In Press: Psychonomic Bulletin & Review

Spacing and the transition from calculation to retrieval

TIMOTHY C. RICKARD, JONAS LAU, AND HAROLD PASHLER
University of California, San Diego

Many arithmetic problems can be solved in two ways: by a calculation involving several steps, and by direct retrieval of the answer. With practice on particular problems, memory retrieval tends to supplant calculation—an important aspect of skill learning. We asked how the distribution of practice on particular problems affects this kind of learning. In two experiments, subjects repeatedly worked through sets of multi-digit multiplication problems. The size of the trained problem set was varied. The smaller set size (with shorter average time between problem repetitions) showed faster responses and an earlier transition to retrieval during training. However, in a test session presented days later, the pattern reversed, with faster responses and more retrieval for the large set size. Evidently, maximizing the occurrence of direct retrieval within training is not the best way to promote learning to retrieve the answer. Practical implications are discussed.

It has been clear for many years that spacing of explicit learning (distributing a fixed amount of study time for certain materials over a longer total time period) can powerfully enhance the probability that these materials can later be recalled. Less well known are the inconsistent effects of spacing on other kinds of learning, in particular those related to acquisition of skill. In the present study, we examine the effects of temporal spacing on a particular form of skill learning: the performance improvement that occurs as people repeatedly do arithmetic calculations. This form of learning differs from the "standard" studies of spacing in the episodic memory literature in at least two respects. First, the information being recalled is not taught to the learner by the experimenter, but is rather self-produced. Second, the most conspicuous change is in speed of responses, rather than accuracy. For reasons to be discussed shortly, effects of spacing in this situation might or might not parallel those found with episodic memory designs.

Spacing Effects

Evidence that spacing can enhance recall probability goes back to Ebbinghaus (1964/1885), and has proven to be quite robust in a variety of tasks involving verbal episodic recall (see Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006, for a recent review). Various studies have documented the fact that spacing enhances probability of success in cued recall and paired associate tasks with long retention intervals (e.g., Glenberg, 1976; Glenberg & Lehman; Rumelhart, 1967; Pashler, Rohrer, Cepeda, & Carpenter, 2007). Spacing can also make children better able to recall newly taught mathematical facts (Rea & Modigliani, 1985).

On the other hand, when one looks within the broad category of "skill learning" or "implicit memory" (tasks where the response does not typically involve explicit recollection), the beneficial effects of spacing are far less clear. For example, spacing effects do not seem to be robust for perceptual identification and word fragment completion tasks (Greene, 1990; Perruchet, 1989). In our own lab, we did not find substantial or robust spacing effects for tasks involving visuospatial categorization learning (Pashler, et al., 2007).

Transitions from Calculation to Retrieval

A particularly prominent consequence of arithmetic skill learning is a gradual increase in the occurrence of direct memory retrieval--directly recollecting the answer in a single step, rather than having to rely on calculation using an explicit algorithm. There has been debate about whether retrieval occurs simultaneously with calculation on any given trial (as suggested by Logan, 1988; Palmeri, 1997) or rather simply comes to supplant calculation (Rickard, 1997; 2004). There is little doubt, however, that with repeated exposure to a

Correspondence concerning this article should be addressed to Timothy Rickard, Department of Psychology, 0109, University of California at San Diego, 9500 Gilman Drive, La Jolla, California 92093-0109. E-mail: trickard@ucsd.edu

given arithmetic problem, retrieval becomes more frequent.

Present Experiments

The present study poses a fairly straightforward question that bridges the topics of spacing and the algorithm-to-retrieval transition. We ask: how does spacing of training on specific problems affect this transition? Spacing of learning is varied within a session, by manipulating the "set size" of arithmetic problems given during training (i.e., the number of problems performed before those problems are repeated). The greater the set size, the greater the average temporal spacing between successive re-presentations of a given problem. This variable has potent effects on learning of new associations (Pashler, Zarow, & Triplett, 2003).

Several hypotheses naturally present themselves. On one hand, one might expect a spacing effect tradeoff similar to those observed in the verbal recall literature, such that spacing results in a slower rate of performance improvement during training but better performance on the test. Schmidt and Bjork (1992) suggest that this tradeoff is common, and they point to spacing as one prominent example of a variable that produces it.

