JOURNAL OF VERBAL LEARNING AND VERBAL BEHAVIOR **16,** 465–478 (1977)

Learning from Tests: Effects of Spacing

WILLIAM B. WHITTEN II

State University of New York at Albany

AND

ROBERT A. BJORK

University of California, Los Angeles

The effects of delaying an immediate test, a rehearsal opportunity, or a second presentation on long-term free recall were investigated in an experiment combining features of the Brown–Peterson and free-recall paradigms. Each trial consisted of: (a) initial presentation of a word double (P_1); (b) 4, 8, or 14 seconds on a digit shadowing task; (c) either a second presentation of the word-pair (P_2), an overt recall test of that pair (T), or an opportunity to covertly rehearse the pair (R); and (d) 13, 9, or 3 seconds of shadowing. A free-recall test was given after each block of 12 trials. The end-of-block recall increased as a function of the second event delay for P_1–P_2 and P_1–T conditions but not for P_1–R conditions. A proposed model assumes that both repetitions and tests promote learning and that the effectiveness of both is a function of the current state of memory at the time of the repetition or test.

If in some situations tests are the functional equivalent of re-presentations of to-be-remembered items, then in these situations the functional effects we observe should be similar for re-presentations and for successful test events. Several free-recall and paired-associate learning experiments purporting to test the equivalence or nonequivalence of test trials and presentation trials have been reported. An overview of these experiments is somewhat confusing, however, since results and conclusions are not consistent. Tulving (1967, Experiment II), Bregman and Wiener (1970), Birnbaum and Eicher (1971), and Donaldson (1971) have all reported free-recall learning experiments in which test trials induced learning equivalent to that induced by study trials. Lachman and Laughery (1968, Experiment I), on the other hand, have concluded that the effect of a test trial depends on the level of training, and Hogan and Kintsch (1971, Experiment II) found that after a 48-hour delay the relative efficiency of test and study trials depends on the method of memory assessment: Test trials were better for delayed free recall, but study trials were optimal for delayed recognition.

The picture is equally cloudy for paired-associate learning experiments. Izawa (1966, 1967, 1970a,b) has concluded that test trials in a modified study–test (or R–T) paradigm function to prevent forgetting but do not induce learning. One interpretation (Izawa, 1970a) is that a test has a "potentiating effect" that somehow increments the conditioning power of a subsequent study trial. In somewhat simpler paired-associate learning experiments, Bregman and Wiener (1970) found test trials to be less effective than study trials, but Allen, Mahler, and Estes (1969)

Research reported here was supported in part by the Air Force Office of Scientific Research under Contract No. F44620-72-C-0038 with the Human Performance Center, Department of Psychology, University of Michigan. Address reprint requests to William B. Whitten, Psychology Department, State University of New York at Albany, 1400 Washington Avenue, Albany New York, 12222.

ISSN 0022–5371

reported that five study trials followed by five test trials resulted in better one-day delayed recall than did ten study trials. Consideration of the above set of studies leads one to suspect that the nature of multiple-trial free-recall learning and paired-associate learning tasks necessarily obscures the effects of *individual* presentation or test trials. These two trial types are certainly not equivalent in the freedom allowed for processing and in the availability of unlearned items for study. In several experiments, differences seemed to balance out so that total learning was similar for various combinations of presentations and tests, but these results may have been a coincidence of the particular experimental situations. In other experiments, differences in the rates of learning appear. But again, these differences may be due to task demands and constraints rather than to a difference in the processes occurring on presentation and test trials. While it is true that the questions relating to the overall rates of learning in situations involving various arrangements of presentations and tests are important and have been answered to some extent, it seems also that the question of equality or inequality of *processes* occurring during presentations and tests has not been directly addressed by these experiments. In these multiple trial situations, one does not know what effects are due to tests, to presentations, or to test–study interactions.

The proposal being made is that a more controlled situation, allowing each item equal processing opportunities and an analysis of the fate of an item after only two such opportunities, is necessary for the evaluation of the equality of processes. One such situation is afforded by the spacing-of-repetitions effect. It is well established in several paradigms that two spaced presentations of simple verbal materials lead to better performance than do two massed presentations. This phenomenon has been demonstrated in the Brown–Peterson paradigm (Peterson, 1963; Pollatsek, Note 1), paired-associate experiments (e.g., Greeno, 1964; Peterson, Hillner, & Saltzman, 1962; Bjork, Note 2; and others), and in the Melton free-recall lag paradigm (Melton, Reicher, & Shulman, Note 3; Melton, 1970).

If test events and re-presentations are reacted to by functionally similar processes, then it follows that one presentation coupled with an initial test should produce a spacing effect similar to that found with two presentations. This prediction must be qualified, however, since increments in final recall which are attributable to increasing the spacing of initial tests must outweigh any decrements in final recall which derive from the lower proportion of initial recall at longer spacing (retention) intervals. Consequently, the general prediction for a final recall would be a function that increases with initial–test delay to some maximum and then, due to initial–test failures after longer delays, decreases. Landauer and Eldridge (1967) and Landauer (1969) have reported several paired-associate recognition experiments and one paired-associate recall experiment that, on the whole, support this prediction. In addition, Young (1971) has reported that an intermediate delay of a paired-associate test results in optimal learning from that test. Young's conclusions, however, are based on a complex experimental design that confounded test effects with test–study interactions, thus limiting their generality.

The present experiment was designed for three purposes. The main purpose was to study the similarity of test–trial and study–trial learning processes by comparing the final free recall function of items which were initially tested after various delays (spacing intervals) with the final recall function of items which were re-presented after those delays. If similar learning processes occur on test and study trials, then similar spacing-effect functions should be obtained for the two conditions. This experiment, therefore, extends the aforementioned research to the free-recall paradigm.

