Optimizing the Long-Term Retention of Skills: Structural and Analytic Approaches to Skill Maintenance

Alice F. Healy
University of Colorado

for

Contracting Officer's Representative
George Lawton

Research and Advanced Concepts Office
Michael Drillings, Acting Director

January 1995

United States Army
Research Institute for the Behavioral and Social Sciences

Approved for public release; distribution is unlimited.
# REPORT DOCUMENTATION PAGE

## 4. TITLE AND SUBTITLE
Optimizing the Long-Term Retention of Skills: Structural and Analytic Approaches to Skill Maintenance

## 6. AUTHOR(S)
Healy, Alice F. (University of Colorado)

## 7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Colorado at Boulder
Campus Box B-19
Boulder, CO 80309

## 9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Research Institute for the Behavioral and Social Sciences
ATTN: PERI-9R
5001 Eisenhower Avenue
Alexandria, VA 22333-5600

## 11. SUPPLEMENTARY NOTES
Contracting Officer's Representative, George Lawton.

## 13. ABSTRACT (Maximum 200 words)
Progress has been made in determining guidelines for optimizing the long-term retention of skills. Studies on learning and retention of color-word interference, schedule components, list components, mental arithmetic, and vocabulary acquisition suggest that optimal retention will result from using procedures during training, relating information to previous experience, making the information distinctive, promoting direct retrieval of the information, and providing refresher or practice tests.

## 14. SUBJECT TERMS
- Long-term retention
- Procedural reinstatement
- Structural approach
- Skill maintenance
- Analytic approach

## 15. NUMBER OF PAGES
46

## 16. PRICE CODE
--

## 17. SECURITY CLASSIFICATION OF REPORT
Unclassified

## 18. SECURITY CLASSIFICATION OF THIS PAGE
Unclassified

## 19. SECURITY CLASSIFICATION OF ABSTRACT
Unclassified

## 20. LIMITATION OF ABSTRACT
Unlimited

NSN 7540-01-280-5500

Form Approved
OBS No 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.
Research accomplished under contract for the Department of the Army

University of Colorado

Technical review by
Peter J. Legree

NOTICES

DISTRIBUTION: This report has been cleared for release to the Defense Technical Information Center (DTIC) to comply with regulatory requirements. It has been given no primary distribution other than to DTIC and will be available only through DTIC or the National Technical Information Service (NTIS).

FINAL DISPOSITION: This report may be destroyed when it is no longer needed. Please do not return it to the U.S. Army Research Institute for the Behavioral and Social Sciences.

NOTE: The views, opinions, and findings in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other authorized documents.
FY90-93, OPTIMIZING THE LONG-TERM RETENTION OF SKILLS: STRUCTURAL AND ANALYTIC APPROACHES TO SKILL MAINTENANCE

EXECUTIVE SUMMARY

Requirement:

The purpose of this research program was to identify the characteristics of knowledge and skill which are most resistant to decay due to disuse. The general goal was to elucidate principles which specify those aspects of a complex skill that resist decay over periods of disuse and how they are distinguishable from more fragile components.

Procedure:

Four features of our program made it unique: First, our assumption was that it is more crucial to optimize performance after a delay interval than to optimize performance during acquisition. Second, relative to most other empirical programs, we used long retention intervals, up to several years. Third, we chose to conduct experiments investigating a wide range of different skills and paradigms. Fourth, we often used nontraditional methods to measure retention.

Findings:

Our research led to the identification and support of several general principles about improving long-term retention or durability of skills. We summarize below three classes of principles and illustrate them with some of our experimental investigations:

The first class of guidelines concerns ways to optimize retention through conditions of training. We discuss three general guidelines in this class. First, superior memory results from the use of cognitive procedures during learning. The procedural reinstatement framework is used to account for the observed superiority of memory for spatial order found in our studies of the retention of scheduling information. Second, retention is aided by prior familiarity. Memory for spatial information of schedules was improved when the information could be related to relevant previous experience. Third, learning is facilitated by distinctiveness of the information, as was evident with the spatial information in our study of list learning.

The second class of guidelines concerns ways to optimize the learning strategies used. We found in our study of mental arithmetic that the strategy used by the subject importantly influences retention and that a direct retrieval strategy leads to faster responding than does a strategy based on the
use of an algorithm. Our study of vocabulary acquisition demonstrated that a direct retrieval strategy may also be achieved in that domain, but mediating associations exert an influence even when retrieval appears to be direct.

The last class of guidelines concerns ways to optimize memory through conditions of retention testing. In our study of vocabulary acquisition we saw remarkable recovery of retrieval speed after a single initial retrieval. Hence, the use of a refresher or practice test before the critical test can have a profound impact on retention performance.

Some of our work also demonstrated the specificity of improvement in performance. Training on specific colors showed limited transfer to new colors in the Stroop color-word interference task. Although our goal in this research program was limited to an examination of the optimization of long-term retention, we learned that optimizing retention does not guarantee generalizability. Hence, future research should be aimed at exploring conditions of training, strategy utilization, and retention testing that simultaneously maximize both generalizability and long-term durability.

Utilization of Findings:

The overriding practical question of this research was how to ensure, through training, that a skilled worker (such as a code recipient, a tank gunner, or an aircraft pilot) has a behavioral tool kit which is just as or nearly as permanently functionable as his or her hardware kit. Our findings should enable those in the military to make relevant recommendations about training routines for long-term maintenance of military knowledge and skill.
OPTIMIZING THE LONG-TERM RETENTION OF SKILLS

Contents

Introduction ......................................................... 1
Specificity Of Training: Color-Word Interference ............... 2
Guidelines for Improving Long-Term Retention ................. 6
  Retention of Components of Schedules ....................... 6
  Retention of Components of Lists ............................ 13
  Mental Arithmetic ............................................ 16
  Direct and Mediated Retrieval in Vocabulary Acquisition .. 21
Summary and Conclusions ........................................ 25
Acknowledgments ..................................................... 27
References .......................................................... 28
Bibliography .......................................................... 30

