

Analysis of search times in paired associates learning

D. J. MURRAY

Queen's University, Kingston, Ontario K7L 3N6, Canada

It was tentatively suggested that a long unsuccessful search on Trial n of a paired associate task might be associated with a higher probability of recall on Trial $n + 1$ than would a short search. The reasons for thinking this included the evidence that several test trials can be associated with good recall and that items found with difficulty in a retrieval task were later well recalled in a free recall task. The prediction was not confirmed, however, in either of two studies. Data are presented on the reduction of search time as learning progresses and on the distribution of subjective feeling-of-knowing states during learning. Subjects appeared either to "know" or "not know" the answer and rarely gave responses indicative of intermediate states of knowing.

In paired associates learning, there are two reasons for thinking that a long search, which does not end in success, nevertheless might result in better acquisition of the response term when the subject is told it. First, Gardiner, Craik, and Bleasdale (1973) have shown that, in a free recall task, a word that is difficult to find might be better recalled, later, from long-term memory. In one of their studies, subjects were given 50 definitions of words plus the initial letters of these words and had to find the appropriate words within 60 sec. After all 50 words had been tested, the subjects were given an unexpected recall test of the words. It was found that words that had been discovered only with difficulty had a higher probability of being recalled in the final free recall. In particular, words that had been associated with a tip-of-the-tongue (TOT) experience, as rated subjectively by participants in the experiment, were recalled. One might try to extrapolate from this free recall study to paired associates learning: If a word is preceded by a long and unsuccessful search, might it not then be better retained than a word not preceded by a long search? Second, Izawa (1967) has shown that even though two test trials are associated with the same accuracy of performance, a study trial following two such test trials can be associated with a higher level of acquisition than can a study trial following only one such test trial. In later work (Izawa, 1970), the effect was successfully extended to cover the case of as many as 19 test trials intervening between study trials. Izawa's explanation for

Preparation of this article was supported by Natural Sciences and Engineering Research Council of Canada Grant AO-126 to the first author. We should like to express our gratitude to Anne Chenier and Fatima Kheraj for research assistance. We are also grateful to Keith Norton for helpful advice and to David Dockendorff for assistance with equipment and programming. Experiment 1 was carried out while the author held Social Sciences and Humanities Research Council Leave Fellowship 451-780198.

the advantage of test trials for future performance was that during a test trial, even though no new items are found, a larger repertoire of possible responses is sampled and thus made available. It seemed reasonable to ask, therefore, whether a long search might not be treated as a miniature set of "tests" all collapsed together—Might not a long search lead to an advantage for retention compared with a short search? The prediction, then, from both of the above arguments is that a long unsuccessful search on Trial n should be associated with a higher probability of being correct for the first time on Trial $n + 1$ than should a short search, for any given item. This prediction was examined in two experiments.

EXPERIMENT 1

Method

General. Subjects worked at their own rate through six trials on a 25-item list of paired associates. They were given either 3 or 6 sec to rehearse each response. Response latencies were recorded.

Procedure. Each of 25 stimulus terms taken from Jenkins' (1970) norms was paired with a rare word taken from Part 2 of the Thorndike-Lorge (1944) lists. Examples are AFRAID-MANGROVE and DREAM-FIESTA. Learning was taken to six trials. Subjects were given either 3 or 6 sec to rehearse each response item as it was shown, but they could search for each response as long as they wished. When the subject had given the response verbally, he or she pressed any key on the keyboard of a PET computer. There is a clock in this computer that operates in units of 60ths of 1 sec ("jiffies"): The time elapsing between stimulus onset and keypress could be measured, thus giving an indication of search time.

On each trial, the screen of the computer showed (1) the word READY for 2 sec, then (2) a blank for 2 sec, (3) a randomly selected stimulus word from the list of 25 pairs, which stayed on until the subject had pressed a key on the keyboard, (4) the correct response term, which stayed on for either 3 or 6 sec, depending on the subject's group, and then (5) the next stimulus word. Steps 1-5 were repeated for the 25 words of a trial, and then (6) the word REST appeared for 30 sec. At the end of the six trials, (7) the computer showed, on the screen, the

complete listing, in the order of testing, of the 25 items of each trial, along with the reaction time associated with each (expressed as a decimal rather than in jiffies). These were copied down by the experimenter.

Design and Subjects. This was a one-factor between-subjects design (the factor was rehearsal time). There were 12 subjects in each of the two groups; the subjects were from introductory psychology classes at Queen's University.