The second purpose was to investigate the similarity of covert rehearsal processes to overt test and repetition (study) processes. To the extent that covert rehearsal requires processes similar to those of an overt test, similar effects of spacing should be observed. The experiment reported below required implicit retrieval for both the covert rehearsal and the overt test conditions; thus, these conditions differed mainly on the covert–overt dimension.

Finally, the third purpose of the experiment was to provide an index of the relative effectiveness of a second presentation, an initial test, and a covert rehearsal opportunity at various spacings.

To investigate these questions, a paradigm was designed that is a cross between the Brown–Peterson paradigm (Brown, 1958; Peterson & Peterson, 1959) and the free-recall paradigm. In this hybrid design, a series of pairs of words (word doubles) are presented with a variable interval between successive word doubles, and all the intervals are filled with a distractor activity designed to minimize rehearsal opportunities. Due to the complexity of the design, the experiment involved two groups. Within a series of Group A, a word double was shown, shadowing was required, and one of two conditions occurred: (1) The word double was shown again; or (2) a cue to covertly rehearse appeared. Varied amounts of shadowing followed either of these events after which another word double was presented. Following a series of 12 such continuous presentations, subjects were cued to free recall. Group B received the same sequence of events within a trial series as Group A, with the exception that Group B had overt tests substituted wherever Group A had covert rehearsal opportunities. Three trial blocks were presented to every subject in each group.

Those trials with two presentations were designed both to verify the spacing effect in this paradigm and to provide a reference condition against which to compare the other conditions. The presentation–rehearsal conditions were designed to test effectiveness of covert rerehearsal at various spacings for long-term retention. Presentation–immediate test conditions served two purposes: (a) to test the facilitation of an overt recall (overt rehearsal) at various spacing intervals for final recall of the trial block and (b) to provide an estimate of the amount of material available to rehearse at each retention interval.

METHOD

Subjects

The subjects were 72 undergraduate and graduate women at the University of Michigan. They were paid for their participation in the experiment.

Materials and Apparatus

The 36 word doubles were constructed from common four-letter nouns under the following constraints: (a) Words in a pair did not begin with the same letter, (b) the words in a pair did not rhyme, (c) each word was used in only one word double, and (d) there were no obvious associations between the words in any one word double.

The apparatus was a high-speed memory drum (change time less than .05 seconds) that was programmed with a timer to produce the various spacing intervals required. All word doubles, digits to be shadowed, and cues to get ready, to rehearse, to recall, or to rest were shown in the window of the apparatus.

For half the subjects a row of Xs highlighted in pink constituted an instruction to covertly rehearse the immediately preceding word double; for the other subjects, this symbol was an instruction to recall that word double. Second presentations of word doubles were overscored with yellow. Free recall of all words from each trial block was signaled by a row of question marks highlighted in green.

Design

Every condition (trial) can be schematized as (P_1), (I_a), (X), (I_b), where P_1 represents the initial presentation of a word double, I_a and I_b represent variable amounts of an interfering shadowing task, and X represents one of three events: a second presentation of the word double (P_2), an opportunity for covert rehearsal of the word double (R), or an overt test for the word double (T). An uninterrupted series of 12 (P_1), (I_a), (X), (I_b) sequences constituted a block of trials. Half the trials of the 36 subjects in Group A included two presentations of a word double (P_1–P_2 conditions); the remaining trials for Group A substituted a covert rehearsal opportunity for the second presentation (P_1–R conditions). The 36 subjects in Group B received P_1–P_2 conditions mixed equally with P_1–T conditions. The latter consisted of a single presentation of a word double followed at various delays by an overt recall test on that word double. Both groups used the same stimulus lists, but Group A was instructed to rehearse covertly (P_1–R conditions) when a symbol appeared, while Group B was instructed to recall aloud (P_1–T conditions) when that symbol appeared. The I_a intervals were 4, 8, or 14 seconds, while the corresponding I_b intervals were 13, 9, or 3 seconds, respectively. Initial presentation (P_1) of each word double was for 2 seconds and the duration of $X = P_2$, R, or T was 3 seconds. Thus, each word double trial within the continuous series lasted 22 seconds and included 17 seconds of shadowing activity.

Each subject, then, was given a total of six different conditions: three P_1–P_2 conditions with 4, 8, or 14 P_1–X spacing intervals and either three P_1–R or three P_1–T conditions with 4, 8, or 14 P_1–X spacing intervals. One instance of each of the six conditions for a given group was randomly positioned in each half of a trial block. Three such blocks were given to each subject. Since the experiment was, in part, a free-recall experiment, conditions were rotated within half-blocks to counterbalance possible serial position effects. This counterbalancing resulted in six different rotations, each of which was used by six of the 36 subjects in both Group A and Group B. In addition, word pairs were rotated to avoid consistent pairings of words with conditions.

Procedure

Subjects were tested individually. Before beginning the experiment subjects were read the instructions and were given an abbreviated block of three continuously presented conditions followed by a free-recall test. The instructions emphasized both the importance of the shadowing task and, for Group B, the importance of the immediate recall test. In addition, subjects were instructed to work only on the current pair of words and not to rehearse others.

The beginning of each trial block was signaled by a 2-second READY signal. In each trial block, subjects received the following uninterrupted sequence: (a) sixteen seconds of shadowing four-digit numbers at a 1-second rate, (b) 12 (P_1), (I_a), (X), (I_b) trial sequences, (c) another 16 seconds of digit shadowing, and (d) a cue to free recall all of the words from the block. A recall interval of 2.2 minutes was followed by a REST signal which indicated a 2.2-minute rest period between blocks. The experimenter, who was seated behind a partition, recorded immediate recall from the Group B P_1–T conditions, and subjects wrote their recall for each block.