List of Tables

1. Results of experiment by Crutcher (1992). Mean correct retrieval time in milliseconds for items trained in the Vocabulary task (V-trained) and items trained in the English subtask (E-trained) as a function of test type (Vocabulary, V, or English, E) and test time. (All means were calculated based on log reaction times and then transformed back to milliseconds by the anti-log function.) ..................... 23

2. Classes of guidelines to optimize long-term retention .... 26

List of Figures

1. Results of the experiment by Clawson, King, Healy, and Ericsson (1993) for the four Stroop interference tests. Mean correct log reaction time on the pretest, posttest, and retention tests as a function of test type. ...................... 3

2. Results of the experiment by Clawson, King, Healy, and Ericsson (1993) for only the Stroop-trained subjects on the four Stroop interference tests, after training. Mean correct log reaction time on the trained and untrained color-word sets as a function of test type. ......................... 4

3. Results of the experiment by Clawson, King, Healy, and Ericsson (1993) for the orthographic Stroop tests. Mean
correct log reaction time on the trained and untrained color-word sets as a function of test time and training condition. ................................................. 4

4. Results of the experiment by Clawson, King, Healy, and Ericsson (1993) for the orthographic Stroop tests, after training. Mean correct log reaction time on the trained and untrained color-word sets as a function of training condition. ................................................. 5

5. Results of Experiment 1 by Wittman (1989). Mean proportion of correct recall of what, who, where, and when information as a function of test time. ................................................. 7

6. Results of Experiment 2 by Wittman (1989) for the map test and the class listing test. Mean proportion of correct recall of what, who, where, and when information as a function of test time. ................................................. 8

7. Results of the experiment by King (1992) for the recall questionnaire test. Mean proportion of correct recall of what, who, when, and where information as a function of test time. ................................................. 9

8. Results of the experiment by King (1992) for the recall questionnaire test. Mean proportion of correct recall of what, who, when, and where information as a function of familiarity of campus. ................................................. 10

9. Results of the experiment by King (1992) for the class listing test. Mean proportion of correct recall of what, who, when, and where information as a function of test time. ................................................. 11

10. Results of the experiment by King (1992) for the class listing test. Mean proportion of correct recall of what, who, when, and where information as a function of familiarity of campus. ................................................. 11

11. Results of the experiment by King (1992) for the map test. Mean proportion of correct recall of what, who, when, and where information as a function of test time. ................................................. 12

12. Results of Experiment 1 by Sinclair, Healy, and Bourne (1993). Mean number of trials to criterion for temporal, spatial, and item information as a function of session. ................................................. 14

13. Results of Experiment 2 by Sinclair, Healy, and Bourne (1993). Mean number of trials to criterion for temporal, spatial, and item information as a function of session. ................................................. 15

14. Results of an experiment by Rickard (1993). Mean proportion of algorithm, retrieval, and other strategy reports as a function of practice block. ................................................. 18
15. Results of an experiment by Rickard (1993). Mean correct log reaction time as a function of log practice block. ....... 18

16. Results of an experiment by Rickard (1993). Mean correct log reaction time for algorithm trials, retrieval trials, and overall as a function of log practice block. ............ 19

17. Results of an experiment by Rickard (1993). Mean correct reaction time in milliseconds for new problems, type-change problems, and no-change problems as a function of test time. (All means were calculated based on log reaction times and then transformed back to milliseconds by the anti-log function.) ........................................... 20

18. Results of an experiment by Rickard (1993). Mean correct reaction time in milliseconds for the three problem types and separately for algorithm and retrieval trials of the no-change problems as a function of test time. (All means were calculated based on log reaction times and then transformed back to milliseconds by the anti-log function.) ........... 21

19. Results of Experiment 1 by Crutcher, Hammerle, and Ericsson. Mean correct retrieval time in milliseconds by Spanish experts as a function of test block, compared with the average retrieval time of extensively trained novices. (All means were calculated based on log reaction times and then transformed back to milliseconds by the anti-log function.) ........................................ 24

20. Results of Experiment 2 by Crutcher, Hammerle, and Ericsson. Mean correct retrieval time in milliseconds by Spanish experts for practiced and new words as a function of test block on a one-week retention test. (All means were calculated based on log reaction times and then transformed back to milliseconds by the anti-log function.) ........... 25
Our research program is aimed generally at understanding and improving the long-term retention of knowledge and skills. Our initial work led us to propose that a crucial determinant of memory concerns the extent to which cognitive procedures acquired during study can be reinstated at test (Healy, Fendrich, Crutcher, Wittman, Gesi, Ericsson, & Bourne, 1992). That is, to demonstrate durable retention across a long delay interval, it is critical that the cognitive procedures used when acquiring the knowledge or skill are reinstated at the later time. Using this work as a foundation, we have tried to develop additional guidelines for promoting superior long-term retention. In a chapter by Healy, Clawson, McNamara, Marmie, Schneider, Rickard, Crutcher, King, Ericsson, and Bourne (1993), we described how the approach we have taken differs from that used in most earlier studies.

Four features of our program are especially important in distinguishing it from earlier research. First, we have been explicitly concerned with optimizing performance after a delay interval rather than assuming retention will be superior given optimized performance during acquisition. Second, relative to most other experimenters, we have used longer retention intervals, usually including tests after at least a week, and in some cases including intervals up to one or two years. Third, we have conducted experiments over a wide range of different types of tasks, because we assumed that our theoretical conclusions may rely heavily on the specific nature of the tasks we studied and in order to capitalize on different processes crucial to memory that could be highlighted in different tasks. Fourth, in many of our studies, we have used nontraditional methods to assess retention, providing training for subjects beyond a fixed accuracy criterion, monitoring component response time measures, or collecting verbal protocols.

This research has led to the support or identification of several guidelines for improving memory for skills (Healy et al., 1993). Here we will focus on three classes of guidelines: those that relate to optimizing the conditions of training, the learning strategy used, and the retention conditions.

