HUMAN FACTORS, 1995, 37(4), 781–803

Practice Effects on Skill Acquisition, Learning Outcome, Retention, and Sensitivity to Relearning

VALERIE J. SHUTE,[1] *Armstrong Laboratory, Brooks Air Force Base, Texas, and* LISA A. GAWLICK, *Galaxy Scientific Corporation, San Antonio, Texas*

This paper presents the results from two experiments that examined practice effects on skill acquisition, learning outcome, retention, and sensitivity to relearning. In Experiment 1, our learning criterion task was an intelligent tutoring system teaching flight engineering knowledge and skills. The system was divided into two main curriculum sections, with two practice conditions per section: abbreviated and extended. Thus there were a total of four practice conditions differing only in the number of practice problems requiring solution across problem sets. Experiment 1 showed that subjects in the more abbreviated conditions completed the curriculum significantly faster than did subjects in more extended conditions, but at the expense of greater errors and latencies during problem solution within the tutor. Despite these acquisition differences, groups performed the same across all learning outcome measures. Experiment 2 examined the effects of practice condition on retention and sensitivity to relearning after two years. Although the sample size was fairly small (34 returning subjects), we found evidence for practice effects on long-term retention but not in the predicted direction. In terms of our sensitivity to relearning measure, the groups did not differ. Implications of these findings for optimizing training are discussed.

INTRODUCTION

Learning, or skill acquisition, represents a change in a person that occurs at a particular time as a function of experience or practice. Because it is not directly observable, learning must be inferred from performance on a memory test, in which the retention interval may be immediate or delayed. How does practice affect an item's probability of being encoded and subsequently remembered over time? In describing the relationship among practice, acquisition, and retention, Ebbinghaus (1964) stated:

> These relations can be described figuratively by speaking of the series as being more or less deeply engraved on some mental substratum. To carry out this figure: as the number of repetitions increases, the series are engraved more and more deeply and indelibly; if the number of repetitions is small, the inscription is but surface deep and only fleeting glimpses of the tracery can be caught; with a somewhat greater number the inscription can, at least for a time, be read at will; as the number of repetitions is still further increased, the deeply cut picture of the series fades out only after longer intervals. (Pp. 52–53)

In the more than 100 years since Ebbinghaus originally presented this analogy, numerous studies have suggested that the relationships among practice, acquisition, and retention are not this straightforward. The literature is

[1] Requests for reprints should be sent to Valerie J. Shute, Armstrong Laboratory (AL/HRTI), 1880 Carswell Ave., Lackland Air Force Base, TX 78236-5507.

© 1995, Human Factors and Ergonomics Society. All rights reserved.

replete with findings that certain acquisition conditions that slow the rate of improvement, or that decrease performance during practice, still yield enhanced posttraining performance relative to a standard practice condition (for a review, see Schmidt and Bjork, 1992). For instance, Shea and Morgan (1979) compared motor skills on three tasks that were learned and tested under random or blocked (fixed-sequence) practice schedules. Retention was compared by condition, and results showed that although the random condition was considerably less effective during the learning process (i.e., it degraded performance during skill acquisition), these subjects performed better on retention tests than did subjects who learned in the blocked condition.

Using a cognitive task, Carlson and Yaure (1990) reported a similar finding. In three studies they showed how random practice schedules yielded poorer acquisition performance but superior retention relative to blocked practice schedules. They attributed the cause of this finding to processing, rather than storage, demands of intertrial activity. That is, working-memory demands were higher for the random trials because subjects had to reload information into working memory more often, impairing the acquisition phase of learning but enhancing the retention and transfer of the acquired skills. In their words, "this intraitem processing view emphasizes fluency in accessing and using component skills rather than procedures for choosing which of several skills to use" (p. 485).

Other kinds of practice schedules also influence acquisition and retention (e.g., Glenberg, 1979; Hintzman, 1974). Some classic examples include the benefits of distributed over massed practice (e.g., Greene, 1989; Proctor, 1980), as well as the benefits of using a variety of example problems during practice and problem solution (e.g., Gick and Holyoak, 1987; Singley and Anderson, 1989). In the latter case, when the variety of example problems is restricted, learning tends to be rapid but transfer tends to be weak. Finally, increasing the time spent exercising a new cognitive skill typically results in improved performance and reduced cognitive load (e.g., Ackerman, 1988; Anderson, 1987, 1993; Bryan and Harter, 1899; Fisk and Rogers, 1992; Sweller, 1988; Woltz, 1988). Rules become strengthened as a result of sustained and successful practice applying them.

What basic mechanisms are presumed to underlie these different practice effects on learning? Although Watkins (1990) passionately argued that use of the memory trace metaphor is counterproductive, we find it useful in characterizing the processes of memory. The probability of establishing a memory trace in the first place (i.e., some knowledge or skill becomes permanently encoded) is related to the number of times an item is reentered into working memory, and memory-trace strength increases as a power function of practice (Anderson, 1987; Carlson and Yaure, 1990; Crowder, 1976). A key issue concerns "reentry" into working memory. In both distributed and variable practice conditions, items go in and out of working memory, and trace strength is accrued for each unique entry. Simply presenting an item over and over again, as in massed or blocked practice conditions, does not guarantee an increment in trace strength because subsequent and original items would coincide in working memory (Glenberg, 1979; Greene, 1989) and, thus, not contribute anything unique to trace activation levels. Once an item has been successfully encoded, the probability of retrieving it (and the time required to do so) is a function of the trace's current level of activation (Anderson, 1983). That is, higher activation levels mean greater probability of retrieval in less time. In terms of retention over time, trace strength has been shown to decrease in accord with a power function decay model (e.g., Wickelgren, 1976; Woltz and Shute, 1995).

These findings suggest that in the acquisition of motor skills and relatively easy cognitive tasks, if there are differing practice schedules, then what an individual ultimately learns may be obscured during the acquisition process because relatively permanent effects become confounded with temporary performance effects, which can disappear after the practice session is finished or when test conditions change. However, if subjects learn from conditions that are based on the same practice schedule and only the number of problems to solve is varied, common sense dictates that solution of more practice problems will lead to effective skill acquisition, greater outcome or transfer performance, better retention, and greater sensitivity to relearning over time.

The purpose of this paper is to examine the notion that amount of practice improves learning and retention. In Experiment 1 we used a consistent practice schedule (blocked problem types) and varied the number of practice opportunities in order to observe the effects on skill acquisition and learning outcome, which were assessed immediately after learning a complex criterion task. Rather than use a typical laboratory criterion task (e.g., paired associates or logic gates), we employed a complex, real-world learning task: a computer-based instructional system for teaching flight engineering knowledge and skills. The program was manipulated to yield contrasting practice conditions, which differed only in the number of problems the learner solved in each of the problem sets. Hence our practice schedule was a blocked design, and our practice conditions varied as a function of number of practice opportunities.

Our learning criterion task was an intelligent tutoring system for teaching flight engineering knowledge and skills. The system was divided into two main curriculum sections: instructing component knowledge and skills, and integrating and applying those components. Each section consisted of two practice conditions: abbreviated and extended. Thus there were four practice conditions, which differed only in the number of practice problems requiring solution across problem sets. The hypotheses tested in Experiment 1 were that (a) subjects assigned to a practice condition consisting of limited practice opportunities (i.e., the abbreviated condition—3 problems per problem set) would take less time, overall, to complete the program because there were considerably fewer problems for them to solve, but that (b) these subjects would not perform as well on the posttest measures as would subjects who had greater practice opportunities (i.e., the extended condition—12 problems per problem set). In addition, (c) subjects in the abbreviated condition would take longer to solve each of their three problems and would manifest more errors during acquisition (given the sparse practice setting) than would the extended-condition subjects on the first 3 of their 12 problems per problem set; and (d) the practice condition for learning the component knowledge and skills would subsequently influence performance on the second part of the tutor curriculum.

In Experiment 2 subjects from Experiment 1 were recalled approximately two years following their participation in the original study, and two memory indices were assessed: retention and sensitivity to relearning. We were interested in seeing how one's original practice condition affected these two measures. In this paper, *retention* is defined as the degree to which subjects remember (or forget) information learned at some point in the past, characterized by a power function decay model (e.g., Anderson, 1983; Wickelgren, 1976; Woltz and Shute, 1995). Our second variable, *sensitivity to relearning*, is defined as the rate at which individuals were able to reacquire knowledge and skills. Ebbinghaus (cited in Adams, 1980, p. 233) originally characterized this as a savings score—the amount of time needed to relearn something relative to the time needed to learn it originally. That is, if it took 1000 s to learn a list originally and only 300 seconds to relearn it after some retention interval, the savings would be 700 s, or 70%. Retention and sensitivity to relearning are related, in that both reflect the degree of decay of the initial learning trace, but retention may be construed as initial performance following a retention interval, and sensitivity to relearning can be viewed as the time it takes to get back to some preretention-interval performance level.

An abridged version of the original computer program (a refresher course) was developed to reteach the same core flight engineering concepts and skills first taught in Experiment 1. We predicted that retention (assessed immediately on returning) and sensitivity to relearning (assessed after the refresher course) would be a function of how well subjects initially acquired knowledge and skills, which in turn was expected to be a function of differential practice opportunities. Thus the hypotheses investigated in Experiment 2 were as follows: If the simplistic "more practice is better" premise is true, all else being equal, then subjects who originally learned in the extended condition would (a) show greater retention (less decay) of material than would subjects who had learned in the abbreviated condition and (b) show greater sensitivity to relearning knowledge and skills, given more initial practice. Furthermore, (c) all returning subjects (regardless of initial condition) would perform better than a control group because of their previous exposure to the material.

EXPERIMENT 1

Method

Subjects. Subjects were 356 participants in a seven-day study on the acquisition of flight engineering knowledge and skills. The sample had a mean age of 22 years and was 76% male and 24% female. All subjects were high school graduates. Subjects were obtained from a local temporary employment agency and were paid for their participation, which consisted of 45 h of testing and learning. None of the subjects had any prior experience or training as flight engineers or pilots.