Results

Although word doubles were presented, the free-recall instructions emphasized that the words of a double need not be recalled together. Data were scored both in terms of word doubles and as individual words. The word-double data parallel the single-word data but are at a lower overall level of performance. The single-word data are presented below.

Recall Data as a Function of Spacing

Performance on P_1–P_2 items. Data were analyzed separately for Groups A and B. The average probability of recall at the end of a

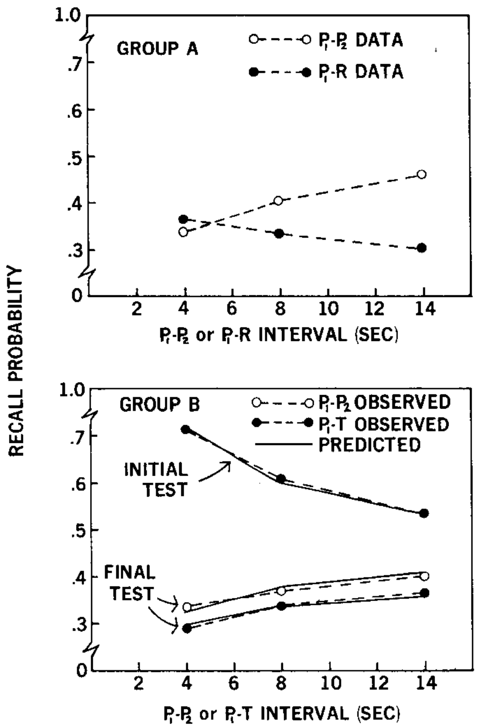

FIG. 1. Observed end-of-block (final) recall for Group A (top panel); observed and predicted initial and end-of-block recall for Group B (lower panel).

continuous trial block is shown as a function of the P_1–X interval (X = P_2, R, or T) in Figure 1. Dashed curves with filled or unfilled circles refer to the data, while the solid curves are predicted values derived from a model that is presented later in this paper. Linear trend analyses showed that recall increased significantly in a mostly linear fashion as the P_1–P_2 interval increased, $F(1, 70) = 7.91$, $p < .01$, $MS_e = 4.93$ for Group A and $F(1, 70) = 4.06$, $p < .05$, $MS_e = 2.63$ for Group B. Thus, the spacing effect with two presentations was verified in this modified paradigm.

Performance on P_1–R items. The performance on P_1–R conditions of Group A is shown in the top panel of Figure 1. An overall decline in recall of .06 is found as the P_1–R retention interval increases from 4 to 14 seconds, but trend analyses provided a nonsignificant linear component, $F(1, 70) = 1.89$, $MS_e = 4.93$. Thus the data fail to show an absolute increase with spacing. At the same time, however, the insignificant decline in recall is not as great as might be expected since, as the P_1–R interval lengthens, the probability of being able to recall the words during the rehearsal interval (R) decreases. An estimate of the amount of material available at the various spacings of rehearsal opportunity is necessary to assess the possibility of a relative (conditional) spacing effect. No estimate of this quantity is available for Group A, but a rough estimate is provided by the immediate test results of Group B.

Performance on P_1–T items. The P_1–T conditions of Group B give both a measure of the items available at the time of the initial test and a measure of final end-of-block recall of P_1–T items. The upper curve in the bottom panel of Figure 1 shows the probability of recall of items at 4-, 8-, and 14-second retention intervals. This measure is directly analogous to the standard Brown–Peterson short-term retention measure and exhibits a decrease in probability of recall of .178 as the retention interval increases from 4 to 14 seconds.

Final recall for P_1–T trials is shown by the lower curve in the bottom panel of Figure 1. Notice that, even though the probability of recall at the immediate test decreases, final recall increases as a function of increasing P_1–T spacing. Analysis of immediate and final test data from the P_1–T conditions showed a strong test (immediate or final) by spacing (4, 8, or 14 seconds) interaction, $F(2, 70) = 17.24$, $p < .001$, $MS_e = 2.28$. In addition, final test P_1–T data exhibited a significant linear trend, $F(1, 70) = 4.25$, $p < .05$, $MS_e = 2.63$. Thus, delay of an overt recall facilitated long-term retention of items given a single presentation. This "spacing effect" due to delayed testing is lower overall but nearly parallel to the standard spacing effect of the two-presentation conditions.

Table 1 shows the conditional probability that a word was correct in final recall when it was correct at an N-second-spaced initial test, $P(C_F|C_N)$. The table also gives the probability that a word was correct at final recall if it was incorrect on the N-second-spaced initial test, $P(C_F|I_N)$. The mean $P(C_F|C_N)$ increased sharply with the delay of the immediate test, while the mean $P(C_F|I_N)$ did not. The data for the average $P(C_F|C_N)$ and $P(C_F|I_N)$ are also shown in Figure 2 where the solid curves again represent theoretical values obtained from the model. It should be noted that, even though these conditional probabilities were calculated for each subject and then averaged, the possible existence of item selection effects dictates a cautious interpretation. These conditional probabilities are included for completeness and because they are predicted by the model. The unconditional probabilities are the data of main importance.

