a given list, thus transferring control of the time manipulation from experimenter to subject.

EXPERIMENT 3

Method

The subjects were 10 students who participated either for introductory psychology credit, money, or a combination of the two. The learning materials were as before in Experiments 1 and 2. There was no preliminary stage; the experimental conditions began immediately.

Each subject memorized 12 lists of words to a criterion of perfect recitation. Three of these lists he studied without interruption, as in the preliminary stage of Experiments 1 and 2, a condition we denote as 100%. We may view this condition as one in which the subject takes 100% of the time he needs to prepare for a perfect recitation. In three other conditions of three lists each, the subject was told to take 25%, 50%, and 75% of the time he thought he would need. When he felt he had reached the designated percentage of time, he turned the list card over and attempted a recitation, after which he resumed his study until he felt he was at the 100% point. He then turned the list card face down again and recited the list. As in Experiment 2, an error meant additional study.

For each subject, the four conditions 25%, 50%, 75%, and 100% were ordered randomly. That subject would then be given three blocks of that order. Thus the sequence of conditions for a given subject might be 75%, 100%, 25%, 50%, repeated twice more, for three replications. When he returned the next day for the second session, the subject relearned all 12 lists in the same order he learned them the day before.

Results and Discussion

The first question is whether the subjects in fact interrupted themselves at different stages of learning. We will look first at the start-to-interruption times they produced in response to the instructional conditions of 25%, 50%, 75%, and 100%. Each subject produced three such times for each condition. Geometric means were computed so that each subject finished with one score per condition. Arithmetic averaging over subjects gave the following start-to-interruption times for the 25%, 50%, 75%, and 100% conditions: 59.6, 78.0, 93.1, and 133.5 sec, respectively [$F(3,27) = 34.612$, $MSe = 285.607$, $p < .001$]. These times, along with their 95% confidence intervals, are plotted as open circles in Figure 4. Thus there can be no doubt that the experimental conditions resulted in systematic differences in interruption times. The dashed line in Figure 4 is the line on which the open circles must fall were the interruption times to match exactly 25%, 50%, and 75% of the 100% time. There appears to have been a tendency to overproduce time for the smaller percentages, a tendency consonant with the notion that early in learning the subject is less confident of what he knows than he need be.[2]

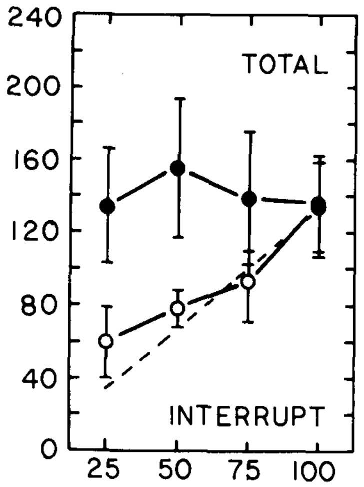

Figure 4. Study time in seconds to interruption (open circles) and to mastery (filled circles) as a function of assigned percentage, averaged over 10 subjects.

The second approach to assessing the subjects' success in testing themselves at different stages of learning en route to mastery is to look at the number of words they were able to produce in the interruption recitation attempt. The means over subjects were 7.4, 8.2, 9.6, and 11.9 words, respectively, out of 12 possible. An analysis of variance yielded $F(3,27) = 19.260$, $MSe = 2.045$, $p < .001$, a wonderfully inflated result, seeing that the mean for the 100% condition had essentially no error variance. Dropping this condition gives $F(2,18) = 5.945$, $MSe = 2.212$, $p = .011$. The conclusion is that the 25%, 50%, 75%, and 100% conditions resulted in interruptions that were at progressively later stages of learning.

The major question is whether the temporal locus of the recitation attempt is a predictor of total time to mastery. The filled circles in Figure 4 are the arithmetic means of the individual-subject geometric means for the 25%, 50%, 75%, and 100% conditions. Also shown are the 95% confidence intervals. We note that a horizontal line is the most sensible descriptor of these data and that each of the four confidence intervals captures all of the means for the other conditions. The corresponding means and confidence intervals for second-day relearning (not shown in Figure 4) were 86.0 (73.553, 98.447), 79.4 (61.658, 97.142), 85.4 (67.826, 102.974), and 81.3 (68.710, 93.890) sec. We note again that the means seem to follow a flat straight line and that each confidence interval captures all the other means.

