References

Clancy, W.J. (in press). Guidon-Manage revisited: A socio-technical systems approach. *Journal of Artificial Intelligence in Education, 4*(1), 5–34.

De Certeau, Michel. (1984). *The practice of everyday life*. (Steven Randall, Trans.). Berkeley, CA: University of California Press.

Haraway, Donna. (1991). *Simians, Cyborgs, and women: The reinvention of nature*. New York, NY: Routledge.

Harding, Sandra. (1991). *Whose science? Whose knowledge? Thinking from women's lives*. Ithaca, NY: Cornell University Press.

Helms, Steve. (1991). *The cybernetics group*. Cambridge, MA: MIT Press.

Jl. of Artificial Intelligence in Education (1993) **4**(1), 61–93

A Macroadaptive Approach To Tutoring

VALERIE J. SHUTE
Armstrong Laboratory, Human Resources Directorate
Brooks Air Force Base, TX 78235, USA

This paper presents a *macroadaptive* approach to tutoring that can produce sizable learning improvements at less cost compared to traditional *micro* approaches. The macroadaptive technique is based on the notion that some learners benefit from instruction provided one way while others learn more if instructed a different way. This notion forms the basis of *aptitude-treatment interaction* (ATI) research. Presented here is a large-scale study that evaluates the efficacy of this approach through the use of an intelligent tutoring system for instruction of flight engineering knowledge and skills. Aptitudes investigated in the study included working-memory capacity and general knowledge. The treatments, or learning environments, compared in this study were *constrained* (few problems per problem set) versus *extended* (many problems per problem set). Findings showed that while there were no differences between the two environments on any outcome measure, there was a significant aptitude-treatment interaction involving working memory, general knowledge, and environment. Subjects with a lot of general knowledge but low capacities performed significantly better if assigned to the constrained environment than to the extended environment. Subjects with less knowledge and high capacities learned significantly more from the extended rather than constrained environment. These findings support the utility of the macroadaptive approach.

During the last decade, cognitive diagnosis has become increasingly prominent in research concerned with the development of intelligent tutoring systems (ITS) (Mandl & Lesgold, 1988; Sleeman & Brown, 1982;

Wenger, 1987). In fact, accurate cognitive diagnoses along with appropriate remediation have been the active ingredients in the transformation of simple tutoring systems into *intelligent* tutoring systems. Currently, the most common approach to cognitive diagnosis occurs at a *micro* or fine-grained level, with remediation fairly quick and precise (e.g., Anderson, Boyle, & Reiser, 1985; Anderson, Conrad, & Corbett, 1989; Sleeman, Kelly, Martinak, Ward, & Moore, 1989; VanLehn, 1990). Computer intelligence, then, may be gauged by the degree to which the system modifies instruction based on an inferred model of the student's current understanding and representation of the particular subject matter (Wenger, 1987). Such intelligence involves a diverse set of knowledge bases as well as inference routines to create instructional interactions dynamically. But there is a cost associated with increasing a system's responsiveness, especially at such a fine-grained level. So it becomes fair to ask the question: Are there alternative ways to achieve significant improvements in learning outcome or efficiency?

One alternative approach is based on the intuitively valid precept that some learners benefit from instruction provided one way while others learn more if instructed a different way. Such relationships between tutor and student are called *aptitude-treatment interactions* (ATI), where *aptitude* is defined in the broadest sense as a person's incoming knowledge, skills, and even personality traits, and *treatment* refers to the condition or environment that supports learning. The point of ATI research is to provide information about learner characteristics that can be used to select the best learning environment for a particular student in order to optimize outcome. Yet, although many studies have been conducted that investigate the presence of ATIs (especially during the 1960s and 1970s), it has been surprisingly difficult to empirically verify learner-by-treatment interactions (see Cronbach & Snow, 1981, for a review of research). One problem with ATI studies conducted in the past has been data *noisiness.* That is, experimental data obtained from classroom studies contained a lot of extraneous, uncontrolled variables such as differing teacher personalities, instructional materials, and classroom environments. Recently, however, there has been renewed interest in examining the ATI issue while using computers as controlled learning environments (see Shute, 1993-a, 1993-b; Snow, 1990).

The macroadaptive approach to tutoring involves the use of ATI methodologies for providing information about learner aptitudes, information that can be applied in *macroadaptive* instruction that involves the selection of the best learning environment for a particular student. The justification for such an approach requires the demonstration that different individuals

learn better (or worse) in different learning environments—a classic ATI issue. In contrast, but not necessarily mutually exclusive, is *microadaptive* instruction (i.e., the more common approach to cognitive diagnosis, mentioned earlier). As the name implies, such instruction involves a low-level analysis of cognitive or learning abilities and is used in response to particular actions (e.g., it is used in the selection of the next small unit of instruction to be presented, based on a specific response history, often at the level of individual productions).

To illustrate the macroadaptive approach, Shute (1993-a, 1993-b) conducted an exploratory study using an intelligent tutoring system (ITS) that provides instruction in the basic principles of electricity. Aptitude and learning-style data were collected from more than 350 subjects who were randomly assigned to one of two learning environments: *rule-induction* (requiring the induction of principles) or *rule-application* (requiring the application of principles provided to the learner). The two learning environments were created from the one tutor by simply altering the feedback to the learner, everything else remaining the same. Results showed the following:

1. When the learning outcome being assessed was *declarative knowledge*, subjects with high measures of associative learning (AL) skills acquired significantly more knowledge from the tutor if they had been assigned to the rule-induction environment, and low AL subjects learned significantly more if they had been assigned to the rule-application environment.
2. When the learning outcome was *procedural skill* acquisition, high AL subjects developed significantly more skill if they had been assigned to the rule-application environment compared to the rule-induction environment, and low AL subjects performed poorly on the more difficult procedural skills tests, overall.
3. Subjects demonstrating exploratory behaviors performed significantly better on all outcome measures if they had been assigned to the rule-induction environment, and less exploratory learners acquired significantly more knowledge and skills from the rule-application environment.

Together these data provide some preliminary support and justification for use of a macroadaptive approach to cognitive diagnosis.

The purpose of this paper is to further explore possible relations among learner aptitudes, learning environments, and various learning outcome measures. To test the efficacy of the macroadaptive approach (i.e., the impact on learning outcome as well as on cost-effectiveness), results are presented from a large-scale experiment that was designed to test for ATIs from a systematically altered tutoring system that teaches flight engineering skills. The tutor was modified to yield two learning environments that differed only in the number of problems the learner needed to solve in each of 23 problem sets. The version with fewer problems was called *constrained*, and the version with many problems was called *extended*. A fixed ratio of 4:1 existed between the two tutor versions. I also measured an array of aptitudes before tutorial instruction, focusing on working-memory capacity (WM) and general knowledge (GK). Finally, I created a variety of learning outcome measures in order to assess a range of knowledge and skills acquired from the tutor. Each of these variables will be discussed in more detail following the presentation of hypotheses.

The simple main-effects hypotheses were straightforward. Subjects assigned to the constrained environment should take less time to complete the tutor because there were considerably fewer problems for them to solve. However, these same subjects were not expected to perform as well on the posttests when compared with subjects learning from the extended environment who would be receiving considerably more practice solving tutor-related problems. Also, it was hypothesized that subjects with higher aptitude scores would perform better on all posttests regardless of environment when compared to subjects with lower aptitude scores.

The interaction hypothesis tested whether learning outcome would be affected if aptitude level was differentially matched to learning environment. Given the two aptitude measures (WM and GK) and the two learning environments (constrained and extended), the interaction hypothesis can be seen in Table 1.

The basis for the interaction hypothesis was the notion that *low aptitude* individuals should profit from the extra practice provided by the extended environment. The additional exercises should foster knowledge and skill acquisition. It was understood that low aptitude subjects learning from the constrained environment might not receive enough problems in order to form proper generalizations and the consequent understanding of the subject matter. In a sense, these subjects would be mentally malnourished in the constrained environment.

High aptitude subjects, on the other hand, were expected to perform equally well on the outcome tests regardless of learning environment. If assigned to the constrained environment, they would have the knowledge and skills needed to extract meaningful information from fewer problems and would require fewer exemplars and problems to extract the gist of the problem set and to weave new knowledge into existing knowledge bases. But high aptitude subjects placed in the extended environment, similar to the lows, might also benefit from the extra practice associated with this environment.

Table 1
Interaction Hypothesis: Matching Aptitude to Environment Enhances Outcome

Aptitude	*Environment*	*Expected Outcome*
Low	Extended	High
Low	Constrained	Low
High	Constrained	High
High	Extended	High

METHOD

Subjects

The subjects in this study consisted of 178 males and females participating in a 7-day study on the acquisition of flight engineering knowledge and skills from an intelligent tutoring system. The gender distribution in the sample was approximately 75% male and 25% female. All subjects were high school graduates (or equivalent) with a mean age of 22 years. Subjects were obtained from a local temporary-employment agency and paid for their participation that consisted of 45 hours of testing and learning. None of the subjects had any prior experience or training as flight engineers or pilots.