TABLE 1

$P(C_F|C_N)$ AND $P(C_F|I_N)$ FOR $P_1 - T$ CONDITIONS

P_1–T interval (seconds)	Block 1	2	3	Mean
	$P(C_F\|C_N)$			
4	.39	.38	.33	.37
8	.56	.46	.47	.50
14	.56	.54	.62	.57
	$P(C_F\|I_N)$			
4	.09	.14	.00	.09
8	.08	.11	.10	.10
14	.07	.14	.16	.11

Serial Position Effects

The design of the experiment was intended to obliterate any serial position effects within a trial block. The initial and final 16 seconds of shadowing were intended to reduce the primacy and recency effects within each trial block. To check the success of these intentions, the probability of recall averaged across all P_1–X intervals was plotted for each group of subjects (A and B) as a function of serial position. These data are shown in Figure 3.

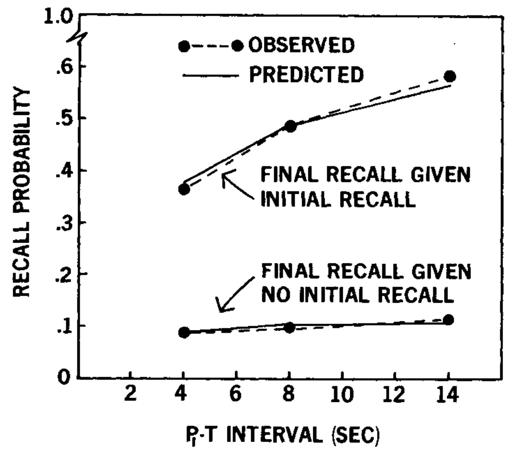

FIG. 2. Observed and predicted conditional probabilities for Group B.

Both words presented in a word double of a trial were considered to be in the same serial position for purposes of this analysis. Thus, in Figure 3, a serial position corresponds to the serial order of one of the 12 sequentially presented trials in a block. The two words presented in each trial were scored independently rather than as completely correct word doubles. Both curves show some evidence of a primacy effect and extraordinary recency effects extend back at least as far as four trials. These recency effects are surprising when it is noted that each trial is 22

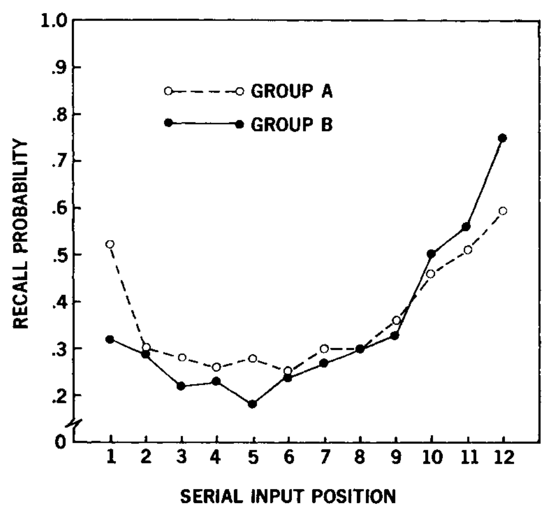

FIG. 3. Final (end-of-block) recall of Group A and Group B as a function of serial input position.

seconds in length and that there is an additional 16 seconds of difficult shadowing activity after the last trial and hence, preceding recall. The last opportunity for a P_2, R, or T event for the fourth trial from the end of the list (ninth serial position) is 85, 91, or 95 seconds before the free-recall test. All of this time was filled with the shadowing activity, or with P_1, P_2, R, or T events for new words. These intriguing long-term serial position effects instigated a series of experiments which have been reported by Bjork and Whitten (1974) and by Whitten and Bjork (Note 4), as well as by Tzeng (1973), Alexy and Gorfein (Note 5), and Elmes and Wood (Note 6). The reader is referred to these papers for a detailed discussion of the long-term recency effects.

Figure 4 shows the proportion recalled at T for each trial block as a function of serial input position. The curves have been smoothed by averaging the data of each serial position with that of the two immediately adjacent serial positions. The endpoints are not averaged. These curves show effects of practice since performance increases over the three successive blocks. In addition, the shape of each curve is exactly what would be expected from a standard series of Brown–Peterson trials: The initial drop in recall probability, usually attributed to proactive interference, is followed by a characteristic slight increase over trials. The slight increases observed, however, are far too small to account for the free recall serial position effects of Figure 3.

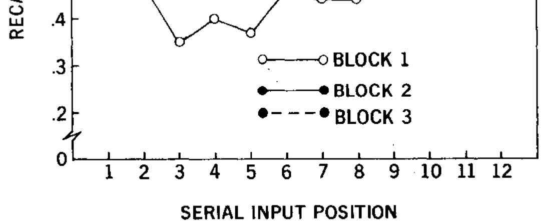

FIG. 4. Initial recall (at T) for each trial block in condition P_1–T shown as a function of serial input order. Data are smoothed (see text).

DISCUSSION

The P_1–P_2 condition spacing effect is similar in magnitude to that found by Peterson (1963), even though the P_2-to-recall interval was much longer in the present study. It is reassuring to find such longer-term effects which coincide with effects reported from often used short-term memory paradigms.

While the P_1–R function of Group A failed to exhibit an increasing spacing function, it is tempting to compare the amount of material available for rehearsal with the final recall at each spacing interval. These comparisons should be made with extreme caution, however. The only estimate of the number of items available for the rehearsal intervals (R) derives from the amounts recalled on the initial tests (T), and these data are from different groups of subjects. In addition, it is possible that item selection effects are present. Keeping these problems in mind, it can be noted that a conditionalized spacing function would monotonically increase for the P_1–R condition as it did for the P_1–T condition.