From the foregoing we conclude that the position of the interruption test in the overall study period does not determine ultimate ease or difficulty of learning. Also, it does not determine long-term retention. These observations, based on group-average data, are fully supported in individual data. Figure 5 presents the first-day learning times (filled circles) and

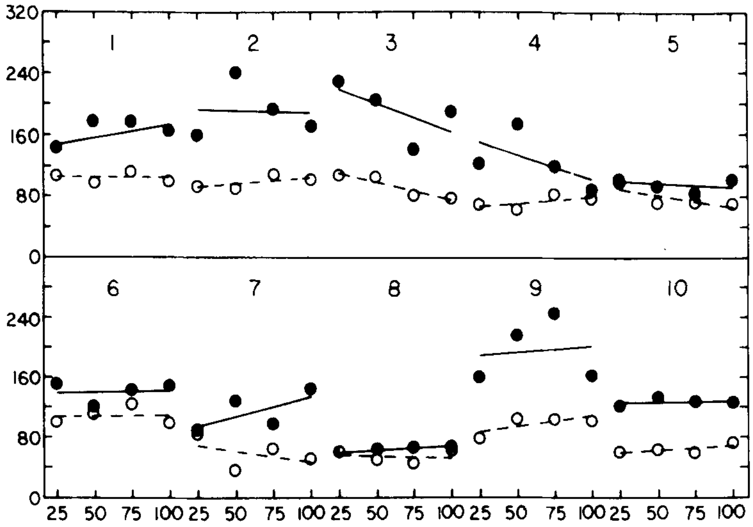

Figure 5. Study time in seconds as a function of assigned percentage for individual subjects in first-day learning (filled circles) and second-day relearning (open circles).

second-day relearning times (open circles) for each of the 10 subjects. Table 3 lists the slopes and standard errors for the least-squares regression lines shown in the figure. All of the 95% confidence intervals capture the zero-slope possibility. The weighted mean slope for the first-day learning times was $\bar{b}$ = .084, with standard error $s_{\bar{b}}$ = .030, and 95% confidence interval .016, .152, which excludes 0. The corresponding statistics for second-day relearning were $\bar{b}$ = .005, $s_{\bar{b}}$ = .043, and –.093, .103, which does not exclude 0.

With respect to the question of whether the simple fact of imposing a test has an effect on learning and retention, regardless of temporal locus, the same data show the answer to be in the negative. Figure 5 suggests that the 100% times in both learning and relearning failed to differ systematically from the 25%, 50%, and 75% times. In Figure 6 the 100% study time is plotted against the arithmetic mean of the 25%, 50%, and 75% study times for each subject, with closed circles for first-day learning and open circles for second-day relearning. The diagonal line is the identity function. The correlation coefficients for the closed and open circles were r(8) = .800 and .842, respectively (ps = .004 and .002).

Table 3
Slopes and Standard Errors for Figure 5

	Day 1		Day 2	
Subject	b	s_b	b	s_b
1	.248	.286	–.040	.141
2	–.048	.802	.168	.131
3	–.728	.643	–.468	.118
4	–.624	.622	.176	.148
5	–.092	.201	–.312	.157
6	.052	.305	.028	.261
7	.536	.423	–.284	.398
8	.092	.032	–.024	.196
9	.152	.917	.288	.190
10	.016	.128	.144	.073

GENERAL DISCUSSION

First we must come to an understanding about accepting a no-effect null result. Consider Figure 6; the line drawn there is the identity function. It represents the absence of an effect of giving a test during the study period: The ordinate is the total study time in the 100% condition, which did not have an interruption test; the abscissa is the mean of the total study times in the 25%, 50%, and 75% conditions, all of which had