On the other hand, given the tenuousness of spacing effects in implicit learning tasks generally, one could hypothesize that spacing effects will be weak or absent. The underlying learning system may be different from that engaged by explicit memory tasks and it may be subject to different temporal dynamics. There is a second reason to suppose that spacing might not benefit arithmetic skill learning: one might suppose that to learn the transition from calculation to retrieval, it is best to actually engage in retrieval (an example of the rather reasonable rule: "if you want to learn X, practice doing X"). If long spacing reduces the probability of using retrieval during training, one might expect it to reduce the learning of the retrieval pathway. This account would predict that shorter spacing is associated with faster performance and greater use of retrieval in both training and test sessions.

Experiment 1

Our task required subjects to multiply a single-digit by a two-digit number (e.g., 6 X 18), a task which few adults will be able to solve using the retrieval strategy prior to training. In Experiment 1, there were two sessions. The first was a training session, in which some multiplication problems were presented with short inter-item spacing and others were presented with a long inter-item spacing. In the test session, all problems were presented in a random order.

Method

Subjects. Thirty-nine subjects from the University of California, San Diego participated for course credit. Ten subjects did not complete the experiment, leaving data from 29.

Stimulus. The experiment involved a total of twenty-four multiplication problems (the problems are listed in the Appendix). Each problem required multiplying a two-digit number by a one-digit number.

Design. Session 1 involved training, and Session 2 involved a test. Every subject was taught all 24 problems within Session 1, receiving 15 exposures to each problem. For every subject, twelve of the problems were practiced in what will be termed the *Set Size Twelve* condition, while the other twelve were practiced in the *Set Size Three* condition.

When a set of problems was taught in the Set Size Twelve condition, the computer simply presented all 12 problems in a random order, then presented the same 12 problems in a new random order, and so forth, until all 12 problems had been presented 15 times.

The 12 problems taught to a given subject in the Set Size Three condition were split into four groups of three (randomly and individually for each subject). Each group of three was practiced 15 times without any other items intervening. The computer presented all three items from a group in a random order, then presented the same group of three in a new random order, and so forth until the group of three problems had been presented 15 times, for a total of 45 presentations. Then the computer moved on to the next group of 3 items, and so forth until all 12 items had been presented 15 times. In both conditions, the constraint was enforced that the same problem could never appear twice in succession (an event that might otherwise have occurred at the boundary between successive presentations of a set).

To insure that Set Size was not confounded with item difficulty or position within the training period, subjects were randomly assigned to one of four counterbalancing conditions. These conditions determined which of two halves of the problem list were assigned to Set Size Three vs. Set Size Twelve (problem groups A and B in the

Appendix), and also determined whether the training on Set Size Three came before or after the training on Set Size Twelve.

Procedure. Each subject was run individually in a moderately illuminated soundproof room. Subjects were told that they would be solving multiplication problems in their head, without any help from pen and paper. More specifically, they were instructed to do these problems in a standard way, i.e., by multiplying the single-digit number with the tens place of the double-digit number, then multiplying the single digit number by the ones place of the double-digit number, and then adding the two products to arrive at the final answer. Subjects were asked to say the answer aloud as soon as they thought they knew it. After the computer picked up the voice, the answer to the problem popped up on the screen. The experimenter pressed one of three buttons to indicate that the response was correct or incorrect, or to indicate that there had been a malfunction (e.g., the voice key tripping off of a subject's cough or throat-clearing, etc.) During both the training and test sessions, if the subject's response was wrong, the correct answer was presented for 1 second, after a delay of 1 second. The next trial then commenced after a further 1-second delay. There was a one-minute pause between the first and the second half of the task.

The second session occurred seven days after the first session. The procedure was as follows: the subject was presented with all 24 problems in a random order; then the same 24 problems in a new random order; and so on for eight runs through the list. Thus, the session consisted of 192 problems, presented without rest breaks, half of which had previously been trained in Set Size Three, and half of which had been trained in Set Size Twelve.