Flight engineering tutor. The computer program was originally developed at the University of Pittsburgh (Lesgold, Bunzo, McGinnis, and Eastman, 1989) and then modified at the Armstrong Laboratory to fit experimental objectives. The system was designed to teach knowledge and skills associated with a flight engineer's job. Job components include collecting and analyzing information about a pending flight and deciding whether various factors (e.g., weather and runway conditions, type of plane) indicate a safe flight. There were two parts to the curriculum: the graph section, which taught basic knowledge and skills used by a flight engineer (e.g., relative wind direction, reading and interpreting graphs), and the takeoff and landing data (TOLD) worksheet, which taught how to complete a worksheet used by actual flight engineers. The second section of the tutor required integration and application of the component skills learned in the first section.

Graph section. The curriculum of the graph section of the tutor consisted of 13 instructional units (or problem sets) teaching progressively more complex skills. Each of the successive problem sets built on information learned from prior sets. For example, subjects were first taught how to interpret data from Cartesian coordinate graphs prior to the more conceptually complex polar coordinate graphs (which introduce rays and arcs in two-dimensional space). We grouped the 13 problem sets of the graph section into four major categories: (a) conversion scale problems (G-1 to G-3), (b) using the Cartesian coordinate grid (G-4 to G-6), (c) using the polar coordinate chart (G-7 to G-10), and (d) interpreting the wind components chart (G-11 to G-13). Relatively easy component skills, such as reading points from or plotting points onto a graph, were taught in the early problem sets (i.e., Categories a, b, and c). More difficult problem sets were presented later in the tutor (Category d) and required interpretation of data from a complex chart for problem solution. For instance, to interpret information from the wind components chart, a learner needed to be able to read from, and plot data onto, both Cartesian and polar coordinate charts.

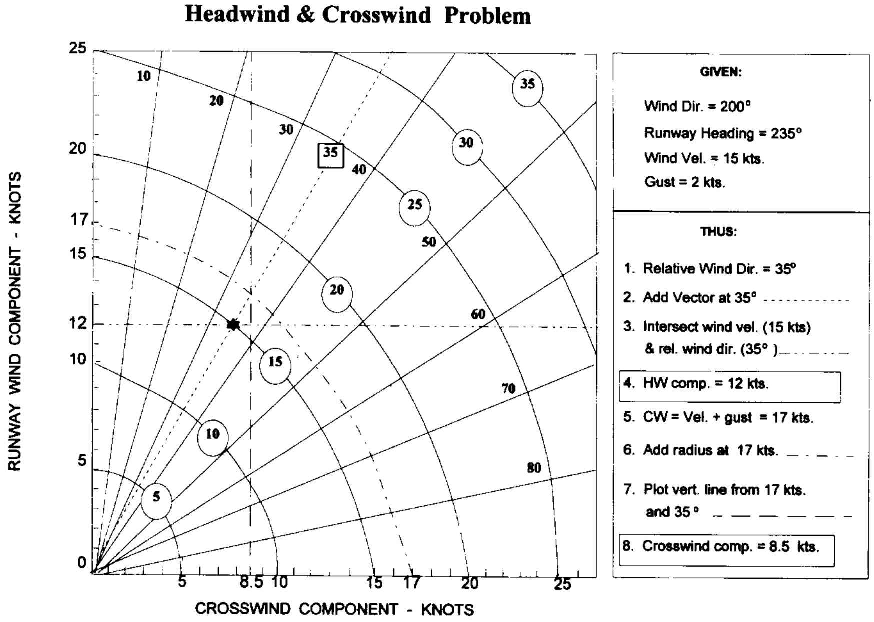
After completing the graph section, subjects proceeded to the TOLD worksheet section, which constituted the more substantive part of the curriculum. It required the learner to integrate all component knowledge and skills learned in the graph section in the solution of the domain-specific problems.

TOLD section. The on-line TOLD worksheet matched the form used by flight engineers. A learner was required to enter relevant data about a flight into a specific form. To accomplish this, one had to extract information from various graphs and enter these data in the proper cells on the worksheet. In the tutor some information was given to learners (e.g., gross weight of the aircraft; wind direction and velocity; runway length, heading, and conditions), whereas other information had to be derived from complex graphs and then entered into the correct cells. For instance, to solve a headwind component problem, a learner needed to know about such things as relative wind direction and wind type, interpreting data from a polar coordinate chart (normalized and quartered), and so on. A detailed description of all problem sets is available on request from the first author.

The curriculum in the TOLD section consisted of nine problem sets, divided into three main categories: (a) computing the maximum allowable crosswind; (b) computing the headwind, tailwind, and crosswind components; and (c) determining whether or not a planned flight was safe to take off. Problem sets within each category were of increasing difficulty. The final three problem sets (Category c) were the most difficult because subjects were required to incorporate all knowledge, skills, and rules learned in prior problem sets (from both the graph and TOLD sections of the tutor) and to use that information to make critical decisions concerning whether or not it was safe to take off.

Figure 1 illustrates a typical problem from the middle of the TOLD section involving the wind components chart. To solve this kind of problem, learners needed to apply rules that were specifically related to the various wind types (headwind, tailwind, and crosswind). An eight-step solution procedure is shown in Figure 1 for determining headwind and crosswind values.

In the TOLD section of the tutor, on-line tools were available to assist the learner in making computations and modifying graphs. These tools enabled learners to draw vertical or horizontal lines on a given chart or graph, add a radius or vector, erase lines, and redisplay a graph. The graph section of the tutor offered the opportunity to seek assistance if the learners were having problems (e.g., to explicitly ask for



Headwind & Crosswind Problem

GIVEN:

Wind Dir. = 200°
Runway Heading = 235°
Wind Vel. = 15 kts.
Gust = 2 kts.

THUS:

1. Relative Wind Dir. = 35°
2. Add Vector at 35°
3. Intersect wind vel. (15 kts) & rel. wind dir. (35°)
4. HW comp. = 12 kts.
5. CW = Vel. + gust = 17 kts.
6. Add radius at 17 kts.
7. Plot vert. line from 17 kts. and 35°
8. Crosswind comp. = 8.5 kts.

Figure 1. *Illustration of an eight-step solution from the TOLD section of the tutor.*

help or to review prior material by previous-paging). Both sections had an on-line dictionary available, which included related information about concepts and procedures. All learning was self-paced.

Practice conditions. Two contrasting practice conditions were developed for both the graph and TOLD worksheet sections of the system: extended (12 problems per problem set: 4 easy, 4 medium, and 4 hard) and abbreviated (3 problems per set: 1 easy, 1 medium, and 1 hard). In the abbreviated condition, the three problems within each problem set were selected from the pool of 12 that existed in the extended condition. For both conditions, difficulty level within each problem set was a function of various dimensions, such as the degree of interpolation required to read data from a graph. A fixed ratio of 4:1 existed between the two versions to provide a reasonable contrast between practice conditions. We established this ratio based on the idea that 12 problems were sufficient to allow proceduralization of new skills, and it is typical of the number of problems found in textbook exercises. Three problems were believed to be the minimum number required to acquire novel concepts and skills presented by the tutor (i.e., one problem would be inadequate to derive a generalization, a second problem could invalidate a developing generalization, but a third problem could resolve conflicts). Because there were two versions for each of the two sections of the program, this resulted in four conditions to which a subject could be assigned: (1) graph-abbreviated and TOLD-abbreviated (A-A), (2) graph-abbreviated and TOLD-extended (A-E), (3) graph-extended and TOLD-abbreviated (E-A), and (4) graph-extended and TOLD-extended (E-E).

Learning outcome measures. The tutor's curriculum covered a wide range of knowledge and skills and was generally graph-intensive. Although some of the introductory graph-related material was familiar to many individuals (constituting a review), the vast majority of the curriculum was related to flight engineering and thus constituted new knowledge and skills. This was further ensured because subjects were initially screened to eliminate anyone from participating in the study who had prior training or experience in the area. Because the material was unfamiliar, we chose not to include a pretest of flight engineering knowledge and skills, as we believed subjects would have scored at about chance levels of performance.

A posttest was created to assess the depth and breadth of flight engineering knowledge and skills acquired from the tutor. There were three parts to this test: declarative knowledge about flight engineering, procedural knowledge about flight engineering, and graph reading/interpretation of Cartesian coordinate grids, maximum allowable crosswind charts, and wind components charts. An example of a procedural-knowledge item from the test is "If the relative wind direction is 100 deg and the wind velocity is 40 knots, what other piece of information is needed to determine the tailwind component? (a) gust increment, (b) runway heading, (c) crosswind component, (d) altitude pressure, (e) runway condition reading, (f) none of these options." (The correct answer is a.) A detailed description of all test items can be found in Shute (1993).

There were 46 items altogether in this posttest, which was a multiple-choice format with six alternative answers to choose from. As already mentioned, pretest data were not collected from subjects in this study, so to establish a needed baseline measure of performance, we collected test data from a different group of subjects who came from the same population as those in the current study (i.e., temporary service employees, matched on cognitive and demographic characteristics to Experiment 1 subjects). The mean pretest score from the control group was 39% (SD = 11.5%, n = 59).

The posttest was administered on the computer at the conclusion of the tutor, and it differed from the tutor learning environment in the following manner. Whereas the tutor contained a variety of tools to assist learners in the extraction of information from the different graphs (e.g., "plot point," "draw vector"), subjects had to solve posttest problems without the aid of tools (i.e., a "near transfer" task).

Procedure. Subjects were tested in groups of approximately 20 persons, and a total of 20 groups were tested. (Some individuals dropped out before completing all parts of the study.) Each group spent up to seven days (six hours per day) in this study, depending on how quickly they were able to complete the tutor. The study began with testing subjects on basic cognitive process measures and collecting demographic data. Then subjects were randomly assigned to one of four conditions (A-A, A-E, E-A, or E-E). The number of problems that subjects had to solve in each of these conditions were 66, 147, 183, and 264, respectively. This distribution was based on the number of problem sets per tutor part (graph = 13 and TOLD = 9) and whether subjects were learning from the abbreviated (3 problems) or extended (12 problems) version of that section. Subjects were not aware that there were different practice conditions, nor were they allowed to discuss their experiences with other subjects. Subjects were paid for completing the study rather than being paid by the hour, so they were motivated to proceed at expeditious rates. Directly following the completion of the tutor, subjects were administered the posttest.