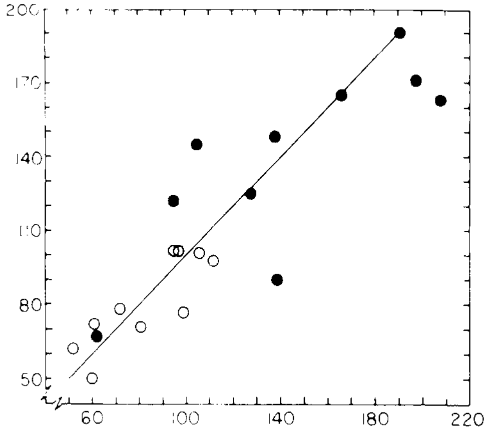

Figure 6. Study time in seconds for 100% condition plotted against average study time in seconds for 25%, 50%, and 75% conditions for individual subjects in first-day learning (filled circles) and second-day relearning (open circles).

interruption tests. This straight line, with slope 1, is the function embodied in a null hypothesis of no effect, had such a hypothesis been made. Probably, most readers would agree that it represents the data in a reasonably honest way, that it is acceptable as a description of the scatterplot. Moreover, our acceptance of this description would not be seriously jeopardized by learning that the Y-on-X linear regression functions for the closed circles (first-day learning) and open circles (second-day relearning) have slopes of .636 and .735, respectively, especially if the 95% confidence intervals are given (.246, 1.027 and .352, 1.119, both of which include the unit slope). In general, whenever we conclude that some set of data is described by some function, we are offering to the consumer an induced, post factum summary that he is expected to accept as not rejectable by the data from which the function was induced. In this kind of argument, the actual value of the slope of the induced function plays no role at all; it could just have well been zero, as when we speak of total study time as a function of point of interruption. Note that the situation under discussion has a different logical status than the situation where an a priori hypothesis is stated, tested, and not rejected.

Preliminary to considering the conclusion that temporal locus of a recitation test does not affect either learning difficulty or long-term retention, we must treat several potential problems. The first has to do with the possible role of warm-up and learning-to-learn effects as the experimental session proceeds. Some observations: Experiments 1 and 2 began with preliminary lists, which would serve to reduce such effects. Moreover, the presentation order of the interruption conditions in these two experiments were of the form ABCDDCBA, thus at least partially compensating for any warm-up and learning-to-learn effects that might remain. Experiment 3 consisted of three successive within-subjects replications, or cycles, of the interruption conditions, which means that all the interruption conditions occurred equally often in early, middle, and late segments of the sessions for each subject. The mean study times for the three cycles, collapsing over interruption conditions and averaging over subjects, were 150.1, 131.8, and 141.4 sec, respectively [$F(2,18) = 1.450$, $MSe = 577.974$, $p = .261$].

The second potential problem has to do with whether or not it is reasonable to represent the individual-subject data in Figures 1, 2, 3, and 5 with straight lines. Except for a few cases the answer is self-evident. The exceptions are the first-day learning times for Subject 5 in Figure 1, Subject 6 in Figure 3, and Subject 9 in Figure 5. The last of these seems the most serious candidate for genuine nonlinearity: Each point is the geometric mean over three cycles through the four conditions, and the study times from each of the three cycles show the same concave downward pattern. In contrast, the points for the other two subjects appear to gain their pattern from chance factors, since in both cases there is little agreement either between the two within-subjects replications of the four conditions on the first-day or between the first- and second-day patterns. Thus we see the straight line to be the most defensible summary of the individual data in Figures 1, 2, 3, and 5, with the reservation that extensive testing might turn up reliable nonlinear functions for some individuals.

The major issue to be taken up is whether it is sensible to conclude that the numerous linear slopes we have reported are but sampling variations about zero. Out of 50 slopes listed in Tables 1, 2, and 3, only 4 have 95% confidence intervals that exclude zero. The split between positive and negative slopes in first-day learning was 13 to 12, while for second-day relearning it was 10 to 15. Using the logic of McNemar's (1947) test, these two splits do not differ from each other. Using the logic of any test, neither split is distinguishable from a 50-50 split. We may continue this line of thought by computing the mean slope and 95% confidence interval for the mean slope, giving the individual slopes equal weight. The first-day learning data ($n = 25$) gave a mean slope of $-.069$, with a confidence interval of $-.246, .109$. The corresponding statistics for the second-day relearning data ($n = 25$) were $-.143$ and $-.290, .004$. All in all, there seems to be no reason for not concluding that a zero or near-zero slope is a fair summary.

