2 Aptitude-Treatment Interactions and Cognitive Skill Diagnosis

Valerie J. Shute
Armstrong Laboratory, Brooks Air Force Base, Texas

Individuals come to any new learning task with differing profiles of knowledge and skills. The "intelligence" in an intelligent tutoring system (ITS) resides in the ability to analyze solution histories dynamically, using principles, rather than preprogrammed responses, to decide what to do next (e.g., Clancey, 1986), and to adapt instruction to different learners (e.g., Sleeman & Brown, 1982; Wenger, 1987). Valid and reliable cognitive diagnoses, then, are essential to computer systems that adapt to their users' needs.

The standard approach to cognitive skill diagnosis represents *emerging* knowledge and skills of the learner. The computer responds to these updated observations with a modified curriculum, adjusted error by error, action by action, minute by minute. Instruction, therefore, is dependent on individual response histories. More sensitive approaches permit even greater tailoring of curriculum to learner characteristics by considering *incoming* as well as *emerging* knowledge and skills in the cognitive diagnosis. This enables the curriculum to adapt to both persistent and momentary performance information as well as the interaction.

One would think that increasing the personalization of instruction would enhance learning efficiency, improving both the rate and quality of cognitive skill acquisition. However, results cited in the literature on learning in relation to increased computer adaptivity are equivocal. In some cases, researchers have reported no advantage of error remediation in relation to learning outcome (e.g., Bunderson & Olsen, 1983; Sleeman, Kelly, Martinak, Ward, & Moore, 1989). In others, some advantage has been reported for more personalized remediation (e.g., Swan, 1983).

Cognitive diagnosis serves two main purposes: classification and explanation

(Snow, 1990). As explained previously, the result of cognitive diagnosis suggests uniquely appropriate curricular paths (classification); however, it also provides an interpretive theory about a learner's performance history (explanation). Snow (1990) has suggested using aptitude-treatment interaction (ATI) methodologies in order to exploit these two functions fully (for background, see Cronbach & Snow, 1977). ATI research provides information about initial states of learners that can be applied in macroadaptive instruction (e.g., selection of a learning environment for a particular student); then microadaptive instruction can be used as a response to particular actions (e.g., selection of the next small unit of instruction to be presented based on a specific response history). In other words, initial states are characterized by an aptitude profile.[1] Then microadaptive instructional systems can either focus on strengths, circumvent weaknesses, or highlight deficits to be remediated.

Obviously there is some cost associated with increasing a system's responsiveness, which raises two important practical questions: (a) How much, and what kind of, information about a learner is required to tailor instruction to his or her needs[2] so as to maximize chances for learning to occur? (b) What is the payoff of increasing a system's adaptability? Sleeman (1987) has argued that "if one takes seriously the findings of the ATI work of Cronbach and Snow (1977), it would appear that there is little likelihood of producing instruction that is uniquely individualized" (p. 242). The key word in this statement is "uniquely." An exhaustive characterization of a learner would probably not warrant the effort and expense in terms of increases in final outcome. However, the empirical question remains: How much is enough? Answers to these cost–benefit questions are discussed at the end of the chapter following a description of the macroadaptive approach and an examination of its theoretical and motivational bases.

Macroadaptive Approach

The approach to cognitive diagnosis taken in this chapter involves conducting a controlled experiment with the purpose of determining individual differences in learning and possible ATIs. Before the experiment, certain critical decisions have to be made. For instance, what aptitudes should be measured before the instruction, which treatment effects should be manipulated, what learning indicators should be recorded to measure learning progress, and what learning outcome and efficiency measures should be used?

The learning skills taxonomy developed by Kyllonen and Shute (1989) can assist in rendering principled decisions to some of these questions. This taxonomy defines a four-dimensional space involving the subject matter, learning environment, desired knowledge outcome, and learner attributes. Interactions among these dimensions are believed to influence outcome performance. For example, no single type of learning environment (e.g., exploratory-discovery) is best for all persons. Rather, aptitude-treatment interactions occur in which certain learner characteristics are better suited to certain learning environments than to others in order to achieve optimal outcome performance (see Shute & Glaser, 1990). Similarly, some domains lend themselves more readily to certain kinds of knowledge outcomes than to others. For instance, nonquantitative fields (e.g., history) emphasize propositions, whereas quantitative fields (e.g., calculus) focus on procedures. And finally, knowledge outcomes covary with instructional method: Propositions are more commonly learned by rote and procedures are more commonly learned by practice.