Results

Before testing our hypotheses, we examined cognitive and demographic data to ensure that subjects assigned to the four practice conditions were similar. This was an important test because, in the absence of pretest data from this particular group of subjects, we needed to ascertain that the groups were comparable in terms of important aptitude and background indices. We compared the groups on the following information: working memory capacity, general knowledge, inductive reasoning skills, fact (or associative) learning skills, and procedural learning skills. (For more information about the computer-administered cognitive tests, see Shute, 1993.) In addition, we looked at age, years of school completed, number of algebra and geometry courses taken, number of programming courses taken, and gender. No significant differences were found among subjects in the four practice conditions on any of these cognitive and demographic variables prior to learning from the tutor. These data may be seen in Table 1. Although this is not shown in the table, we also analyzed only the quantitative sections of the cognitive test data, but even this refined analysis failed to show any differences among groups.

Next, we examined subjects' total time-on-

TABLE 1

Cognitive and Demographic Data, Separated by Practice Condition

Variable	*A-A* *(n = 81)*	*A-E* *(n = 91)*	*E-A* *(n = 81)*	*E-E* *(n = 83)*
Working memory	61.97 (17.52)	63.83 (18.73)	65.58 (18.01)	62.84 (17.52)
General knowledge	72.14 (12.16)	71.38 (13.08)	72.44 (12.40)	69.90 (12.47)
Inductive reasoning	57.54 (16.90)	59.82 (17.10)	58.72 (16.07)	57.61 (16.24)
Fact learning	68.33 (08.40)	68.77 (09.60)	69.94 (09.15)	67.76 (10.29)
Skill learning	78.47 (10.10)	78.26 (11.25)	79.69 (10.75)	77.83 (11.89)
Age	22.10 (2.80)	21.39 (2.49)	22.00 (2.91)	22.26 (2.99)
School (years)	12.88 (1.47)	12.89 (1.33)	12.89 (1.33)	12.58 (1.45)
Algebra & geometry courses	2.34 (1.29)	2.61 (1.27)	2.31 (1.47)	2.49 (1.39)
Programming courses	0.49 (1.06)	0.50 (0.98)	0.43 (1.01)	0.38 (1.05)
Gender (prop. female)	0.21 (0.41)	0.31 (0.47)	0.20 (0.40)	0.26 (0.44)

Notes. The first letter of each group refers to the condition of the graph section of the tutor (A = abbreviated and E = extended). The second letter refers to the condition of the TOLD section (same abbreviations). Standard deviations are given in parentheses. *N*'s may differ among tables because of missing data.

TABLE 2

Mean Time on Tutor (h), Separated by Practice Condition

Graph Section	*TOLD Section*	*Mean*	*SD*	*n*
Abbreviated	Abbreviated	7.07	3.16	91
Abbreviated	Extended	9.50	3.94	92
Extended	Abbreviated	8.96	4.56	90
Extended	Extended	11.68	4.75	91

tutor data by the four practice conditions. The average tutor completion times are presented in Table 2. An ANOVA computed on these data revealed significant group differences, $F(3,360) = 18.99$, $p < 0.001$. As can be seen in Table 2, the largest difference was between the most contrastive practice conditions: A-A and E-E (with mean completion times of 7.1 and 11.7 h, respectively). However, the ordering of tutor completion times was not strictly a function of number of problems to solve. That is, the A-E and E-A groups received 147 and 183 problems, respectively, yet the A-E group required 9.5 h and the E-A group only 9 h to complete the tutor. This difference was not significant, $F(1,180) = 0.73$, $p = 0.394$.

Before addressing how well the respective groups performed on the posttest, we needed to determine whether the posttest items were reliable indices. Results of an item analysis computed on the posttest data showed that there were no exceptionally poor items—that is, there were no items that all subjects got right or wrong; most had around 50% accuracy (lowest item accuracy was 11% and highest was 91%). Moreover, the items within the test were reliable: odd-even reliability = 0.91 overall. The data were normally distributed, with a mean = 57.80% and *SD* = 20.66%. As stated earlier, the control group's pretest mean was 39% (*SD* = 11.5%); thus most of the subjects did appear to learn something from the tutor (estimated gain was almost 20%).

We next tested whether subjects in the extended condition would demonstrate higher outcome scores than would subjects learning in the abbreviated condition. Table 3 shows the mean scores and Table 4 shows the mean latencies for each individual test, as well as the overall posttest score (or latency), separated by the four versions of the tutor.

These results were surprising. There were no significant differences among the practice conditions on any of the posttest scores or latency data. In fact, the criterion data among the four groups were strikingly similar.

The first set of questions investigated group differences at the end of learning, or how practice affected what learners took with them. The next set of questions addressed differences among groups during the acquisition process, testing whether subjects in the abbreviated condition required more time to complete problems and committed more errors than did extended-condition subjects. We examined data on the problem set level and discuss each section of the tutor separately.

TABLE 3

Mean Percentage Correct Scores on Posttests, by Practice Condition

Posttest	*A-A (n = 91)*	*A-E (n = 92)*	*E-A (n = 86)*	*E-E (n = 87)*	*F*	*Signif.*
Declarative knowledge	61.44 (23.76)	60.77 (25.25)	60.57 (26.74)	61.02 (24.05)	0.02	*ns*
Procedural knowledge	46.92 (22.59)	47.83 (25.02)	49.30 (27.69)	48.62 (24.17)	0.15	*ns*
TOLD sheet performance	62.81 (19.28)	65.96 (19.23)	62.60 (22.13)	65.75 (19.39)	0.73	*ns*
Overall average	57.06 (19.23)	58.18 (20.37)	57.49 (23.19)	58.46 (20.06)	0.10	*ns*

Notes. The first letter of each group refers to the condition of the graph section of the tutor (A = abbreviated and E = extended). The second letter refers to the condition of the TOLD section (same abbreviations). Standard deviations are given in parentheses. The *F* tests are all based on 3 and 352 degrees of freedom.

TABLE 4

Mean Latency per Posttest Section (s), by Practice Condition

Posttest	*A-A* *(n = 91)*	*A-E* *(n = 92)*	*E-A* *(n = 86)*	*E-E* *(n = 87)*	*F*	*Signif.*
Declarative knowledge	228.80 (70.9)	226.45 (74.3)	216.00 (93.2)	211.96 (73.3)	0.92	*ns*
Procedural knowledge	311.85 (128.5)	307.51 (103.2)	281.56 (120.0)	288.42 (102.2)	1.39	*ns*
TOLD sheet performance	660.10 (214.4)	639.43 (182.1)	616.53 (203.6)	652.90 (218.6)	0.73	*ns*
Overall average	1282.58 (379.4)	1254.87 (338.44)	1173.47 (388.1)	1230.48 (375.0)	1.38	*ns*

Notes. The first letter of each group refers to the condition of the graph section of the tutor (A = abbreviated and E = extended). The second letter refers to the condition of the TOLD section (same abbreviations). Standard deviations are given in parentheses. The *F* tests are all based on 3 and 352 degrees of freedom.

Graph Section: Component Skill Acquisition

Figure 2 shows problem-solving latency and accuracy data (i.e., proportion of correct responses divided by the total number of attempts) across each of the 13 problem sets in the graph section of the tutor. We will examine only two conditions in the graph section because the TOLD condition would not have affected graph section learning. Problem set clusters are joined by lines in the figure. For instance, G-4, G-5, and G-6 represent graph problem sets 4, 5, and 6, which are related to solving Cartesian coordinate grid problems. For each condition (i.e., abbreviated and extended) problem-solving data for the first three problems were summed (for latency) or averaged (for accuracy) and then compared.

When we computed ANOVAs on the latency and accuracy data, we found significant differences between conditions, but only for the later, more difficult problem sets within the tutor (i.e., G-9 to G-13). These results are summarized in Table 5. In all cases of significant differences, subjects in the abbreviated condition were less accurate and required more time than the extended-condition subjects.

A more detailed view of these acquisition data (i.e., learning curves) is shown in Figure 3, where the two groups' latencies are depicted across trials (3 vs. 12) for each of the 13 problem sets. What can be clearly seen is the trial number at which these data appear to stabilize. For the abbreviated-condition group, the data typically show a continuing downward trend across the three problems (i.e., learning is still in progress), whereas data for the extended-condition group seem to asymptote around Trial 4 or 5 for most of the problem set data, indicating the point at which proceduralization commences.

TOLD Section: Performance Measures

We first tested the hypothesis that differential practice opportunities for learning component skills (i.e., graph section condition, abbreviated or extended) would influence TOLD performance. We computed a TOLD performance measure for each subject, multiplying standardized latency and error data (where bigger numbers reflected poorer performance). Next, we computed an ANOVA comparing this TOLD performance index by both graph and TOLD practice conditions. Results showed that prior graph practice did not significantly affect TOLD section performance, $F(1,350) = 0.17$ (*ns*), $p = 0.679$, but practice within the TOLD condition did: $F(1,353) = 5.99$, $p < 0.02$. The interaction was not significant, $F(1,350) = 0.04$, $p = 0.852$. The following analyses of TOLD data are thus discussed in terms of only two conditions (TOLD section, abbreviated or extended), rather than the four possible conditions (A-A, A-E, E-A, and E-E).