Results

As expected from previous research (e.g., Millward, 1964; Postman & Kaplan, 1947), as learning progressed, search times decreased across trials. On the very first trial, most subjects pressed the key immediately to obtain the response term, but thereafter, they searched for the correct response before pressing the key. On the 3-sec rehearsal time condition, the mean search times were: Trial 2, 10.213 sec; Trial 3, 9.218 sec; Trial 4, 7.232 sec; Trial 5, 5.831 sec; Trial 6, 4.994 sec. On the 6-sec rehearsal time conditions, the mean search times (in seconds) were: Trial 2, 7.257; Trial 3, 7.379; Trial 4, 6.139; Trial 5, 4.627; Trial 6, 3.838. Analysis of variance showed that the decrease in search time across trials was highly significant [$F(5,110) = 24.26$, $MSe = 5.217$]. Neither the effect of rehearsal condition nor the interaction Trials by Rehearsal Conditions was significant. As for the number of items correct, the mean number correct out of 25, averaging over trials, was 11.217 for the 3-sec rehearsal condition and 13.4 for the 6-sec rehearsal condition. This difference, however, was not significant, nor was the interaction with trials significant. Increasing rehearsal time had, therefore, only a small positive effect.

The main question, however, was whether longer search times on Trial n led to an increased probability of correct recall on Trial $n + 1$. To establish this, the mean search time of each subject on each trial was first ascertained. Then the items he had searched for were divided into two kinds: L items, which were incorrect items associated with search times longer than the subject's average for all incorrect items for that trial, and S items, which were incorrect items associated with search times shorter than the subject's average for all incorrect items for the trial. Then the probability that the L items would be correct for the first time on the next trial was calculated, as was the probability that the S items would be correct for the first time on the next trial. The general prediction was that this probability would be higher for L items than for S items. These probabilities were calculated separately for Trials 3, 4, 5, and 6 and for the two rehearsal conditions. They are shown in Table 1. A general analysis of variance for the probabilities in question showed that these probabilities tended to increase significantly over trials [$F(3,66) = 6.57$, $MSe = .06$, $p < .001$]. This was to be expected, given that the number of L and S items decreased across trials as more and more responses were correctly given. But there was no indication that L items, which yielded

Table 1
Probability Correct on Each Trial in Experiment 1 as a Function of Whether Search Time for the Equivalent Item on the Previous Trial Was on Average Long (L Items) or Short (S Items)

		Trial 3	Trial 4	Trial 5	Trial 6	Mean
3 Sec	L	.194	.256	.422	.437	.327
	S	.171	.371	.300	.402	.311
6 Sec	L	.303	.256	.581	.304	.361
	S	.260	.321	.467	.340	.347

Note—See text for details.

a mean probability correct on the next trial of .344, were associated with a significantly higher probability correct on the next trial than were S items, which yielded a mean probability correct on the next trial of .329. Nor was the effect of rehearsal condition significant, although there was a tendency, significant at the .05 level, for the probability correct after an L or S item to increase more steeply over trials for the 3-sec rehearsal condition than did that for the 6-sec rehearsal condition [$F(3,66) = 2.77$, $MSe = .06$]. No other interactions were significant.

Discussion

There was no evidence that incorrect items associated with longer than average search times were more likely to be correct for the first time on the next trial. But the original Gardiner et al. (1973) research had indicated that, when difficult retrieval was associated with better later recall, it was associated not only with a longer search time, but also with reports of a TOT state. Furthermore, the use of only one kind of list in the present experiment, a list involving rare response terms, might have not been very conducive to TOT states. In the next experiment, therefore, lists of varying difficulty, which might be expected to yield a wider range of reported TOT states, were used. Ratings of subjective feelings, concerning one's knowledge of the correct responses, were also obtained.

EXPERIMENT 2

Method

General. Subjects learned one of three lists, "easy," "moderate," or "difficult." On each trial, they were allowed to search as long as they wished for the item; when they gave their response, they also gave a confidence report. Search times were recorded.

Procedure. Each subject learned one list for a maximum of six trials: There were three lists of 25 items each. The stimulus terms in each list were the same, namely, those used in Experiment 1 (e.g., AFRAID, DREAM). But the response terms differed among the three lists, being rated fairly high, medium, or low on a scale of 1 to 7 in "associative strength." The mean rated associative strength for the "easy" lists was 3.52 (examples: AFRAID-HAPPY, DREAM-SCENE), for the "moderately difficult" lists it was 2.33 (examples: AFRAID-BREAK, DREAM-GROWING), and for the "difficult" lists it was 1.50 (examples: AFRAID-PURPOSES, DREAM-DROPPED). The mean Kučera-Francis (1967) word frequency of the response terms was approximately constant across the three lists (131.69). When each stimulus term came on the screen, the subject was instructed to search for the item and, when he found it, to give it (or guess at an answer) and also give a confidence rating on a 4-point scale. The written instructions on this matter were as follows:

"To indicate how near you think you are to the answer: give a number from one to four. The lower the number, the harder you find it to get the answer. 'One' means you have little or no idea of the correct answer. 'Two' means you're trying hard but the answer is still elusive. 'Three' means it's on the tip of your tongue. 'Four' means you found the answer easily."