The data of Group B are interpretable in a more straightforward way. Even though the proportion of material recallable at immediate test decreases, the proportion recalled during end-of-block free recall increases as the P_1–T interval increases. Thus, the best description of these results is as a spacing-of-practice or distributed-practice effect where practice consisted of the opportunity to recall aloud the to-be-remembered words. More precisely, then, the effect may be termed the "spacing-of-tests effect." The analysis of the spacing effect found for the P_1–T conditions relative to the amount of material retrievable at the immediate test is even more striking. This conditionalized comparison is again open to the criticism of item selection effects but is included for the reasons cited in the previous section.

Why did recalling aloud (T) in the P_1–T condition produce an absolute spacing effect while covert rehearsal (R) in the P_1–R condition did not? There are two obvious possibilities. First, it is possible that overt retrieval and covert rehearsal involve qualitatively different processes. Little data exists to support or deny this conclusion. Second, it is possible that the differences between P_1–T and P_1–R results reflect different amounts of experimental control in the two conditions. Group B subjects were strongly constrained to overtly recall and hence rehearse the current pair of words since the experimenter was scoring this performance. Even though Group A subjects were instructed to work only on the current words, it was possible for them to practice older words or do nothing since rehearsal was done covertly. This possibility underscores an inherent problem of studying covert processes. It remains to be shown whether covert rehearsal always leads to a decreasing unconditional spacing-of-tests function as in the present experiment, or whether this result is dependent on inadequate experimental control.

The reported P_1–T spacing-of-tests function was monotonically increasing over the P_1–T intervals investigated. All obvious conceptualizations predict that the spacing function will reach a maximum and then decline at longer spacing intervals: After a very long delay, any possible benefits of the delayed test should be overshadowed by decrements due to initial-recall failures. Additional research will be needed to specify the optimal test-spacing interval.

Theoretical implications of the spacing-of-tests effect are not clear. Both consolidation theory (e.g., Landauer, 1969, 1974) and variable encoding theory (e.g., Martin, 1968) may apparently be extended to account for the phenomenon. Consolidation theory suggests that recall at T would reactivate the original trace that was due to P_1 and, as the P_1–T interval increases, the reactivation needed would be more complete. The total neural activity experienced would therefore be greater for the longer P_1–T interval items. This notion predicts a maximum followed by a decline in the spacing-of-test effect. When the initial consolidation asymptotes or ceases altogether, the level of reactivation needed for further consolidation of the original trace should be a constant. At this point, further spacing of the test would not improve the $P(C_F|C_N)$, but would produce a down-turn in the absolute spacing-of-tests effect curve since few items would be recalled at T.

Differential encoding theory could similarly account for the spacing-of-tests effect. If it is assumed that redundant encoded stimulus patterns (retrieval routes), each of which is sufficient for recall, are established at P_1, then the reasoning is similar to standard stimulus fluctuation notions (Estes, 1955a, b). If the availability of retrieval routes or aspects of retrieval routes fluctuate over time, then on the average, fewer would be available at longer P_1–T intervals. If whenever an item is recalled, the subject elaborates on the retrieval route, this elaboration would more likely be nonoverlapping with the original when some of the previously conditioned aspects of the retrieval route were unavailable. The total elaboration of retrieval routes and hence the probability of long-term recall would therefore be greater for spaced-test items.

In addition to these two theories, the multiple-copy random-storage model of Landauer (1975) might predict the spacing-of-tests effect. This model would predict that a new storage of an item would occur whenever that item is retrieved at T. The probability of a second storage would decrease as the P_1–T interval increases, but the value of the second storage would increase because it would probably occur at a greater distance from the original encoding. In this regard, this spatial-analog model is conceptually related to the differential encoding explanation. For both models, the degree that the second encoding overlaps with the original encoding decreases as the test–spacing increases.

Some recent research concerning the effectiveness of type of rehearsal for immediate and delayed recall tests has produced results which seem strikingly parallel to those reported in this paper. Jacoby and Bartz (1972) and Jacoby (1973, Experiment II) have found that initial recall of five-word lists was better after a 15-second silent delay than after a 15-second filled delay. Final recall, however, was better for items initially recalled after the filled delay. Thus, as in the spacing-of-tests effect, items that are more difficult to recall initially are more likely to be recalled finally.

Jacoby and Bartz offered two possible explanations for these effects. The first is that different types of processing are being invoked during list study, the implication being that subjects expecting interference during the delay choose to process the items in some more substantial manner. Although this is a possible explanation of their results, it is not readily applicable to the results of the present paper since, due to the within-subject design, the subjects' trial-to-trial expectancies were not differentiable. In addition, subjects of our experiment were expecting the final recall test and should therefore treat all to-be-remembered items similarly.

The second explanation offered by Jacoby and Bartz is that there is a differential effect of initial tests depending upon whether the items are being recalled from short-term memory (STM) or from long-term memory (LTM). A related hypothesis that may also be offered is that the effect of a test on subsequent memory may depend on the level-of-processing or quality of information used during the test (Craik & Lockhart, 1972; Whitten, Note 7). Both notions incorporate the assumption that the effectiveness of the tests depends on the state of memory at the time of the test. In the next section a model is presented which incorporates this idea as one of its basic assumptions. While the STM–LTM terminology is used in the model, it is entirely reasonable to translate the model into the levels-of-processing terminology.

Toward a Model of Test Events as Learning Trials

A model is presented in this section that describes a possible set of psychological processes which account for initial and final recall of the P_1–T conditions, free recall of the P_1–P_2 conditions, and the probabilities of being correct on the final test conditional on being correct or not on the initial test, $P(C_F|C_N)$ and $P(C_F|I_N)$, respectively. The purpose of this model is to demonstrate that test trial and study trial processes can be conceptualized as functionally identical. The model describes data of Group B only since, as noted above, it is not yet clear how to interpret the P_1–R data.