Data on problem-solving latency and number of errors committed are presented in Figure 4 for each of the nine problem sets in the TOLD worksheet section of the tutor. Similar to Figure 2, these data were summed (for latency) and averaged (for accuracy) across the first three trials per problem set and separated by practice condition. It is interesting to note that by the end of the graph section problem sets (shown in Figure

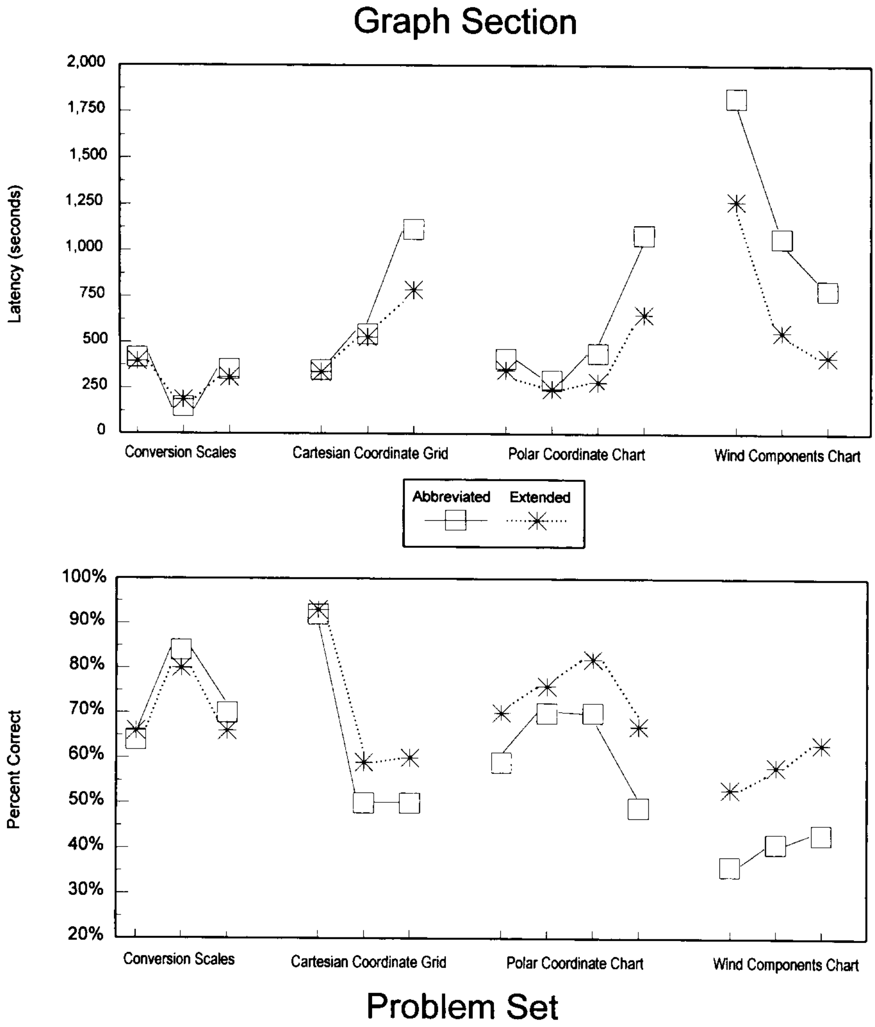


Figure 2. *Graph section latency and accuracy data, collapsed across 13 problem sets.*

2), there were large performance differences between conditions, but at the beginning of the TOLD section, these differences (shown in Figure 4) had disappeared. The two groups seemed to acquire the subject matter comparably until about the middle of the TOLD section curriculum. As in Figure 2, problem set clusters are joined by lines (e.g., T-7, T-8, and T-9 represent performance measures of integrating all previous knowledge and skills and deciding on whether or not it is safe to take off).

Table 6 summarizes the results of ANOVAs computed on the latency and accuracy data per problem set, by practice condition. As shown in the table, significant differences between conditions emerged for the more difficult problem sets within the tutor. Again, subjects in the abbreviated condition spent more time solving these problem sets and committed more errors than did those in the extended condition.

In addition to the collapsed data, we present the specific acquisition data in Figure 5. The two groups' learning curves are plotted across trials (3 vs. 12) for each of the nine problem sets. Again, similar to the graph section data, we see that the abbreviated-condition groups'

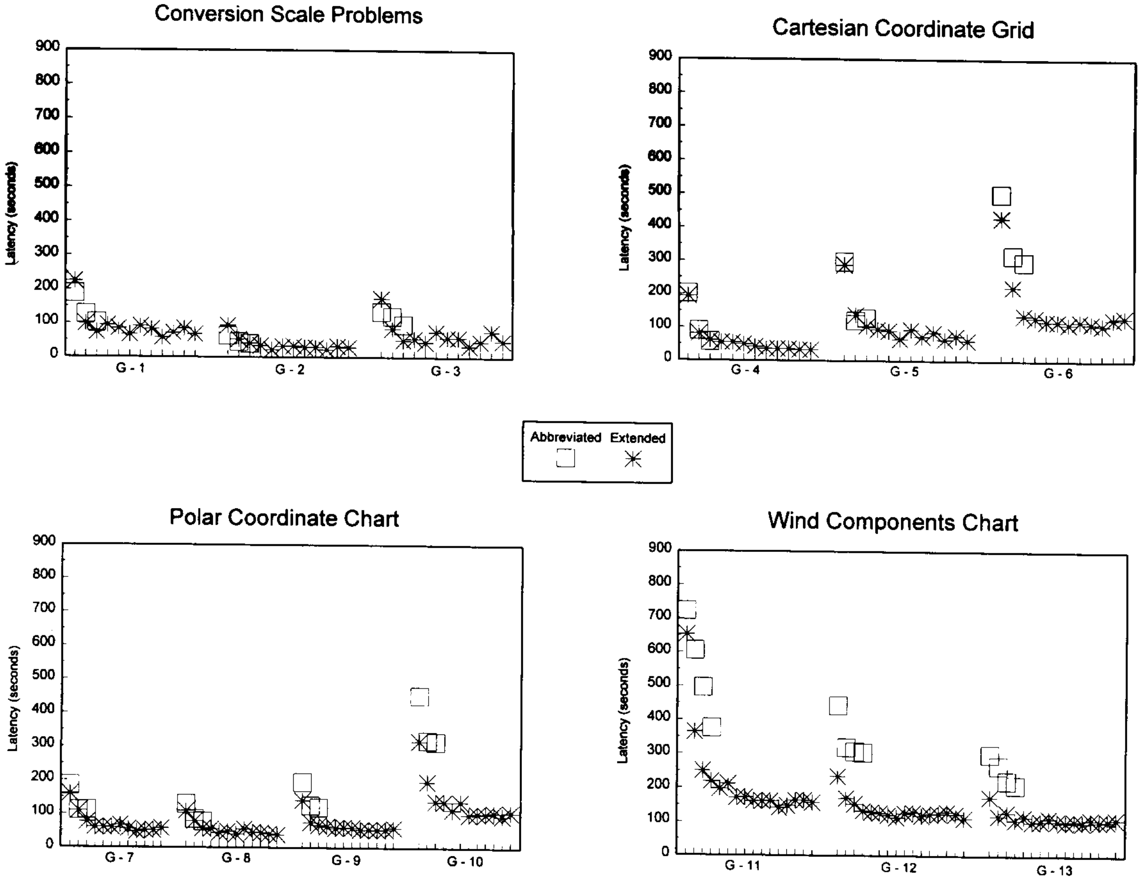


Figure 3. *Graph section learning curves; all problems across 13 problem sets.*

acquisition latencies continue to decrease across their three trials, whereas the extended-condition groups' data appear to asymptote somewhere between Trials 4 and 7, particularly for the most difficult problem sets.

Because the posttest assessed tutor-instructed knowledge and skills, and subjects in both groups performed comparably, abbreviated-condition subjects appeared to acquire knowledge and skills from the tutor without the benefits of the extra practice afforded to the extended-condition group. Other than spending more time (and making more errors) solving the more complex problem sets, did the abbreviated-condition subjects do anything different from the extended-condition subjects during the learning process that would serve ultimately to attenuate outcome differences? We decided to investigate subjects' utilization of the help features of the tutor (i.e., a post hoc analysis).

Self-Help Activities

The computer automatically tallied time-stamped data on subjects' use of the elective, self-help activities across all problem sets. We computed proportions from these data by dividing the time spent asking for help, previous-paging, and using the dictionary by the total time spent on the tutor. Proportions were necessary because time on tutor differed for everyone, within and between conditions. We tested whether the abbreviated-condition groups' greater problem-solving latencies were attributable to their spending a greater proportion of time using the system's help features compared with the extended-condition group. Table 7

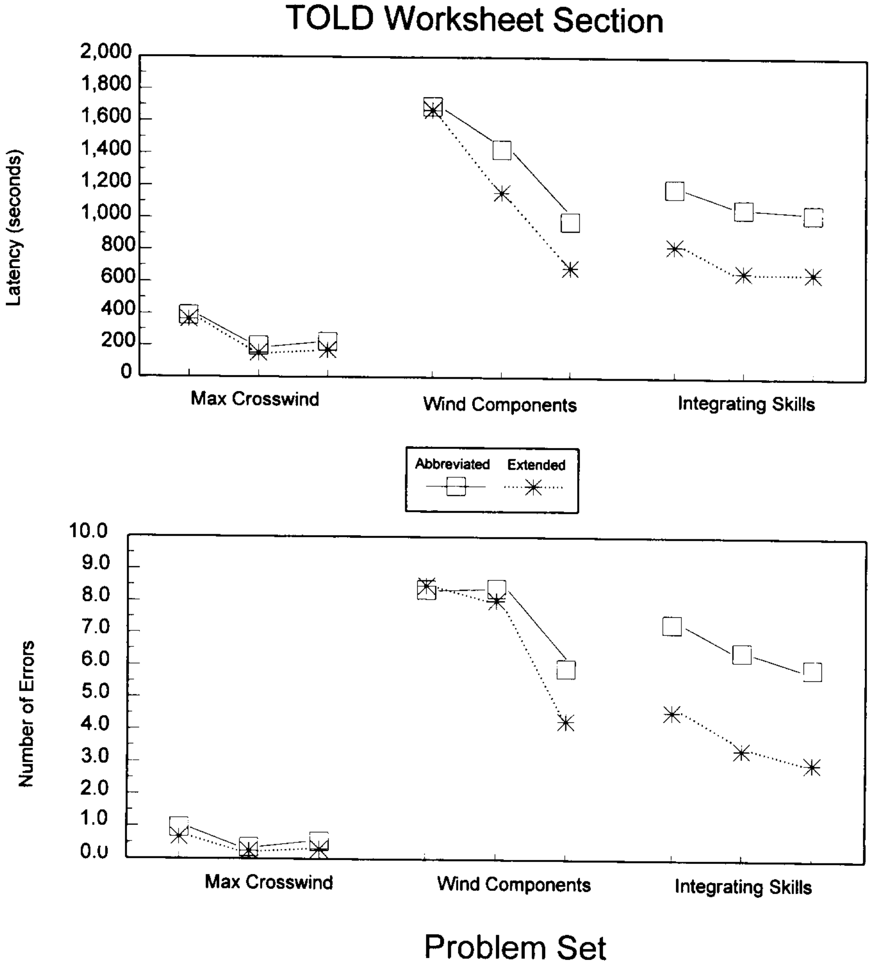


Figure 4. *TOLD section latency and error data, collapsed across nine problem sets.*

shows these data (percentage of self-help time in relation to total learning time) separated by practice condition for each section of the tutor. Clearly, the abbreviated-condition group devoted a significantly higher percentage of time seeking assistance from the tutor than did the extended-condition group. In fact, overall, they spent almost twice (>1.8) the percentage of time engaged in self-directed, supplemental instruction compared with the extended-condition group.