Flight Engineering Tutor

The tutor was originally developed at the University of Pittsburgh (Lesgold, Bunzo, McGinnis, & Eastman, 1989) and then modified at the Armstrong Laboratory to fit experimental objectives. The tutor was designed to teach the knowledge and skills associated with a flight engineer's job. Relevant job components included the collection and analysis of information about pending flights and decisions about whether various factors (e.g., weather and runway conditions, type of plane) warrant or preclude safe flights. The body of the tutor consisted of two main parts: *Graph Reading* and the *TOLD* (Take Off and Landing Data) sheet.

The Graph Reading portion of the curriculum consisted of 14 problem sets that related to reading and understanding progressively more complex graphs. Problem sets from this part of the tutor focused on scale conversions, linear relationships, and the interpretation of polar and Cartesian coordinate charts. Figure 1 illustrates a simple problem concerning scale conversion—the relationship between the Fahrenheit and Celsius scales.

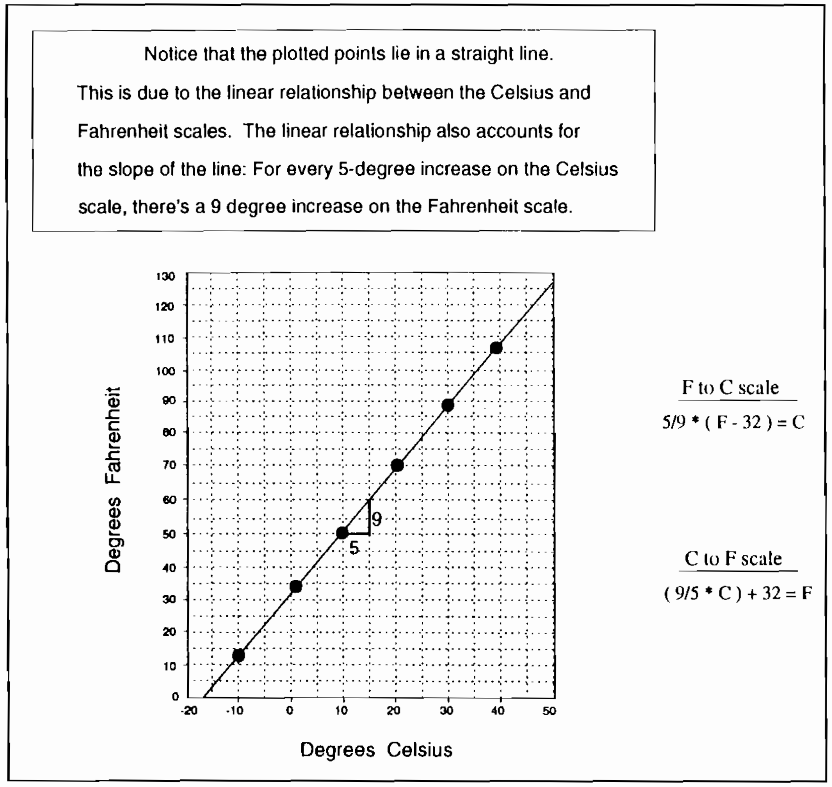


Figure 1. Scale conversion problem from graph part of the tutor: Fahrenheit and Celsius scales

All problem sets involved teaching both declarative and procedural knowledge. For instance, some declarative knowledge included the following:

> When the wind is blowing against the back of the plane moving down the runway, this is called a *tailwind*. If it is blowing against the nose of

> the plane, this is a *headwind*. When it is blowing across the runway, against the side of the plane (left to right, or right to left), this is called a *crosswind*.... The *relative wind direction* refers to the direction of the wind relative to the runway heading....

Procedural knowledge included rules to be learned. For instance, after learning the declarative knowledge presented above, learners were taught:

1. To compute the relative wind direction, subtract the wind direction from the runway heading (or the runway heading from the wind direction), always subtracting the smaller value from the larger one.
2. If the relative wind direction is less than 90 degrees or greater than 270 degrees, then there is a headwind.
3. If the relative wind direction is greater than 90 degrees but less than 270 degrees, then there is a tailwind.
4. At 90 or 270 degrees, there is a crosswind.

Learners then applied the rule while solving a series of problems dealing with the specific topic.

The TOLD Sheet part of the curriculum consisted of 9 problem sets and involved implementation of the graph reading skills taught earlier in order to fill in the TOLD worksheet. There were three main parts to this section:

1. computing the maximum allowable crosswind,
2. computing the headwind, tailwind, and crosswind components, and
3. integrating procedures with the given information to determine whether to take off or not.

The online TOLD sheet was designed to match the actual worksheet used by flight engineers (see Figure 2).

On this worksheet, information must be assembled and entered into the proper cells. Some information is given to subjects, information such as the gross weight of the aircraft, wind direction and velocity, the length, heading, and conditions of the runway, and obstacles in the flight path. Other information needs to be figured out by using the given information in conjunction with various charts and graphs. For instance, the wind-components chart is used to compute the headwind, tailwind, and crosswind components and consists of two superimposed charts: polar and Cartesian coordinates. In Figure 3, the chart is used to determine the headwind and crosswind components in an 8-step procedure. As can be seen, this is a nontrivial exercise.

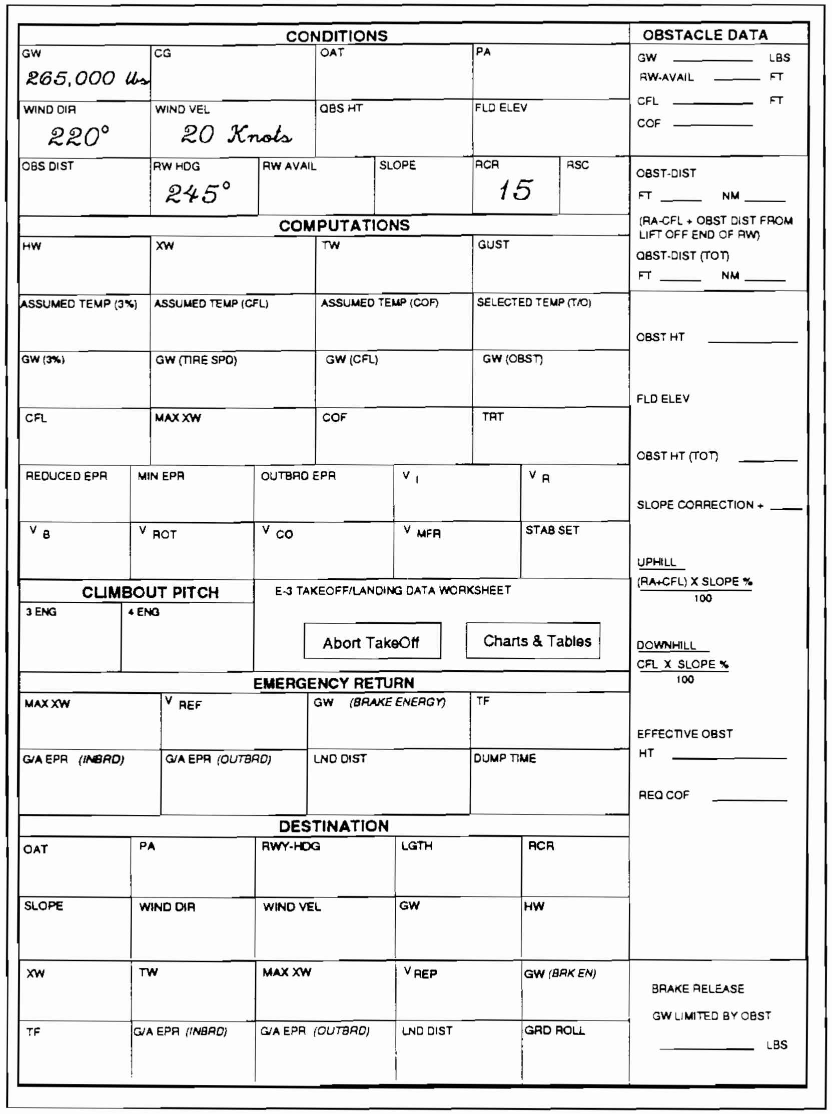


Figure 2. TOLD sheet used in the tutor and by flight engineers

Online tools are available to assist the learner in making the necessary computations. Some of the tools that can be used in conjunction with the

charts include tools for drawing vertical or horizontal lines, adding a radius or vector, erasing lines, redisplaying graphs, and so forth. In addition, a *Help* window allows the learner to read related information about the topic under study. Problem sets within each category were progressively more difficult. All learning was self-paced.