So, to begin answering questions concerning the most appropriate aptitude, outcome, and efficiency measures to use, one must consider the effect(s) of possible combinations of the four dimensions comprising the taxonomy. Aptitudes to assess before instruction should be relevant to the subject matter, the desired knowledge outcome, and the type of learning environment in which instruction will take place.

Theoretical Basis for Macroadaptive Approach

The Learning Abilities Measurement Program (LAMP)[3] conducts basic research on the nature of human learning abilities. In the past, LAMP studies have examined relationships between aptitude measures and performance on simple learning tasks. Recently, large-scale studies have been conducted validating the computerized aptitude tests against more complex learning from intelligent tutoring systems (e.g., Shute, 1990, 1991). Major research now in progress examines whether learning can be predicted from basic cognitive process measures or aptitudes.

The theoretical model of learning underlying LAMP has been influenced by Anderson's ACT* model (see Anderson, 1983; Kyllonen & Christal, 1989). Basically, it posits three stages of learning (i.e., declarative knowledge, procedural skills, and automatic skills) and two sets of learning predictors: enablers (i.e., what one already knows and can transfer to new situations) and mediators (i.e., cognitive processes determining what one can acquire, such as working-memory capacity and information-processing speed). Relations among the learning stages, enablers, and mediators show how, as learning progresses, enablers become elaborated, and working-memory capacity and processing speed become functionally larger and faster, respectively (e.g., Chase & Simon, 1973; Chi, Glaser, & Rees, 1982; Siegler & Richards, 1982).

[1] I define *aptitude* in this chapter as the incoming knowledge and cognitive abilities possessed by individuals arriving at a new learning task. Personality variables, classified by some researchers as aptitudes, are not included in this definition.

[2] *Needs* are defined in this context as the individual differences measures (i.e., aptitudes) believed to impact learning outcome and efficiency.

[3] LAMP is a project at the Armstrong Laboratory, Brooks Air Force Base, TX.

Justification and Motivation for This Approach

The justification for such a broad approach requires evidence that individuals do, in fact, perform better or worse under different learning conditions (or environments). As noted earlier, this is usually referred to as aptitude-treatment interaction research (Cronbach & Snow, 1977). ATI research was very popular in the 1960s and 1970s; then popularity waned. One of the major reasons contributing to the decline was that the older ATI research typically involved studies conducted in classroom environments. Data were confounded by many extraneous variables (e.g., personality of the teacher, instructional materials, classroom dynamics) making ATIs hard to find and difficult to interpret. A second factor contributing to the decline was the realization that we did not understand the processing requirements underlying performance on the various aptitude measures. This motivated process-oriented analyses, using elementary cognitive tasks as tests. A second generation of ATI research using theoretically derived aptitude measures and controlled learning environments is discussed.

Several factors motivated this approach to cognitive skill diagnosis. First, the learning skills taxonomy (with its four interactive dimensions) provided a framework for the systematic design and evaluation of intelligent tutoring systems. Second, after testing over 800 subjects on Smithtown,[4] it was clear that some individuals thrived in this type of guided-discovery environment, whereas others did not. This finding prompted the identification of characteristics of individuals who succeeded (and failed) in such a learning environment (see Shute & Glaser, 1990; Shute, Glaser, & Raghavan, 1989). Finally, findings reported in two recent studies reported no effect of different instructional treatments on learning outcome.

In the first, Sleeman et al. (1989) investigated the effects of different remediation techniques on high school students' learning of algebra. They concluded, "Three studies suggest that when initial instruction and remediation are primarily rule-based and procedural, remedial reteaching appears to be as effective as MBR (model-based remediation). From this it follows that classical CAI [computer-assisted instruction] would be as effective as an ITS" (p. 563).

In the other article, Anderson, Conrad, and Corbett (1989) reported results from various manipulations made to the LISP tutor environment. They concluded that "well-designed feedback can minimize the time and pain of learning, but has no effect on final instructional outcome" (p. 498).

