AL/HR-TP-1995-0032

INTERFACE, INSTRUCTIONAL APPROACH, AND DOMAIN LEARNING WITH A MATHEMATICS PROBLEM-SOLVING ENVIRONMENT

ARMSTRONG LABORATORY

Robert M. Yadrick
J. Wesley Regian

HUMAN RESOURCES DIRECTORATE
TECHNICAL TRAINING RESEARCH DIVISION
7909 Lindbergh Drive
Brooks AFB, Texas 78235-5352

Linda Robertson-Schule
Catherine Connolly Gomez

Galaxy Scientific Corporation
Bldg. 9020
1880 Carswell Avenue
Lackland AFB TX 78236-5507

November 1995

Interim Technical Paper for Period October 1993-December 1993

AIR FORCE MATERIEL COMMAND
BROOKS AIR FORCE BASE, TEXAS

NOTICES

When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely Government-related procurement, the United States Government incurs no responsibility or any obligation whatsoever. The fact that the Government may have formulated or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication, or otherwise in any manner construed, as licensing the holder, or any other person or corporation; or as conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

The Office of Public Affairs has reviewed this paper, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.

This paper has been reviewed and is approved for publication.

ROBERT M. YADRICK
Project Scientist

R. BRUCE GOULD, Ph.D.
Technical Director

JAMES B. BUSHMAN, LT.COL., USAF
Chief, Technical Training Research Division
Human Resources Directorate

Please notify this office, AL/HRPP, 7909 Lindbergh Drive, Brooks AFB TX 78235-5352, if your address changes, or if you no longer want to receive our techinical reports. You may write or call the STINFO office at DSN 240-3853 or commercial (210) 536-3853.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)

2. REPORT DATE
November 1995

3. REPORT TYPE AND DATES COVERED
Interim Technical Paper Oct 93 - Dec 93

4. TITLE AND SUBTITLE

Interface, Instructional Approach, and Domain Learning with a Mathematics Problem-Solving Environment

5. FUNDING NUMBERS

C - F41622-92-0-0006
PE - 61102F
PR - 2313
TA - T2
WU - 18

6. AUTHOR(S)

Robert M. Yadrick
Linda Robertson-Schule
J. Wesley Regian
Catherine Connolly Gomez

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Galaxy Scientific Corporation
1880 Carswell Ave, Bldg 9020
Lackland AFB TX 78235-5507

8. PERFORMING ORGANIZATION

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Armstrong Laboratory
Human Resources Directorate
Technical Training Research Division
Brooks AFB TX 78234-5352

10. SPONSORING/MONITORING

AL/HR-TP-1995-0032

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

We conceptualize Computer-Based Instruction (CBI) as involving both a primary task, learning to use a particular domain, and a secondary task, learning to use instructional approach embodied in a computer system and to manipulate its interface. We argue that the complexity and compatibility of the instructional approach and the interface separately and interactively influence the ease with which students are able to learn the subject matter. Examples from previous research show how looking at CBI in this way provides a framework for better understanding the results of some previous studies of CBI and CBI-related interface design. We also describe an experimental methodology for dissociating and measuring the separate relative effect on learning of using a CBI system for mathematics word problems. Finally, we report on an experiment with low-ability, remedial subjects that used this methodology to examine pretest-posttest differences due to each component.

14. SUBJECT TERMS
Computer-based instruction
Mathematics learning
Interface evaluation

15. NUMBER OF PAGES
25

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

Table of Contents

List of Tables

List of Figures

PREFACE

This report describes one of several experiments conducted in the TRAIN Cooperative Laboratory from October 1993 to March 1994. Funds for this research were provided by the U.S. Air Force Office of Scientific Research and the Armstrong Laboratory TRAIN Project, AL/HRTI, Brooks AFB, TX, Dr. Wes Regian, Director.

INTERFACE, INSTRUCTIONAL APPROACH, AND DOMAIN LEARNING WITH A MATHEMATICS PROBLEM SOLVING ENVIRONMENT

INTRODUCTION

It seems likely that several factors contribute to the overall effectiveness of Computer-Based Instruction (CBI), including the instructional approach, the specific domain and content, the interface design, a variety of student characteristics, and whether the CBI is the sole, the primary, or a supplemental means of instruction. Although the relative contributions of these factors, and the relationships and interactions between them, are not well understood, improving effectiveness may involve considering interactions between all of them. This is a tall order and, unfortunately, little guidance is available to the CBI developer.

A study by Schuerman and Peck (1991) provides an example of the kinds of surprises that may result from this lack of understanding. Schuerman and Peck considered how menu design interacts with the course of instruction by influencing the usage patterns of learners. Among other things, they found that the availability of pull-down menus did not encourage subjects to randomly access instructional components. Instead, those to whom this capability was available tended to proceed sequentially, much like others to whom this capability was not available. Thus, an interface feature intended to affect the flow of the course of instruction had no apparent effect.

Why would students not take advantage of a potentially useful and interesting feature? It is possible that they were, of necessity, more interested in staying on a simple and understandable instructional track than they are in jumping across instructional components. This points up a fundamental difference in purpose between application software and CBI software. Typically, when one learns to use application software, the primary task at hand is to learn how to use the software to accomplish a task that is itself already understood. For example, most people who learn to use a word processing package already know how to write, type, and edit what they have written, and it remains for the user to figure out how to use the system to do these well-understood tasks. On the other hand, the usual primary task of someone receiving CBI is to learn the domain, while learning to use the software is a secondary task, a means to an end.

The full implications of this are not clear. Studies show that decrements typically occur in primary task learning and performance when a secondary task is learned or performed simultaneously (e.g., Tirre & Pena, 1992; Ware, Bonner, Knight, & Cater, 1992), because performing the secondary task divides attention and increases working memory requirements, the so-called "cognitive load". An exception occurs if the secondary task has been learned to the point of automaticity, in which case the additional cognitive load is minimal or nonexistent, and primary task learning or performance proceeds unhindered (Schneider, Dumais, & Schiffrin, 1984.)