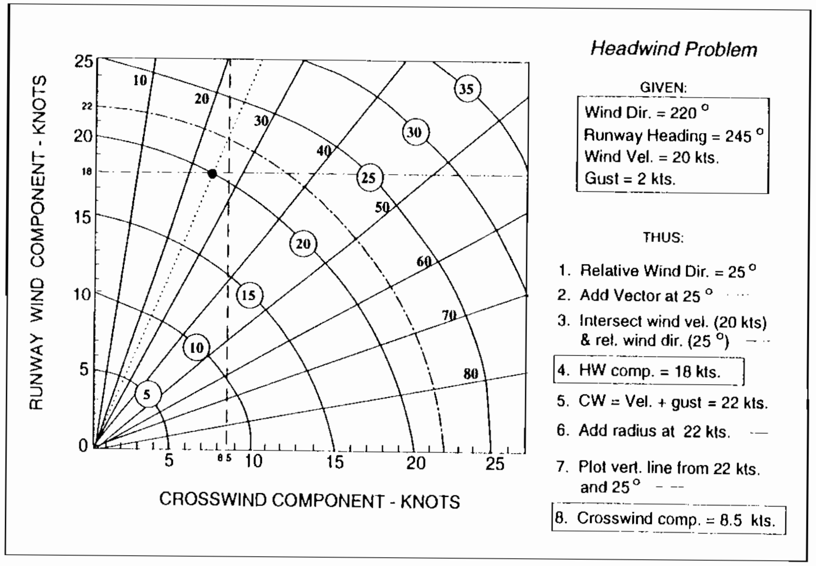


Figure 3. Solution of headwind and crosswind problem by using the wind components chart

Learning Environments

When we[1] received the tutor, each of the 23 problem sets required the correct solution of numerous problems (not consistent in number but typically greater than 10). However, such a problem-solving requirement seemed somewhat excessive. For instance, after solving just a few problems (e.g., converting inches to feet or Fahrenheit to Celsius), simple procedures would easily be acquired (i.e., a direct mapping between relevant scales).

Some of the 10 pilot subjects ($N = 6$, author included) reported that it was very frustrating to be forced to solve so many additional problems after having already learned the concept or procedure in question. For the others, this was not reported as a problem because they needed the extra practice to learn the concepts. So, even within the pilot-testing of the tutor, there existed differences among learner characteristics, differences that suggested at possible aptitude-treatment interactions. Motivated by this idea, I created learning environments that differed only in the number of practice problems given to the subjects.

Accordingly, the extended learning environment consisted of 12 problems per problem set and represented a range of values. By comparison, the constrained version consisted of only 3 select problems per problem set. As mentioned earlier, a 4:1 ratio was maintained between the extended and constrained versions of the tutor (12 to 3), and the constrained environment contained a subset of items from the extended environment. This ratio was established to provide a reasonable (yet maximal) contrast between practice environments. Twelve problems constituted a large enough number to examine learning curves while three problems were believed to be the minimum number of problems required to acquire novel concepts and skills presented by the tutor.

Both main parts of the tutor (Graph Reading and the TOLD sheet) were divided in this manner. That is, the graph portion as well as the TOLD sheet portion each consisted of two learning environments (extended and constrained). Before instruction began for the graph portion of the tutor, subjects were randomly assigned to either the extended or constrained version. Likewise, prior to instruction for the TOLD sheet portion, subjects were randomly assigned to either the extended or the constrained version. Thus, there were actually four different learning environments corresponding to whether the two main tutor parts were short or long:

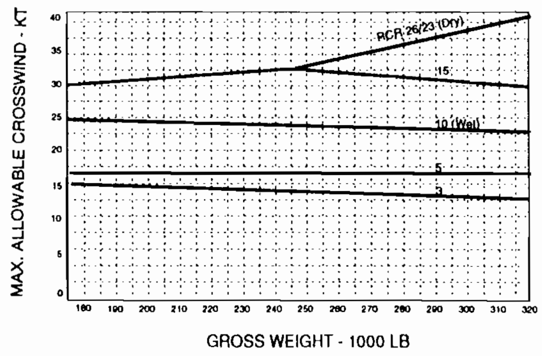
1. short, short (constrained),
2. long, short (emphasis on graph learning),
3. short, long (emphasis on TOLD sheet), and
4. long, long (extended).

For this paper, I focus on only the two extreme contrasts: constrained and extended.

Figure 4 illustrates a relatively easy problem set from the tutor: figuring out the *maximum allowable crosswind*, given the gross weight of the aircraft and the runway condition reading. For an aircraft weighing 265,000 pounds and a runway condition reading of 26/23 (nice and dry runway), the maximum allowable crosswind is 35 knots. Prior to the problem solving phase of each problem set (seen in the figure as empty boxes to

be filled in), concepts or procedures related to the problem set were discussed and illustrated by the tutor in detail. For example, the tutor explicitly demonstrated how to figure out the maximum allowable crosswind before presenting any problems. First, the computer drew a line upwards from the x axis (corresponding to the given gross weight). This line was shown intersecting the RCR. Then, another line was displayed moving across to the y axis to yield the crosswind value. Text always accompanied the visual (dynamic) displays (e.g., "The MAXIMUM ALLOWABLE CROSSWIND FOR TAKEOFF chart is shown below. To use the chart, values from two variables are needed: 1. Gross weight of the aircraft and 2. Runway condition reading [RCR]. RCR is a measure of the tire-to-runway coefficient of friction...").

Gross Weight (x1000)	RCR	Max. XWind
265	26/23	☐
220	10	☐
320	5	☐
190	26/23	☐
285	3	☐
240	15	☐
205	15	☐
310	10	☐
255	26/23	☐
215	5	☐
290	3	☐
225	26/23	☐



Given gross weight and RCR (runway condition reading), determine the maximum allowable crosswind for takeoff using the chart above and the tools from the tool menu.

Figure 4. Computing the maximum allowable crosswind, extended version (12 problems)

The tutor, on the whole, tended to be very graph intensive—a flight engineer needs to be able to read and interpret complex graphs proficiently because misreading a graph can have severe consequences. Data derived from various graphs provide the basis for important decisions concerning a safe takeoff and landing. So, what incoming aptitudes predict success in this area?

Aptitude Measures

Two aptitudes were examined in this paper: working-memory capacity (WM) and general knowledge (GK). These represent a subset of measures assessed in a larger battery of computerized cognitive tests, the CAM-4 battery (Kyllonen et al., 1990). These tests have been administered to thousands of subjects over the past several years and constitute valid and reliable measures.

Working memory generally refers to either the temporary storage or the activation level of information being processed during a cognitive task (e.g., Anderson, 1983; Baddeley, 1986). Furthermore, working-memory capacity has repeatedly been shown to be a strong predictor of learning across many and varied learning tasks (e.g., Ackerman, 1988; Kyllonen & Christal, 1990; Shute, 1991; Woltz, 1988). In this study, four working-memory tests were administered in each of three domains (verbal, quantitative, and spatial), yielding 12 WM tests altogether (with matched test paradigms in each of the domains). Figure 5 shows an example working-memory test across domains, the *4-Term Order Test*. To illustrate, in the verbal domain, subjects see three statements, displayed one at a time, at the top of the computer screen. The first two statements describe the order in which four key words are to appear (*dog* comes before *cat*; *desk* comes after *chair*). The third statement describes the sequence of the other two by using category names (*furniture* comes before *animals*). The correct response for this example is: chair desk, dog cat. Subjects must simultaneously store old information and process new information in order to succeed on this task. Eight numbered alternatives appear on the computer screen and consist of all possible sequences of the words. Subjects are allowed 15 seconds to type in one of the numbered responses after which they are told whether it was correct or not. With incorrect responses, subjects are allowed to review the three sentences and alternative answers to understand why they made a mistake. This test contains 24 items, and this test paradigm was also administered in the quantitative and spatial domains using numeric and spatial stimuli, respectively.

General knowledge is a measure of the breadth of common knowledge possessed by, and accessible to, an individual. In tests measuring GK, subjects are required to answer questions by typing in the first two letters of the answer (e.g., What is the name of the largest mountain range in South America? Answer: AN <CR>, for *Andes*). To measure general knowledge in the spatial domain, one test measured subjects' knowledge of United States geography with questions concerning the direction from one major city to another (e.g., If you are in Nashville, which direction would you travel to get to Seattle?). Eight numbered alternatives appeared on the screen along with the question until subjects typed in their response (e.g., 8 = Northwest). A compass also appeared in the upper right corner of the screen for subjects to use as a reference. This test contained 30 items. Figure 6 shows some sample items from the verbal and spatial domains.

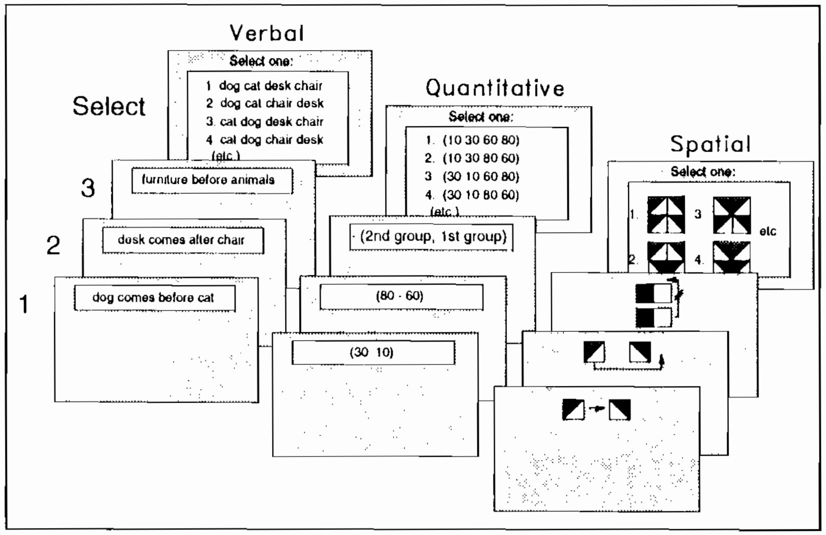


Figure 5. Working memory test example, across domains: Four Term Order Test

Learning Outcome Measures

The tutor's curriculum covered a wide range of new knowledge and skills and was generally very graph-intensive. While some of the basic graph-related material was familiar to many individuals (thus constituting a review), all of the material related to flight engineering constituted new knowledge and skills. This was assured because all subjects were screened to ensure that they had no prior related training or experience in the area. In addition, the nature of the material itself was fairly esoteric.