There are at least two alternative explanations for both of these findings. The most obvious one is that the respective modifications were not distinct enough to impact learning outcome. A second explanation is that perhaps there was an effect of the manipulations (i.e., remediation and feedback), but to find it would require considering some additional variable(s).

[4] This is the name of a somewhat guided but mostly discovery environment for learning principles of microeconomics. The coach addresses specific scientific inquiry skills, such as changing one variable at a time while holding others constant. The coach does not address economic principles.

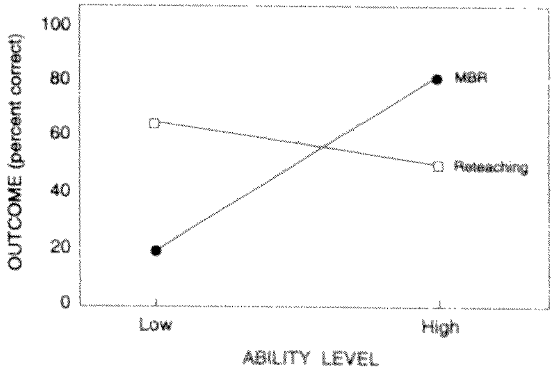


FIG. 2.1. Hypothetical depiction of Sleeman et al. (1989) data.

A hypothetical depiction of the Sleeman et al. (1989) data appears in Fig. 2.1, illustrating the second explanation. Although these data show no main effect between treatment condition on outcome measure (both about 50%) with aptitudes in the equation, we do see differential outcome effects due to treatment (or aptitude-treatment interaction). In this figure, high-aptitude individuals (possessing good reasoning and verbal skills, broad general knowledge, large working-memory capacities, and so on) benefit from the more elaborate explanations offered by the model-based remediation (MBR) approach. This approach addresses specific errors made during the solution process. On the other hand, low-aptitude individuals (possessing less of the same attributes) perform better in the reteaching condition. This approach simply demonstrates the relevant procedure without addressing the learner's error. There is some evidence for this proposition in the ATI literature. More elaborate explanations were found to help high-aptitude subjects, but less elaborate explanations were more effective for the low-aptitude subjects, "Elaboration that takes the form of systematic explanations places a burden of comprehension on the learner, which tends to help Highs" (Cronbach & Snow, 1977, p. 501).

Similarly, a hypothetical depiction of the Anderson et al. (1989) data appears in Fig. 2.2. One study they reported contrasted outcome (quiz) performance based on whether the student or the tutor controlled the feedback. The same logic applies here as with the previous illustration. Subjects with high aptitudes could, theoretically, benefit more by taking an active, independent approach during the learning process. They would have the necessary capabilities to direct the course of their own learning. But the low-aptitude subjects could perform better if the tutor guided them through the curriculum.[5] There is also some support for this

[5] Actually, John Anderson said that he did look for, but did not find, any ATI's in this data (see discussion section following this chapter). This was attributed to the restricted range of aptitude levels in the sample of university students used as subjects in the study.

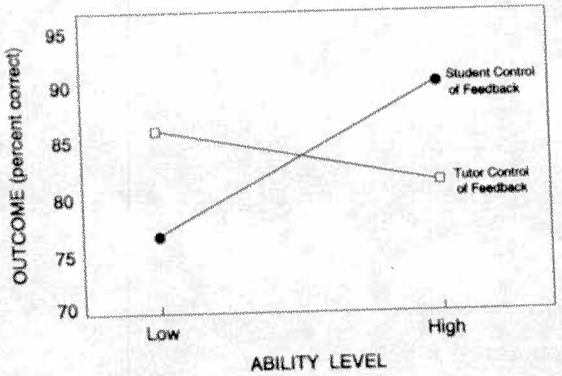


FIG. 2.2. Hypothetical depiction of Anderson et al. (1989) data.

premise in the ATI literature. Campbell (1964) contrasted two learning environments and found the "self-direction" condition was better for the high-aptitude group, whereas "programmed instruction" was better for the low-aptitude subjects. In addition, Cronbach and Snow (1977) reported that high-aptitude subjects profit from the opportunity to process the information in their own way, whereas low-ability subjects tend to be handicapped: "we see the Highs doing better when given greater freedom to proceed in their own manner, when thrown more upon their own resources. And we see regression slopes becoming flatter when more of the intellectual work is done for the learner" (p. 503).