General Assumptions of the Model

(1) Upon presentation an item is stored in both short-term memory (STM) and long-term memory (LTM) with probability STORE. Items not stored are assumed to be stored in but lost from a rapidly decaying sensory store.

(2) STM decays if not "recycled."

(3) LTM consists of highly correlated but not identical "core encodings" and "retrieval routes." Core encodings are conceptualized as those connections or nodes which are central to the meaning of the items while the retrieval routes are structures which allow access to the core encodings and, hence, contain such information as temporal and spatial order of presentation and context. LTM retrieval routes of individual items fluctuate between an available state (A) and

an unavailable state (U) at a rate given by the matrix

$$\begin{array}{c|cc} & A_{N+1} & U_{N+1} \\ \hline A_N & 1-\delta & \delta \\ U_N & \delta' & 1-\delta' \end{array}$$

where δ is the rate at which items go from A to U, and δ' is the rate at which items fluctuate from U to A with $\delta > \delta'$. As N becomes large, the probability that a retrieval route is available reaches an asymptote (AA) of $\delta'/(\delta + \delta')$.

Assumptions for P_1–T Conditions

At the initial test (T) one of the following four events occurs:

(1) Items still in STM are output from STM with no reference to and, therefore, no effect on LTM. Thus, LTM fluctuation rates are unchanged.

(2) If an item was stored at P_1 but is gone from STM and is available in LTM at T it is retrieved from LTM. This exercise of the retrieval route stabilizes availability by slowing the retrieval route fluctuation rate. The result is functionally similar to traditional consolidation notions. The slower fluctuation rates are given by

$$\begin{array}{c|cc} & A_{N+1} & U_{N+1} \\ \hline A_N & 1-\alpha & \alpha \\ U_N & \alpha' & 1-\alpha' \end{array}$$

where $\alpha > \alpha'$ and $\alpha < \delta$. As N becomes large, this matrix predicts an availability asymptote (AA') of $\alpha'/(\alpha + \alpha')$.

(3) If an item was stored at P_1 but is lost from STM and unavailable from LTM at T, no change is made by the test opportunity and the initial retrieval route fluctuation rates apply.

(4) Items not stored at P_1 are never available at T.

Assumptions for P_1–P_2 Conditions

(1) If the item is still in STM at P_2, LTM is not referenced so that P_2 has no effect on the fluctuation rates. A new encoding is not stored either.

(2) If the item was stored at P_1 but is lost from STM and is available in LTM at P_2, the retrieval route fluctuation rate is altered as in the P_1–T conditions.

(3) If the item was stored at P_1 but is lost from STM and unavailable in LTM at P_2, a new encoding is stored with probability STORE.

(4) If nothing was stored at P_1, the item is stored with probability STORE at P_2.

Derivations and Predictions of the Model

The model was written as a computer subroutine with four free parameters in the equations. The best simultaneous estimates of these parameters were obtained by using the subroutine STEPIT (Chandler, Note 8) whose function is to choose parameter values in such a way as to minimize the value of a statistic which compares the model predictions with the observed data. The statistic used in this case was χ^2. The estimated parameters were: (1) D, the probability of decay or loss from STM; (2) STORE, the probability of storing an item in STM and LTM; (3) AA, the asymptotic value of the probability that an item is available in LTM; (4) A_4, the probability of an item being available in LTM after 4 seconds. The derivations of the model are as follows:

The probability that an item is in STM after N seconds is given by

$$P(\mathrm{STM}_N) = (1 - D)^N. \qquad [1]$$

The probability of immediate recall from either STM or LTM after an interval of N seconds is given by

$$P(\mathrm{C}_N) = \mathrm{STORE}\cdot P(\mathrm{STM}_N) + \mathrm{STORE}\cdot \mathrm{A}_N\,[1 - P(\mathrm{STM}_N)], \qquad [2]$$

where A_N is the probability that an item is available in LTM after N seconds.

The probability that an item is recalled finally when not recalled after N seconds is given by

$$P(\mathrm{C_F}|\mathrm{I}_N) = S[P(\mathrm{C_F}|\text{not stored})\cdot P(\text{not stored}|\mathrm{I}_N) + P(\mathrm{C_F}|\mathrm{U}_N)\cdot P(\mathrm{U}_N|\mathrm{I}_N)], \qquad [3]$$

where U_N means "unavailable in LTM at N seconds" and where S is a forgetting parameter that adjusts downward the overall performance level of final recall as compared to initial recall. This adjustment does not alter the relation between values of $P(C_F|I_N)$ for various values of N, but it does allow for overall decrements in recall which are due to memory search-set size or output interference.

The conditional probability of correct final free recall given correct initial recall is provided by

$$P(C_F|C_N) = P(C_F|STM_N) \cdot P(STM_N|C_N) + P(C_F|A_N \cap \overline{STM_N}) \cdot P(A_N \cap \overline{STM_N}|C_N). \quad [4]$$

Notice that $P(C_F|STM_N)$ must equal the asymptotic value of the LTM availability parameter adjusted for output interference, or $S \cdot AA$. Also, $P(STM_N|C_N) = STORE\,[P(STM_N)/P(C_N)]$. The $P(C_F|A_N \cap \overline{STM_N})$ is equal to $S' \cdot AA'$, the asymptotic value of the new (higher) probability that an item is available in LTM, again adjusted for forgetting. And, note that $P(A_N \cap \overline{STM_N}|C_N) = 1 - P(STM_N|C_N)$. Making the proper substitutions, Eq. [4] becomes

$$P(C_F|C_N) = S \cdot AA \cdot STORE \cdot [P(STM_N)/P(C_N)] + S' \cdot AA' \cdot \{1 - STORE[P(STM_N)/P(C_N)]\}. \quad [5]$$

The values of $(S \cdot AA)$ and $(S' \cdot AA')$ were estimated from the data.