One can view CBI as requiring the simultaneous learning of multiple competing, although partially-dependent and interacting skills. These skills eventually come to complement and support each other, but learning to coordinate and integrate them takes time. Learning the domain is the nominal primary task, but understanding the instructional approach constitutes a secondary task. Understanding the instructional approach, in turn, may involve such components as gaining conceptual understanding of the purpose and rationale of the method, such as a particular strategy for solving problems, and procedural understanding of the steps involved in following the method. Moreover, students must also learn the interface manipulations by which to implement the steps. As a simple example, consider the process of correcting a minor interface manipulation error, such as backtracking to correct a mistaken menu selection made during the course of solving a problem. To do this, a student must hold in memory the action he or she intended to perform, as well as the action's purpose within a sequence of actions directed at satisfying the requirements of the instructional approach. The student must do this while remembering or figuring out how to cancel the accidental selection, pulling the menu down again, and locating and selecting the correct item. Finally, the student must do all this while remembering the purpose for the entire sequence of actions that included the incorrect selection. The problem may be made worse if it takes the student a while to realize in the first place that something has gone wrong and that the cause must be that the initial menu selection was faulty. Beginners unfamiliar with either the domain or the CBI system are probably most likely to make this sort of error. In some circumstances, therefore, learning to use the system may even rival domain learning in importance to the student. Eventually, however, with continued practice, the instructional approach will be understood and the interface manipulations will be learned to the point where using the CBI system presumably no longer hinders domain learning. The situation might be considered analogous to learning to read, albeit in a temporally-condensed form. In a theory advanced by Chall (1979), at one point a young child who can read fluently may still be unable to learn from reading because the high processing demands of word identification leave little room for acquiring new information. Further, different instructional approaches, such as visually-based retrieval or phonetic recoding, may speed or hinder the process of automatizing lexical access.

In theory, good interface design can minimize the likelihood that the incorrect menu pick will occur in the first place. The analysis of task structure and subsequent design and evaluation of computer system interfaces has been a primary focus of ergonomics research and a staple work area for human factors specialists for years. One result is that several comprehensive sets of interface design principles, guidelines, and specifications are now available (e.g., Mayhew, 1992; Schneiderman, 1987; Smith & Mosier, 1986). Only a handful of published reports, however, has directly examined the issue of interface design for CBI (e.g., Bolton & Peck, 1991; Clark, 1986; Schuerman & Peck, 1991), and virtually no specific principles or guidelines exist. Perhaps this is not a problem, and for the most part the CBI developer can simply follow his or her own intuitions or extrapolate directly from interface design principles established for various kinds of application software. This implies that an instructional approach should not limit the applicability of good interface design principles. Ideally, the interface and instructional design processes should be conducted simultaneously and each should inform the other.

Extending this line of reasoning leads to several assumptions which can be subjected to experimental examination. A trivial implication is that a relatively low cognitive load should result from using a simple CBI system to learn a simple domain. Another, more important implication is that a system which produces a cognitive load beyond some critical level will diminish the benefits of CBI or even retard learning relative to a non-CBI-based method of delivery. As domain complexity rises, this threshold presumably will go lower, while at the same time a complex interface and instructional approach may be needed to present the material in a complex domain adequately.

It also appears likely that a heavy cognitive load may affect low-achieving or less-talented students more than others (Woltz, 1988.) Mayes (1992) provided some evidence for this notion, finding that low-achieving secondary-school students learned better with a non-CBI approach to teaching problem-solving, while medium-achieving students were helped by using both the CBI and non-CBI approaches in conjunction with one another, and high-achieving students learned well with or without the CBI. In accordance with the spirit of this paper, Mayes (1992) explained his results by suggesting that "Students on this level may be overwhelmed by the joint problem-solving and computer treatment due to lower initial mathematics knowledge" (p. 247).

Finally, the effects of high cognitive load may not be uniformly distributed across all aspects of a domain. This may be one reason why Funkhouser and Dennis (1992) found that students given computer-augmented instruction showed more improvement on mathematical content than on problem-solving ability, although their results may also reflect the inherent difficulty of teaching problem solving.

This paper describes a methodology designed to decompose the effects on learning of the components described above, and also reports on a study based on the methodology. This study had limited goals; it examined only portions of the hypotheses we've outlined, focusing on low-ability subjects and using as a testbed the Word Problem Solving Environment (WPSE), a computer-based system which provides instruction and support for solving ninth-grade-level mathematics word problems. Further, the study was semi-naturalistic, lacking some of the control elements of a true laboratory experiment, and the number of subjects was relatively small. In these respects, it lies partway between a psychological experiment and the kind of naturalistic classroom-based studies frequently reported in the educational technology literature. Still, the results suggest that the effects of competing CBI component tasks can be examined and gauged. We would argue that practical usability criteria for CBI systems should focus on learning outcomes, or at least take them into account, and we regard this study as a step in that direction.

The Word Problem Solving Environment

The WPSE was developed at the Air Force's Armstrong Laboratory using the Toolbook software construction set (Toolbook 1.0, 1989). The instructional approach and problem pool were developed by mathematics teachers from San Antonio area middle schools and high schools.

The WPSE was an appropriate selection for our purpose. Pilot work showed that many of our subjects found both the structured, non-standard instructional approach and the interface somewhat difficult to understand and work with comfortably at first. Also, the basic problem-solving instructional approach could have been implemented in a number of different ways, so there is no necessary mapping between the approach and WPSE's particular implementation. Moreover, during pilot work it became apparent that in general the subjects who encountered the most difficulty were those who had little previous experience with computers and/or whose current mathematics skills were relatively poor. One use that has been proposed for the system, however, is remedial skills training for both civilian high school students and Air Force recruits who require it.

The system offers substantial instructional capacity and support, although the user must learn a problem-solving strategy and a particular series of steps to implement the strategy. Executing these steps, in turn, involves learning to execute what at times is a fairly complex sequence of interface manipulations. The system allows considerable user control in some ways. For example, users pick the next step to perform, determine when the current step is completed, decide when to ask for help, when to review the lesson, when to look up unit conversions, etc. The system also allows the subject to err in a number of ways. For example, the subject may select steps out of sequence, only to find eventually that he or she can't finish the current step because a previous step was not finished. The subject must then return to the previous step, work through it, then return to the current step.

The WPSE was developed as a computer-laboratory supplement to regular in-class mathematics curricula, and the system is presently being tested on-site at schools in New Mexico, New York, Ohio, and Texas. The curriculum is modular. Each module is loaded separately and concerns a different topic, such as proportions, percents, algebraic equations, geometric equations, ratios, etc. Each begins with a self-paced lesson which describes the concepts and principles of the topic, including basic formulas and simple worked example problems, using graphics and animation keyed to the text to illustrate important points. Figure 1 shows a sample instructional screen from a module on geometric equations.