Results and Discussion

Figure 1 shows the mean reaction times (RTs) for correct trials as a function of condition, session, and block number, where a block is a sequence of one presentation of each item in the set. As expected, the figure shows a steady decrease in RT over training. The decrease in RT was substantially greater, however, for Set Size Three. This pattern was reversed on the test, where Set Size Twelve shows substantially enhanced performance compared to Set Size Three. This cross-over interaction (Set Size Three being faster in training, slower on test) was confirmed by a within-subjects analysis of variance (ANOVA) with a 2 (condition) x 2 (session) factorial design, $F(1,28)=70.6$; $p < .0001$.

Error results were analogous. In the training session, mean error rates were .055 and .096 for Set Size Three and Set Size Twelve, respectively. In the test session, the pattern reversed: the error rate for Set Size Three was .081, whereas that for Set Size Twelve was .064.

The results are clearly in line with the view suggested by Schmidt and Bjork (1992). Greater set sizes (greater spacing) reduce the rate of performance improvement during training but improve performance on the delayed test.

Experiment 2

Although the results of Experiment 1 provide a fine example of the generalization suggested by Schmidt and Bjork (1992), they do not provide any clear information on the way that spacing may have modulated the transition from calculation to retrieval. Drawing on the prior literature (e.g., Rickard 2004) we hypothesize that the patterns in Experiment 1 reflect, to a large extent, different patterns of shift to retrieval for the two conditions. For the Set Size Three condition, the shift to retrieval may have happened relatively quickly during practice. It appears, however, that the shift did not reflect stable long-term learning, and therefore that subjects were often forced to revert to use of the slower algorithm strategy during the test. In the Set Size Twelve condition, the reverse appears to have happened. The transition to retrieval may have occurred for a smaller percentage of problems during training in that condition, but for the shifts that did occur there may have been more stable long-term learning. Hence, on the test, a higher percentage of retrievals occurred for problems in that condition. Experiment 2 is designed to test this account of the cross-over interaction by using strategy probing.

Method

Subjects. A total of 22 subjects participated in a 3 session experiment. Of these, 21 were paid to participate, while one subject participated in two sessions in return for course credits and was paid for the last session.

Materials, Design, & Procedure. These aspects of Experiment 2 were identical to those of Experiment 1, except as noted here. There was a two-day interval between training sessions 1 and 2. On every trial of the last five blocks of Session 2, the subject was asked to indicate whether he or she had arrived at their answer by calculating, retrieving from memory, or using other means. The same strategy probing was also done on every trial of the test session. To make their strategy choice, subjects pressed one of three buttons on the button box. The exact wording used was as follows: *How did you arrive at your answer? Please press C for "Calculation", D for "Direct Retrieval" or O for "Other".*

Results and Discussion

Figure 2 shows the mean RTs for correct trials as a function of condition, session and block number. The results for session 1 mirror those of Experiment 1. On the first block of session 2 there was a temporary reversal, such that subjects responded significantly faster in the Set Size Twelve condition than in the Set Size Three condition (3632 msec vs. 2989 msec), $t(21) = 3.78$, $p < .01$. Set Size Three RTs were faster throughout the remainder of session 2. Throughout the test session, subjects performed better on the problems trained in the Set Size Twelve condition, just as in Experiment 1.

The error rates mirrored the RTs. For Set Size Three problems, errors rates were .057 and .020 for the first and second sessions, respectively, and .076 on the test. For Set Size Twelve these same values were .095, .050, and .069.

The strategy probing results for the last five blocks of the session 2 and for the test session are shown in Figure 3. In session 2, subjects were generally relying on direct retrieval in the Set Size Three condition, but were doing so only about half the time in the Set Size Twelve condition. This pattern reversed in the test session, with direct retrieval being reported more frequently for the Set Size Twelve problems.

To explore the possibility that the superior performance on the test in the Set Size Twelve condition was driven primarily or solely by the increased rate of retrieval in that condition, we computed mean RTs on the test for each condition and separately by strategy report ("algorithm" or "retrieval"; the relatively small number of "other" reports were excluded). Five subjects who did not report using both strategies in both conditions were excluded from this analysis. The overall analysis for this subset of subjects (collapsing over strategy) confirmed the advantage for Set Size Twelve that was reported in the analyses of the full set of subjects (means of 2384 msec and 2772 msec for Set Size Twelve and Set Size Three, respectively), $t(1, 16) = 3.27$, $p < .01$.