These two graphs present alternative perspectives on the reported findings, but are hypothetical. To make the case more authentic requires empirical evidence from controlled research using a large, heterogeneous sample in order to allow the hypothesized ATIs to emerge. The following study was designed to provide such evidence.

Comparing Two Learning Environments

A study was conducted employing an intelligent tutoring system instructing basic principles of electricity (Ohm's and Kirchhoff's laws) as the complex learning task. Research questions examined in this experiment related to the influence of different learning environments on learning outcome and efficiency measures. Other research issues looked at the relationships among individuals' associative learning skills, learning environment, and learning outcome and efficiency measures.

I tested 282 individuals[6] (84% males, 16% females) participating in a 7-day (45-hour) study on the acquisition of basic principles of electricity. All subjects were high school graduates (or equivalent), with a mean age of 22 years. A restriction on this sample was that individuals could have no prior electronics training or formal instruction. Subjects were obtained from a temporary employment service and paid for their participation.

Experimental cognitive aptitude tasks were administered on Zenith 248 microcomputers (AT-compatible) with standard keyboards and EGA color video monitors. The intelligent tutoring system was administered on Xerox 1186 computers with standard keyboards and high resolution monochromatic displays on 19 in monitors. Software was written in InterLISP-D and LOOPS.

Subjects were tested in groups of 15–20 at Lackland Air Force Base, Texas, in the Complex Learning Assessment (CLASS) laboratory. They occupied individual testing carrels, and instructions, testing, and feedback were computer administered with proctors available to answer questions. On the morning of the first day, subjects were given a brief orientation to the entire study and then randomly assigned to one of two learning conditions. They were then administered half of the battery of cognitive ability tasks. On subsequent days, subjects were provided with instructions and practice problem solving involving electrical circuits delivered by the ITS. Upon completion of the tutor, subjects were given a criterion posttest battery and then completed the other half of cognitive ability tests.

[6] Approximately 320 subjects actually participated in the study, but only 282 completed all of the testing and learning activities reported in this chapter.

Cognitive Ability Tasks

A comprehensive battery of computerized tests was administered to all subjects to assess their incoming knowledge and cognitive skills. A full discussion of this battery (Kyllonen et al., 1990) is beyond the scope of this chapter. The present focus is on just one of the cognitive process measures—associative learning (AL) skills. Because of the exploratory nature of this study, I wanted to investigate a fundamental learning ability. One such parameter involves the rate and quality of forming associations when learning something new. The notion that associative learning skills are general and important to knowledge and skills acquisition is certainly not new. Rather, the literature offers ample support for this proposition (e.g., Anderson, 1983; Kyllonen & Tirre, 1988; Malmi, Underwood, & Carroll, 1979; Underwood, 1975; Underwood, Boruch, & Malmi, 1978).

Three computerized tests were administered in each of the verbal, quantitative, and spatial domains for a total of nine tests on this measure. Examples of three AL tests are described here, one from each of the verbal, quantitative, and spatial domains (note matching test paradigm).

Verbal AL Test. Subjects are required to learn eight pairs of words displayed in two rows at the top of the computer screen. The word pairs consist of an occupation directly above a piece of furniture (e.g., lawyer/table, carpenr/couch). The eight pairings remain the same throughout the test, although the airs' positions vary with each new question. For example, while lawyer would lways be paired with table, it may come either before or after carpenter/couch in ıe listing at the top of the screen. Questions appear one at a time at the bottom of ıe screen and consist of either a match or mismatch to the word pairs being earned. After the subject enters a response (typing "L" for like or "D" for ifferent), another question is displayed. Subjects are asked to remember the vord pairs as quickly and accurately as possible so that they will not have to keep ooking up at the top of the screen to confirm a match. At the end of each set of juestions (one set = 32 items), subjects are informed of their accuracy and atency on that set. There are 10 sets of items in the entire test (320 items), and or the first eight sets of items, word pairs remain on the screen. For the last two ets, word pairs do not appear on the screen and subjects are tested on how many pairs they successfully memorized during the preceding trials. Accuracy and atency data (milliseconds) are recorded. Odd–even reliability is .98.