In our earlier studies, we were impressed with the remarkable degree of long-term retention that subjects were able to achieve in a number of perceptual, cognitive, and motor tasks, including target detection (Healy, Fendrich, & Proctor, 1990), data entry (Fendrich, Healy, & Bourne, 1991), and mental arithmetic (Fendrich, Healy, & Bourne, 1993; Healy et al., 1992). Our more recent research has identified one important limit of this durable retention phenomenon, namely the specificity, or lack of generalizability of the attained improvements in performance (Rickard & Bourne, 1992; Rickard, Healy, & Bourne, in press). First, we will present some new evidence for such specificity, and then we will discuss our optimization guidelines.
Specificity Of Training: Color-Word Interference

Our new evidence for specificity involves the Stroop effect (Stroop, 1935). In the Stroop color-word interference task, subjects are asked to name the color of the ink in which color words are displayed. The ink color and word do not correspond. For example, given the word purple printed in red ink, the appropriate response is "red." Our study (which is discussed in greater detail by Clawson, King, Healy, & Ericsson, 1993) involved training in two different color-naming situations: The patches training condition involved practice in simply naming color patches; the Stroop training condition involved practice in naming the colors of incongruent color words.

Specifically, the study provided four subjects with 12 sessions of training either on the Stroop task itself or on simple color-patch naming; two subjects in a control condition received no training. All six subjects were tested in a pretest prior to training as well as in a posttest after training and in a retention test after a month-long delay. Included in each test session was a set of four tests related to Stroop interference: one test each on word naming and on simple color-patch naming plus a Stroop test and a test with Stroop stimuli but requiring word-naming responses (which we call "reverse Stroop"). Two additional orthographic manipulation tests consisted of a Stroop test in which the letters of the color words were bracketed by asterisks (e.g., *p*u*r*p*l*e* in red letters) and one in which the letters were all uppercase (e.g., PURPLE in red letters) as opposed to the lowercase letters used in training. These orthographic manipulations provided an indication of specificity to the word form. Another index of specificity was provided by the use of two different color-word sets. Although the experimental subjects trained on only one color-word set (with the set counterbalanced across subjects), all subjects were tested on both sets.

If there was specificity of training, then there should have been less improvement on the orthographic manipulations than on the normal Stroop test and less improvement on the untrained color-word set than on the trained set. On the basis of our previous studies showing extremely good retention of procedural skills (Healy et al., 1992), we also predicted relatively little evidence of forgetting across the one-month delay interval. Of greatest interest was whether any specificity effects persisted across this long retention interval.

The results, averaged across the three training conditions, are summarized in Figure 1 in terms of log correct reaction times at the pretest, posttest, and retention test for the four Stroop interference tests. As in previous research, reaction times were faster for the test types involving word naming than those involving color naming and were slower for the test types involving incongruous stimuli than for those that did not. Note that the test types in order of fastest to slowest were: word naming, reverse Stroop, color-patch naming, and Stroop. The effects of training are evident by the fact that subjects were faster overall on the posttest than on the pretest. Note that, as in the previous
studies of procedural skills (Healy et al., 1992), there was no significant forgetting evident from the posttest to the retention test.

Figure 1. Results of the experiment by Clawson, King, Healy, and Ericsson (1993) for the four Stroop interference tests. Mean correct log reaction time on the pretest, posttest, and retention tests as a function of test type.

The specificity of training is reflected by three related observations. First, reaction times decreased from the pretest to the posttest more for the color-word set on which subjects had trained (pretest M = 2.765, posttest M = 2.698) than for their untrained color-word set (pretest M = 2.744, posttest M = 2.698). Second, as shown in Figure 2 which presents data only from the Stroop-trained condition after training (i.e., on the posttest and retention test), subjects who were trained to name colors and ignore the words were faster on the trained set than on the untrained set when naming colors, that is in the Stroop test and in the color-patch naming test, but not when naming words, that is in the word naming test and in the reverse Stroop test. For these last two tests, reaction times were actually faster for the untrained set. Third, this advantage for the trained set on color naming responses and the advantage for the untrained set on word naming responses were only found for subjects in the Stroop training condition, not for subjects in either the color-patch training or control conditions.
Figure 2. Results of the experiment by Clawson, King, Healy, and Ericsson (1993) for only the Stroop-trained subjects on the four Stroop interference tests, after training. Mean correct log reaction time on the trained and untrained color-word sets as a function of test type.

Figure 3. Results of the experiment by Clawson, King, Healy, and Ericsson (1993) for the orthographic Stroop tests. Mean correct log reaction time on the trained and untrained color-word sets as a function of test time and training condition.
The results of the orthographic manipulation tests revealed no effect of orthographic test type with reaction times nearly identical for the three orthographic test types (standard $M = 2.807$, asterisks $M = 2.810$, uppercase $M = 2.808$). Importantly, this finding suggests that the effects of training were not specific to the word form. In contrast, specificity of color-word set was again evident by two observations: First, as shown in Figure 3, the greatest decrease in reaction time from the pretest to the posttest occurred for the Stroop training condition with the trained color-word set. Second, as illustrated in Figure 4, after training (i.e. on the posttest and retention test), only the Stroop training condition yielded faster reaction times for the trained set than for the untrained set. Because the same pattern was found for all three orthographic test types, the results are consistent with the hypothesis that training is specific to the colors and words employed but not to the orthographic form of the color words.

![Graph showing reaction time (log msec) for different training conditions](image)

Figure 4. Results of the experiment by Clawson, King, Healy, and Ericsson (1993) for the orthographic Stroop tests, after training. Mean correct log reaction time on the trained and untrained color-word sets as a function of training condition.

In summary, we found clear evidence for lasting specificity of training effects in the Stroop task, suggesting that there are limits to the generalizability even of well- retained improvements in performance.
Guidelines for Improving Long-Term Retention

We turn now to our research on the general optimization guidelines outlined earlier, starting with the class of guidelines regarding optimization of the conditions of training.