RTs as a function of strategy and set size are shown in Figure 4. A 2 (strategy) by 2 (condition) within subjects ANOVA confirmed the strong affect of strategy, $F(1, 16) = 51.8$ $p < .001$, but there was no longer a significant effect of condition, $F(1, 16) = 2.06$ $p = .17$, and there was no strategy by condition interaction, $F(1, 16) = 1.42$ $p = .25$. These results indicate that the condition difference in the overall test analysis was driven primarily by the increased use of the retrieval strategy in Set Size Twelve.

Although the strategy reports discussed above are of course correlated with RT, there are several factors that support the idea that they provide a generally valid index of actual strategy use. First, on transfer tests, subjects revert back to reporting algorithm usage for new problems while continuing to report retrieval for old (previously practiced) problems (Rickard, 1997), showing that subjects do not simply report more retrieval use over the course of the practice as a habit or to satisfy perceived demand characteristics (see also Rickard, 2004). Second, arithmetic algorithms are believed to involve sub-vocal intermediate steps whereas direct retrieval does not. The distinction between algorithm and retrieval is thus an exemplary case in which subjects are expected to have access to memories of their mental processes that are diagnostic of strategy use (Ericsson & Simon, 1993). Third, a result from the current experiments supports the validity of the strategy probing. For session 1 of Experiment 2, the mean RTs on the first training block, which reflect use of the algorithm strategy, were 4600 msec for the first set of problems trained in the Set Size Three condition, and 4299, 4102, and 4195 msec, for the second, third, and fourth sets trained, respectively. These results suggest some general improvement in algorithm efficiency between the first and third problem sets but not thereafter. Supporting validity of the algorithm reports, the means for the third and fourth sets are in the same range as the 3851 msec mean for the Set Size Three problems on the test when the algorithm was reported.

General Discussion

In the Introduction, it was pointed out that while spacing effects are ubiquitous in studies examining recall accuracy for newly acquired associations, it is not so clear that spacing has a beneficial effect of the tuning that goes on during skill learning, which may rely more upon “implicit memory” for which spacing effects seem not so robust (Greene, 1990; Perruchet, 1989). To explore this issue, we investigated whether greater spacing of arithmetic problems would promote—or impede—the speedup that comes with repetition training. The results showed that *during training* greater spacing resulted in slower response latencies, less accuracy, and a much smaller likelihood of switching from calculation to memory retrieval. However, in the test sessions (in which time gaps between reoccurrences of any given problem were equated across condition) as well as on the first block of the second training session in Experiment 2, these effects were reversed. These results show that arranging training in a way that maximizes the frequency of direct retrieval within a session (or that simply optimizes performance) is probably not the best way to arrange practice.

The results fit well with the general observation of Schmidt and Bjork (1992) that manipulations facilitating performance during training will often reduce the degree or quality of learning. At a practical level, the results imply that spacing should probably be incorporated into drilling on arithmetic facts, even though this will produce less apparent fluency during training.

Like many (but not all) manipulations of spacing, the current Set Size manipulation may have affected not only the time elapsed since previous encounters with a given problem, but also the likelihood that a problem was still represented in working memory. It may be the case, as some models of spacing have suggested (Young, 1966), that associations are strengthened more when the answer is retrieved from long-term memory, rather than working memory. In principle, it would be possible to manipulate both timing and the number of intervening problems separately, and thus determine which of these variables is most responsible for the effects observed here.

Another well-known account of spacing attributes the effect to changes in the context that is present at the time of encoding. As spacing between repetitions is increased, there is more time for the encoding context to have drifted, resulting in a greater expected difference between contexts. On certain assumptions, this could make it more likely that the context at test is similar to the context present during at least one of the encoding events (Glenberg, 1979; Howard & Kahana, 2002; Whitten & Bjork, 1977). This account of the present results would seem to be a potentially viable. However, encoding variability models have difficulty accounting for certain findings in the literature (e.g., Ross & Landauer, 1978).