Quantitative AL Test. This test is identical to the verbal AL test just described, except that in this test, subjects are required to learn eight pairs of *numbers* located at the top of the screen, one row above the other (e.g., 41 over −2, 95 over 6, 89 over −9). The number pairings remain the same throughout the test while the pairs' positions may vary from question to question. Instructions, number of items, and goals are the same as presented above for the verbal AL test. Odd–even reliability is .96.

Spatial AL Test. This test is the same as the verbal and quantitative AL tests, except that subjects must learn eight pairs of simple geometrical shapes located in two parallel rows at the top of the screen (e.g., arrow above L, triangle above +). Odd–even reliability is .96.

The other AL tests are similar, requiring connections to be established between verbal, quantitative, or spatial stimuli. Individual differences in associative learning skills have been shown to predict complex learning (see Kyllonen & Tirre, 1988). The complex learning task used as a criterion in the current study involved basic electricity content.

Complex Learning Task (Electricity ITS)

The electricity tutor was originally developed at the Learning Research and Development Center, University of Pittsburgh (Lesgold, Bonar, Ivill, & Bowen, 1989) and then modified at the Armstrong Laboratory. In particular, we created two learning environments, developed and coded a variety of learning indicators, established mastery criteria, refined principles, definitions and feedback, and modified the system's interface. Learning from the tutor resulted from working problems, reading definitions of concepts (hypertext structure), and exploring circuits (e.g., taking meter readings and changing values of components).

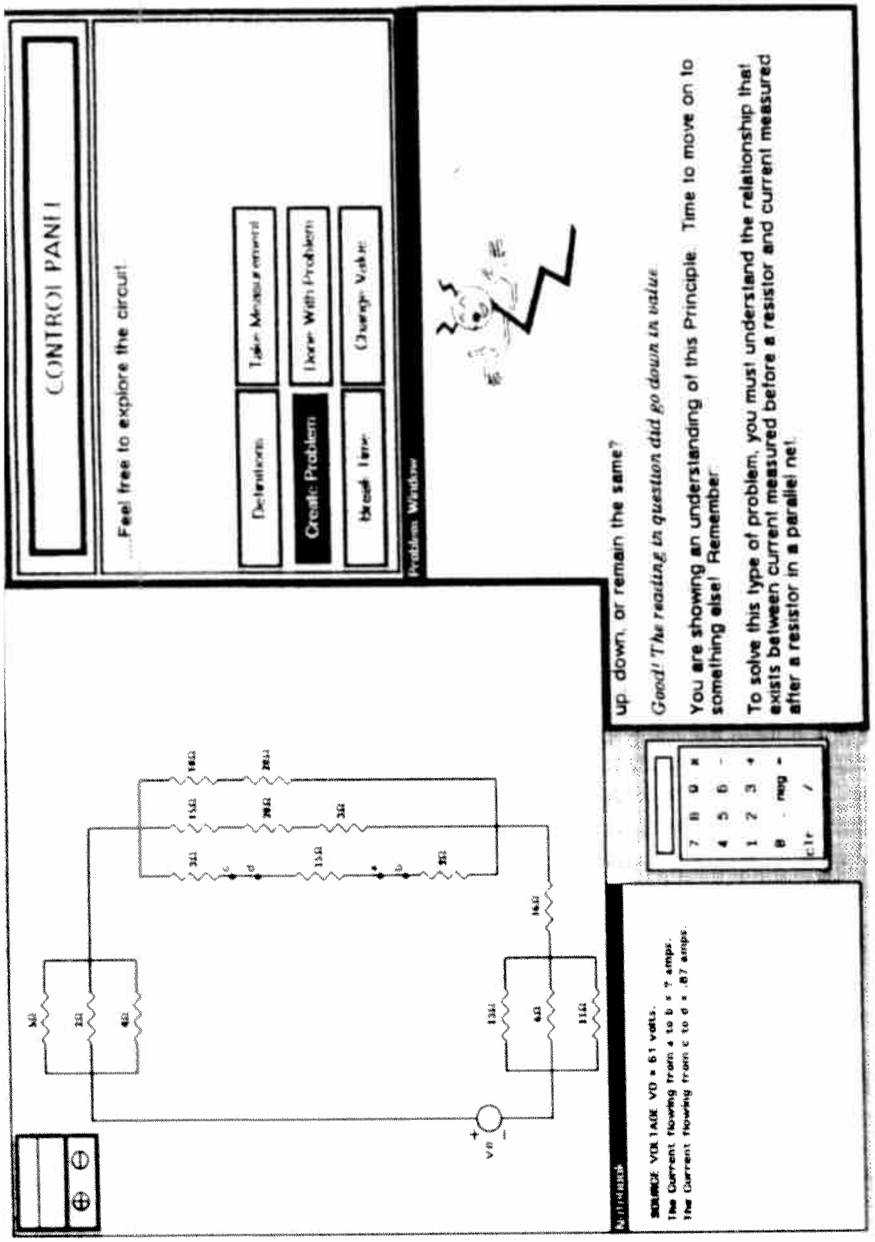
I created the two learning environments specifically to investigate possible ATIs in learning. To differentiate the two learning environments (rule-application and rule-induction), the ITS was manipulated by altering the nature of the feedback to the learner, all else being equal. After completing a problem, subjects in each group received feedback concerning whether their answer was correct. Moreover, the principle (Ohm's or Kirchhoff's laws) that was relevant to the problem was addressed in one of two ways. In the *rule-application* environment, feedback clearly stated the variables and their relationships for a given problem. This was communicated in the form of a rule such as, "The principle involved in this kind of problem is that current before a resistor is equal to the current after a resistor in a parallel net." Subjects then proceeded to apply the rule in the solution of related problems.