Retention of Components of Schedules

Researchers have described what we retain in memory as a composite of three qualitatively separate components based on spatial, temporal, and item information (e.g. Healy, 1974, 1975, 1982; Healy, Cunningham, Gesi, Till, & Bourne, 1991; Lee & Estes, 1981). More specifically, the spatial component involves knowledge about spatial relations, distances, and locations of physical objects in addition to knowledge about how to proceed through the environment. The temporal component includes knowledge of dates and times and the relative order of events. Memory for item information includes knowledge of specific facts and names.

Research has provided evidence that spatial, temporal and item information are retrieved in different ways (King, 1992; Wittman, 1989). For example, Wittman (1989) tested undergraduate students' recall of four different types of course schedule information: course times (temporal), the building in which the course was held (spatial), the title of the course (item), and the name of the course instructor (item). These four types of information can be referred to as when, where, what, and who information. In Wittman's first experiment, during each of three tests separated by six-month intervals, subjects were given a recall questionnaire using a cuing technique with a map to ask about their individual courses from a previous semester. As shown in Figure 5, Wittman found a large degree of forgetting of course schedule information despite repeated exposure and natural learning. An important finding was the superiority of recall for spatial information (where) over item and temporal information (who, what, and when). Wittman proposed that subjects learned where information by repeating the procedure of walking to the classroom for each class session. Conversely, the learning of course title, instructor name, and class time did not involve analogous procedures.

In a second experiment, Wittman (1989) had subjects learn other individuals' course schedules in a laboratory setting. During the study phase, subjects were provided with both a campus map and a course schedule including the same four types of information as in the first experiment. Subjects received nine training trials in addition to two tests, the first test one week later and the second (retention) test after five additional weeks. The training trials consisted of studying the class schedule and map, followed by a recall task. Students were tested using the recall questionnaire as well as two new methods, a map test and a class listing test. In both of these new tests, subjects were required to provide the same type of item information (course title and instructor's name). However, the tests differed in the type of temporal and spatial information required. For the map test,
subjects provided the temporal order of class occurrence during the school week and, as in the recall questionnaire, the building location of each class on the campus map. In contrast, for the class listing test, subjects provided the start time and the building name for each class. Thus, the map test was designed to resemble the natural procedures used in retrieving course locations, whereas the class listing test was designed to remove that procedural component from the recall of course locations.

![Bar chart](chart.png)

**Figure 5.** Results of Experiment 1 by Wittman (1989). Mean proportion of correct recall of what, who, where, and when information as a function of test time.

As in the natural learning experiment, comparison of the one-week test and six-week retention test revealed an overall forgetting of course schedule information, as shown in Figure 6. Retrieval of *where* information was again superior; however, this superiority occurred only on the map tests. On the class listing tests, retention of *where* information was not superior on the one-week test and showed significant loss on the six-week retention test. These results support the notion that the superior memory performance is due to the use of procedures during learning.
Figure 6. Results of Experiment 2 by Wittman (1989) for the map test and the class listing test. Mean proportion of correct recall of what, who, where, and when information as a function of test time.
A further course schedule study by King (1992) separated procedural experience from the use of a map in order to explore the role of procedural knowledge in spatial memory superiority. To separate these two issues, subjects' memory for fictitious course schedules was tested in two separate situations: one in which subjects had previous procedural experience with the campus and one in which subjects were without such experience. If the retention advantage of spatial information is due to procedural experience, then we would expect a retention advantage for spatial information only in the familiar condition, in which subjects had previous procedural experience.

Undergraduate students from either the University of Colorado (CU) or Colorado State University (CSU) participated as subjects. All subjects were unfamiliar with the other campus. Four different fictitious course schedules were constructed, two using a CSU directory of classes, and two using a CU directory. Half of the students were assigned to the familiar condition in which schedules were from their own campus; the other half of the students, in the unfamiliar condition, were assigned to schedules from the other campus. The experiment used Wittman's (1989) testing procedure with the three types of tests (recall questionnaires, class listing tests, and map tests).

![Proportion Correct](image)

**Figure 7.** Results of the experiment by King (1992) for the recall questionnaire test. Mean proportion of correct recall of what, who, when, and where information as a function of test time.
The results of the recall questionnaire are summarized in Figure 7 as a function of test time and information type. There was a significant degree of forgetting across the approximately one-month interval from the one-week test to the retention test. Recall performance differed among the four types of information. This effect of information type was, however, modulated by familiarity, as shown in Figure 8. Performance was better for the familiar condition than for the unfamiliar condition, but only on where information. These results support the prediction that where information would be superior only for the familiar condition.

![Graph showing recall performance by information type and familiarity](image)

**Figure 8.** Results of the experiment by King (1992) for the recall questionnaire test. Mean proportion of correct recall of what, who, when, and where information as a function of familiarity of campus.

The results of the class listing test, in which the where information consisted of building names rather than locations, are summarized in Figure 9. Forgetting was evident; performance on the one-week test was superior to performance on the retention test. Performance varied across information types, with performance on what information highest and performance on where information lowest. Further, there was differential loss from the one-week test to the retention test, with the greatest amount of loss for the when and where information. As shown in Figure 10, performance on where information was better for the familiar than for the unfamiliar condition, as in the recall questionnaire. Again, the effects of familiarity were not significant for the other types of information.
Figure 9. Results of the experiment by King (1992) for the class listing test. Mean proportion of correct recall of what, who, when, and where information as a function of test time.

Figure 10. Results of the experiment by King (1992) for the class listing test. Mean proportion of correct recall of what, who, when, and where information as a function of familiarity of campus.
Figure 11 summarizes the results of the map test, in which, like the recall questionnaire, the where information consisted of building locations rather than building names. Again, forgetting was evident, but here there was a where advantage for both the one-week test and the retention test.

![Graph showing proportion correct for test and retention test across information types (What, Who, When, Where)]

**Figure 11.** Results of the experiment by King (1992) for the map test. Mean proportion of correct recall of what, who, when, and where information as a function of test time.

In summary, on both the recall questionnaire and the class listing test, there was an effect of familiarity on where information but not on the other types of information. These results support our hypothesis that the spatial advantage was due to procedures, because procedural experience with the campus enhanced spatial recall.