The present results make it clear that robust spacing effects can occur in skill learning situations in which latency is the critical variable. Thus, the boundary between the many situations in which spacing effects are found, and those cases in which they are not (several of which were described in the Introduction to this article), is one that needs to be charted in future research. Characterizing this boundary should be important for translational applications of learning science, and may also assist in better understanding the distinction between different underlying memory systems.

Author Note

This work was supported by the Institute of Education Sciences (US Department of Education, Grants R305H040108 and R305H040108 to H. Pashler) and National Science Foundation (Grant BCS-0720375; H. Pashler, PI; and Grant #SBE-0542013 to G. W. Cottrell, PI). David Cun provided expert programming assistance in this project.

References

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J., and Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, 132, 354-380.

Dempster, F. N. (1988) The spacing effect: A case study in the failure to apply the results of psychological research. *American Psychologist*, 43, 627-634.

Ebbinghaus, H. (1964/1885). *Memory: A contribution to experimental psychology*. New York: Dover (Originally published, 1885).

Ericsson, K. A., & Simon, Herbert A. (1993). *Protocol Analysis: Verbal Reports as Data*. MIT Press, Cambridge, MA.

Glenberg, A. M. (1976). Monotonic and nonmonotonic lag effects in paired-associate and recognition memory paradigms. *Journal of Verbal Learning and Verbal Behavior,* 15, 1-16

Glenberg, A. M. (1979). Component-levels theory of the effects of spacing of repetitions on recall and recognition. *Memory and Cognition*, *7*, 95-112.

Glenberg, A. M., and Lehman, T. S. (1980). Spacing repetitions over 1 week. *Memory & Cognition*, 8, 528-538.

Greene, R. L. (1990). Spacing effects on implicit memory tests. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 161004-1011.

Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology, 46*, 269-299.

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review*, 95, 492-527.

Palmeri, T. J. (1997). Exemplar similarity and the development of automaticity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 23, 324-354.

Pashler, H., Rohrer, D., Cepeda, N., & Carpenter, S. (2007). Enhancing learning and retarding forgetting: Choices and consequences. *Psychonomic Bulletin & Review.*

Pashler, H., Zarow, G., and Triplett, B. (2003). Is temporal spacing of tests helpful even when it inflates error rates? *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 29, 1051-1057.

Perruchet, P. (1989). The effect of spaced practice on explicit and implicit memory. *British Journal of Psychology*. 80, 113-130.

Rea, C. P., and Modigliani. V. (1985). The effect of expanded versus massed practice on the retention of multiplication facts and spellings lists. *Human Learning: Journal of Practical Research and Applications*, 4, 11-18.

Rickard, T. C. (1997). Bending the power law: A CMPL theory of strategy shifts and the automatization of cognitive skills. *Journal of Experimental Psychology: General*, 126(3), 288-311.

Rickard, T. C. (2004). Strategy execution in cognitive skill learning: an item-level test of candidate models *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 29(3), 1381-1386.

Ross, B. H., & Landauer, T. K. (1978). Memory for at least one of two items: Test and failure of several theories of spacing effects. *Journal of Verbal Learning and Verbal Behavior*, *17*, 669-680.

Rumelhart, D. E. (1967). The effects of interpresentation intervals on performance in a continuous paired-associate task. Technical Report

116, Institute for Mathematical Studies in the Social Sciences, Stanford University.

Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science*, 3, 207-217.

Young, J. L. (1966). Effects of intervals between reinforcements and test trials in paired-associate learning (Doctoral dissertation, Stanford University, 1966). *Dissertation Abstracts International*, *27*, 3699.

Whitten, W. B., II, & Bjork, R. A. (1977). Learning from tests: Effects of spacing. *Journal of Verbal Learning and Verbal Behavior*, *16*, 465-478.

Figure Captions

Figure 1. Mean reaction time (RT) as a function of set size, session, and block for Experiment 1. The vertical dashed line separates the training and test sessions. Error bars are standard errors corresponding to matched t-tests performed separately for each block.