In the *rule-induction* environment, the tutor provided feedback that identified the relevant variables in the problem, but the learner had to induce the relationships among those variables. For instance, the computer might give the following feedback: "What you need to know to solve this type of problem is how current behaves, both before and after a resistor, in a parallel net." Subjects in the rule-induction condition, therefore, generated their own interpretation of the functional relationships among variables comprising the different rules. Subjects were randomly assigned to one of the two environments for the entire study.

The computer presented all problems (under both learning conditions) by showing different electrical circuits and asking questions about them. Figure 2.3 shows an example of the main screen. On the screen's left, a parallel circuit depicts various component values. The upper right of the screen shows the main options (e.g., look at definitions, take a measurement on the circuit). Problems were presented in the lower right quadrant of the screen with feedback given in the same window. A notebook in the lower left of the screen allowed students to store information from their explorations and manipulations. Finally, an on-line calculator was always available for computing solutions to more complex, quantitative problems.

Figure 2.4 shows an example of a definition. When the "View Definitions" option was selected, the screen cleared and a menu of items appeared. In the main definition window, bold-faced words implied connections between the immediate word and related concepts. Choosing a bold-faced word with the mouse resulted in that concept's appearance on the screen. In some cases, dynamic simulations, such as comparing current flow in series versus parallel circuits, were available to the learner.

The electricity curriculum consisted of 15 principles (see Table 2.1). Problems were generated by the computer based on those principles. Each problem was unique (i.e., generated "on the fly," not preprogrammed), based on the particular learner's response history. If a student needed more work on current flow in parallel circuits, for instance, the system would generate a problem satisfying specific constraints, such as, it must be a parallel circuit problem involving current, perhaps a more difficult quantitative solution required, and so forth.



...Feel free to explore the circuit.

up, down, or remain the same?

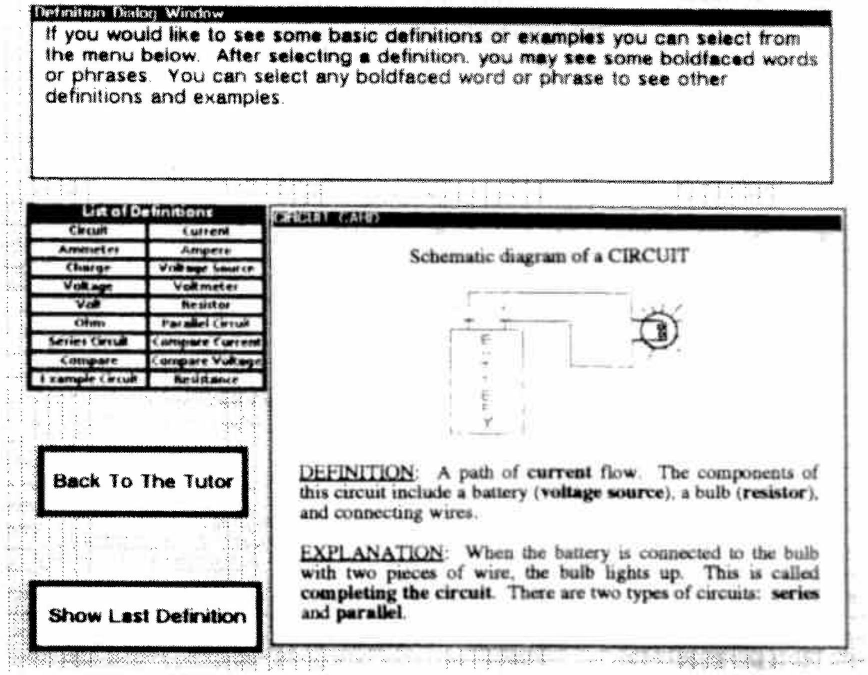
Good! The reading in question did go down in value.