For the class listing test, what, who, and when information was recalled better than where information. Thus, when the test required the building names rather than their locations, there was no spatial advantage over temporal and item information. Conversely, as expected, there was a spatial advantage on the map test, which required the building locations.

The opportunity to use learned procedures in answering location questions and to relate information to previous experience led to an advantage for the spatial aspect over the item and temporal aspects of course information.
Retention of Components of Lists

In the study of memory for course schedules, the who, what, where, and when questions necessarily differed from each other along a number of dimensions other than whether they involved temporal, spatial, or item information; for example, in the recall questionnaire and map tests the where questions were a type of recognition test whereas the who, what, and when questions were recall tests. Two laboratory experiments by Sinclair, Healy, and Bourne (1993) controlled for those other dimensions. The objective was to determine whether a spatial advantage would occur under these more controlled conditions. Because there was no procedural component in these experiments, no spatial advantage was predicted.

In the first experiment, subjects learned a list of 20 common nouns, each beginning with a different consonant from the alphabet. The words were presented one at a time in a vertical array on a computer terminal, with each word occurring for 2 seconds in a different location within the array. At the termination of the list presentation, subjects recalled the words by writing them on a sheet of paper. A trial, thus, consisted of one presentation and one recall attempt.

Three groups of subjects recalled the words in an order determined by either the temporal, spatial, or item information in the list. The first group of subjects was required to recall the words according to the temporal sequence of presentation; for these subjects the spatial arrangement of the words was alphabetical. The second group of subjects was required to recall the words according to the words’ spatial locations within the vertical field during presentation; for these subjects the temporal sequence of the words was alphabetical. The third group of subjects was required to choose the 20 words that had been presented from an alphabetically organized 210-word list including the critical words intermixed with similar distractor words. For this last group of subjects, both the temporal and spatial arrangements of the words were alphabetical. For all subjects, after each recall, another trial was started with the same words being presented in exactly the same sequence and locations. This process continued until the subject achieved a criterion of correct recall on three successive trials.

Subjects returned after a one-week delay and were asked to recall the 20 words as they had during the first session. After this initial retention test, the presentation and recall trials were resumed as in the first session, and continued until the criterion of correct recall on three successive trials was achieved again.

The initial retention test yielded the greatest proportion of correct responses for item information (.969) and substantially lower proportions for temporal (.769) and spatial (.750) information. The mean number of trials to criterion for each information type and session are summarized in Figure 12. Note that learning was most difficult in the spatial condition and least difficult in the item condition, and that first-session learning required more trials than did second-session relearning. Figure 12
illustrates the interaction between information type and session. Although initial learning proceeded more slowly in the spatial condition than in the temporal and item conditions, relearning of information in the spatial condition, once initial learning was achieved, was similar to that of the other conditions.

Figure 12. Results of Experiment 1 by Sinclair, Healy, and Bourne (1993). Mean number of trials to criterion for temporal, spatial, and item information as a function of session.

It is likely that the higher degree of learning difficulty observed in the spatial condition of Experiment 1 was due partially to the subjects' inability to discriminate effectively one spatial location from another. Central locations in the vertical array contained no unique information to distinguish them from neighboring locations. Hence, in the second experiment, a new array of 18 word locations arranged in two three-by-three matrices replaced the old vertical array of 20 word locations used in Experiment 1. Each spatial location was thus made unique and easily distinguishable from every other location within the new array. Half of the subjects had a retention period of 1 week and the others had a retention period of 6 weeks to elucidate the time course of forgetting from long-term memory.

Subjects showed substantially lower proportions of correct responses on the retention test after six weeks (temporal, .236; spatial, .347; item, .882) than after one week (temporal, .785; spatial, .875; item, .986). The mean number of trials to criterion for each information type and session are summarized in Figure 13.
Most interesting is the observation that performance was better on
the spatial than on the temporal information. Thus, the ordering
of the temporal and spatial conditions was the reverse of that in
Experiment 1, in which performance on temporal information was
better than that on spatial information. As expected, simply
changing the presentation array so that each of its component
positions provided unique spatial information facilitated learning
in the spatial condition. Also note that although initial learning
proceeded at very different rates for the three types of
information, their relearning was again similar. That is, the
initial learning rates for the three types of information varied
more than did their relearning rates.

Figure 13. Results of Experiment 2 by Sinclair, Healy, and Bourne
(1993). Mean number of trials to criterion for temporal, spatial,
and item information as a function of session.

The number of weeks intervening between original learning and
second-session relearning affected recall greatly. The trials to
criterion required for relearning in Session 2 were greater after a
six-week delay (Session 1 M = 5.54 trials, Session 2 M = 4.75
trials) than after a one-week delay (Session 1 M = 5.71 trials,
Session 2 M = 3.71 trials). It is clear, however, that even in the
six-week condition some information from the first session was
retained because the number of trials required for relearning was
less than that required for learning in the first session.

There are three conclusions that can be made on the basis of
these findings. First, although there was considerable long-term
retention evident for all three types of information, there was nonetheless significant forgetting for temporal and spatial information. This forgetting was already evident across the one-week retention interval and was even greater across the six-week retention interval. The forgetting observed for the declarative information studied here was generally consistent with that found in our study of course schedules but was in marked contrast to the substantial retention we found in our earlier studies examining the long-term retention of procedural skills (see Healy et al., 1992), such as the study of the Stroop task (see Clawson et al., 1993). Second, there were large differences in the learning of temporal sequence, spatial arrangement, and item identity, with smaller differences in the relearning of the three types of information. Third, learning spatial information was more difficult than learning temporal information when the spatial positions were hard to differentiate, but the opposite pattern of results was found when each spatial position was made distinctive. Thus, making the to-be-learned spatial information distinctive facilitated learning.

Mental Arithmetic

Next we turn our attention to the guidelines regarding the optimization of learning strategies. The first study in this section concerns mental arithmetic, an area of our research that is further discussed by Rickard and Bourne (1993).