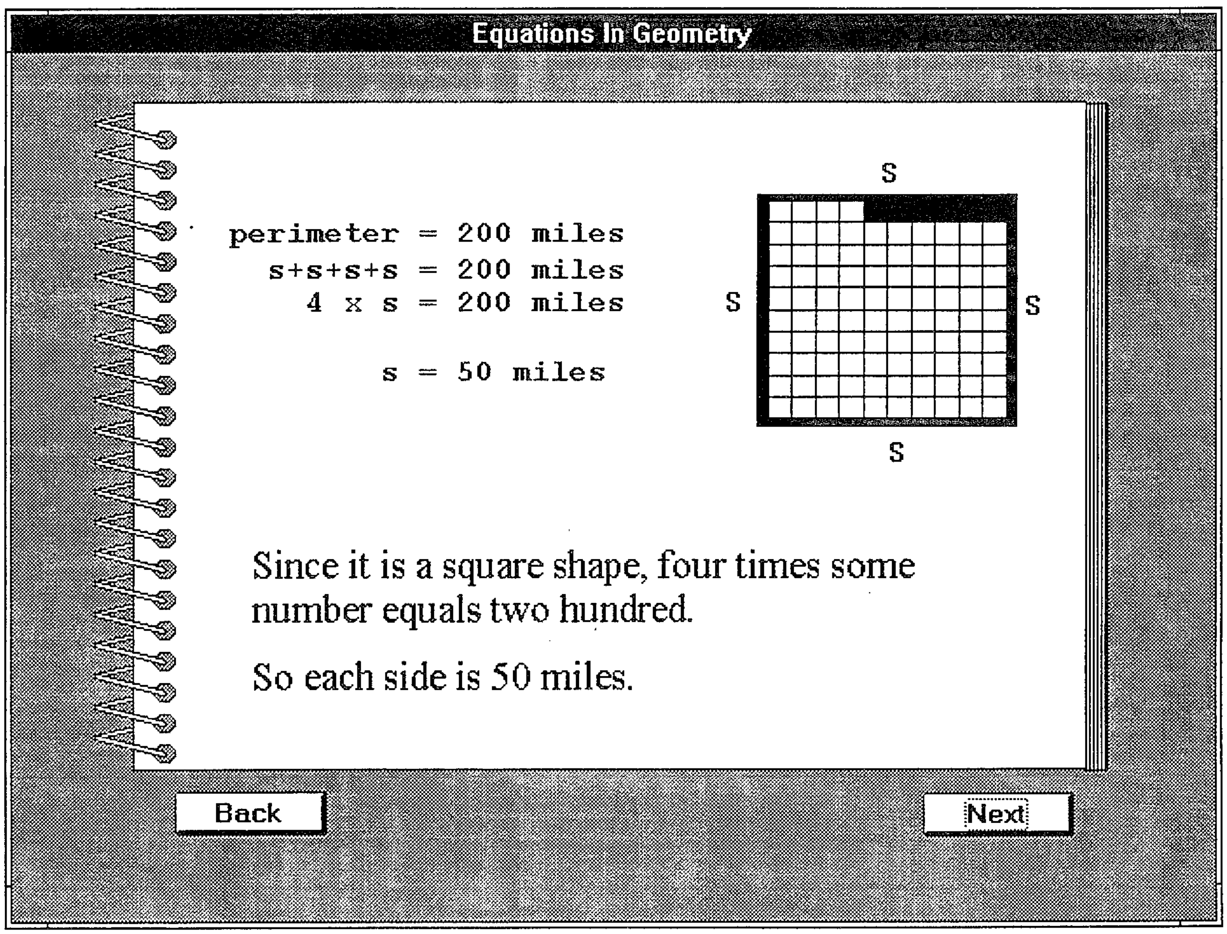

Figure 1. Sample WPSE instructional screen

Each lesson is followed by a set of exercises (problems to solve), arranged such that there are a few problems at each of several ascending difficulty levels. In a typical module, there are about 20-25 problems at 6-7 levels. Figure 2 shows the basic screen for the problem-solving environment, showing a level-2 difficulty problem from the geometric equations module (out of 7 levels for this particular module). The screen shows a menu bar at the top and several windows. The problem statement appears in the Problem Window.