Discussion

When a criterion task represents an unfamiliar, complex domain, students are not expected to know much about it. We did not expect any of our subjects to have flight engineering knowledge; in fact, we screened subjects to make sure they did not. Consequently, we believed that if we administered an unsolvable pretest, it might negatively predispose learners to the experiment. However, we need to point out that the posttest-only design used in this study presents a potential problem in interpreting results. In the absence of a pretest, any group differences (or lack thereof) on the posttest could be attributed either to treatment conditions or to group selection differences. Although it is impossible to separate out treatment and selection effects in this design, one suggested remedy is to measure

TABLE 5

Graph Section ANOVAs across Problem Sets (Latency and Accuracy)

	Latency		*Accuracy*	
Problem Set	*F test (df = 1,361)*	*Signif.*	*F test (df = 1,362)*	*Signif.*
G-1	0.15	*ns*	0.46	*ns*
G-2	2.44	*ns*	3.21	*ns*
G-3	2.64	*ns*	2.69	*ns*
G-4	0.21	*ns*	0.29	*ns*
G-5	0.10	*ns*	11.37	$p < 0.001$
G-6	13.38	$p < 0.001$	9.20	$p < 0.005$
G-7	3.59	*ns*	14.61	$p < 0.001$
G-8	2.21	*ns*	4.73	$p < 0.05$
G-9	13.24	$p < 0.001$	16.34	$p < 0.001$
G-10	26.67	$p < 0.001$	32.27	$p < 0.001$
G-11	24.23	$p < 0.001$	36.36	$p < 0.001$
G-12	38.25	$p < 0.001$	34.78	$p < 0.001$
G-13	31.74	$p < 0.001$	71.79	$p < 0.001$

"proxies" for pretest assessments (Cook, Campbell, and Peracchio, 1990), which may be measures of demographic data, cognitive data, or other relevant information. As mentioned earlier, we collected and compared both cognitive and demographic data and found no significant differences among our four conditions; thus we were fairly confident that our groups were comparable. We realize that using proxies is a weak substitute for an actual pretest of knowledge relevant to the domain, and that is why we later collected pretest data from a control group of subjects who were matched cognitively and demographically with our original subjects.

In this study practice effects were investigated in relation to overall learning time, learning outcome, and parameters of skill acquisition (i.e., latency and accuracy data within the tutor). Our original prediction was straightforward: Abbreviated-condition subjects would take less time to complete the tutor than would extended-condition subjects but would perform worse on the outcome test. The first part of the hypothesis was supported, but the second part was not. That is, subjects in the abbreviated condition did require significantly less time, overall, to complete the tutor than did subjects in the extended condition, but that was no surprise,

TABLE 6

TOLD Section ANOVAs across Problem Sets (Latency and Accuracy)

	Latency		*Accuracy*	
Problem Set	*F test (df = 1,358)*	*Signif.*	*F Test (df = 1,358)*	*Signif.*
T-1	0.45	*ns*	3.58	*ns*
T-2	18.53	$p < 0.001$	1.62	*ns*
T-3	20.03	$p < 0.001$	5.00	$p < 0.05$
T-4	0.06	*ns*	0.05	*ns*
T-5	11.64	$p < 0.001$	0.31	*ns*
T-6	32.98	$p < 0.001$	7.92	$p < 0.01$
T-7	49.21	$p < 0.001$	18.55	$p < 0.001$
T-8	73.12	$p < 0.001$	24.56	$p < 0.001$
T-9	80.63	$p < 0.001$	26.82	$p < 0.001$

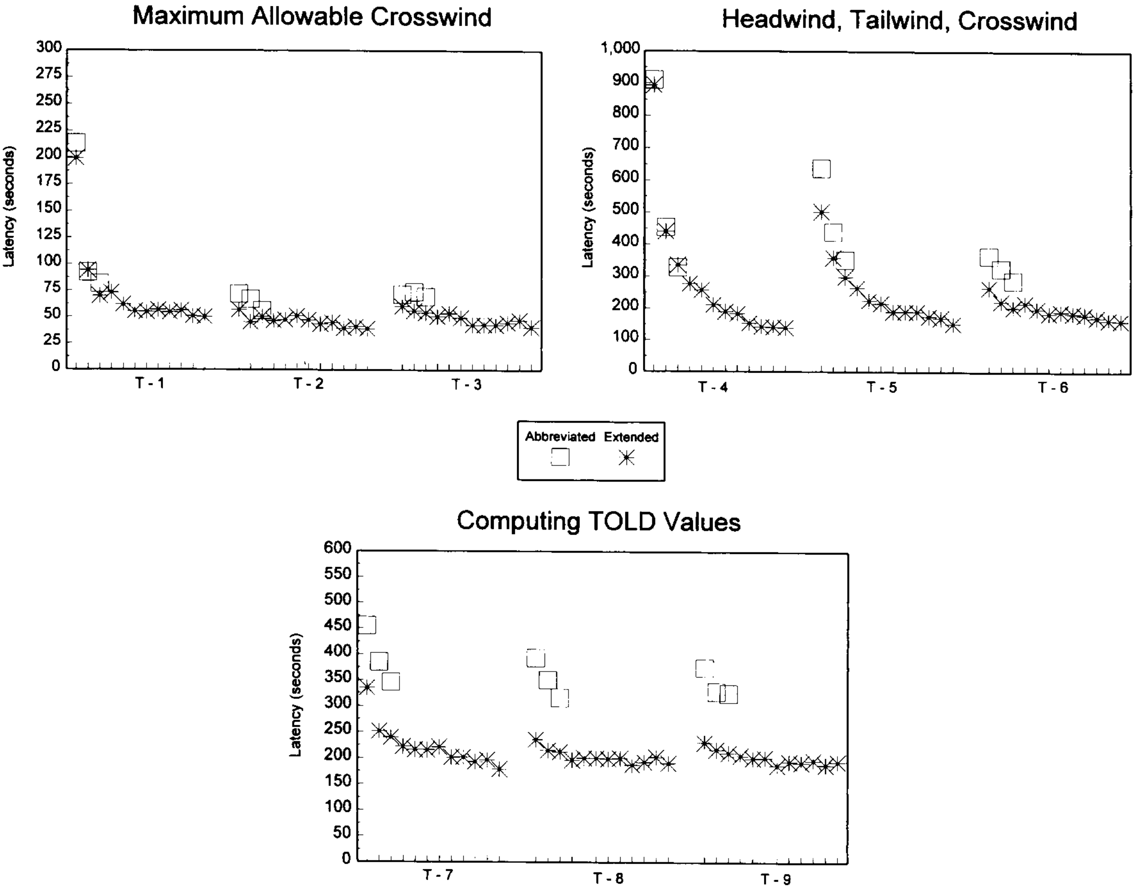


Figure 5. *TOLD section learning curves; all problems across nine problem sets.*

considering that they had fewer problems to solve. What was surprising was that the four groups performed the same on all posttest measures (accuracy and latency data).

These findings can be clarified if we liken our extended and abbreviated conditions to massed and distributed practice schedules, respectively. Enhanced memory performance has been well documented for items whose presentations are more distributed than massed (see Glenberg, 1979). Our extended condition maps well onto a massed practice schedule (i.e., solving 12 contiguous problems per problem set), but mapping the abbreviated condition onto a distributed

TABLE 7

Mean Time on Tutor Seeking Assistance (%), by Practice Condition

Tutor Assistance	*Abbreviated*	*Extended*	*F Test*	*Signif.*
Graph section	(*n* = 181)	(*n* = 183)		
Help	2.23 (1.80)	1.39 (1.24)	26.25	$p < 0.001$
Previous page	5.43 (3.90)	2.99 (2.23)	53.56	$p < 0.001$
Dictionary	0.15 (0.34)	0.07 (0.17)	7.03	$p < 0.01$
TOLD section	(*n* = 181)	(*n* = 183)		
Dictionary	1.79 (1.70)	0.78 (0.86)	18.96	$p < 0.001$

practice schedule requires a consideration of their problem-solving activities. That is, abbreviated-condition subjects (A-A) received one-quarter the problems of the extended-condition group (E-E), but they did not complete the tutor in one-quarter the time; rather, they required more than half (0.61) the extended-condition group's time to complete the tutor. This finding is consistent with studies showing that massed items receive fewer overt rehearsals than distributed items (Ciccone and Brelsford, 1974; Rundus, 1971).

Moreover, a greater proportion of the abbreviated-condition group's time was spent obtaining supplemental assistance from the tutor, and their extracurricular activities could have resulted in a roughly distributed practice schedule as they interspersed problem-solving episodes with previous-paging through the curriculum, accessing the on-line dictionary, and so on. These activities, in turn, would have resulted in more elaborative processing, per problem, which produces redundant traces and, hence, better memory performance (Anderson, 1983; Crowder, 1976; Glenberg, 1979). Abbreviated-condition subjects appeared to have self-regulated their learning behaviors in response to deficient instruction, and this paid off in outcome performance levels that were comparable to those of subjects enjoying the benefits of more extended practice.