To assess the acquisition of tutor-specific knowledge and skills and of graph-reading and interpretation skills, three categories of posttests were created:

1. Basic Graph Knowledge and Skills,
2. Complex Graph Knowledge and Skills, and
3. Tutor-Specific Knowledge and Skills.

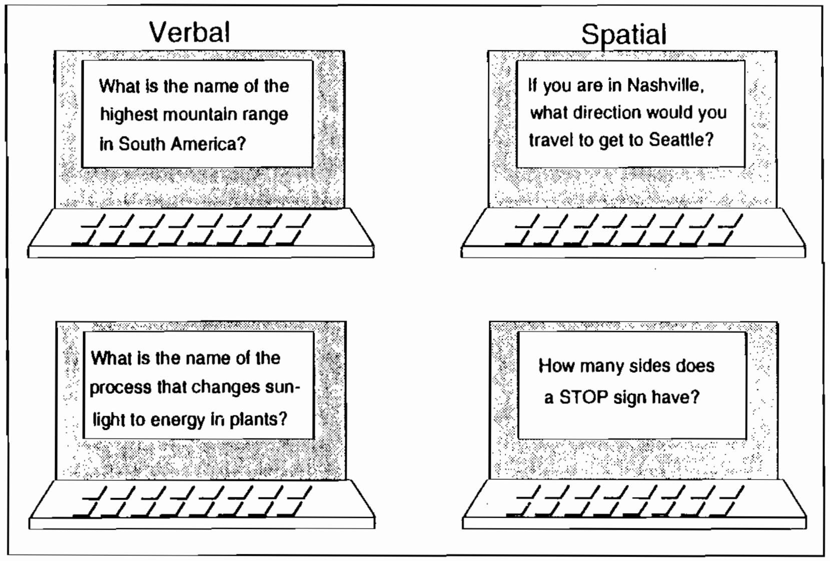


Figure 6. General knowledge test example, across domains

Each of these three categories consisted of several tests for measuring the acquisition of declarative knowledge and procedural skills. Test construction for graph reading and interpretation skills (categories a and b) was based on findings from research in the area. For example, Soule (1990) classified four hierarchical levels of graph reading skill: factual, inferential, relational, and synoptical. These levels are explained by the types of questions asked about a graph. At the factual level, questions involve selection of specific points and verification of the position according to labels (and vice versa). Inferential questions involve extrapolation from given information in order to predict points. The relational level concerns comparisons between two or more facets of the graph (e.g., overall trends). The synoptical question addresses the evaluation of global trends within a graph.

Another way to classify graphs is in terms of the features of the graph: local versus global and quantitative versus qualitative (Leinhardt, Zaslavsky, & Stein, 1990). By interpreting graphs according to a local (pointwise) approach, one is required to analyze the exact points of the graph. The global approach requires interpretation of the general features of a graph (e.g., shape, increasing and decreasing intervals) regardless of whether it represents a specific situation or an abstract functional relationship (Leinhardt et al., 1990). Quantitative interpretation requires analysis of specific values of the graph, and qualitative interpretation requires analysis of relationships and patterns on the graph as a whole.

These two classification systems guided the development of the five subtests in Posttest 1, Basic Graph Knowledge and Skills:

1. Declarative Graph Knowledge,
2. Quantitative Graph Reading—Factual,
3. Qualitative Graph Reading—Relational,
4. Quantitative Graph Interpretation—Points, and
5. Qualitative Graph Interpretation—Global.

Posttest 1 items were designed to assess fundamental graph knowledge and skills (e.g., understanding of general concepts such as what an *x*-axis is and how a linear trend appears). These tests were presented online in a multiple choice format with 6 alternatives[2] to choose from. There were 50 items altogether. See Figure 7 for sample items from some of these tests.

Posttest 2, Complex Graph Knowledge and Skills, was designed to assess performance on more diverse graph-related problems and included three subtests: (a) Understanding Functions, (b) Story Problems, and (c) Interpreting Complex Graphs. Understanding functions was regarded as an important principle because graphs and functions tend to be complementary (Leinhardt et. al., 1990). Although the tutor did not explicitly instruct algebraic functions, items were constructed where subjects could solve problems with given graphs or produce graphical representations mentally. Story problems were also included in this posttest and were created to be analogous to some of the problems encountered in the Flight Engineering tutor but without any of the vernacular. Subjects needed to solve items with graphical representations or formulas, thus extending the idea that functions and graphs are relative. Some of the problems contained graphs while others did not. Finally, the Complex Graphs section asked questions that followed the qualitative/quantitative, functional/relational, and local/global categorization discussed above. For this test, complex graphs were constructed for the subjects to interpret (e.g., scatter plots, multidimensional graphs). In Posttest 2, the items were designed to measure sophisticated

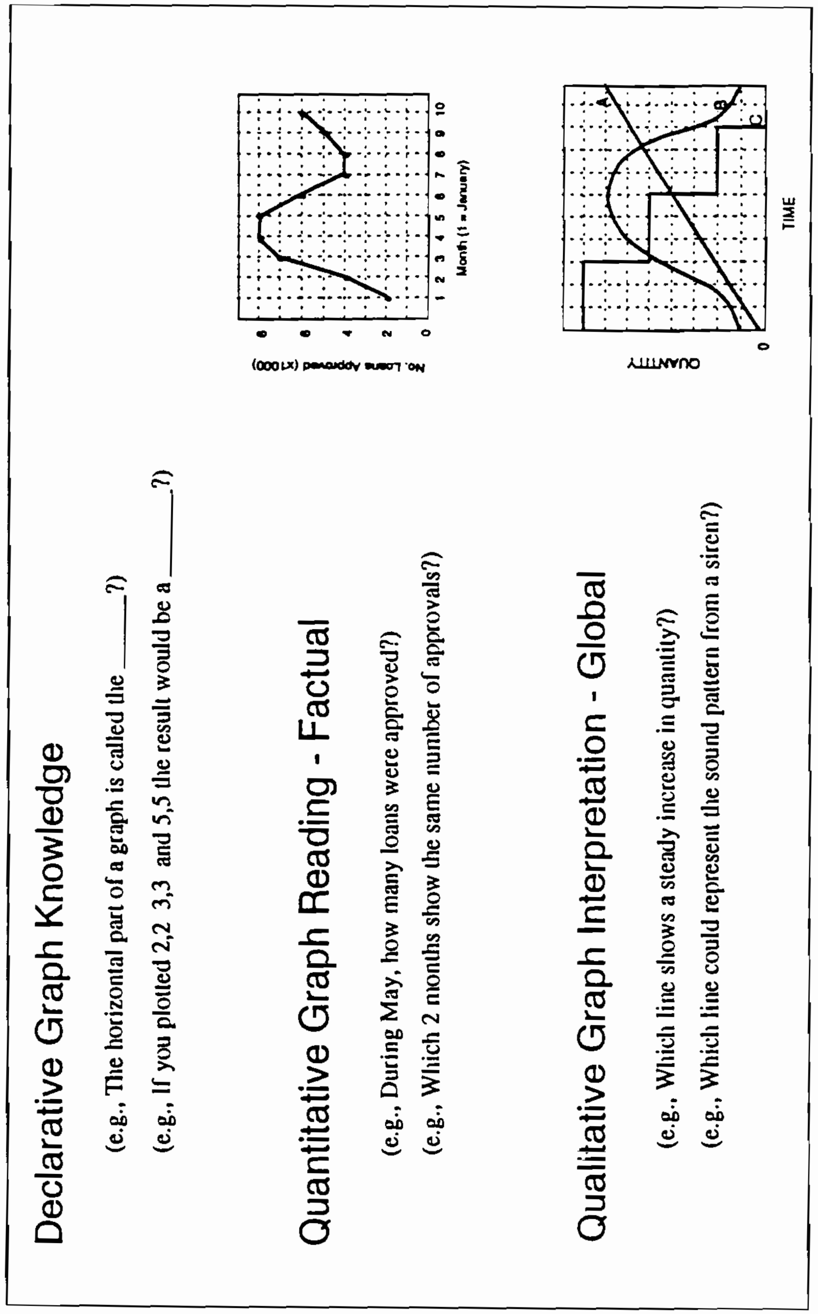


Figure 7. Posttest 1 sample items: basic graph knowledge and skills

graph interpretation skills. This test was administered online in a multiple choice format with (typically) 6 alternatives to choose from. There were 30 items altogether in this test. See Figure 8 for samples of test items.

Posttest 3, Tutor-Specific Knowledge and Skills, was designed to assess the depth and breadth of knowledge and skills acquired from the tutor. All three parts comprising this posttest were tutor-specific. The three parts measured

1. declarative knowledge about flight engineering,
2. procedural knowledge about flight engineering, and
3. graph reading and interpretation of flight engineering charts and graphs (i.e., Cartesian coordinate grid, Maximum Allowable Crosswind chart, and Wind Components chart).