In reporting the results of the three experiments, we summarized the individual-subject slopes in each experiment with a weighted mean, where the weights that were applied to the individual-subject slopes were the reciprocal variances of those slopes. This procedure deemphasizes the contribution of individual-subject slopes in direct proportion to their unreliability and yields the minimum variance mean slope (Hald, 1952, p. 243-245). These results are brought together in Table 4. As can be seen, the second-day mean slope from Experiment 1 was significantly negative and the first-day mean slope from Experiment 3 was significantly positive. Several questions arise. How is it, for example, that in Experiment 3 none of the individual-subject first-day learning slopes (Table 3) differed significantly from zero, yet the weighted mean slope (.084 in Table 4) did so differ? To answer this question, consider the individual-subject weights, $1/s_b^2$. As can be computed from Table 3, they varied from $1/(.917)^2 = 1.189$ for Subject 9 to

Table 4
Weighted Slopes

	Day 1	Day 2
Experiment 1	−.218	−.508*
Experiment 2	.430	.001
Experiment 3	.084*	.005

**95% confidence interval excludes 0.*

$1/(.032)^2 = 976.562$ for Subject 8. The weights for all subjects totalled 1,098.661. Thus the slope for Subject 8 comprised 976.562/1,098.661 = .889, or 89%, of all contributions to the mean slope. Further computations show that the four negative slopes (Subjects 2-5 in Table 3) contributed only 3% compared to the 97% contribution of the six positive slopes. It is not our intention to repudiate this weighting system, but rather to point out a feature of its behavior, namely, its sensitivity to highly reliable contributions. In any event, we consider the progression from Experiment 1 to 2 to 3 to be one of increasing precision for detecting an effect of temporal locus of test interruption, thus making the near-zero slopes from Experiment 3 the most credible.

There are two apparently inevitable responses to our results. The first has to do with "power," meaning the ability of our research to detect a discrepancy from an a priori hypothesis. But as we noted in the opening paragraph of this discussion, we had no a priori hypothesis; rather we set out to induce a relation, whatever it might be. As indicated by the final sentence in the preceding paragraph, the result of our inductive considerations is a flat straight line. A peculiar thing about a fairly induced function is that there is no way to statistically reject it using the data from which it was induced, no matter how many observations comprise the data. Thus the question of power cannot arise. It is permissible, though, for someone to hypothesize any other function he likes and use our data to evaluate that hypothesis. We have assisted all such someones by reporting confidence intervals. Thus, for example, the result from two paragraphs earlier (where $\bar{b} = -.069$, with a 95% confidence interval of $-.246, .109$ for 25 subjects) may be paraphrased by saying that all hypotheses about populational slope β with values $\beta < -.246$ and $\beta > .109$ are statistically rejectable at at least the .025 level in favor of a directional alternative closer to the value $\beta = 0$. As for whether we have enough data for a fair induction, it is not clear what a sensible stopping rule might be, as the exhausting and dismal history of this problem (generally known as Hume's problem) shows. We provide at least some compensation for this uncertainty both by introducing procedural variations from experiment to experiment and by displaying the data of every subject individually so that the reader may satisfy himself that there are no averaging artifacts.

The other inevitable response to our results is to ask about the amount of time the subjects spent in their recitation attempts, the idea being that perhaps these times vary with temporal locus of interruption and should be counted as part of the overall learning time. We regret to report that we did not foresee this inevitability. Recitations were indeed recorded on tape, but for the immediate purpose of verifying the experimenter's record of what was recited, which means that few such recordings are still on hand. The only observation we can offer is that since, in Experiment 3, the study times for the 100% condition (in which the subjects were not interrupted for a test) did not differ from the study times in the 25%, 50%, and 75% conditions (in which interruptions did in fact occur), it follows that whatever the relation might be between duration of a recitation attempt and point of interruption, it can only signal the degree to which the subject's time was wasted by the recitation attempt.