On each trial, (1) the word READY appeared on the computer screen for 2 sec, (2) the screen went blank for 2 sec, (3) a randomly selected stimulus term appeared on the screen until the subject pressed a key on the keyboard of the computer, (4) the response term appeared for 3 sec, and (5) the next stimulus term then appeared. Stages 3 and 5 were repeated for the 25 pairs of a trial, and then (6) the word REST appeared for 30 sec. After six trials, (7) the original pairs, in their order of appearance, were shown on the screen, along with the response times recorded. The experimenter copied these down. During the experimentation itself, the experimenter also noted the response given and the confidence ratings.

Subjects. There were 12 subjects in each of the three groups. They were volunteers from introductory psychology classes at Queen's University.

Results

If we examine the total number of items correct, we find that learning became easier as the lists became easier: The mean number correct per trial on the "difficult" lists was 9.117, on the "moderate" lists it was 14.217, and on the "easy" lists it was 20.35. The effect of list difficulty was significant at the .001 level [$F(2,33) = 23.14$, $MSe = 82.02$], as was the effect of trials [$F(4,132) = 171.3$, $MSe = 5.38$]; the List Difficulty by Trials interaction was significant at the .01 level [$F(8,132) = 2.62$, $MSe = 5.38$]. However, eight subjects got the complete list correct in fewer than six trials on the "easy" condition, with the result that there were many missing data for reaction times on this list. The data available, however, showed that, as performance improved on the "easy" condition, reaction times decreased across the five trials in which searching was involved, as follows: 5.29, 4.14, 3.57, 2.85, and 2.66 sec. Reaction times on the "moderate" lists had a mean of 5.887 sec and fell across the five trials as follows: 7.243, 6.968, 5.553, 4.818, and 4.855 sec. Those on the "hard" lists had a mean of 7.766 sec and fell across the five trials as follows: 8.648, 8.624, 8.268, 6.886, and 6.402 sec. The difference between the two means was significant at the .05 level [$F(1,22) = 4.71$, $MSe = 22.5$]. The falloff over trials was significant at the .001 level [$F(4,88) = 6.06$, $MSe = 4.54$]; the interaction of List Difficulty by Trials was not significant.

But, again, the main question was whether L items (as defined in Experiment 1) would be associated with greater accuracy on the subsequent trial than would S items. Table 2 shows a breakdown of the probabilities that L items would be correct for the first time on the next trial. Treating the "easy" list separately because of the problem of missing data, we see that the L items were associated with a slightly higher probability of being correct on the subsequent trial than were the S items overall, but on no single trial was the difference significant by a *t* test. Applying analysis of variance

Table 2
Probability Correct on Each Trial in Experiment 2 as a Function of Whether Search Time for the Equivalent Item on the Previous Trial Was on Average Long (L Items) or Short (S Items)

		Trial 3	Trial 4	Trial 5	Trial 6	Mean
Easy	L	.622	.685	.512	.750	.642
	S	.683	.722	.636	.500	.635
Moderate	L	.316	.377	.309	.546	.387
	S	.392	.429	.444	.635	.475
Difficult	L	.139	.159	.364	.432	.273
	S	.208	.260	.347	.339	.288

Note—See text for further details.

to the data for the "moderate" and "difficult" groups yielded the following: The probabilities increased over trials as in Experiment 1 [$F(3,62) = 9.437$, $MSe = .048$], and the probabilities were nearly significantly higher on the "moderate" lists than on the "difficult" lists [$F(1,22) = 3.71$, $MSe = .29$, $p = .06$]. The overall difference in the probabilities associated with L items vs. S items had a probability of $.1 > p > .05$ [$F(1,18) = 3.0$, $MSe = .043$]. There were no significant interactions. In fact, the S items were associated with slightly higher probabilities of being correct on the next trial than were the L items, contrary to expectation.