Declarative knowledge items required subjects to show an understanding of the *definitions* of important concepts learned from the tutor. Procedural knowledge questions required subjects to demonstrate an understanding of *rules* learned from the tutor. For the Graph Interpretation section, subjects were given specific graphs and were required to use them in solving tutor-related problems (see Figure 9). There were 45 items altogether in this posttest, administered online in a multiple choice format with 6 alternatives to choose from.

Procedure

Subjects were tested in groups of approximately 20 persons, and there were 20 groups tested, total.[3] Each group spent 7 days (6 hours per day) in this study. Subjects began the study by being tested on basic cognitive-process measures and were then administered a pretest consisting of items parallel to those found in Posttest 1 (i.e., the Basic Graph Knowledge and Skills posttest). Following this testing, subjects were randomly assigned to one of the learning environments of the tutor. At the end of the tutoring session, subjects were administered the online battery of posttests.

Experimental cognitive-aptitude tasks as well as the pretest and posttests were administered on Zenith 248 microcomputers (AT-compatible) with standard keyboards and EGA monitors. The flight engineering tutor was administered on Xerox 1186 computers with standard keyboards and high resolution monochromatic displays on 19-inch monitors. Software was written in Interlisp-D and LOOPS. All computers were located in the same laboratory at Lackland Air Force Base, Texas.

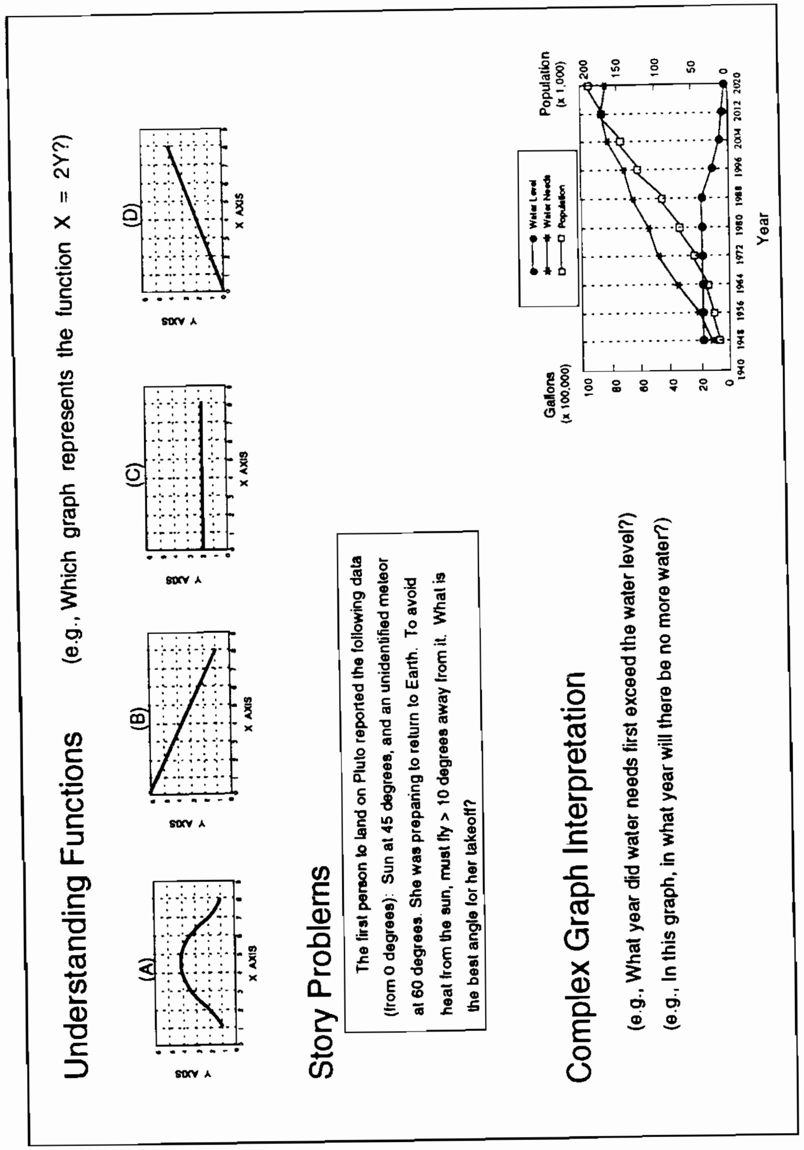


Figure 8. Posttest 2 sample items: complex graph knowledge and skills

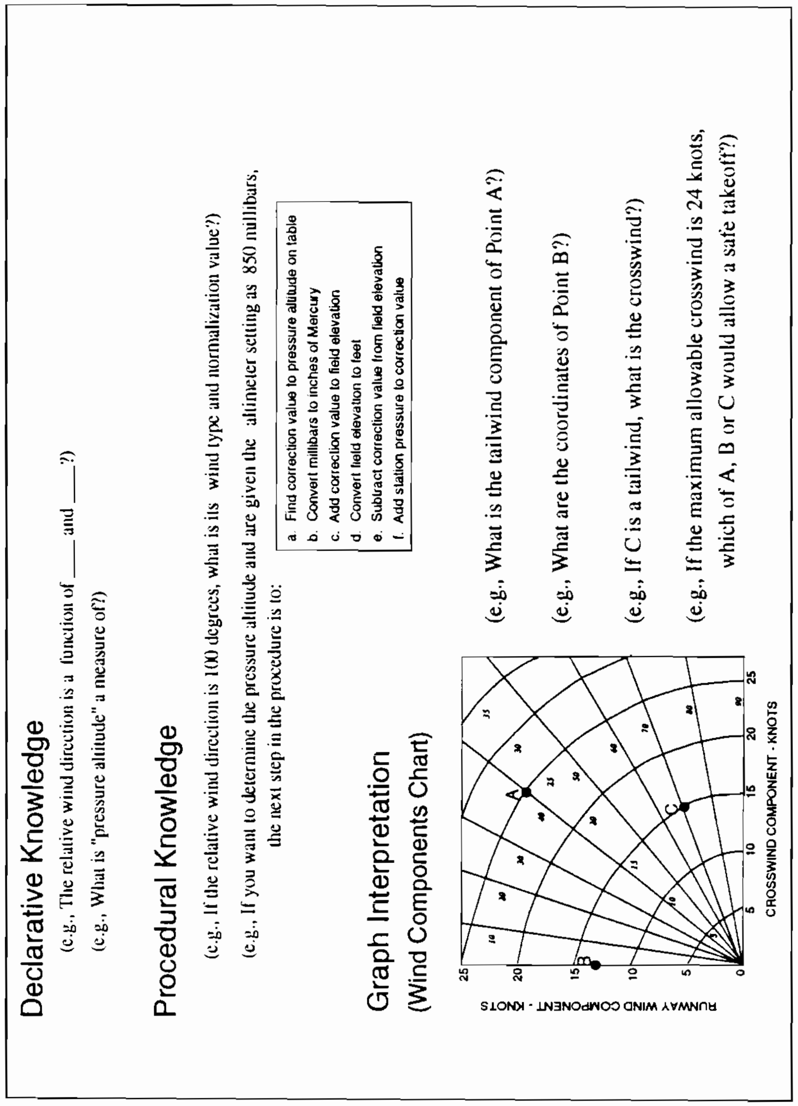


Figure 9. Posttest 3 sample items: tutor-specific knowledge and skills

Subjects were neither aware that there were different versions of the tutor nor allowed to discuss their experiences with other subjects. In addition, subjects were paid for completing the study rather than being paid by the hour, so subjects were motivated to proceed at expeditious rates regardless of experimental condition.

RESULTS

Main Effects

When the data were examined separately by environment for each posttest and time on tutor, the hypotheses concerning simple main effects of learning environment were not supported. These data can be seen in Table 2. As can be seen in Table 2, there were no differences on *any* of the posttests for individuals in the two contrasting environments. However, subjects in the extended environment took significantly more time to complete the tutor (over 4.5 hours more, on average) than subjects assigned to the constrained environment. So, one conclusion that could be drawn (albeit prematurely as will be demonstrated later) is the following: Given this *no difference* finding on the outcomes, all subjects should be placed in the constrained environment because subjects completed the tutor much faster in that environment with no adverse effects on outcome performance.

The second main effects hypothesis tested whether individuals with higher aptitude measures (WM and GK) would perform better on all outcome measures when compared to subjects with lower aptitudes. I computed median splits of the WM and GK data, and results are shown below in Table 3. While this hypothesis was supported, it was not at all surprising because high aptitude individuals were expected to perform better on outcome measures than low aptitude subjects.

Aptitude Measures

To test whether WM and GK comprised one general factor (e.g., g), two distinct factors, or some other factor structure, I computed a factor analysis (principle components with varimax rotation) on the data from the 12 WM tests and 5 GK tests. Two orthogonal factors were extracted, accounting for 54% of the variance: WM factor and GK factor. Factor scores were saved for each person and later used in a regression analysis predicting learning outcome.