How to relate our results to the existing literature on test effects is not clear. The importance of an early test phase in paired associate learning finds no counterpart in our data, and the interchangeability of study and test phases in free recall learning is denied to our task. Thus neither of the major results of testing in the two traditional paradigms generalizes to the present learning situation.

REFERENCES

Allen, G. A., Mahler, W. A., & Estes, W. K. Effects of recall tests on long-term retention of paired associates. *Journal of Verbal Learning and Verbal Behavior*, 1969, **8**, 463-470.

Birnbaum, I. M., & Eichner, J. T. Study versus test trials and long-term retention in free-recall learning. *Journal of Verbal Learning and Verbal Behavior*, 1971, **10**, 516-521.

Bregman, A. S., & Wiener, J. R. Effects of test trials in paired-associate and free-recall learning. *Journal of Verbal Learning and Verbal Behavior*, 1970, **9**, 689-698.

Brooke, J. B., & MacRae, A. W. Error patterns in the judgment and production of numerical proportions. *Perception & Psychophysics*, 1977, **21**, 336-340.

Derks, P. L. The length-difficulty relation in immediate serial recall. *Journal of Verbal Learning and Verbal Behavior*, 1974, **13**, 335-354.

Donaldson, W. Output effects in multitrial free recall. *Journal of Verbal Learning and Verbal Behavior*, 1971, **10**, 577-585.

Hald, A. *Statistical theory with engineering applications*. New York: Wiley, 1952.

Hogan, R. M., & Kintsch, W. Differential effects of study and test trials on long-term recognition and recall. *Journal of Verbal Learning and Verbal Behavior*, 1971, **10**, 562-567.

Hudson, R. L., Solomon, M. L., & Davis, J. L. Effects of presentation and recall trials on clustering and recall. *Journal of Verbal Learning and Verbal Behavior*, 1972, **11**, 356-361.

Izawa, C. Function of test trials in paired-associate learning. *Journal of Experimental Psychology*, 1967, **75**, 194-209.

Izawa, C. Optimal potentiation effects and forgetting prevention effects of tests in paired-associate learning. *Journal of Experimental Psychology*, 1970, **83**, 340-344.

Lachman, R., & Laughery, K. R. Is a test trial a training trial in free recall learning? *Journal of Experimental Psychology*, 1968, **76**, 40-50.

LaPorte, R., & Voss, J. F. Paired-associate acquisition as a function of number of initial nontest trials. *Journal of Experimental Psychology*, 1974, **103**, 117-123.

LaPorte, R. E., & Voss, J. F. Retention of prose materials as a function of postacquisition testing. *Journal of Educational Psychology*, 1975, **67**, 259-266.

Martin, E., Fleming, F. G., Hennrikus, D. J., & Erickson, E. A. Studies of the length-difficulty relation in serial memorization. *Journal of Verbal Learning and Verbal Behavior*, 1977, **16**, 535-548.

McNemar, Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 1947, **12**, 153-157.

Rosner, S. R. The effects of presentation and recall trials on organization in multitrial free recall. *Journal of Verbal Learning and Verbal Behavior*, 1970, **9**, 69-74.

Tulving, E. Subjective organization in free recall of "unrelated" words. *Psychological Review*, 1962, **69**, 344-354.

Tulving, E. The effects of presentation and recall of material in free-recall learning. *Journal of Verbal Learning and Verbal Behavior*, 1967, **6**, 175-184.

NOTES

1. If X is positively skewed such that log X is normal, then X is said to be distributed lognormally. For X lognormal, the geometric mean of X is identical to the median of X, just as for log x or any other symmetrically distributed variable the arithmetic mean is identical to the median.

2. In a magnitude production task where the subject is verbally given a proportion or a percentage to actively produce in a physical array, there is overproduction at percentages between 0% and 50% and underproduction at percentages between 50% and 100% (Brooke & MacRae, 1977). Our subjects were similarly given verbal percentages and asked to produce something, namely, appropriate study interruption times. The open circles along the dashed line in Figure 4 are suggestive of a sinusoid with respect to that line, in agreement with the over- and underproduction phenomena in the more usual psychophysical situation.

(Received for publication October 3, 1977;
revision accepted January 23, 1978.)