In conclusion, results from Experiment 1 suggest that extended practice environments may not always result in greater learning outcomes, at least not when immediately assessed or when self-help options are available. To determine the extent and limitations of these findings, we next discuss practice effects on retention following an extended time. Would subjects who learned the subject matter in the abbreviated condition still show performance equal to that of extended-condition subjects at some point in the future, or would we begin to see the cost of decreased practice manifest itself in terms of poorer retention of knowledge and skills and/or decreased sensitivity to relearning? Experiment 2 was conducted to answer these questions.

EXPERIMENT 2

The purpose of Experiment 2 was to test long-term effects of practice on retention of skills and sensitivity to relearning. To accomplish this we recalled subjects who had participated in Experiment 1, invited them back, and then retested them on various outcome measures before and after a refresher course. The main hypothesis was that subjects who had originally learned in the extended condition, although not showing any immediate outcome advantages, would ultimately show better retention of, and greater sensitivity to, relearning knowledge and skills acquired two years before when compared with abbreviated-condition subjects. Determining the factors that influence the degree to which subjects retain knowledge and skills, and also the rate by which old information can be relearned, has relevance to instructional methods for students and trainees across a wide array of settings (e.g., academic, military).

Method

Subjects. Subjects were 92 men and women participating in a two-day study on long-term retention of flight engineering knowledge and skills. All subjects from Experiment 1 who had not moved or changed phone numbers were contacted and invited to return for this follow-up study. Of these, 34 were able to return and participate in Experiment 2. The subjects in this returning group were distributed among the former practice conditions as follows: A-A (n = 13), A-E (n = 10), E-A (n = 6), and E-E (n = 5). We examined cognitive and demographic data of these subjects to determine whether they were representative of the population from Experiment 1. No significant differences were found between these returning subjects and the entire sample from Experiment 1 in terms of our measures of cognitive skills (e.g., working-memory capacity, fact-learning skills, and general knowledge) or demographic data (e.g., age, years of school), nor were there any cognitive differences among the four conditions within this sample of returning subjects.

From a local temporary employment agency, we also obtained 58 subjects who served as a control group. These subjects represented the same population as those employed in Experiment 1. The control subjects (like returning subjects) were all high school graduates with a mean age of 23 years and no prior experience or training as flight engineers or pilots. All subjects were paid for their participation, consisting of 16 h of testing and learning. The sample was 53% male and 47% female.

Refresher course. A shortened (2-h) version of the tutor used in Experiment 1 was developed at the Armstrong Laboratory (Shute and Walker, 1992) and designed as a review of previously instructed flight engineering material. It was written in Visual Basic 2.0 and delivered on Compaq 486-33 microcomputers (800 × 600 resolution SVGA Nec/MultiSync 4DS monitors). The original tutor's main goal was to enable learners to interpret and correctly fill out a TOLD worksheet. Thus the curriculum underlying the refresher course focused on completing an on-line TOLD worksheet. It consisted of 10 problem sets (compared with 22 in the full tutor), starting with material from problem set G-7 in the original tutor (relative wind direction and wind type) and continuing to the end of the original curriculum (T-9). The final problem set of the refresher course was comprehensive, requiring subjects to integrate all knowledge and skills from earlier problem sets, as in the final problem set cluster of the original tutor.

Unlike the original tutor, the refresher course had only one practice condition—that is, each subject had the same number of practice problems to solve (four) for each of the 10 problem sets. Subjects had to answer all problems correctly before proceeding to the next problem set. If a subject missed a problem after three attempts, the computer provided the correct answer and the subject had to apply it. Furthermore, no explicit Help function was available, as there was in the original tutor, but on-line tools were available to assist problem solution (e.g., horizontal ruler, vertical ruler, and plot vector), and subjects could use a Previous Page option to go back and read previous parts of the curriculum.

Learning outcome measures. For the criterion test in Experiment 2 we created isomorphic items (in duplicate) for each of the three parts of the posttest used in Experiment 1: (a) declarative knowledge about flight engineering, (b) procedural knowledge about flight engineering, and (c) graph reading and interpreting the maximum allowable crosswind and wind components charts. These isomorphic items were carefully created to be as similar as possible to the original items, differing only in specific content (e.g., the numbers to be manipulated).

Because we failed to collect pretest data in Experiment 1, we were unable to analyze pretest-to-posttest changes in learning within that study. Consequently, in this experiment we administered parallel forms of the test (A and B) as pretest and posttest using the duplicate, isomorphic items, mentioned earlier. These parallel forms were administered prior to and at the end of the refresher course.

Procedure. Subjects were tested in five groups of approximately 20 persons each. (Several subjects dropped out before completing the study.) Each group spent two days (about 8 h per day) in this study. Subjects began the study being tested on flight engineering knowledge and skills, completing one of the two forms of the test, which were counterbalanced (i.e., half the subjects received Form A as a pretest and Form B as a posttest and the other half received the tests in the reverse order). Next, subjects spent approximately 2 h learning from the refresher course (which, for the control subjects, was not a refresher but all new material). Immediately after the refresher course, subjects were administered the posttest.

Results

Because no main effect was attributable to test form (A and B), all ensuing data analyses were collapsed across form. The first hypothesis related only to the returning subjects, so we tested whether those who had originally received more extended practice opportunities would show greater retention of information, compared with subjects who had learned from a more abbreviated practice condition. *Retention* was defined as the degree to which subjects remembered information learned two years previously, and it was operationalized as the difference between Experiment 1 posttest score and Experiment 2 pretest score. Specifically, we computed a new retention variable, which was the Experiment 2 pretest score, holding the posttest score from Experiment 1 constant (i.e., included as a covariate in the equation to control for any differences on the original posttest score). We expected the following ordering of conditions by degree of retention: A-A < A-E and E-A < E-E. (We combined the mixed groups' data, as we had no a priori reason to expect any differences between their scores. This decision was validated by the fact that their adjusted pretest scores were equivalent: mean A-E = 52.4% and mean E-A = 52.9%.) Table 8 shows the returning subjects' posttest scores from Experiment 1 and their Experiment 2 pretest and posttest scores by former practice condition, as well as the control group's pretest data. Figure 6 plots these data by condition.

First, the subset of returning subjects, like our full sample from Experiment 1, showed no significant group differences on original posttest data, $F(2,30) = 0.82, p = 0.450$. This was encouraging, as it supported our premise that the returning group was representative of the larger sample from which they came. Next, we computed an analysis of covariance (ANCOVA) on the returning subjects' retention measure—Experiment 2 pretest data controlling for Experiment 1 posttest data. Results showed a significant difference, overall, among groups, $F(2,30) = 4.16, p < 0.03$, but the pretest means for the three groups (with posttest score partialed out) were not in the predicted direction: A-A = 45.01%, A-E and E-A = 52.54%, and E-E = 44.14%. That is, the mixed group retained more knowledge and skills compared with both the E-E group, $F(1,18) = 5.30, p < 0.04$, and the A-A group, $F(1,26) = 6.19, p < 0.02$. There was no significant difference in retention between the

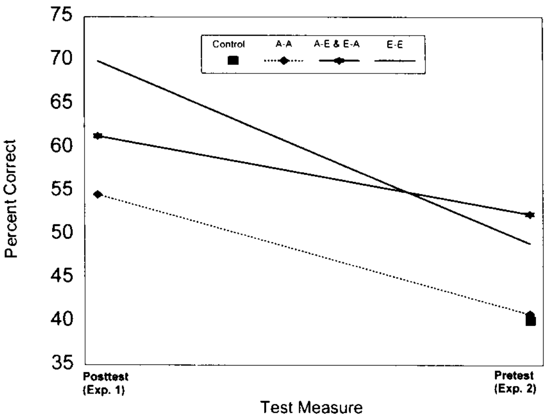


Figure 6. *Plot of returning subjects' retention data, by condition.*

TABLE 8

Retention: Actual Mean Scores by Group for Posttest (Experiment 1), Pretest, and Posttest (Experiment 2)

Test % Correct	*A-A (n = 13)*	*A-E & E-A (n = 16)*	*E-E (n = 5)*	*Control (n = 59)*
Posttest, Exp. 1	54.64 (26.1)	61.39 (20.2)	69.26 (14.4)	—
Pretest, Exp. 2	40.86 (16.3)	52.37 (12.7)	48.89 (12.9)	39.10 (11.5)
Posttest, Exp. 2	59.32 (23.3)	69.93 (20.3)	69.33 (16.5)	58.19 (17.3)

Notes. Data in this table reflect actual, not adjusted, mean scores from the various tests. We defined a new variable, *elapse*, representing the number of days intervening between the end of Experiment 1 and start of Experiment 2 ($M = 635$ days, $SD = 90$, min. = 525, and max. = 821). Although elapse was not needed as a covariate in posttest data (Experiment 1), it was used in one of the two ANCOVAs computed on the pretest data (Experiment 2, matched sections). As a covariate, elapse was not significant in the ANCOVA discussed in the text; that is, there was no main effect attributable to this variable, $F(1,30) = 0.001$, *ns*.

A-A and E-E groups, $F(1,15) = 0.02$, $p = 0.889$, so subjects in the most extreme practice conditions (A-A and E-E) retained about the same extent of knowledge and skills following a two-year period (and controlling for posttest score), whereas subjects who had previously learned under the mixed-practice conditions (A-E and E-A) retained the greatest degree of knowledge and skill. Thus our hypothesis concerning the delayed effects of amount of practice on retention was not supported.

With regard to our second hypothesis, we tested whether subjects who had received extra practice while learning flight engineering knowledge and skills would show greater sensitivity to relearning after exposure to the refresher course. We defined *sensitivity to relearning* as the time it took learners to reacquire knowledge and skills back to preretention-interval performance levels (i.e., obtaining at least the same posttest scores as obtained in Experiment 1, following refresher course instruction). Experiment 2 posttest data are presented in Table 8. In all cases, outcome scores were the same (or slightly higher) following the refresher course compared with Experiment 1 posttest data, so all groups achieved our criterion of returning to preretention-interval performance levels.