Another aim in Experiment 2 was to determine whether items associated with a report of TOT state on Trial *n* were more likely to be correct on Trial *n* + 1. Figure 1 shows the distribution of reports of confidence over the trials for the three kinds of lists: TOT states were associated with a report of "3." It can be seen that reports were predominantly "1" or "4," that is, subjects only rarely used reports "2" or "3." Moreover, reports of "2" or "3" were not distributed in any consistent manner across trials or condition of list difficulty. Reports of "4," indicating that the answer was found easily, increased over trials, whereas reports of "1" diminished. Nevertheless, an analysis of reports of "3" was undertaken with the aim of determining whether a reported TOT state on an incorrect item was likely to be associated with a higher probability of that item's being correct on the next trial. If there was no such effect, the probability of being correct on the next trial should be .5. For "easy" lists, the associated probabilities for Trials 3, 4, and 5, respectively, were .6, .5, .6 (mean = .56), for "moderate" lists, the probabilities for Trials 3, 4, 5, and 6, respectively, were .3, .66, and .5 (mean = .53), and for "difficult" lists, the probabilities were .4, .33, .5, and .5 (mean = .4325). The overall probability was .5075. It would be wrong, therefore, to assert that a TOT state was associated with a relatively high probability of being correct on the next trial. There was no evidence that L items were associated with a higher probability of being correct on the next trial than were the S items for cases rated "2" alone, or for cases rated "3."

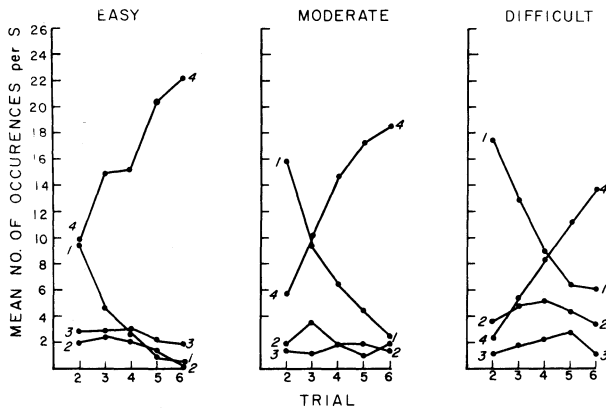


Figure 1. An analysis of the way the four kinds of response rating changed across trials for the three levels of associative strength in Experiment 2. Each curve gives the result for the rating of the number shown: 1 = little or no idea of the correct answer, 2 = trying hard but answer is elusive, 3 = tip of the tongue, 4 = found answer easily.

Discussion

As in Experiment 1, there was no evidence that longer than average search on Trial n increased the likelihood of that item's being correct on Trial $n + 1$; nor was there good evidence that subjects frequently achieved TOT states in the present situation, or that, when these occurred, they were likely to facilitate recall on the next trial. That subjects searched to some extent for difficult items, rather than just quickly responding with the rating "1" and pressing the button to receive the answer, is shown by the following mean search times for incorrect responses: "easy" lists, 6.44 sec; "moderate" lists, 8.79 sec; "difficult" lists, 8.34 sec. (There was, incidentally, no very evident trend for search times on incorrect items to increase or decrease systematically across trials.) In general, both rating data and the analysis of search times supported the idea that a response is either "known" or "not known" in this particular paradigm.

It seems reasonable to ask why the Gardiner et al. (1973) finding that long search facilitated subsequent retrieval was not obtained here. In their study, the subjects knew that they "knew" the correct word that was being cued: It was a matter of finding it. In the early trials of the present study, subjects, it seems, knew that they did not "know" the correct response, at least not well enough to find it. Perhaps responses must have a certain strength or be produced by subjects themselves, before it can be shown that a long search facilitates subsequent retrieval. Similarly, the analogy between a long search and a succession of collapsed test trials breaks down because in the latter instance the subject is reinforced several times, whereas he is reinforced only once in the long search. These remarks may indirectly suffice to show the importance of the timing of the administration of the correct response in paired associates learning.

REFERENCES

- GARDINER, J. M., CRAIK, F. I. M., & BLEASDALE, F. A. Retrieval and subsequent recall. *Memory & Cognition*, 1973, 1, 213-216.
- IZAWA, C. Functions 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*, 1970, 83, 340-344.
- JENKINS, J. J. The 1952 Minnesota word association norms. In L. Postman & G. Keppel (Eds.), *Norms of word association*. New York: Academic Press, 1970.
- KUČERA, H., & FRANCIS, W. N. *Computational analysis of present-day American English*. Providence, R.I: Brown University Press, 1967.
- MILLWARD, R. Latency in a modified paired-associate learning experiment. *Journal of Verbal Learning and Verbal Behavior*, 1964, 3, 309-316.
- POSTMAN, L., & KAPLAN, H. L. Reaction time as a measure of retroactive inhibition. *Journal of Experimental Psychology*, 1947, 37, 136-145.
- THORNDIKE, E. L., & LORGE, I. *The teacher's word book of 30,000 words*. New York: Teachers College, Columbia University, Bureau of Publications, 1944.

(Received for publication November 19, 1980.)