Table 2
Posttest Scores (Percent Correct) and Time on Tutor, by Learning Environment

Criteria	*Constrained (N = 91) MEAN (SD)*	*Extended (N = 87) MEAN (SD)*	*F-ratio*	*Signif*
1. Basic Graphs (AVG)	59.7 (18.9)	57.6 (20.2)	0.54	NS
Declarative Knowledge	61.5 (27.4)	58.2 (25.5)	0.73	NS
Quantitative Graph Reading	52.7 (25.3)	50.6 (24.5)	0.34	NS
Qualitative Graph Reading	69.9 (22.8)	64.5 (25.8)	2.20	NS
Quantitative Graph Interp.	61.0 (19.7)	63.0 (22.8)	0.39	NS
Qualitative Graph Interp.	53.5 (22.8)	51.7 (23.3)	0.27	NS
2. Complex Graphs (AVG)	57.0 (19.4	57.6 (18.5)	0.05	NS
Understanding Functions	28.4 (33.6)	28.4 (34.5)	0.00	NS
Story Problems	57.5 (17.8)	58.1 (18.6)	0.06	NS
Complex Graph Interp.	69.5 (23.9)	70.4 (21.8)	0.06	NS
3. Tutor Specific (AVG)	59.0 (18.7)	60.9 (19.4)	0.43	NS
Declarative Knowledge	61.4 (23.8)	61.0 (24.1)	0.01	NS
Procedural Knowledge	46.9 (22.6)	48.6 (24.2)	0.23	NS
Turor Graph Skills	62.8 (19.3)	65.7 (19.4)	1.02	NS
4. Time on Tutor (Hours)	7.1 (3.2)	11.7 (4.8)	59.19	p<.0001

Notes. Percent correct scores are included for each posttest and standard deviations are included in parentheses. The time on tutor data reflect the required number of hours to complete the tutor.

Because of the extraction technique employed (i.e., principal components analysis), the correlation between these two factors was 0. This is important because these two variables are normally correlated (between r = .60 to .70). If I had used the raw data, many subjects might have fallen into the extreme categories (namely, high-WM/high-GK and low-WM/low-GK), and the two interesting *crosses* (involving WM and GK) would be poorly represented by individuals. This turned out not to be a problem because with the orthogonal factors, subjects covered a range of aptitude profiles that included the two extreme groups as well as those characterized by high-GK/low-WM and low-GK/high-WM. Furthermore, when subjects' data (N = 159) were divided into aptitude categories based on median splits of WM and GK, there was a fairly equal distribution: low-WM/low-GK (N = 37), low-WM/high-GK (N = 33), high-WM/low-GK (N = 41), and high-WM/high-GK (N = 48).

Table 3

Posttest Scores Separated by Aptitude Level: Low v. High WM and GK

Criteria	*Low WM (N = 79)*	*High WM (N = 78)*	*F-ratio*	*Signif*
1. Basic Graphs (AVG)	50.5 (19.3)	69.0 (14.7)	46.15	p<.001
Declarative Knowledge	49.5 (27.5)	73.0 (19.8)	37.60	p<.001
Quantitative Graph Reading	42.3 (22.6)	63.6 (22.2)	35.52	p<.001
Qualitative Graph Reading	58.9 (25.8)	77.4 (18.7)	26.68	p<.001
Quantitative Graph Interp.	56.7 (22.6)	69.2 (17.3)	15.17	p<.001
Quantitative Graph Interp.	45.3 (23.3)	61.5 (19.7)	22.18	p<.001
2. Complex Graphs (AVG)	43.5 (16.6)	63.7 (17.6)	54.73	p<.001
Understanding Functions	16.2 (24.1)	42.6 (37.8)	29.16	p<.001
Story Problems	51.3 (18.8)	67.1 (11.7)	39.75	p<.001
Complex Graph Interp.	63.1 (22.6)	80.5 (16.8)	30.17	p<.001
3. Tutor-specific(AVG)	50.6 (19.8)	66.9 (15.8)	32.60	p<.001
Declarative Knowledge	53.7 (26.7)	70.3 (17.3)	21.20	p<.001
Procedural Knowledge	41.5 (22.2)	55.9 (23.1)	15.82	p<.001
Tutor Graph Skills	56.4 (18.8)	74.6 (14.5)	45.81	p<.001
	Low GK (N = 79)	*High GK (N = 78)*	*F-ratio*	*Signif*
1. Basic Graphs (AVG)	48.9 (18.2)	70.3 (14.1)	67.71	p<.001
Declarative Knowledge	48.2 (25.1)	73.9 (21.6)	47.55	p<.001
Quantitative Graph Reading	41.7 (22.3)	63.9 (22.0)	39.57	p<.001
Qualitative Graph Reading	56.9 (24.3)	79.1 (18.8)	41.09	p<.001
Quantitative Graph Interp.	54.9 (23.2)	70.9 (15.0)	26.45	p<.001
Qualitative Graph Interp.	42.9 (21.0)	63.7 (20.1)	39.79	p<.001
2. Complex Graphs (AVG)	43.5 (16.8)	63.7 (17.4)	64.56	p<.001
Understanding Functons	18.5 (25.1)	41.3 (38.7)	19.06	p<.001
Story Problems	50.6 (18.5)	67.7 (11.3)	48.58	p<.001
Complex Graph inters.	61.4 (21.9)	82.2 (15.7)	46.18	p<.001
3. Tutor-specific (AVG)	47.3 (17.6)	69.9 (14.6)	76.39	p<.001
Declarative Knowledge	49.4 (23.8)	74.3 (16.5)	58.26	p<.001
Procedural Knowledge	37.1 (19.8)	60.1 (21.6)	48.55	p<.001
Tutor Graph Skills	55.5 (19.1)	75.2 (13.2)	56.50	p<.001

Notes. Percent correct scores are included for each posttest and standard deviations are included in parentheses. The sample sizes of 79 and 78 are slightly lower than the other sample sizes reported in this paper (e.g., Table 2 data) as not all subjects finished the 12 working memory and 5 general knowledge tests. This resulted in missing values.

Outcome Measures

The three posttests were found to be reliable measures. Odd-even reliabilities were computed across all items in the respective subtests, and these reliabilities were as follows—Posttest 1: $r = .91$, Posttest 2: $r = .90$, and Posttest 3: $r = .91$. Thus, the tests reliably assessed what they were intended to measure: Posttest 1 = Basic graph knowledge, Posttest 2 = Complex graph knowledge, and Posttest 3 = Tutor-specific knowledge and skills. In addition, these tests turned out to be highly correlated with one another: $r_{(1,2)} = .84$; $r_{(1,3)} = .83$, and $r_{(2,3)} = .82$.

Because of the exploratory nature of this study and to simplify the data, I computed a factor analysis (principle axis factoring) on the three posttest scores. A single factor was extracted (thus no rotation necessary), accounting for 89% of the variance. This outcome factor was labeled *Postfac* and denoted general as well as specific knowledge-and skill-acquisition. The respective weights per test on the outcome factor were as follows—Posttest 1: .92, Posttest 2: .91, and Posttest 3: .91. Thus, all tests showed equal and high loadings on the general outcome factor. Factor scores were saved for each person and used as the dependent variable in the regression analysis that tested the interaction hypothesis.

Interaction: Aptitude by Environment

To test the interaction hypothesis, a regression analysis was computed that predicted the outcome score (*Postfac*) from the independent variables: WM factor, GK factor, Envir, WM × Envir, GK × Envir, WM × GK, and WM × GK × Envir. With all variables in the equation, $R^2 = .81$. Table 4 shows the results from the regression equation for these data.

The first observation from this table is that both WM and GK were very strong predictors of knowledge and skill acquisition. Not surprising, individuals with high WM and GK measures scored significantly higher on the outcome measures when compared with individuals with lower aptitude measures (see also Table 3). Next, there was no main effect of environment on learning outcome (i.e., there was no significant benefit from the extra practice in the extended environment). While this was a surprise, of most interest to this study was the three-way interaction involving WM × GK × Environment as a significant predictor of learning outcome.

To illustrate this three-way interaction, results from the regression equation are plotted separately by learning environment for four groups of subjects. These four groups were established based on WM and GK factor scores, where *low* refers to 1 standard deviation below the mean, and *high* refers to 1 standard deviation above the mean, per aptitude measure. Figure 10 depicts a plot of *expected outcomes* (based on the obtained regression equation using the b-weights shown in Table 4) for each of the two learning environments. What we see in Figure 10 is that the disordinal interaction occurs for the two groups of subjects that have *mixed* aptitudes (i.e., high-WM/low-GK and low-WM/high-GK). These two groups appear to have differential learning-environment requirements. The boxed area in Figure 10 has been enlarged for closer inspection of this interaction.

Table 4
Multiple Regression Solution Predicting Postfac
(Multiple $R = .90$, $R^2 = .81$, $N = 155$)

Variable	*B-Weight*	*DF*	*Sum of Squares*	*F-Test*	*Signif (P Value)*
WM Factor	0.600	1	45.01	260.22	.001
GK Factor	0.637	1	54.37	314.37	.001
Environment	0.025	1	0.09	0.54	NS
WM x Envir	0.064	1	0.51	2.95	NS
GK x Envir	-0.067	1	0.61	3.52	NS
WM x GK	-0.005	1	0.00	0.02	NS
WM x GK x Envir	0.078	1	0.88	5.10	.025

Cost-Benefit Analysis

So, what is the cost of this macroadaptive approach to tutoring? This question has two parts: the cost associated with aptitude measurement and the cost associated with ITS modification. Also, what is the payoff of this approach in terms of learning outcome?