Given the design of the refresher course used in Experiment 2, subjects had to demonstrate mastery of each problem set before advancing to the next one. Therefore, the rate at which one attained mastery (and, hence, completed the self-paced refresher course) indicated reacquisition efficiency, or how quickly one remembered and could apply previously instructed knowledge and skills. We expected reacquisition to be a function of differential practice opportunities provided in Experiment 1. Specifically, we predicted that time to complete the refresher course would be ordered as follows: E-E < A-E and E-A < A-A < control. We further predicted that all returning subjects would complete the refresher course faster than a control group. Table 9 contains the data for completion times by group.

There were no significant differences among groups in the amount of time needed to complete the refresher course, $F(2,31) = 0.65$, $p = 0.530$. We also computed a *percentage-of-savings* score for each group, defined as the relationship between new learning time and original acquisition time—that is, [(original time − new time)/original time] × 100. This new savings variable was quite large in all cases and did show significant differences in the predicted direction: A-A = 78.30% < mixed = 81.98% < E-E = 87.55%, $F(2,31) = 6.68$, $p < 0.005$. Furthermore, there was no difference between the two conditions comprising the mixed group on this savings index (A-E = 81.7% and E-A = 82.3%), so combining their data was again warranted. Finally, as expected, the returning subjects, overall, completed the refresher course faster

TABLE 9

Mean Time to Complete Refresher Course, Original Tutor, and Percentage Savings, by Condition

Acquisition Variable	*A-A (n = 13)*	*A-E & E-A (n = 16)*	*E-E (n = 5)*	*Control (n = 59)*
Refresher course hours:mean (*SD*)	1.60 (0.67)	1.42 (0.43)	1.32 (0.32)	1.97 (0.78)
Original tutor hours:mean (*SD*)	7.38 (4.17)	7.88 (2.63)	10.60 (5.08)	—
Percentage savings (new/old)	78.30% (6.05)	81.98% (5.71)	87.55% (2.96)	—

Notes. Data representing the original tutor completion times (above, in hours) may differ slightly from the data presented in Table 2 because in this table we report only the data from our subset of returning subjects. Percentage savings was computed as follows: Original time to complete the tutor, minus time to attain preretention interval mastery levels (or higher) on the refresher course, divided by original hours, and multiplied by 100 (see Adams, 1980).

than the control group, $F(1,90) = 11.04$, $p < 0.01$.

Discussion

In Experiment 2 we tested the hypothesis that practice effects, though not immediately discernible, would show up as long-term differences in retention and/or sensitivity to relearning. We recalled subjects from Experiment 1 and compared their original posttest data scores to pretest data from Experiment 2, looking for differences in degree of forgetting (or retention) as a function of original practice condition. We predicted that subjects in the extended condition would show greater retention of the material than would abbreviated-condition subjects following a two-year interval. Significant differences were found, but not in the predicted direction. That is, the A-A group's degree of retention was comparable to that of the E-E group (in terms of adjusted pretest means), and another unexpected finding was that the mixed-practice group (A-E and E-A) showed the greatest retention of all.

The small sample participating in Experiment 2 (especially the E-E group, 5) presented a potential problem in terms of data analysis. We first ascertained that the returning subjects were representative of our sample from Experiment 1 in terms of comparable cognitive and demographic data, tutor completion times, and Experiment 1 posttest performance measures. Next, given the small sample size and our finding of significant group differences on the retention variable, we needed to address the issue of the power of our statistical tests to detect these differences. Overall, the power was low (0.20). Thus it is possible that the small sample influenced our obtained effect size (with a mixed-group advantage), which resulted in our low power. With this in mind, we decided to accept our statistical findings with the understanding that, had our sample size been larger, this may have resulted in even more dramatic effects. Consequently, what is needed is a replication of these findings; another retention study is planned using a larger sample (see the section on future research).

We also tested whether extended-condition subjects would show a greater sensitivity to relearning than would abbreviated-condition subjects. This hypothesis was only weakly supported: There were no significant group differences in total time to complete the refresher course, but group differences were found in regard to the percentage-of-savings scores. Savings scores were all high and ordered in the predicted direction. However, even though all groups required about the same amount of time to complete the refresher course, the most extended-condition group would naturally show the highest savings score because it is only an artifact of original learning time. That is, the E-E group originally required about 11 h to complete the tutor because they had so many problems to solve, so their relative savings score would be much higher in relation to those of the other groups (and the converse would be true for the A-A subjects).

These findings together suggest that for the cognitive skill instructed in this investigation, practice appears to affect retention in a somewhat counterintuitive way—that is, the extreme practice conditions (A-A and E-E) showed the least retention and the mixed-practice condition showed the most. A cost-benefit analysis suggests that the mixed-practice conditions (A-E and E-A) are optimal if long-term retention is the goal and, further, there is no extra cost in terms of reacquisition time—all groups required about the same amount of time to get back to preretention-interval performance levels.

CONCLUSIONS

In this paper we examined practice effects on complex cognitive skill acquisition—during the learning process, immediately following learning, and after a significant period of time had elapsed. We predicted that, given the same practice schedule (i.e., blocked problem sets) and varying only the number of items per problem set, solving more items should result in superior skill acquisition, transfer, retention, and sensitivity to relearning. In summary, we found that subjects learning under an abbreviated practice condition in Experiment 1 required less time to complete the tutor than did subjects in the extended condition; however, this was an artifact of having to solve fewer problems, and it was also at the expense of problem-solving accuracy within the tutor. Despite their less accurate acquisition performance, the abbreviated-condition subjects managed to end up with the same learning outcome scores as the extended-condition subjects. This initially surprising null finding was not a function of posttest ceiling or floor effects, and it served to motivate Experiment 2, in which we retested a subset of these subjects following a two-year interval.

In the follow-up study, we found evidence for practice effects on long-term retention, but not in the predicted direction. That is, subjects in our mixed conditions (switched midway through the curriculum from one practice schedule to the other) showed significantly greater retention than did subjects assigned to either of the two extreme practice conditions. In terms of our sensitivity to relearning measure, the groups did not differ in actual reacquisition times, but our computed percentage-of-savings scores did show an effect of former practice condition (albeit artifactual).

Experiment 1 findings. One way to interpret the results from Experiment 1 is in terms of differential proceduralization (Anderson, 1987). That is, once some knowledge or skill becomes proceduralized, domain-specific declarative information no longer has to be retrieved from long-term memory and placed into working memory. Instead, the products of these operations are built into new productions, resulting in faster, more accurate performance. Subjects learning from abbreviated practice conditions (solving only three problems before being moved on to a different topic) presumably continued to interpret knowledge and skills declaratively because of insufficient opportunities to proceduralize new information. Thus their acquisition phase appeared slower and less accurate than that of subjects who had more extended practice opportunities and who were able to proceduralize new skills, typically by Trials 4 to 7 (see Figures 3 and 5, extended condition).

Despite acquisition differences, subjects having abbreviated practice opportunities were able to perform the same on all posttest measures as were subjects in extended conditions, most likely because of the extra cognitive effort they exerted during the acquisition process. That is, abbreviated-condition subjects ultimately became facile at (a) retrieving relevant information from long-term memory and placing it into working memory (an especially apt skill at posttest time) and also (b) initiating searches for supplemental sources of information from the tutor, which resulted in redundant traces on items, making them easier to retrieve, as needed. In addition, subjects in the extended practice condition were required to solve problems beyond the point at which their performance levels had asymptoted. This could easily lead to frustration and decreased motivation to attend to the material, which in turn could attenuate outcome performance, bringing it down to the range of the abbreviated-condition group (whose own measures, as discussed, may have been elevated by their increased cognitive efforts during learning). Similarly, Greene (1989) posited a theory of distributed practice that contends that massed items receive less rehearsal and are therefore processed much less than distributed items. The same can be applied here, in that more is not necessarily better (i.e., the law of diminishing returns).

Subjects assigned to one of the two mixed-practice conditions (abbreviated condition in one part of the tutor and extended in the other) apparently reaped the benefits of both conditions (i.e., had greater facility in retrieval processes and proceduralization) without the full impact of the associated disadvantages (inadequate practice opportunities and frustration). The net result would then be the same outcome scores evidenced by all four groups, which is exactly what we found.

Experiment 2 findings. Subjects who had originally learned from mixed-practice conditions showed the greatest retention, after two years, of flight engineering knowledge and skills, with retention scores on average almost 10% higher than those of subjects in either of the two extreme conditions. One way to interpret this finding is in terms of different practice schedules and how they affect performance and retention. The extreme practice conditions (A-A and E-E) may be characterized as a blocked (and for the E-E group, massed) practice schedule, with no respite from the perpetual presentation of either 3 or 12 problems for each of the 22 problem sets in the curriculum. Conversely, about halfway through the curriculum, the mixed groups (A-E and E-A) were changed from one practice condition to the other. Thus the mixed-practice condition is similar to a variable-practice schedule, in contrast to the blocked schedules of the A-A and E-E groups. As discussed earlier, under variable-practice schedules, subjects typically perform poorly during acquisition but much better on retention and transfer tasks as they become adept at accessing component skills across various problem-solving situations (e.g., Carlson and Yaure, 1990; Shea and Morgan, 1979).

Consider the advantages and disadvantages associated with each of the two extreme practice conditions. The A-A group was relatively more active during the learning process (e.g., seeking supplemental instruction) and may have realized that it was beneficial to stay alert, given their invariably restricted practice opportunities. The disadvantage of this condition was learners' inability to proceduralize new skills before being moved on to a different topic. The main benefit of being in the E-E condition was the ample opportunity to practice—and, consequently, proceduralize—developing knowledge and skills. However, the deleterious effects of massed practice have been well documented in the literature (e.g., fatigue, boredom, inattentiveness). Overall, for these extreme groups the advantages and disadvantages are believed to have balanced out in the long run (in terms of general activation levels of stimuli), resulting in no great advantage for either condition and comparable decay rates over time.