Figure 2. Mean reaction time (RT) as a function of set size, session, and block for Experiment 2. The vertical dashed lines separate the training and test sessions. Error bars are standard errors corresponding to matched t-tests performed separately for each block.

Figure 3. Proportion of direct retrieval reports as a function of set size, session, and block in Experiment 2. These data are from the last five blocks of the second training session and from the entire test session. The vertical dashed line separates the training and test sessions.

Figure 4. Mean response time (RT) as a function of set size and strategy in session 3 of Experiment 2.

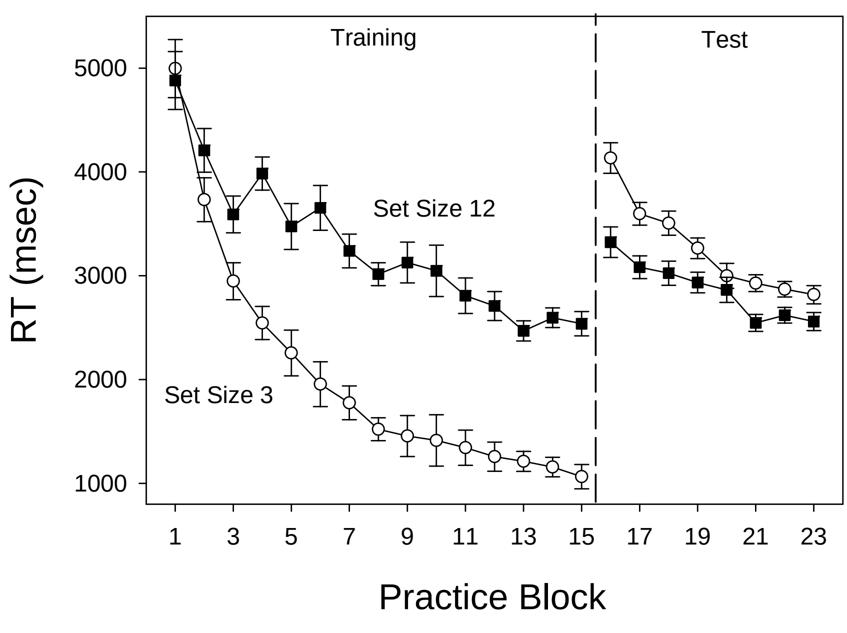

Figure 1.

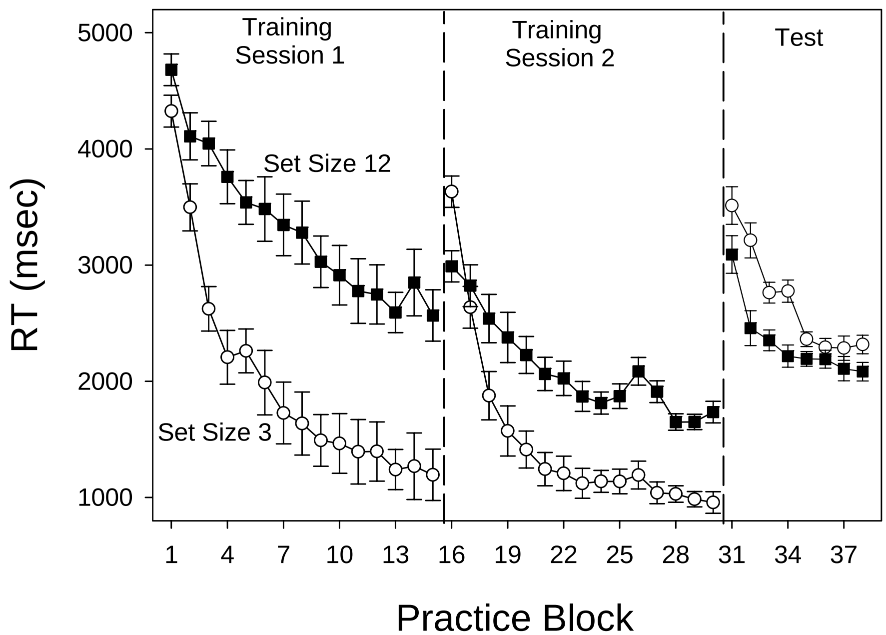

Figure 2.