You are showing an understanding of this Principle. Time to move on to something else! Remember:

To solve this type of problem, you must understand the relationship that exists between current measured before a resistor and current measured after a resistor in a parallel net.

SOURCE VOLTAGE VO = 61 volts.
The Current flowing from a to b = ? amps.
The Current flowing from c to d = .87 amps.

FIG. 2.3. Examples of electricity tutor screen.



If you would like to see some basic definitions or examples you can select from the menu below. After selecting a definition, you may see some boldfaced words or phrases. You can select any boldfaced word or phrase to see other definitions and examples.

List of Definitions	
Circuit	Current
Ammeter	Ampere
Charge	Voltage Source
Voltage	Voltmeter
Volt	Resistor
Ohm	Parallel Circuit
Series Circuit	Compare Current
Compare	Compare Voltage
Example Circuit	Resistance

Schematic diagram of a CIRCUIT

DEFINITION: A path of **current** flow. The components of this circuit include a battery (**voltage source**), a bulb (**resistor**), and connecting wires.

EXPLANATION: When the battery is connected to the bulb with two pieces of wire, the bulb lights up. This is called **completing the circuit**. There are two types of circuits: **series** and **parallel**.

FIG. 2.4. Example of electricity tutor on-line dictionary.

Problem types ranged from easy to difficult (Levels 1, 2, 3) and included qualitative problems (requiring responses of up, down, or stay the same), relative problems (requiring responses of higher, lower, or equal to), and quantitative ones (requiring calculations and numeric input). Learners "mastered" a principle once they had answered correctly three consecutive problems per principle.

Learning Outcome Measures

A four-part criterion test battery was developed measuring knowledge and skills acquired from the tutor. This battery was administered on-line after the student completed the tutor. The first two tests in the battery were also administered at the beginning of the experiment.

Part 1 of the criterion battery (pretest and posttest) assessed declarative knowledge understanding of different components and devices involved in basic electronics: ammeter, ampere, charge, circuit, current, ohm, parallel circuit, resistance, resistor, series circuit, volt, voltage, voltage source, voltmeter. Some example true/false questions included: A voltmeter is used to measure a voltage drop across two points in a circuit; there is a standard number of volts used in a parallel circuit; a resistor is designed to store electricity.

TABLE 2.1
Principles Underlying the Electricity Tutor

Kirchhoff's Laws

1. The current at one point in an uninterrupted wire is equal to the current at another point in an uninterrupted piece of wire.
2. The current is the same before and after a voltage source.
3. The current is the same before and after a resistor.
4. The current before a resistor is equal to the current after a resistor in a parallel net.
5. The current in the branches of the parallel net sums to the current in the entire net.
6. The current in a component is lower than the current for the entire net.
7. Voltage drop is lower across any single component of a series net than across the whole net.
8. Voltage drop is lower across any single component of a series net than across the whole net.
9. Voltage drop is the same across parallel components.
10. Voltage drop is the same across any component as across the whole parallel net.

Ohm's Laws

11. Voltage is equal to the current multiplied by the resistance ($V = I \times R$).
12. When the current goes up/down and the resistance stays the same, this implies that the voltage will go up/down.
13. Current is equal to voltage divided by resistance ($I = V / R$).
14. When the voltage goes up/down and the resistance stays the same, this implies that current will go up/down.
15. Resistance is equal to voltage divided by current ($R = V / I$).