Joint probabilities were calculated by

$$P(CC) = P(C_N) \cdot P(C_F|C_N) \quad [6]$$
$$P(CI) = P(C_N) \cdot [1 - P(C_F|C_N)] \quad [7]$$
$$P(IC) = [1 - P(C_N)] \cdot P(C_F|I_N) \quad [8]$$
$$P(II) = [1 - P(C_N)]\,[1 - P(C_F|I_N)] \quad [9]$$

Final recalls for the P_1–T conditions were predicted by

$$P(P_1 - T_N \text{ Final Recall}) = P(C_F|C_N) \cdot P(C_N) + P(C_F|I_N)\,[1 - P(C_N)]. \quad [10]$$

Final recalls for P_1–P_2 conditions were predicted by:

$$P(P_1\text{–}P_{2,N} \text{ Final Recall}) = STORE \cdot S \cdot AA \cdot (1 - STORE) + S \cdot AA \cdot STORE \cdot P(STM_N) + S' \cdot AA' \cdot STORE \cdot A_N[1 - P(STM_N)] + \{(1 - STORE) \cdot S \cdot AA + STORE[2 \cdot S \cdot AA - (S \cdot AA)^2]\} \cdot \{STORE[1 - P(STM_N)]\,(1 - A_N)\}. \quad [11]$$

The last values needed for the chi-square goodness-of-fit statistic are given by

$$P(\text{Incorrect on } P_1\text{–}P_{2,N} \text{ Final Recall}) = 1 - P(P_1\text{–}P_{2,N} \text{ Final Recall}). \quad [12]$$

A chi-square statistic compared the three predictions from each of Eqs. [6–9], [11], and [12] with the 18 corresponding observed values for each set of parameter values assigned by STEPIT. The best estimates fit the data within 1% for most comparisons that entered the χ^2 calculations giving $\chi^2 = 2.41$ ($df = 8$, $p > .95$). Observed and predicted values for these 18 calculations are given in Table 2. The estimated parameter values were: $D = .202$; STORE = .888; AA = .581; $A_4 = .689$. A third column of Table 2 gives values for an alternative model which is described below.

Predictions of immediate and final P_1–T recall probabilities and of final P_1–P_2 recall probabilities are shown in the lower panel of Figure 1. The $P(C_F|C_N)$ and $P(C_F|I_N)$ predictions are shown in Figure 2. Predicted values of $P(STM_N)$ were .406, .165, and .043 for $N = 4$, 8, and 14 seconds, respectively, and the values of A_N were .689, .609, and .585 for $N = 4$, 8, and 14 seconds.

While the model's assumption (1) for the P_1–P_2 conditions is not uncommon, it may seem unreasonable to some investigators that a repetition is postulated to have no effect whatsoever when the repeated items are still in STM. An alternative model that was identical in all respects except for assumption (1) of the P_1–P_2 conditions was tested. The new assumption was as follows: "If an item is still in STM at P_2, LTM is not referenced and another encoding is stored with probability STORE. This encoding is not the same as the

TABLE 2

CORRESPONDENCE BETWEEN OBSERVED DATA AND THE PREDICTIONS OF THE ORIGINAL AND ALTERNATIVE MODELS FOR QUANTITIES USED BY STEPIT

Equation number	Quantity	P_1–P_2 or P_1–T interval (seconds)	Observed value	Original model predictions	Alternative model predictions
6	P(CC)	4	.262	.270	.257
6	P(CC)	8	.299	.292	.294
6	P(CC)	14	.310	.304	.314
7	P(CI)	4	.454	.453	.458
7	P(CI)	8	.315	.305	.296
7	P(CI)	14	.222	.231	.230
8	P(IC)	4	.025	.024	.028
8	P(IC)	8	.037	.041	.041
8	P(IC)	14	.053	.051	.046
9	P(II)	4	.259	.252	.257
9	P(II)	8	.350	.361	.370
9	P(II)	14	.414	.414	.411
11	P_1–P_2 recall	4	.336	.326	.342
11	P_1–P_2 recall	8	.368	.380	.383
11	P_1–P_2 recall	14	.400	.408	.403
12	P_1–P_2 error	4	.664	.674	.658
12	P_1–P_2 error	8	.632	.620	.617
12	P_1–P_2 error	14	.600	.592	.597

P_1 encoding and does not change the fluctuation rate of the P_1 retrieval route." The second term of Eq. [11] is replaced by: $\{(1 - \text{STORE}) \cdot S \cdot AA + \text{STORE}[2 \cdot S \cdot AA - (S \cdot AA)^2]\}$. The best fit of this alternative model yielded $\chi^2 = 3.66$ ($df = 8$, $.90 > p > .75$). Parameter estimates for this model were: $D = .234$; STORE $= .999$; $AA = .533$; $A_4 = .566$. The predicted values of this alternative model are given in the right column of Table 2.

The major assumption underlying these models is that learning can occur on test trials and that the processes of learning on test trials are essentially identical to processes occurring on presentation trials. The initial model assumes the fewest differences between the P_1–P_2 and P_1–T conditions and the excellent correspondence between the predictions and the data provides a strong argument for equality of processes. The alternative model, allowing a second presentation to affect learning even when the items are still in STM, is not rejected, however. Even so, the argument for identical processes underlying learning on both test and presentation trials is not seriously weakened since it is reasonable to assume that a successful recall of items from STM may also be followed by additional processing, if time allows. Within the confines of the present experiment, this possibility is minimized, thus making the differential assumption for P_1–P_2 and P_1–T conditions reasonable.