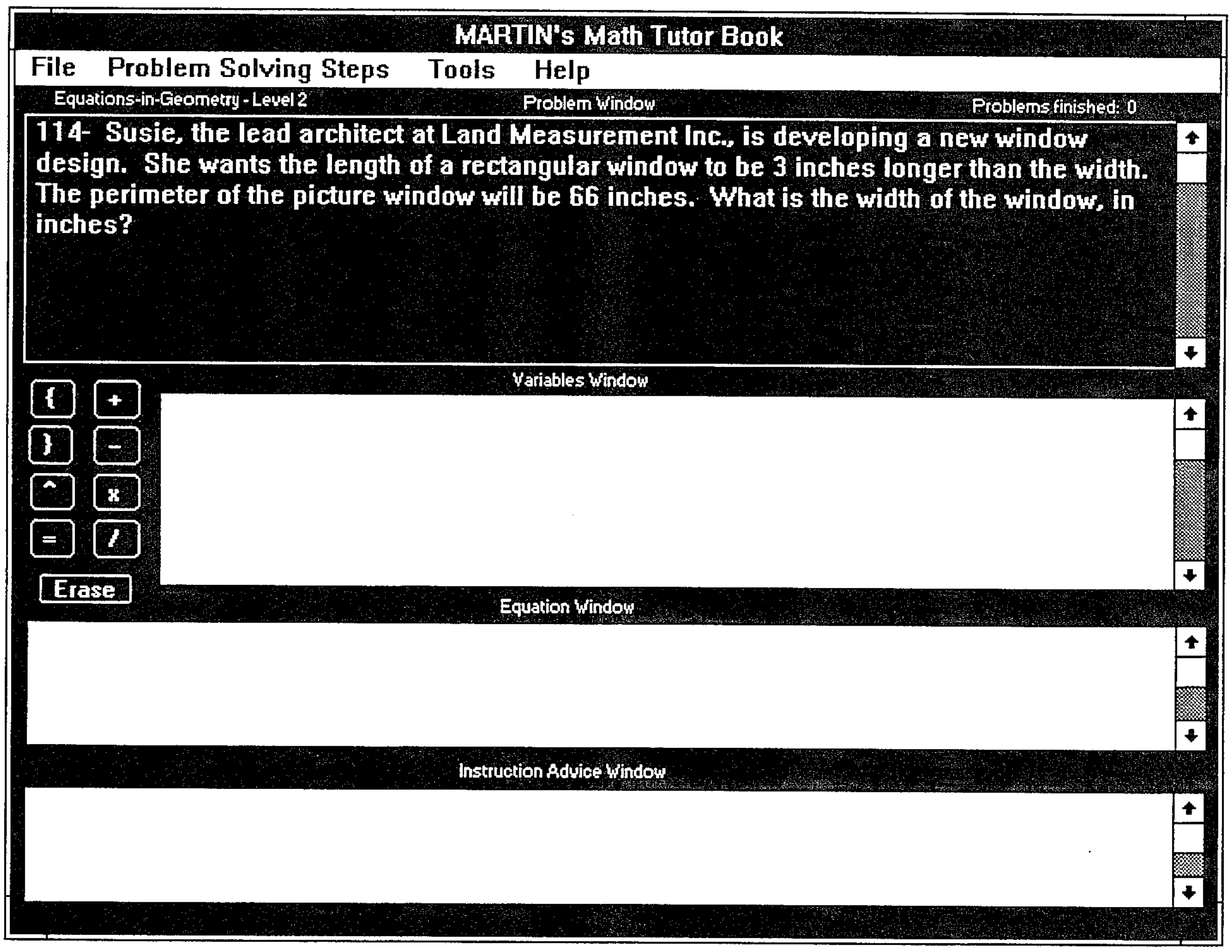

Figure 2. WPSE basic environment screen

The instructional approach involves practicing five problem-solving steps in a particular order. The developers' intent was to focus on the process whereby students understand a problem and build an equation to represent it. Clicking on "Problem Solving Steps" on the menu bar causes a pull-down menu to appear, listing the following steps: "Identify Goal", "Make Variables", "Make Equation", "Solve Equation", and "Answer Question". The user must first identify the goal by clicking on any word in the goal sentence ("What is the width..."). Next, he/she must provide verbal labels for necessary variables (e.g., "Perimeter", "Longer than Width") and assign them values by clicking on numbers in the Problem Window or by entering numbers. The next step involves constructing an equation in verbal form by assigning an "equation label" to represent the quantity being sought ("Width" for the example problem), then clicking in turn on variables in the Variables Window and operators on the keypad to the left of the Variables Window. The equation appears in the equation window as it is constructed. For the problem in Figure 2, such an equation might read "Perimeter = (2 X Width) + (2 X (Width + Longer than Width))".

The next step makes the solution to the equation appear in the Equation Window, so that users have an opportunity to decide if the solution seems reasonable before proceeding. The last step is to answer the question by entering the numerical answer and units ("15 inches") in a window that appears when the step is selected. The Instruction/Advice Window at the bottom of the screen retains all the help the subject receives from the system, so that by scrolling he/she can review whatever hints, formulas, definitions, etc., have been presented previously.

Clicking on "Help" on the menu bar produces a menu that allows selection of a weights & measures conversion table, a table of basic formulas, a glossary, interface help, or hints. Repeated requests for hints are answered with successively more precise and concrete hints, tailored to the currently active problem solving step. For example, a request for a hint during the "Identify Variables" stage is answered at first by the rather nonspecific advice to "reread the question and determine what variables are important for solving the problem", but in response to a second request the system suggests a name for one of the needed variables. A third request produces the value for that variable, and so on, until all the variables have been given. Successive hints for the "Make Equation" step attempt to guide the subject toward the correct arrangement of variables, and eventually offer an acceptable equation for solving the problem. Subjects were shown how to receive and use successively more specific hints during a tutorial session described later in this paper.

Finally, clicking on "Tools" produces a menu that allows selection of the Notebook and the Plan, two features not used for this study, as well as a "lesson review" feature which allows one to stop working on a problem at any time, jump back to the instructional session that begins the module, browse around and find particular information, then return to finish the problem. Subjects were also introduced to this feature during the tutorial.

METHOD

Participants

A total of 56 subjects, 27 males and 29 females, completed the study. All subjects were high-school graduates or had a GED and were between the ages of 18 and 30. They were recruited through local temporary employment agencies and paid $5.00/hour for their participation. The groups into which they were divided will be described later in this paper, at which time group sizes will be given as well.

Subjects were selected from a larger pool of subjects according to their performance on a screening test which is described in detail later in this paper. The purpose of the screening test was to identify remedial subjects at a relatively low level of mathematics ability at the time the study began. As high school graduates or the equivalent in Texas, all subjects had at some time completed at least one mathematics course which covered the concepts (e.g, ratios, percentages) included in this study. None, however, were able any longer to work problems reliably.

Subject attrition was relatively high. In addition to the 56 who finished, a total of 15 other subjects began but dropped out of the study before finishing the full three days. Dropouts were not concentrated in any particular group, and were replaced randomly the next week. We contacted the appropriate employment agency and tried to determine why each dropout did not return. Most said they did not return because of legitimate reasons, such as car trouble or a child's illness. A few said candidly that, despite being paid, they disliked spending the day working math problems. One additional subject was dropped for reasons described later.

Materials and Equipment

The WPSE was hosted on Compaq 486/33L computers with NEC/Multisync VGA monitors, standard keyboards, and Logitech three-button MouseMan computer mice.

The Tutorial -- A 7-page tutorial booklet walked subjects through the process of solving three problems selected from a module on volumes, which was not used again in the study. The tutorial was pedantic and comprehensive. It gave specific instructions on exactly what to do to solve the problems, such as where to click, what names to give variables, and so on. The process of solving one of the problems included steps to show how to correct errors, use all the essential help features and get hints from the system, etc. Solutions to the other two problems were relatively straightforward and simply showed how to work the problem efficiently. Subjects kept this booklet throughout the study so they could refer to it as needed.

Practice Modules -- Each subject received a total of three practice modules. The three modules were selected from among the many modules available with the WPSE.

Module 1 consisted of 20 problems on percentages. Module 2 was actually a combination of two different short modules and included two lessons, one about ratios and one about writing algebraic equations, along with a total of 19 problems. Module 3 included 19 problems and covered elementary geometric equations. Problems within each module were arranged by difficulty levels from easiest to most difficult.

Tests -- Each subject took a screening test, a pretest, and a posttest. The screening test was administered before the study proper began. There were three parts. Part 1 consisted of calculation problems in addition, subtraction, multiplication, and division, to test whether prospective subjects could perform very basic mathematical operations accurately. It also included very simple algebraic equations such as solving the equation "8x = 24" for x. There was a total of eight problems on Part 1, which subjects answered by filling in blanks. Part 2 presented a total of five word problems. Each problem in this part was selected from among the level-1 problems in the modules used in the study, that is, they were similar to the easiest problems that subjects would work with later. Each problem was followed by two multiple-choice questions, so that the maximum score for Part 2 was ten points. One of the multiple-choice questions asked subjects to select the correct equation to solve the problem from among four alternatives, and the other asked them to select the correct answer to the problem. Part 3 was structured like Part 2, that is, there were five problems with two multiple-choice questions concerning each problem. However, problems were selected from the pool of middle-difficulty problems that the subject would work with later. Subjects were allowed a maximum of 30 minutes to complete this test.

In order for a potential subject to qualify for the study, he/she had to answer at least six of eight Part 1 problems and at least four of ten Part 2 questions correctly, but could not answer more than four of ten questions in Part 3 correctly. These criteria were decided upon because pilot work showed them to be satisfactory for the purpose of selecting subjects who could not work the majority of problems in the problem set, but who had the prerequisite reading and mathematics skills to learn to solve at least some additional problems. Overall, approximately 60% of potential subjects screened were not selected because they had too many correct answers on Part 3 problems, and about 5% more were not selected because they lacked the skill to answer problems in Parts 1 and 2 satisfactorily. These subjects were assigned to participate in other studies. When the screening test was administered, subjects were unaware of these criteria and did not know what sort of subjects were being sought for the study.