The mixed group, with a more variable practice schedule, somehow ended up with the greatest retention of previously learned material. We speculate that these subjects derived the benefits of both abbreviated and extended practice conditions without the associated disadvantages (i.e., they were not continually faced with inadequate practice opportunities or fatigue and inattention); thus, overall, there was a net gain compared with the two extreme conditions. For example, A-E subjects started out learning component knowledge and skills in the abbreviated condition, becoming facile at loading items into and out of working memory, and experiencing increased activation attributable to elaborative processing associated with supplemental instruction. They also showed increased arousal/attention given their sparse practice environment. When they were switched to the extended condition for the more substantive part of the tutor, the somewhat fragile component skills were able to become proceduralized during the course of additional practice opportunities. In addition, as their learning environment suddenly became unpredictable (e.g., they did not know whether they would receive 3 or 12 problems to solve in the next problem set), their arousal levels would be elevated, and thus they would not experience the boredom and inattentiveness believed to accompany the E-E condition.

The E-A group began learning component skills from the extended condition (without a lot of extra effort or supplemental instruction). Just as they settled into a routine, however, their practice schedule switched, and they subsequently had to integrate newly acquired skills in an abbreviated practice condition. As with the other mixed group, their comfortable practice schedule was suddenly disrupted, and they were forced to be more alert and to increase their cognitive effort as the new schedule offered few practice opportunities. In other words, the unstable environment served to increase working memory demands and arousal levels and, consequently, enhanced retention over time (Carlson and Yaure, 1990).

These findings have several implications for the design of adaptive training or instruction, particularly for teaching complex skills when practice opportunities are important. Briefly, we found that (a) there seems to be an optimal amount of practice, beyond which there are diminishing returns; (b) mixing practice conditions, at least minimally, appears to enhance retention; and (c) instruction should be pilot tested to determine how much practice to provide. Conventional wisdom holds that more practice is better than less practice. In contrast, these results show that with regard to practice opportunities, some intermediate amount may be optimal, especially in conjunction with a variable practice schedule (even one as minimally varied as in this design).

This is supported by the success of the mixed groups in terms of retention. In general, three problems may be too few and 12 may be in excess of what is needed to optimize learning. Thus an intermediate number may be optimal when instructing this type of complex cognitive skill. This argument is further supported by data shown in Figures 3 and 5 from Experiment 1, in which proceduralization occurs between Trials 4 and 7 (i.e., learning curves begin to flatten). More support for this middle position is the fact that the refresher course contained an intermediate number of practice problems (four), and a control group of subjects was able to attain, in only 2 h, the same outcome level as abbreviated-condition subjects, who had originally spent more than 7 h learning the same material (in addition to spending 1.5 h on the refresher course).

How can we tell how much practice is enough, per person, and then make systems maximally responsive to these idiosyncratic needs? We could establish latency and accuracy criteria per problem. When learners attain asymptotic performance, that would suggest they are ready to tackle the next problem set. Furthermore, after some minimum number of practice problems is completed (e.g., 3), subjects could simply be asked if they want or need more problems. However, this option requires that subjects be aware of cognitive strengths or weaknesses. That is not always the case, as some subjects are unaware and others are overly confident in electing to take on more problems. By adopting a policy of iterative pilot testing, developers can avoid unnecessary time and effort in providing too many (or too few) practice opportunities. Such a policy would also avoid investing too much of the students' time in accomplishing unnecessary practice opportunities. In addition, undue tedium during training almost certainly has other negative consequences, such as generalized loss of motivation, reduced time for other training needs, and increased wash-out rate.

Future research. Our finding concerning greater retention of material by the mixed-condition subjects compared with the extreme-condition subjects has been explained in relation to the efficacy of variable over blocked practice schedules. However, one unresolved aspect of this finding is that in our design, the mixed groups were varied only once (from abbreviated to extended, or vice versa). Thus a salient question for future research is, What are the parameters of this finding? That is, if the mixed groups show greater retention following such minimally varied practice, how does increasing the variability schedule relate to retention? This relationship may be linear (i.e., more variability leads to greater retention), or it may be optimally fit by some other trend (e.g., too much variability hurts retention, a smaller amount is beneficial, but an intermediate degree of variability is best).

Because our sample of returning subjects was so small (and consequently decreased the power of our tests), our findings need to be replicated with a larger group of subjects. In response to this, a second retention study is scheduled that will attempt to replicate the findings reported here. In our upcoming experiment we will use a statistics tutor (Stat Lady—Shute and Gluck, 1994) and vary the number of problems to solve across each of the problem sets. Furthermore, the replication study will consist of five instead of four practice conditions: A-A, A-E, E-A, E-E, and *learner's choice* (in which subjects control the number of problems to solve). Subjects will be randomly assigned to one of these practice conditions (80 subjects per condition) and will take a pretest, learn the material (roughly 3-10 h of instruction, depending on condition), and then take a matched posttest. We will recall subjects after six months to participate in a follow-up study; they will return and complete another parallel form of the test for retention assessment. Of particular interest will be whether or not retention following the lag will again show a mixed-practice condition advantage. Additionally, we will be able to assess the impact on learning from self-selecting practice opportunities.

ACKNOWLEDGMENTS

We offer special thanks to Wes Regian, Barry Goettl, Lee Gugerty, Malcolm Ree, and Kevin Gluck, who provided valuable suggestions about various aspects of this paper; Rich Walker and Trace Cribbs, who handled all of the programming tasks and associated support; and Cathy Gomez, Wayne Crone, and Linda Robertson-Schüle, who provided great assistance in the massive data collection job. The research reported in this paper was conducted as part of the Armstrong Laboratory, Human Resources Directorate. This study represents basic research funded by the Air Force Office of Scientific Research. The opinions expressed in this article are those of the author and do not necessarily reflect those of the U.S. Air Force.

REFERENCES

Ackerman, P. L. (1988). Determinants of individual differences during skill acquisition: Cognitive abilities and information processing. *Journal of Experimental Psychology: General, 117,* 288–318.

Adams, J. A. (1980). *Learning and memory: An introduction* (Rev. ed.). Homewood, IL: Dorsey.

Anderson, J. R. (1983). *The architecture of cognition.* Cambridge, MA: Harvard University Press.

Anderson, J. R. (1987). Skill acquisition: Compilation of weak-method problem solutions. *Psychological Review, 94,* 192–210.

Anderson, J. R. (1993). Problem solving and learning. *American Psychologist, 48,* 35–44.

Bryan, W. L., and Harter, N. (1899). Studies on the telegraphic language: The acquisition of a hierarchy of habits. *Psychological Review, 6,* 345–375.

Carlson, R. A., and Yaure, R. G. (1990). Practice schedules and the use of component skills in problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 484–496.

Ciccone, D. S., and Brelsford, J. W. (1974). Interpresentation lag and rehearsal mode in recognition memory. *Journal of Experimental Psychology, 103,* 900–906.

Cook, T. D., Campbell, D. T., and Peracchio, L. (1990). Quasi-experimentation. In M. D. Dunnette and L. M. Hough (Eds.), *Handbook of industrial and organizational psychology: Vol. 1* (pp. 491–576). Palo Alto, CA: Consulting Psychologists.

Crowder, R. G. (1976). *Principles of learning and memory.* Hillsdale, NJ: Erlbaum.

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

Fisk, A. D., and Rogers, W. A. (1992). The application of consistency principles for the assessment of skill development. In J. W. Regian and V. J. Shute (Eds.), *Cognitive approaches to automated instruction* (171–194). Hillsdale, NJ: Erlbaum.

Gick, M. L., and Holyoak, K. J. (1987). The cognitive basis of knowledge transfer. In S. M. Cormier and J. D. Hagman (Eds.), *Transfer of learning* (pp. 9–46). New York: Academic.

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

Greene, R. L. (1989). Spacing effects in memory: Evidence for a two-process account. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 371–377.

Hintzman, D. L. (1974). Theoretical implications of the spacing effect. In R. L. Solso (Ed.), *Theories in cognitive psychology: The Loyola symposium* (pp. 77–99). Hillsdale, NJ: Erlbaum.

Lesgold, A., Bunzo, M. S., McGinnis, T., and Eastman, R. (1989). *Windy, the flight engineering tutor.* Unpublished computer software program, University of Pittsburgh, Learning Research and Development Center.

Proctor, R. W. (1980). The influence of intervening tasks on the spacing effect for frequency judgments. *Journal of Experimental Psychology: Human Learning and Memory, 6,* 254–266.

Rundus, D. (1971). Analysis of rehearsal processes in free recall. *Journal of Experimental Psychology, 89,* 63–77.

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

Shea, J. B., and Morgan, R. L. (1979). Contextual interference effects on the acquisition, retention, and transfer of a motor skill. *Journal of Experimental Psychology: Human Learning and Memory, 5,* 179–187.

Shute, V. J. (1993). A macroadaptive approach to tutoring. *Journal of Artificial Intelligence in Education, 4,* 61–93.

Shute, V. J., and Gluck, K. A. (1994). *Stat Lady: Descriptive Statistics Module* (Unpublished computer program). Brooks Air Force Base, TX: Armstrong Laboratory.

Shute, V. J., and Walker, R. (1992). *Flight Engineering Refresher Course* (Unpublished computer program). Brooks Air Force Base, TX: Armstrong Laboratory.

Singley, M. K., and Anderson, J. R. (1989). *The transfer of cognitive skill.* Cambridge, MA: Harvard University Press.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12,* 257–285.

Watkins, M. J. (1990). Mediationism and the obfuscation of memory. *American Psychologist, 45,* 328–335.

Wickelgren, W. A. (1976). Network strength theory of storage and retrieval dynamics. *Psychological Review, 83,* 466–478.

Woltz, D. J. (1988). An investigation of the role of working memory in procedural skill acquisition. *Journal of Experimental Psychology: General, 117,* 319–331.

Woltz, D. J., and Shute, V. J. (1995). Time course of forgetting exhibited in repetition priming of semantic comparisons. *American Journal of Psychology, 108,* 499–525.

Date received: August 2, 1993
Date accepted: October 11, 1994