***Cost of aptitude measurement*.** The *developmental* costs were nil because I used the computerized Cognitive Abilities Measurement (CAM-4) battery of tests (Kyllonen et al., 1990) to assess incoming knowledge and cognitive skills. This battery has been tested on thousands of subjects, and the internal consistency and test-retest reliabilities are high. The *administrative* cost of assessing the aptitudes in this study was about 2–3 hours. The average time per WM test was 11 minutes, and the average time per GK test was 5 minutes. But not all the tests used in this experiment would need to

be administered to come up with measures of WM and GK. Rather, tests with the highest factor loadings could be selected as the best representative tests for those particular factors, thereby cutting aptitude-testing time significantly. I chose to administer all the tests, given the exploratory nature of this study.

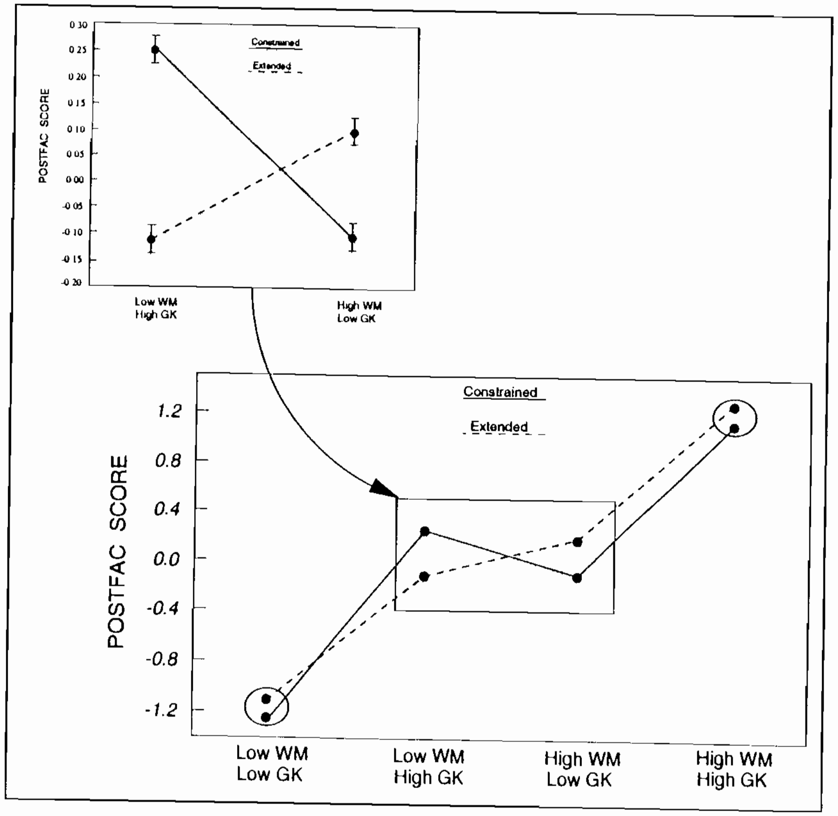


Figure 10. Regression equation plotted by learning environment predicting outcome score

***Cost of ITS modification*.** The *developmental* cost associated with creating the learning environments was minor. For the extended environment, it took about 1–2 hours to select 12 appropriate problems for each of the 23 problem sets. That required, in a few cases, the addition or deletion of problems. For the constrained environment, it also took between 1–2 hours to select a subset of 3 relevant problems from the pool of 12 problems in the extended version (per problem set). The *administrative* cost of the tutor manipulation was nil because in this exploratory study, subjects were randomly assigned to environment. But theoretically, a decision rule could be invoked to select the appropriate environment, given information about a student's aptitude. From the results of this exploratory study, the decision rule would be

If	low WM and high GK,
then	constrained environment,
Else	extended environment

The rationale underlying this decision rule is that while there is no significant main effect of environment on outcome performance, when subjects' aptitudes are considered, there seem to be differential effects of aptitude by treatment that influence learning outcome. While most groups of subjects seem to benefit from the extended environment to various degrees (see Figure 10), one group of subjects, characterized by low working-memory capacity but high general knowledge, performed significantly better in the constrained environment. The actual increases in outcome attributable to the matching of aptitude-to-environment are discussed next.

Benefit of this approach. Overall, the mean Postfac score = 0 (SD = .94, N = 159), so average outcome performance hovers around zero when subjects are randomly assigned to one of two environments. To determine the amount of outcome improvement using the decision rule, I computed four new variables for each subject. Using the b-weights from the regression equation in Table 4 in conjunction with the appropriate variable values, I computed

1. predicted outcome scores using subjects' actual (randomly-assigned) environment (*pred outcome*),
2. predicted outcome scores when environment was set to constrained (*pred constrained*),
3. predicted outcome scores when environment was set to extended (*pred extended*), and
4. predicted outcome scores assigning subjects to environment based on the decision rule determined from this study's data (*pred rule*).

Table 5 shows a summary of these new variables. The following equations were used (note: constrained environment = -1, extended = 1):

- pred outcome = (CONSTANT = .081) + (WM × .5995) + (GK × .6368) + (ENVIR × .0248) + (WM × ENVIR × .0638) + (GK × ENVIR × -.0673) + (WM × GK × -.0051) + (WM × GK × ENVIR × .0777).
- pred constrained = (CONSTANT = .081) + (WM × .5995) + (GK × .6368) + (-1 × .0248) + (WM × -1 × .0638) + (GK × -1 × -.0673) + (WM × GK × -.0051) + (WM × GK × -1 × .0777).
- pred extended = (CONSTANT = .081) + (WM × .5995) + (GK × .6368) + (1 × .0248) + (WM × 1 × .0638) + (GK × 1 × -.0673) + (WM × GK × -.0051) + (WM × GK × 1 × .0777).
- pred rule was computed as follows:
  If low WM and low GK, then pred rule = pred extended
  If low WM and high GK, then pred rule = pred constrained
  If high WM and low GK, then pred rule = pred extended
  If high WM and high GK, then pred rule = pred extended

Table 5
Predicted Outcome Values for All Subjects: Random Assignment, Constrained, Extended, or Decision Rule Environment

Variable	*N*	*Mean*	*SD*	*Min*	*Max*
Pred Outcome	159	0.05	0.83	-2.13	1.66
Pred Constrained	159	0.01	0.86	-2.41	1.66
Pred Extended	159	0.06	0.84	-1.94	1.94
Pred Rule	159	0.12	0.82	-1.94	1.94

If all subjects were randomly assigned to a learning environment as they were in this study, the mean predicted outcome score = .05 (SD = .83). The mean predicted outcome score for subjects learning from the constrained environment (pred constrained) = .01 (SD = .86) and from the extended environment (pred extended) = .06 (SD = .84). Finally, the predicted outcome score for subjects assigned to environment based on aptitude profile = .12 (SD = .82). Thus, random assignment or placement in one or the other environment produces fairly average performance. However, placement in an environment based on aptitude data shows the largest predicted gain. But how do we interpret these values?

One way to answer that question is to compute *effect sizes*. This provides a standardized basis for comparisons between conditions. Effect sizes are computed as (mean treatment – mean control)/SD control. *Treatment* refers to the condition to be tested in relation to some other baseline or *control* condition. Effect sizes were computed to examine differences between the following conditions:

1. constrained versus random (pred constrained - pred outcome)/.83,
2. extended versus random (pred extended - pred outcome)/.83,
3. pred rule versus random (pred rule - pred outcome)/.83,
4. pred rule versus constrained (pred rule - pred constrained)/.86, and
5. pred rule versus extended (pred rule - pred extended)/.84.

Table 6 shows a summary of these values.

Table 6
Effect Sizes for Different Condition Comparisons:
Constrained, Extended, and Rule-Based

Effect-Size Comparison	*N*	*Mean*	*SD*	*Min*	*Max*
Constrained vs. Random	159	-0.05	0.18	-0.64	0.69
Extended vs. Random	159	0.01	0.24	-1.91	0.49
Rule vs. Random	159	0.08	0.15	-0.15	0.69
Rule vs. Constrained	159	0.13	0.15	-0.33	0.62
Rule vs. Extended	159	0.07	0.21	-0.04	1.89

DISCUSSION

In summary, I created contrasting learning environments from one tutor: extended and constrained. These environments differed only in terms of the number of problems needed to be solved per problem set (4:1 ratio). Findings showed that overall there were no differences between these two environments on any of the outcome measures. However, subjects learning from the constrained environment did take significantly less time to complete the tutor when compared with subjects learning from the extended environment. The fact that the extended environment did not directly enhance learning outcome was surprising given the premise that "practice makes perfect." But when the data were analyzed with aptitudes in the equation, environments were shown to be differentially effective for individuals.

Findings extended the simple interaction hypothesis listed earlier (see Table 1). That is, an aptitude *profile* was found to interact with learning

environment rather than being simply *high* or *low* on a general aptitude measure. When aptitude *profiles* were matched to environment, learning outcome was increased. The optimal matches were low-WM/low-GK in the extended environment, low-WM/high-GK in the constrained environment, high-WM/low-GK in the extended environment, and high WM/high-GK in either environment. Furthermore, results from a regression equation showed that 81% of the learning outcome variance could be accounted for by these few variables: WM, GK, environment, and associated interactions, representing a large amount of predictive validity.