There were two forms (labeled Form A and Form B) of the pretest/posttest. Half (28) of the subjects received Form A as their pretest and Form B as their posttest, while the others received these tests in the reverse order. The two forms were composed of the same number of problems at each difficulty level from each practice module. Each consisted of 13 problems selected from among the medium-difficulty (levels 3, 4, and 5) problems for each practice module. There were five multiple-choice questions for each problem, and each question was followed by four alternative answers. The multiple-choice questions involved the skills developed through practice on the five-step problem-solving process. The first question asked what the subject was to find, and the correct alternative paraphrased the goal of the problem. The second required identification of a piece of extraneous information given in the problem, and the foils were all pieces of information which were needed to establish constraints and assign values to variables. The third required identification of a correct equation for the problem, and the fourth asked for the answer. The fifth asked for the correct unit (e.g., gallons, miles) for the answer. Subjects were provided with scratch paper and calculators for the tests, but were not allowed to use notes or any other supporting materials.

Throughout the rest of this paper, the word "problem" will be used to refer to entire word problems, while "item" will refer to each of the multiple-choice questions that followed each problem. Finally, "item type" will refer collectively to a particular sort of item over all problems; for example, all questions regarding the correct equation constitute an item type.

Design and Procedure

The study was conducted over the course of eight weeks, with groups of 8-12 subjects selected to participate each week. The study lasted three days of each week. Subjects were allowed 1 1/2-hour lunch periods and 10-minute rest breaks at the end of every hour of work.

Subjects completed the screening test as part of a set of first-day screenings and intake exercises. Those selected were assigned randomly to a group (unknown to them until after the pretest and tutorial) and took the pretest. They were allowed up to 60 minutes to complete the pretest.

For standardization purposes, all subjects, even those who would not use the computer again during the study, began the study proper by logging on to a computer and going through the tutorial. Most spent between two and 3 hours (total apart from breaks) on the tutorial, and they were dismissed for the day once they had finished.

Each subject was then assigned to one of four groups and spent the bulk of the remaining two days working on the three practice modules, which were administered in the fixed order described previously. Subjects could work on each module for a maximum of 3 hours, excluding break time. Subjects who finished a module before the time was up remained at their stations and were free to rest or to read magazines or books. The posttest was administered on day 3, after all subjects had finished all the modules. As with the pretest, a maximum of 60 minutes was allowed.

All subjects were given the same tests and the same practice module problems to work. The only difference between groups was the way in which the practice sessions were conducted. Subjects in the WPSE group received all their instruction from and worked problems using the WPSE, and received no additional instructional or supporting materials other than a calculator and scratch paper. Within the framework of this paper, subjects in this group were required simultaneously to learn the problem-solving approach inherent in the WPSE, how to implement the problem-solving process in the WPSE's problem-solving steps, the WPSE interface, and the mathematics domain. We assumed that the resulting cognitive load would be highest in this group and that learning would therefore be poorest.

Subjects in the Worked Examples group were given the problems to work in paper booklets, and did not use the computer again following the tutorial. They were also given a second set of booklets containing worked example problems. Each worked example problem was equivalent (that is, had the same underlying structure and similar cover stories; see Reed, Dempster, and Ettinger, 1985) to a single practice problem, and served as a guide to working the problem. In terms of this paper, these subjects learned mathematics using an alternative (worked examples) instructional approach which did not involve problem solving, learning the WPSE problem-solving steps, or using the WPSE interface. Sweller (1989) has argued persuasively that the cognitive load for this instructional approach is low in comparison to some other approaches, including working-forward approaches such as that used in the WPSE. With a relatively easy instructional approach and no competing tasks to learn, we predicted that these subjects would learn the material better than those in any other group.

Subjects in the Enter Worked Examples group were given the same booklet of worked examples as subjects in the Worked Examples group, but worked problems using the WPSE. In other words, this group represents a hybrid of the Worked Examples and WPSE groups. These subjects did not need to learn the problem-solving approach, but did need to learn the WPSE steps in order to translate the formulation they arrived at by studying the worked example, and also needed to learn the interface and the domain. With three competing tasks to learn, their cognitive load should be nearly, but not quite, as high as that for subjects in the WPSE group.

Subjects in the Paper WPSE group were given workbooklets which were intended to reproduce the WPSE on paper as closely as possible. All of the WPSE instructional screens, problems, and hints were included, and subjects in this group saw no worked examples. In terms of this paper, they learned mathematics using the problem-solving approach, but were not required to learn the WPSE interface or to follow the particular sequence of WPSE solution steps. We predicted that the cognitive load would be low, but not as low as that for the Worked Examples group.

These groups and our assumptions concerning the simultaneous learning requirements impinging on each group are summarized in Table 1. Our prediction, in essence, was that the groups' mathematics learning performance would be an inverse function of the number of competing tasks that must be learned.

Table 1
Hypothetical task learning requirements for each group

	LEARNING REQUIREMENTS			
GROUP	Problem Solving	Steps	Interface	Mathematics
WPSE	X	X	X	X
Worked Examples				X
Enter Worked Examples		X	X	X
Paper WPSE	X			X

One undesirable treatment difference was unavoidable. Both of the computer groups (WPSE and Enter Worked Examples) received immediate feedback. Once a subject gave an answer, he or she was immediately told whether the answer was or was not correct. Apart from informing subjects of incorrect answers, the WPSE gives no clue as to what the nature of the problem might be.

Giving immediate feedback to paper-and-pencil groups (Worked Examples and Paper WPSE) was not feasible, since it would have been disruptive and time-consuming for proctors to check each problem as each subject finished it. Instead, subjects finished as many problems in the booklet as they could and went on break while the booklet was scored. Incorrect answers were clearly indicated, without hints or help regarding the nature of the error. Subjects finished a second pass and turned in the booklets again, for the final time.

Subjects could ask proctors questions if they did not understand a problem statement, but no other help or information was given.

A total of 14 subjects completed the study in the Worked Examples group, 15 in the Paper WPSE group, 14 in the WPSE group, and 13 in the Enter Worked Examples group.

RESULTS

All results reported here were obtained using SPSS for Windows, version 6.09 (SPSS, 1993) in which repeated measures Analysis of Variance (ANOVA) and Analysis of Covariance (ANCOVA) are computed using a Multiple Analysis of Variance (MANOVA) procedure. All F-values reported are for unique sums of squares, and an alpha level of .05 was used for all tests.

Pretest Performance

There were no statistically significant differences between groups on overall (i.e., summing across the five types of multiple-choice questions) pretest scores. A one-way Analysis of Variance (ANOVA) between the four groups yielded $F(3, 52) = .61$, $MSE = 70.25$, $p = .61$. There were 65 possible points, with 16.25 correct expected by chance. Means and standard deviations for the groups were as follows: Worked Examples $M = 28.42$, $SD = 11.02$; Paper WPSE $M = 24.33$, $SD = 7.54$; WPSE $M = 26.36$, $SD = 6.99$; and Enter Worked Examples $M = 25.46$, $SD = 7.32$.