Our previous work on mental arithmetic has uncovered several important facts about the acquisition, transfer, and retention of skill (Fendrich et al., 1993; Rickard et al., in press). First, in accordance with most other research on skill acquisition, speed-up with practice on simple multiplication and division problems follows the power law of learning (Newell & Rosenbloom, 1981; Rickard et al., in press). Second, in accord with our work on the Stroop task, arithmetic skill is almost entirely specific to the problems practiced, suggesting that adult subjects store each problem separately in some form of fact memory (Ashcraft, 1992; Rickard et al., in press). Even complementary multiplication and division problems, such as 4 x 7 and 28 ÷ 7, are represented by adults as independent facts. Third, we have shown that speed-up from practice is maintained over retention intervals of a month or longer without significant decrement (Fendrich et al., 1993).

Although our research, as well as a substantial amount of other research in the literature (reviewed by Ashcraft, 1992), suggests that adult performance on simple arithmetic mostly reflects retrieval of facts from memory, it has been demonstrated by Siegler (1986, 1988) that children often rely on algorithms to calculate the solutions to arithmetic problems. In order to study similar processes in adults, we developed a novel mental arithmetic task which, initially at least, requires the application of a general algorithm (just as in the case of children beginning to learn arithmetic), but with sufficient practice should be performable by retrieving answers directly from memory (Rickard, 1993). This task allowed us to test the generalizability of our findings with simple arithmetic, for which practice simply strengthens access to already existing facts, to a task for which practice results in a transition from algorithm to retrieval.
Adult subjects were trained on two types of problems, based on a novel, arbitrary operation symbolized by the pound sign (#). On Type I problems, subjects were given two elements from a simple arithmetic progression, and were required to generate the third (next) element. The generic progression that we used was one in which the third element is the second element plus the difference between the first and second elements, plus 1. For example, the answer to \( 7 \# 15 = \_ \) is computed as \( 15 \% (15 - 7) + 1 = 24 \). Type II problems were based on the reverse algorithm, the answer being the second element of the series (e.g., \( 7 \# \_ = 24 \)). Across five sessions subjects received 90 blocks of problems; a block consisted of a single presentation of each of 12 unique problems: 6 Type I problems and 6 Type II problems. At the end of the fifth session subjects were given a transfer test, on which they were retested on the practice problems (no-change problems), and were also tested on practice problems with the missing element changed (type-change problems, e.g., a Type I problem became a Type II problem) and on unpracticed problems (new problems). Finally, subjects were given the same test six weeks later to measure retention.

During practice, subjects were probed on one third of the trials to determine whether they used the algorithm that they were taught, retrieved the answer directly from memory, or used some other, unspecified approach. During both the immediate and the delayed tests, subjects were probed after every trial. On probe trials, subjects signified "algorithm", "retrieve", or "other" by pressing labeled buttons on a response console.

The strategy probing results from practice are shown in Figure 14. Note that practice was successful in creating a transition from the algorithm to direct retrieval. By about Block 60, direct retrieval was the reported strategy on nearly all trials. The transition from algorithm to retrieval was virtually complete for all subjects. After Block 60, few if any problems required intermediate stages for solution.

Log reaction time averaged across subjects and across correctly solved problems is displayed in Figure 15, plotted as a function of log block. The average reaction time for Block 1 was about 13 seconds. By Block 90, the average reaction time was about 1 second. The line drawn through the data corresponds to the best fitting power function. Note the clear deviation from linearity evident in the data. This pattern, combined with the evidence from the strategy probing data, suggested to us that the power law may actually be strategy specific, and may not hold during the transition from algorithm to retrieval (cf. Logan, 1988).

To test this hypothesis, additional reaction time analyses were performed only for trials on which strategy probes were collected. Considering only algorithm trials, there should be power function speed-up with practice; that is, the data should plot linearly in log-log coordinates. Similarly, when considering only retrieval trials, there should be power function speed-up, and thus the data should plot linearly in log-log coordinates. However, the two power functions would be unlikely to share the same parameters. Figure 16 shows the log reaction time data for strategy probing trials overall, for algorithm responses only, and for retrieval responses only. As
predicted by our hypothesis, when the data were separated by strategy, they conformed very nicely to two linear but different power functions.

Figure 14. Results of an experiment by Rickard (1993). Mean proportion of algorithm, retrieval, and other strategy reports as a function of practice block.

Figure 15. Results of an experiment by Rickard (1993). Mean correct log reaction time as a function of log practice block.
Figure 16. Results of an experiment by Rickard (1993). Mean correct log reaction time for algorithm trials, retrieval trials, and overall as a function of log practice block.

The average reaction time results for the immediate and delayed tests are shown in Figure 17. On both tests, performance was much faster on no-change problems than on either new or type-change problems. The strategy probing data help to explain this difference. Subjects reported using direct retrieval on nearly all no-change problems of the immediate test, and on roughly half the no-change problems of the delayed test. In contrast, the algorithm was the reported strategy on nearly all new and type-change problems on both tests. These results show that the skill acquired with practice was highly specific to the individual problems that were practiced. Indeed, even when there was only a type-change from practice to test (e.g., 3 # 17 = _ to 3 # _ = 32), there was no transfer of learning.

Reaction times for no-change problems on the delayed test were about half-way between reaction times for no-change and new problems on the immediate test, indicating some skill retention. Nevertheless, the substantial increase in reaction time for no-change problems on the delayed test indicated a much greater loss in skill across the retention interval than we had observed in our previous work on simple arithmetic (e.g., Fendrich et al., 1993; Rickard et al., in press). To investigate this finding further, we plotted the reaction times for no-change problems on the delayed test separately by strategy (algorithm or retrieval) as shown by the dotted lines in Figure 18. When retrieval was the reported strategy for no-change problems on the delayed test, the reaction times were almost exactly the same as for the no-change problems on the immediate test. When the algorithm was
the reported strategy, the reaction times were nearly exactly the same as those for new and type-change problems. This result suggests that the effects of the retention interval were only to decrease the probability with which the retrieval strategy was used, without changing the time required to execute that retrieval strategy when it was used. Thus, a training procedure that promotes the use of an optimal strategy for a given task appears to contribute to the maintenance of training levels of performance on later tests of retention.