Computing effect size demonstrates how one treatment fares in relation to another (control) in terms of some variable such as outcome score. What should be the control condition? Although effect sizes were computed for the three treatments (constrained, extended, and decision rule) in relation to *random assignment* as the control, this does not represent a reasonable control condition because random assignment would not be warranted except in exploratory studies (like this). Assignment to environment should be based on efficiency or outcome considerations. Thus, a better baseline (or control) condition is the constrained environment because subjects take less time to go through the curriculum and do so with no adverse effects on outcome. When decision-rule assignment of subjects was compared to constrained assignment, the effect size = .13. In other words, subjects' predicted outcome scores are .13 standard units *higher* using the decision rule in contrast with simply assigning subjects to the shorter environment. Alternatively, one could compare the decision-rule versus extended environment assignment. The rationale underlying this contrast is the premise that "practice makes perfect"; thus all subjects should, by default, be placed in that environment. The effect size of this contrast = .07. Again, the decision rule is superior to either condition alone.

The cost associated with implementing the decision rule is low. Working memory and general knowledge questions could be presented online prior to the start of the tutor. If a subject scored below average on the working memory items but above average on the general knowledge questions, the subject would then be assigned to the constrained environment and others assigned to the extended condition. The investment of about one hour of aptitude testing could increase outcome performance by .13 or .07 standard units when compared with placing all subjects in either constrained or extended versions of the tutor.

Assigning all subjects to the constrained version would require less training time but at the cost of learning outcome for some groups of subjects. Assigning all subjects to the extended version would require considerably more training time but again would not be optimal for all subjects. Obviously, more research is needed to determine boundary and intermediate conditions to see how many practice problems are necessary and sufficient for different aptitude profiles. That is, while three problems may represent the lowest limit, and 12 a possible upper limit, an intermediate number of problems may provide the benefit of less extreme aptitude profiles. *Average* individuals may find some number in between 3 and 12 problems optimal. Also, consideration of other aptitudes in the equation may enrich the decision rule. In addition to measuring working memory and general knowledge, CAM-4 computerized cognitive tests measure information processing speed, associative learning, procedural learning, and inductive reasoning. These aptitudes were not investigated in this study, but other aptitude *profiles* may have differential learning requirements.

In conclusion, the potential utility of this approach is great: The cost is low, and the payoff is relatively high. Results from this research provide some preliminary information about which learning environments are more suitable for which learners and why. In this study, we saw that individuals with low incoming knowledge but high working-memory capacities profited from the extra practice afforded by the extended environment. These individuals had the capacity as well as the need (i.e., knowledge deficiency) to take on more information. However, individuals with a lot of incoming knowledge but low capacity were better off learning from a constrained environment. Their rich knowledge structures allowed them to interleave new knowledge into existing databases and enabled them to require fewer problems to get the point of a problem set. Furthermore, given relatively low WM capacities, they would not be as cognitively taxed in the constrained environment as in the extended environment. For others (i.e., high-WM/high-GK and low-WM/low-GK), learning environment did not influence outcome. These subjects performed well and poorly on the outcome tests, respectively.

The research discussed earlier employing the macroadaptive approach (Shute, 1993-a, 1993-b) identified other matches between aptitudes and learning environments. And, as in the present study, other research has found ATIs involving aptitude profiles where pairs of aptitudes function interactively to influence outcome differentially by environment. For example, Snow (1989) reported a study that contrasted two learning conditions: a well-structured, filmed presentation of physics problems versus an ill-structured, live demonstration of the same physics problems. The findings were that individuals with low knowledge and high aptitude (quantitative abilities) showed better learning from the open-ended live presentation

than from the more structured (film) condition. However, subjects with high knowledge and low aptitude measures performed better under the structured condition than under the live condition. But much more research is required to validate these exploratory studies. In all cases, subjects were randomly assigned to environment, and data were analyzed *after the fact* to determine the presence of ATIs. A confirmatory test of these findings is needed to validate the approach. In other words, subjects need to be assigned a priori to learning environment based on their aptitude profiles. Plans for conducting such a study are underway, and a confirmatory test of the macroadaptive approach using the electricity tutor will start this year.

In conclusion, the tutor used in this study was not intelligent in the sense defined earlier (i.e., the degree to which the system modifies instruction based on an inferred model of the student's current understanding). Rather, it simply presented the flight engineering curriculum (23 problem sets), and learners solved either 3 (constrained) or 12 (extended) problems per problem set. But no attempt was made to have the system *intelligently* manage the number of practice problems. On the contrary, in this exploratory study, the number of practice problems one received was fixed (and randomly determined). The point was to show how a relatively *unintelligent* tutoring system can promote learning improvements using an alternative, macroadaptive approach to tutoring—by determining optimal matches between aptitude profiles and learning environments.

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.

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

Anderson, J.R., Boyle, C.F., & Reiser, B.J. (1985). Intelligent tutoring systems. *Science, 228*, 456–462.

Anderson, J.R., Conrad, F.G., & Corbett, A.T. (1989). Skill acquisition and the LISP tutor. *Cognitive Science, 13*(4), 467–505.

Baddeley, A. (1986). *Working memory*. Oxford: Clarendon Press.

Cronbach, L.J., & Snow, R.E. (1981). *Aptitudes and instructional methods: A handbook for research on interactions*. New York, NY: Irvington.

Kyllonen, P.C., and Christal, R.E. (1990). Reasoning ability is (little more than) working-memory capacity?! *Intelligence, 14*, 389–433.

Kyllonen, P.C., Christal, R.E., Woltz, D.J., Shute, V.J., Tirre, W.C., & Chaiken, S. (1990). *Cognitive Abilities Measurement (CAM) Test Battery, Ver-*

sion 4.0. Unpublished computerized test battery, Armstrong Laboratory, Brooks Air Force Base, TX.

Leinhardt, G., Zaslavsky, O., & Stein, M.K. (1990). Functions, graphs, and graphing: Tasks, learning, and teaching. *Review of Educational Research, 60*(1), 1–64.

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

Mandl, H., & Lesgold, A. (Eds.). (1988). *Learning issues for intelligent tutoring systems*. New York, NY: Springer-Verlag

Shute, V.J. (1991). Who is likely to acquire programming skills? *Journal of Educational Computing Research, 7*, 1–24.

Shute, V.J. (1993-a). Aptitude-treatment interactions and cognitive skill diagnosis. In J.W. Regian & V.J. Shute (Eds.), *Cognitive approaches to automated instruction*. Hillsdale, NJ: Erlbaum.

Shute, V.J. (1993-b). A Comparison of learning environments: All that glitters.... In S.P. Lajoie & S.J. Derry (Eds.), *Computers as cognitive tools*. Hillsdale, NJ: Erlbaum.

Sleeman, D., & Brown, J.S. (1982). *Intelligent tutoring systems*. London: Academic Press.

Sleeman, D., Kelly, A.E., Martinak, R., Ward, R.D., & Moore, J.L. (1989). Studies of diagnosis and remediation with high school algebra students. *Cognitive Science, 13*(4), 551–568.

Snow, R.E. (1989). Cognitive-conative aptitude interactions in learning. In R. Kanfer, P.L. Ackerman, & R. Cudeck (Eds.), *Abilities, motivation, and methodology* (pp. 435–474). Hillsdale, NJ: Erlbaum.

Snow, R.E. (1990). Toward assessment of cognitive and conative structures in learning. *Educational Researcher, 18*(9), 8–14.

Soule, C.S. (1990). *A cognitive skills analysis of a graph reading task.* Unpublished doctoral dissertation, University of Georgia, Athens, GA.

VanLehn, K. (1990). Mind Bugs: The origins of procedural misconceptions. Cambridge, MA: MIT Press.

Wenger, E. (1987). Artificial intelligence and tutoring systems. Los Altos: Morgan Kaufmann.

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

Notes

1. *We* refers to the Learning Abilities Measurement Program (LAMP), Armstrong Laboratory, Brooks Air Force Base, TX. The tutor was developed for LAMP as part of a research project examining the predictive validity of cognitive process measures and using intelligent tutoring systems as dependent measures (or complex learning tasks).
2. For just a few items, there were only four alternatives to choose from (e.g., Qualitative Graph interpretation—Global had four choices: curve a,b,c, or none).
3. Some individuals (about 45 out of 400) dropped out before completing all parts of the study. In addition, there were four conditions in the larger experiment, but I am focusing on just two of them. Thus the N in this paper (178) represents a subset (about one half) of all subjects actually tested.

Acknowledgements

I would like to thank Pat Kyllonen, Bill Alley, Susan Embretson, Wes Regian, and Lisa Gawlick-Grendell for their valuable suggestions, conceptual and analytical, during the conduct and summary of this research. Rich Walker, Trace Cribbs, and Jamie Burns provided excellent programming support for the tutor and outcome tests while Wayne Crone and Linda Robertson-Schule provided enormous assistance in the massive data collection job.

The research reported in this paper was conducted as part of the Armstrong Laboratory, Human Resources Directorate—Learning Abilities Measurement Program (LAMP). 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 does not necessarily reflect those of the Air Force.