The high standard deviation for the Worked Examples group was cause for concern. Examination of the data showed that a male subject in the Worked Examples group had a pretest score (53) that was considerably higher than that for any other subject (the next highest was 43, for a subject in the Paper WPSE group), and sufficiently high to bring into question the subject's suitability for the study, despite their screening test performance. We decided to exclude this subject from further analyses, which brought the final number of subjects to 55 and reduced the mean for the Worked Examples group to 26.54, with a standard deviation to 8.80. Rerunning the one-way ANOVA on pretest scores without this subject yielded an $F(3, 51) = .25$, $MSE = 58.88$, $p = .86$.

The two pretest forms were of equal difficulty. A 2 X 2 ANOVA using pretest form and group as factors showed neither a main effect for form, $F(1, 47) = .33$, MSE = 61.01, $p = .57$, nor an interaction between form and group, $F(3, 47) = .61$, $p = .61$. Mean correct was 26.25, $SD = 7.14$ on one form, and 25.00, $SD = 7.96$ on the other form.

There were no pretest score differences by sex, independent $t(53) = -.66$, $p(\text{2-tailed}) = .51$. The mean score for males was 24.92, $SD = 7.67$, and the mean for females was 26.28, $SD = 7.44$.

Overall Pretest-Posttest Differences

On average, subjects correctly worked more posttest problems than pretest problems. Posttest means and standard deviations for the groups were as follows: Worked Examples $M = 37.31$, $SD = 11.61$; Paper WPSE $M = 30.07$, $SD = 7.55$; WPSE $M = 29.79$, $SD = 9.03$; and Enter Worked Examples $M = 30.15$, $SD = 9.64$.

More importantly, groups differed appreciably with regard to how much their scores improved, and, as predicted, improvement appears inversely related to the number of hypothetical learning requirements shown in Table 1. The Worked Examples group mean improved by 10.77 items, or about 40.6%, while the WPSE group mean improved by about 3.4 items, or a little over 13%. The other groups fell between these two. The Paper WPSE group improved by 5.74 items, or 23.6%, and the Enter Worked Examples group improved by 4.69 items, or 18.4%. Figure 3 shows overall pretest-posttest differences by group.

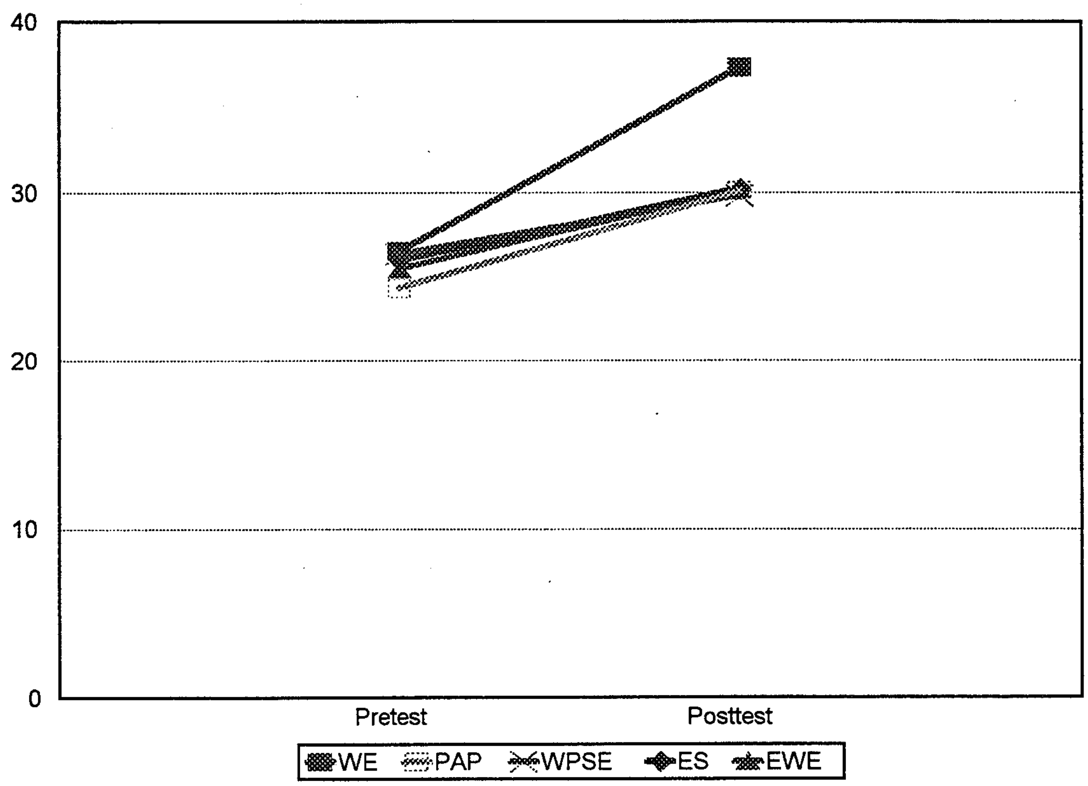

Figure 3. Overall pretest-posttest differences

Although no pretest differences between groups were significant, some differences were sizeable relative to pretest/posttest improvement levels, and pretest and posttest scores were significantly correlated, $\underline{r}(55) = .58$, $\underline{p} < .001$. We decided, therefore, to analyze these data using a repeated-measures ANCOVA which examined posttest score differences using pretest scores as a covariate. Although the pattern of results is consistent with our predictions, the group x repeated measure interaction result only weakly approached statistical significance, F(3, 51) = 2.22, MSE = 30.84, p = .097. Note that standard deviations for the posttest were higher than for those for the pretest, especially for the Worked Examples group. This shows that not everyone in any group benefitted equally (or at all) from their instruction and practice. In general, increases in means were accompanied by increases in variability.

There were no significant sex differences in overall improvement. The mean posttest score for males was 32.15, SD = 11.05, and the mean for females was 31.34, SD = 8.58. Repeated-measures ANCOVAs using pretest scores as a covariate showed no interaction between the repeated measure and sex, F(1, 53) = .97, MSE = 32.95, p = .33.

Pretest-Posttest Differences by Item Type

Overall scores represent a composite of five very different item types, which may partly account for the failure to find clear group differences on overall scores. We conducted an examination at the item-type level for this reason.

There were no differences between groups either in their ability to discriminate between information that was necessary or unnecessary to solve problems (item type 2) or in their ability to identify the correct unit (item type 5). A repeated-measures ANCOVA on the number of correct posttest type 2 items using number of type 2 correct pretest items as a covariate yielded F(3,51) = 1.67, MSE = 3.85 p = .185. The corresponding analysis on type 5 items yielded F(3,51) = .43, MSE = 8.50, p = .733.