![Graph showing reaction time in milliseconds for new problems, type-change problems, and no-change problems as a function of test time.](image)

**Figure 17.** Results of an experiment by Rickard (1993). Mean correct reaction time in milliseconds for new problems, type-change problems, and no-change problems as a function of test time. (All means were calculated based on log reaction times and then transformed back to milliseconds by the anti-log function.)
Figure 18. Results of an experiment by Rickard (1993). Mean correct reaction time in milliseconds for the three problem types and separately for algorithm and retrieval trials of the no-change problems as a function of test time. (All means were calculated based on log reaction times and then transformed back to milliseconds by the anti-log function.)

Direct and Mediated Retrieval in Vocabulary Acquisition

Retrieval strategies were also a focus of our studies concerning foreign vocabulary acquisition (discussed more fully by Crutcher & Ericsson, 1993). Learning vocabulary items in a foreign language is in many ways an ideal everyday task for the study of retention under controlled conditions, due to the independent and often arbitrary nature of its required associations. In most of our earlier research (Crutcher, 1990; Crutcher & Ericsson, 1992) students unfamiliar with Spanish learned approximately 40 Spanish vocabulary items, after which their retention was tested one week, one month, and even one year later. The current set of studies extends our findings to significantly more practice and makes comparisons to retrieval by experts (that is, advanced students of Spanish). Before turning to these new findings, let us briefly review the procedures and general results of our previous work.

For the vocabulary items we used, the Spanish word was completely unrelated to its English translation (e.g., doronico and "leopard"). To facilitate learning we instructed subjects in the
use of the keyword method. In the keyword method, the Spanish word is first related to a similar-sounding English word (the keyword) provided by the experimenter (e.g., doronico and "door"). The keyword is then associated to the English translation by forming an interactive image. This method of learning provided a great deal of control over the mediating processes, thus assuring a very similar encoding structure across subjects. After subjects had acquired all vocabulary items, we examined their retrieval speed and accuracy in three ways: using the Spanish word as a cue to retrieve the English translation (vocabulary task), using the Spanish word as a cue to retrieve the keyword (keyword subtask), and using the keyword as a cue to retrieve the English translation (English subtask).

The results from these tests immediately following acquisition showed that retrieval of the English translation when cued by the Spanish word involved access and mediation of the keyword (Crutcher, 1990). Retention testing at a one-week or one-month delay showed two further results. First, accuracy of retrieval was reduced (especially after a month) and inability to retrieve the English translation given the Spanish word was almost perfectly predicted by inability to retrieve the English translation given the keyword. Second, for a given item, retrieval speed on the first block of the retention test was considerably slower than on the immediate test after training, but on the second block of the retention test the speed was comparable to that on the immediate test. That is, after a memory trace was successfully accessed the first time at retention, its strength appeared to be completely recovered.

In more recent studies, subjects practiced retrieving the English translation for 80 training blocks (Crutcher, 1992; Crutcher & Ericsson, 1992). Half the items were consistently cued by only the Spanish word (the vocabulary task) and the other half of the items were cued by only the keyword (the English task). Subjects were then tested on both retrieval tasks for all items, first immediately after the extended practice and again one month later. On the immediate test, for items trained with the English subtask, the retrieval times on the vocabulary task were longer, consistent with sequential access mediated by the keyword. However, for items trained with the vocabulary task, the retrieval times on the English subtask were longer, implying the emergence of direct access. Retrospective reports provided convergent evidence for direct retrieval with extended practice on the vocabulary task.

In the experiment giving subjects extended practice (Crutcher, 1992), new analyses of retention after a one-month delay showed a very interesting pattern. Although accuracy of retrieval was uniformly high on both retrieval tasks, there was a robust interaction of retrieval task and training condition. For items trained with the English subtask, recall proportion was reliably worse on the vocabulary task (M = .90) than on the English subtask (M = .97), which suggested a loss of the association between the keyword and the Spanish word. Training with the vocabulary task yielded worse performance for retrieval on the English subtask (M = .95) than on the vocabulary task (M = .99). For about 4 percent of
the items, the English translation could be retrieved using the Spanish word as a cue without the subjects' being able to retrieve the English translation using the keyword as a cue. At the same time, the keywords remained effective cues for 95 percent of the items, although the keywords had hardly been presented since the original acquisition of the items.

Table 1 presents mean retrieval times for the immediate test after practice and for the retention test's Block 1 (the first encounter of an item) and Block 2 (the second encounter). For both blocks of the retention test, as well as for the immediate test, there was an interaction between training condition and retrieval task. Furthermore, retrieval speed on the first block of the retention test was much slower than it had been one month earlier on the immediate test, but retrieval speed on the second block was virtually indistinguishable from that on the immediate test. Hence, this study with extensively practiced items replicated our earlier findings (Crutcher, 1990) with forgetting of responses due to loss of the connecting associations and a remarkable recovery of the entire pattern of retrieval times after the first exposure of the retrieval task at a long delay.

Table 1. Results of experiment by Crutcher (1992). Mean correct retrieval time in milliseconds for items trained in the Vocabulary task (V-trained) and items trained in the English subtask (E-trained) as a function of test type (Vocabulary, V, or English, E) and test time. (All means were calculated based on log reaction times and then transformed back to milliseconds by the anti-log function.)