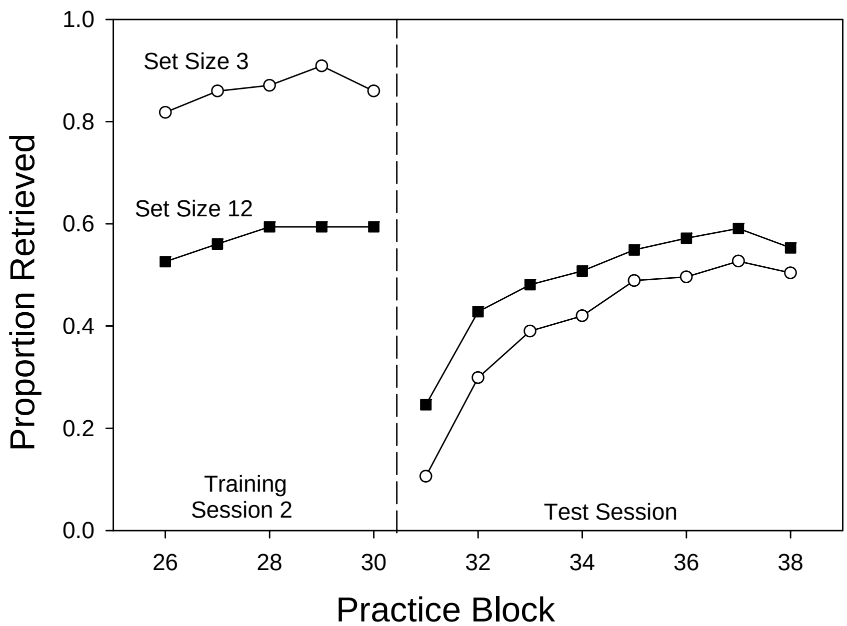

Figure 3.

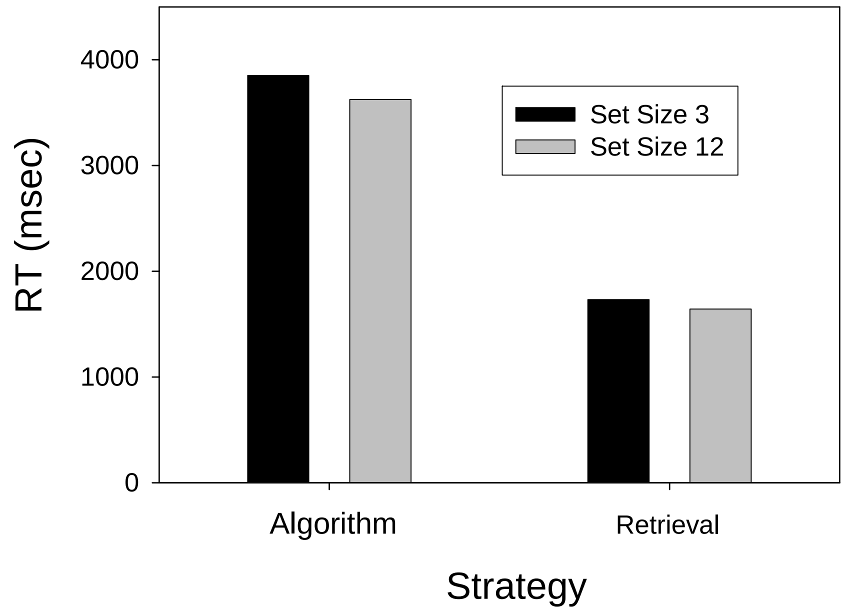

Figure 4.

Appendix

Problem Group A.	Problem Group B.
2 X 19	2 X 26
2 X 23	3 X 14
3 X 22	3 X 27
4 X 13	4 X 17
4 X 24	5 X 19
5 X 16	5 X 21
6 X 18	6 X 24
6 X 27	7 X 18
7 X 26	7 X 23
8 X 14	8 X 16
8 X 21	9 X 13
9 X 17	9 X 22