However, the difference between groups on their posttest ability to identify correct equations (item type 3) approached significance, F(3,51) = 2.44, MSE = 3.15, p = .075, while there were significant differences between the groups on item types 1 (identifying the goal), F(3, 51) = 4.06, MSE = 2.44, p = .012 and 4 (identifying the correct answer), F(3, 51) = 3.92, MSE = 1.72, p = .014. Table 2 gives means and standard deviations for groups by item type, while Figure 4 shows pretest-posttest improvement on item type 4 by group. We highlight the data for item type 4 because being able actually to solve problems and identify correct answers, we would argue, is the most important of the various learning measures we tested.

We calculated difference scores (posttest score - pretest score) for data on item types 1 and 4, in order to examine differences between particular groups. We believe that use of difference scores is appropriate under the circumstances, in part because of the generally low mean pretest performance and relatively large individual differences on the pretest. In addition, we would argue that difference scores accurately reflect subjects' improvement during training and that an analysis based on difference scores is consistent with the theory underlying analysis of covariance.

For these difference scores for item type 1, F(3, 51) = 4.06, MSE = 4.87, p = .0115. Cohen's D for this analysis was 1.22. A Newman-Keuls test showed that the paper and pencil groups (worked examples and paper WPSE) differed significantly from the computer groups (WPSE and Enter Worked Examples), but there were no other differences among groups.

The corresponding analysis for item type 4 yielded F(3, 51) = 3.92, MSE = 3.44, p = .0136, while Cohen's D for this analysis equalled 1.17. Another Newman-Keuls test showed that the worked examples group differed from the other three, which did not differ among themselves.

Table 2
Group pretest and posttest means and standard deviations by item type

	GROUP			
	Worked Examples	Paper WPSE	WPSE	Enter Worked Examples
PRETEST				
Item Type 1 Mean	3.92	4.40	4.50	4.85
SD	1.80	2.03	2.03	1.34
Item Type 2 Mean	5.23	5.33	5.50	6.38
SD	1.96	1.80	2.21	1.98
Item Type 3 Mean	3.54	4.20	3.79	3.92
SD	2.18	1.37	1.42	2.69
Item Type 4 Mean	4.69	4.27	4.36	4.31
SD	1.89	1.71	2.17	1.84
Item Type 4 Mean	6.92	6.13	7.50	6.46
SD	2.75	2.29	1.70	2.18
POSTTEST				
Item Type 1 Mean	6.54	5.00	5.14	4.54
SD	2.26	1.36	2.17	2.10
Item Type 2 Mean	7.77	6.53	6.43	6.54
SD	3.09	2.39	2.34	2.93
Item Type 3 Mean	6.54	4.93	4.57	5.08
SD	2.93	1.79	2.21	2.10
Item Type 4 Mean	7.69	5.53	5.00	5.69
SD	2.36	1.92	2.18	2.46
Item Type 5 Mean	8.77	8.07	8.64	8.31
SD	2.52	2.12	2.59	2.25

Practice Module Performance

Table 3 shows means and standard deviations for each of the three practice modules. Both first- and second-pass statistics are given for paper and pencil groups, while statistics for computer groups are listed under the "second pass" heading. The practice module data for one subject in the WPSE group were lost due to an unrecoverable disk failure.

Table 3
Means and standard deviations for practice performance by group, module, and pass

	GROUP			
	Worked Examples	Paper WPSE	WPSE	Enter Worked Examples
MODULE 1				
Pass 1 Mean	17.08	12.33		
SD	2.06	4.22		
Pass 2 Mean	19.64	16.60	11.31	12.31
SD	0.63	2.82	5.22	7.80
MODULE 2				
Pass 1 Mean	15.46	13.60		
SD	3.64	2.82		
Pass 2 Mean	18.92	18.47	12.77	9.54
SD	2.40	2.45	5.54	3.89
MODULE 3				
Pass 1 Mean	16.46	14.60		
SD	3.53	4.07		
Pass 2 Mean	17.43	17.80	14.69	16.92
SD	3.20	1.74	4.48	4.05

We were surprised that practice module 2 performance for the Enter Worked Examples group was so poor, relative to the other groups. We checked the data and calculations several times and found them to be correct, although we have no satisfactory explanation. It is difficult to imagine why subjects in this group should have more trouble with this module than with either of the others, unless the data reflect some unusual interaction between the particular topics used for module 2 (ratios and algebraic equations) and the process of translating from the worked examples to a form suitable for the WPSE. This seems unlikely, however, and the performance of the other groups shows that the module 2 practice problems were inherently no more difficult than those in the other two modules.

One-way ANOVAs performed on the "second pass" data given in Table 3 showed significant group differences for practice module 1 performance, $\underline{F}(3, 50) = 8.55$, $\underline{MSE} = 23.15$, $\underline{p} = .0001$. Cohen's $\underline{D}$ for this analysis equalled 1.46. A Newman-Keuls test showed that the Worked Examples and Paper WPSE groups, which used paper and pencil, differed significantly from the WPSE and Enter Worked Examples groups, which used the computer. Neither the paper and pencil groups nor the computer groups differed from each other, however. The corresponding analysis for practice module 2 yielded $\underline{F}(3, 50) = 22.03$, $\underline{MSE} = 10.54$, $\underline{p} < .0001$. Cohen's $\underline{D}$ for this analysis was 1.76. Another Newman-Keuls test again showed that the paper and pencil groups, Worked Examples and Paper WPSE, differed from the computer groups, WPSE and Enter Worked Examples. In addition, the WPSE and Enter Worked Examples groups differed from each other. For practice module 3, however, there were no significant differences between groups, $\underline{F}(3, 50) = 2.07$, $\underline{MSE} = 12.22$, $\underline{p} = .12$.

DISCUSSION

In general, the results of this study are consistent with previous research in finding that low-ability subjects may not learn from computers as well as they learn from non-computer approaches (Mayes, 1992), and with the findings of Sweller and his colleagues (Sweller, 1989) in showing better performance by subjects who learn from worked examples as opposed to a problem-solving approach. The results also reinforce, extend, and clarify Mayes' explanation of his results in terms of low-ability students' being "overwhelmed" when they learn how to solve problems using the computer. Although group comparisons are not always statistically significant for these relatively small groups, they are in important respects consistent with our conceptual dissociation, summarized in Table 1, of the effects of the different aspects of the problem-solving instructional approach, learning to manipulate the interface, and learning the domain. The results were clear and statistically significant with respect to finding the correct answer to a problem, which we consider to be the most important measure from among the five item types, and with respect to identifying a restatement of the goal of the problem. The results approached significance with respect to identifying a correct equation, another very important skill involved in translating text into symbolic form. No group differences were found for identifying the correct unit for the answer or for identifying unneeded information. The mean pretest and posttest scores for these item types were higher than for the other item types, however. It appears that these two skills were relatively easy, compared to the other three, and that comparatively little learning was required or occurred.