The spacing-of-tests effect has now been demonstrated in both paired-associate and free-recall paradigms. This phenomenon has both practical and theoretical significance. It suggests that, in practical applications, learning may be optimized when retrieval practice is delayed for a brief interval. The theoretical significance of the spacing-of-tests effect is that the learning derived from retrieval practice produces functions which are qualitatively similar to spacing-of-repetition functions. This result supports the notion that a successful retrieval increments memory in the same way as a repetition. The success of the above model lends further support to that idea.

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.

Bjork, R. A., & Whitten, W. B. Recency-sensitive retrieval processes in long-term free recall. *Cognitive Psychology*, 1974, **6**, 173–189.

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.

Brown, J. Some tests of the decay theory of immediate memory. *Quarterly Journal of Experimental Psychology*, 1958, **10**, 12–21.

Craik, F. I. M., & Lockhart, R. S. Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, 1972, **11**, 671–684.

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

Estes, W. K. Statistical theory of spontaneous recovery and regression. *Psychological Review*, 1955a, **62**, 145–154.

Estes, W. K. Statistical theory of distributional phenomena. *Psychological Review*, 1955b, **62**, 369–377.

Greeno, J. G. Paired-associate learning with massed and distributed repetitions of items. *Journal of Experimental Psychology*, 1964, **67**, 286–295.

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.

Izawa, C. Reinforcement-test sequences in paired-associate learning. *Psychological Reports*, 1966, **18**, 879–919.

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

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

Izawa, C. Reinforcement-test-blank acquisition programming under the unmixed-list design in paired-associate learning. *Psychonomic Science*, 1970b, **19**, 75–76.

Jacoby, L. L. Encoding processes, rehearsal, and recall requirements. *Journal of Verbal Learning and Verbal Behavior*, 1973, **12**, 302–310.

Jacoby, L. L., & Bartz, W. H. Rehearsal and transfer to LTM. *Journal of Verbal Learning and Verbal Behavior*, 1972, **11**, 561–565.

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

Landauer, T. K. Reinforcement as consolidation. *Psychological Review*, 1969, **76**, 82–96.

Landauer, T. K. Consolidation in human memory: Retrograde amnestic effects of confusable items in paired-associate learning. *Journal of Verbal Learning and Verbal Behavior*, 1974, **13**, 45–53.

Landauer, T. K. Memory without organization: Properties of a model with random storage and undirected retrieval. *Cognitive Psychology*, 1975, **7**, 495–531.

Landauer, T. K., & Eldridge, L. Effect of tests without feedback and presentation-test interval in paired-associate learning. *Journal of Experimental Psychology*, 1967, **75**, 290–298.

Martin, E. Stimulus meaningfulness and paired associate transfer: An encoding variability hypothesis. *Psychological Review*, 1968, **75**, 421–441.

Melton, A. W. The situation with respect to the spacing of repetitions and memory. *Journal of Verbal Learning and Verbal Behavior*, 1970, **9**, 596–605.

Peterson, L. R. Immediate memory: Data and theory. In C. N. Cofer & V. S. Musgrave (Eds.), *Verbal behavior and learning: Problems and processes.* New York: McGraw-Hill, 1963.

Peterson, L. R., Hillner, K., & Saltzman, D. Supplementary report: Time between pairings and short-term retention. *Journal of Experimental Psychology*, 1962, **64**, 550–551.

Peterson, L. R. & Peterson, M. Short-term retention of individual items. *Journal of Experimental Psychology*, 1959, **58**, 193–198.

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.

Tzeng, O. J. L., Positive recency effect in a delayed free recall. *Journal of Verbal Learning and Verbal Behavior*, 1973, **12**, 436–439.

Young, J. L. Reinforcement-test intervals in paired-associate learning. *Journal of Mathematical Psychology*, 1971, **8**, 58–81.

Reference Notes

1. Pollatsek, A. W. *Rehearsal, interference, and spacing of practice in short-term memory* (Tech. Rep. No. 16). University of Michigan, Human Performance Center, 1969.
2. Bjork, R. A. *Learning and short-term retention or paired-associates in relation to specific sequences of interpresentation intervals* (Tech. Rep. No. 106), Stanford University, Institute for Mathematical Studies in the Social Sciences, 1966.
3. Melton, A. W., Reicher, G. M., & Schulman, H. G. *A distributed practice effect on probability of recall in free recall of words.* Paper read at the Meeting of the Pyschonomic Society, St. Louis, October 1966.
4. Whitten, W. B., II, & Bjork, R. A. *Are long-term recency effects caused by output interference*? Paper presented at the Meetings of the Psychonomic Society, St. Louis, 1976.
5. Alexy, M., & Gorfein, D. S. *A context interpretation of recency-sensitive retrieval processes in long-term free-recall.* Paper presented at the Annual Meeting of the Midwestern Psychological Association, Chicago, May 1973.
6. Elmes, D. G., & Wood, J. A. *Rehearsal and imagery effects on long-term recency.* Paper presented at the Meetings of the Psychonomic Society, Denver, November 1975.
7. Whitten, W. B., II. *Retrieval "depth" and retrieval component processes: A levels-of-processing interpretation of learning during retrieval* (Tech. Rep. No. 53). Ann Arbor, Michigan: University of Michigan, Human Performance Center, 1974.
8. Chandler, J. P. *Subroutine STEPIT* (Program QCPE 66). Bloomington, Indiana: Indiana University, Quantum Chemistry Program Exchange, 1965.

(Received October 4, 1976)