<table>
<thead>
<tr>
<th>Item Type</th>
<th>Immediate V</th>
<th>Immediate E</th>
<th>Delayed, Block 1 V</th>
<th>Delayed, Block 1 E</th>
<th>Delayed, Block 2 V</th>
<th>Delayed, Block 2 E</th>
</tr>
</thead>
<tbody>
<tr>
<td>V-trained items</td>
<td>830</td>
<td>963</td>
<td>1207</td>
<td>1332</td>
<td>883</td>
<td>952</td>
</tr>
<tr>
<td>E-trained items</td>
<td>1376</td>
<td>813</td>
<td>1636</td>
<td>1236</td>
<td>1146</td>
<td>888</td>
</tr>
</tbody>
</table>

The finding that the keyword remained accessible at delay even when only the direct connection between the Spanish word and its English translation had been practiced was consistent with other results obtained in our laboratory. For example, we found that after extended practice leading to apparent direct access of the English translation, it was still possible to interfere selectively with the speed of retrieval by having subjects memorize new associates to the keywords (Crutcher, 1992). It would, thus, appear that the original encoding of an item during learning continues to influence retrieval after extended practice even when other evidence points to unmediated direct retrieval.
Finally, some of us (Crutcher, Hammerle, Ericsson) have further explored our findings by studying experts in vocabulary retrieval, namely advanced students of Spanish. In our initial expert study, experts simply took the previously used vocabulary test, restricting the test to retrieving English translations for the Spanish words. To minimize warm-up effects, we had the subjects read both the Spanish words and keywords as fast as possible for two blocks. Figure 19 shows subsequent retrieval speed of vocabulary items for the five blocks of testing.

Figure 19. Results of Experiment 1 by Crutcher, Hammerle, and Ericsson. Mean correct retrieval time in milliseconds by Spanish experts as a function of test block, compared with the average retrieval time of extensively trained novices. (All means were calculated based on log reaction times and then transformed back to milliseconds by the anti-log function.)

There were two interesting findings. First, by the fifth block the experts were able to retrieve the vocabulary items in about 630 ms, which was 200 ms faster than the retrieval speed for our novice subjects after extended training. Second, there was a considerable speed-up in retrieval for the experts during the five blocks.

In a second experiment, subjects were tested initially on only half of the words, but returned one week later for a retention test. Figure 20 shows retrieval times at retention both for the old words that had been practiced and for the unpracticed, or new, words. Subjects maintained their original speed-up on the words
they had practiced, but this speed-up did not transfer to the new words. Retrieval times for the new words improved dramatically during the retention test, an improvement similar to that for old words during the original training.

![Diagram](image)

**Figure 20.** Results of Experiment 2 by Crutcher, Hammerle, and Ericsson. Mean correct retrieval time in milliseconds by Spanish experts for practiced and new words as a function of test block on a one-week retention test. (All means were calculated based on log reaction times and then transformed back to milliseconds by the anti-log function.)

In summary, performance of the Spanish experts improved greatly after a few exposures to the items being tested. The improvement in retrieval speed for practiced words was essentially perfectly retained during the delay, but did not generalize to new, unpracticed words.

**Summary and Conclusions**

In closing, we review the three classes of guidelines we found to optimize long-term retention, as summarized in Table 2. The first class of guidelines concerns ways to optimize the conditions of training. We discussed three general guidelines in this class. First, superior memory results from the use of reinstatable procedures during learning. The procedural reinstatement framework was used to account for the observed superiority of memory for
spatial order found in our studies of the retention of course schedule information. Second, retention can be aided by prior familiarity. Memory for spatial information of course schedules was improved when the information could be related to previous experience. Third, learning is facilitated by distinctiveness of the information, as was evident with the spatial information in our study of list learning.

Table 2. Classes of guidelines to optimize long-term retention.

<table>
<thead>
<tr>
<th>Guideline</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Optimize conditions of training:</td>
<td>Schedule components</td>
</tr>
<tr>
<td>(a) Use procedures during learning.</td>
<td>Schedule components</td>
</tr>
<tr>
<td>(b) Relate information to previous experience.</td>
<td>List components</td>
</tr>
<tr>
<td>(c) Make the to-be-learned information distinctive.</td>
<td></td>
</tr>
<tr>
<td>(2) Optimize the learning strategy:</td>
<td>Mental arithmetic</td>
</tr>
<tr>
<td>Direct retrieval is best.</td>
<td></td>
</tr>
<tr>
<td>(3) Optimize retention conditions:</td>
<td>Vocabulary acquisition</td>
</tr>
<tr>
<td>Provide refresher or practice tests.</td>
<td></td>
</tr>
</tbody>
</table>

The second class of guidelines concerns ways to optimize the learning strategies used. We found in our study of mental arithmetic that the strategy used by the subject importantly influenced retention performance and that a direct retrieval strategy led to faster responding than did a strategy based on the use of an algorithm. Our study of vocabulary acquisition demonstrated that a direct retrieval strategy may also be achieved in that domain, but mediating associations may still exert an influence even when retrieval appears to be direct.

The last class of guidelines concerns ways to optimize retention conditions. In our study of vocabulary acquisition we saw remarkable recovery of retrieval speed after the initial warm-up retrieval. Hence, it appears that the use of a refresher or practice test before the critical test may have a profound impact on retention performance.

We began this report by summarizing some of our work demonstrating the specificity of improvement in performance. Stroop training on specific colors and words showed excellent retention across a month-long delay interval but limited transfer to new colors and words. Although our original goal in this research program had been limited to an examination of the optimization of long-term retention, we have learned that optimizing retention does not guarantee generalizability.

Overall in our research program, we have been able to demonstrate conditions that lead to highly durable performance over time. Our basic interpretation of this result invokes the
principle of procedural reinstatement, that is, retention performance will be best when the mental procedures required at test match those employed during training. Our more recent results have highlighted a limitation of this principle. Durable retention is associated with highly specific skill, that is, retention performance will be poor when the mental procedures required at test do not match exactly those employed during training. The relationship between durability and specificity in memory will be the focus of our future research.

Acknowledgments

This research was supported by Army Research Institute Contract MDA903-90-K-0066 to the Institute of Cognitive Science at the University of Colorado. Robert J. Crutcher is now at the University of Illinois, Chicago; K. Anders Ericsson is now at Florida State University, Tallahassee.
References


Conference Papers and Colloquia


Ericsson, K.A. (1991, November). Prudent practice makes perfect. Invited address at the Nutrition Summit Conference "Genetics,
how much of mother nature can be overridden?" Newport Beach, California.
presented at the U.S. Army Research Institute Contractors' Meeting. Winter Park, Florida.