Although the WPSE and Enter Worked Examples groups still lagged behind the others, the means for practice module 3 suggest that interference between tasks was subsiding as subjects worked on module 3. This is not reflected in posttest performance, but may be obscured because the contribution of module 3 material to posttest scores was overshadowed by the combined contribution of the other modules. Alternatively, it may be that even after subjects in these groups learned to use the system reasonably well, using it still took enough of their attention that they didn't learn very much about the domain. Moreover, the difference between the two computer groups for module 2 is both interesting and noteworthy, because it suggests that interference in the Enter Worked Examples group, which in our analysis involves learning three simultaneous tasks (see Table 1), begins to subside sooner than that in the WPSE group, which we assume involves learning four simultaneous tasks. These results suggest that the initial learning decrement that can arise from using a CBI system eventually subsides, but that this may not occur over the course of at least the first several hours of work. Unfortunately, this may be when much basic learning should occur. The possibility exists that students can advance through the basic material without the full understanding that serves as a foundation for advanced learning.

A few additional comments are in order. First, we concentrated on low-ability subjects. This was necessary because only the low-ability subjects in our pool were unable initially to work most of the problems in the WPSE problem set. It was also desirable in that we expected inter-task interference to affect low-ability subjects the most, and because remedial students are one proposed target user population for the system. It may also have been unfortunate, however. For one thing, we would reasonably expect the poorest learning in this population as well, which may have resulted in relatively limited improvement in all groups. We can speculate that clearer intergroup differences might have emerged, allowing us better to assess the relative strength of interference from each of the competing tasks listed in Table 1, if a wider range of subject abilities had been included. Also, screening out all but low-ability subjects, in conjunction with attrition and time limitations on our use of the laboratory, served to restrict the final sample size.

Second, nothing we've said should be taken as suggesting that the WPSE is an intrinsically poor system. The ultimate value of the system is a separate empirical question beyond the scope of this paper, although it seems that student characteristics and how it is used may play an important role.

Third, there was the possible problem involving the differential provision of feedback between groups. As we mentioned previously, computer groups received immediate feedback about their answers and the paper and pencil groups did not. The effects of delayed feedback on learning are unclear; the traditional view that immediate feedback aids learning has been challenged over the past few years. Some reports indicate that delay of feedback has no effect (e.g., Salmoni, Schmidt, & Walter, 1984), while other reports indicate that immediate feedback can be detrimental to learning (e.g., Swinnen, Schmidt, Nicholson, & Shapiro, 1990). For the most part, however, studies of delayed feedback have involved relatively simple motor or verbal tasks and delays on the order of seconds. It is therefore difficult to assess the extent to which their results are pertinent to our study. Although at least one article (Simmons & Cope, 1993) has reported negative effects of immediate feedback from a computer-based system, system and subject differences still make direct comparison of the Simmons and Cope results with our present results problematical. In any event, multiple-task learning appears to have affected the paper WPSE group in much the same way it affected the computer groups, even though the subjects in the paper WPSE group did not receive immediate feedback. This suggests that our results were not affected by feedback conditions.

As a final note, some researchers (e.g., Schmidt & Bjork, 1992) suggest that there is a fundamental difference between conditions that tend to produce short-term gains and those that tend to produce long-term gains in performance and transfer. They contend that, at least in some circumstances, conditions that make initial skill learning difficult and adversely affect early performance may lead to longer retention, better transfer, and improved later performance. One possible implication of this is that a difficult but information-rich approach, such as that used for the hybrid Enter Worked Examples group in the present study, might, after extended practice, actually result in better long-term learning for users of varying ability levels. Of course, this is sheer speculation at present.

REFERENCES

Bolton, J. W., & Peck, K. L. (1991). Menu sequence, learner control, and achievement in CAI. Interactive Learning International, 7, 95-100.

Chall, J. S. (1979). The great debate: Ten years later, with a modest proposal for reading stages. In L. B. Resnick & P. A. Weaver (Eds.), Theory and practice of early reading. Hillsdale, NJ: Erlbaum.

Clark, D. (1986). Aspects of menu design. Journal of Computer-Assisted Learning, 2, 172-177.

Funkhouser, C., & Dennis, J. R. (1992). The effects of problem-solving software on problem-solving ability. Journal of Research on Computing in Education, 24, 338-347.

Mayes, R. L. (1992). The effects of using software tools on mathematical problem solving in secondary schools. School Science and Mathematics, 92, 243-248.

Mayhew, D. J. (1992). Principles and guidelines in software user interface design. Englewood Cliffs, NJ: Prentice-Hall, 1992.

Nathan, M. J., Kintsch, W., & Young, E. (1992). A theory of algebra word comprehension and its implications for the design of learning environments. Cognition and Instruction, 9, 329-389.

Reed, S. K., Dempster, A. & Ettinger, M. (1985). Usefulness of analogous solutions for solving algebra word problems. Journal of Experimental Psychology: Learning, Memory, and Cognition, 11, 106-125.

Salmoni, A. W., Schmidt, R. A., & Walter, C. B. (1984). Knowledge of results and motor learning: A review and critical appraisal. Psychological Bulletin, 95, 355-386.

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

Schneider, W., Dumais, S. T., & Schiffrin, R. M. (1984). Automatic and control processing and attention. In R. Parasuraman, R. Davies, and J. Beatty (Eds.), Varieties of attention (pp. 1-27). New York: Academic Press.

Schneiderman, B. (1987). Designing the user interface. Reading, MA: Addison-Wesley.

Schuerman, R. L., & Peck, K. L. (1991). Pull-down menus, menu design, and usage patterns in computer-assisted instruction. Journal of Computer-Based Instruction, 18, 93 - 98.

Simmons, M., & Cope, P. (1993). Angle and rotation: Effects of different types of feedback on the quality of response. Educational Studies in Mathematics, 24, 163-176.

Smith, S. L., & Mosier, J. N. (1986). Guidelines for designing user-interface software. Bedford, MA: Mitre Corporation.

Sweller, J. (1989). Cognitive technology: Some procedures for facilitating learning and problem solving in mathematics and science. Journal of Educational Psychology, 81. 457-466.

Swinnen, S. P., Schmidt, R. A., Nicholson, D. E., & Shapiro, D. C. (1990). Information feedback for skill acquisition: instantaneous knowledge of results degrades learning. Journal of Experimental Psychology: Learning, Memory, and Cognition, 16, 706-716.

Tirre, W. C., & Pena, C. M. (1992). Investigation of functional working memory in the reading span test. Journal of Educational Psychology, 84, 462 - 472.

Toolbook 1.0 [Computer Program] (1989). Bellevue, WA: Asymetrix Corporation.

Ware, C., Bonner, J., Knight, W., & Cater, R. (1992). Moving icons as a human interrupt. International Journal of Human-Computer Interaction, 4, 341 - 348.

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