Part 2 of the battery tested qualitative understanding of Ohm's and Kirchhoff's laws. These questions did not require any computations to be performed. Instead, the subject needed to understand the important variable relationships corresponding to the different principles. An example test item can be seen in Fig. 2.5.

Part 3 assessed the degree to which procedural skills were acquired from the tutor. Subjects needed to *apply* Ohm's and Kirchhoff's laws in the solution of different problems. Because test items required computations in their solution, an on-screen calculator was provided. A typical problem presented a circuit, and the subject had to figure out what the reading was (at some point or points) for some component. An example question is shown in Fig. 2.6.

The last test in the criterion battery, Part 4, measured a subject's ability to *generalize knowledge and skills* beyond what was explicitly instructed by the tutor. The subject was required to generate or design circuits to do specific things. An example item from this test is included in Fig. 2.7.

In summary, the four tests measured different aspects of electronics knowledge and skill acquisition. Test 1 measured declarative knowledge understanding; test 2 assessed qualitative knowledge of variable relationships (mental model without procedural skills); test 3 measured quantitative understanding and ability to apply Ohm's law (procedural skills); and test 4 gauged transfer or generalization of skills (mental model with procedural skills).

Is the current from point a to b higher, lower, or equal to the current from point c to d?

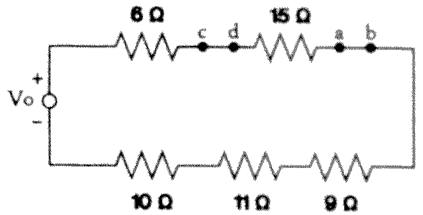


FIG. 2.5. Example item for posttest 2.

In this circuit, the voltage source is 11 volts and the voltage across a to b is 2.14 volts. What is the current flowing from b to c?

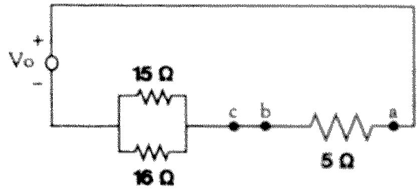


FIG. 2.6. Example item for posttest 3.

What are the resistor values for R_1 and R_2 that will yield a current from a to b of 2.5 amps and a voltage drop from c to d of 14 volts? The voltage source is 24 volts.

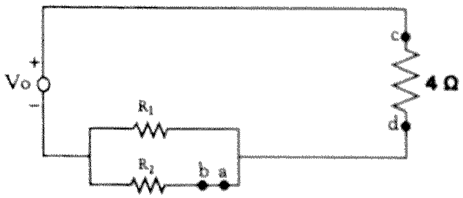


FIG. 2.7. Example item for posttest 4.

Learning Efficiency Measure

Another learning measure used in this study was defined as the total time spent completing the ITS curriculum. This "learning efficiency" measure involved both speed and accuracy because subjects could not proceed to the next principle until they had mastered the current one. Again, the mastery criterion was three consecutive correct responses per principle.

Research Questions

The research questions involving main effects included: (a) Is there an effect of learning environment on subsequent learning outcome? and (b) Is there an effect of learning environment on learning efficiency? I hypothesized that subjects in the rule-application environment would complete the tutor faster, but would not do as well on the posttests compared to subjects in the rule-induction environment. The basis for this belief was that the rule-application environment, by providing subjects explicitly with the relevant principle, was more straightforward and hence easier to get through. On the other hand, the more active participation required by the rule-induction environment was hypothesized to involve more of a time investment but to result in greater learning outcome (see Shute & Glaser, 1990).

The next question concerned the interaction between learning environment and aptitude affecting either learning outcome or learning efficiency. I hypothesized that high-ability subjects (i.e., those with above average associative learning skills) would benefit from the rule-induction environment because it provides more learner control (i.e., independence) compared to the rule-application environment. However, I hypothesized that lower-ability subjects would perform better on the outcome measures if they had learned from the rule-application environment because it provides more structure and support during the learning process than the rule induction environment.

RESULTS

A MANOVA was computed on the four posttest scores as dependent variables, the two pretest scores as covariates (to control statistically for incoming, related knowledge),[7] and environment as an independent variable (coded 0, 1 for rule-application, rule-induction, respectively). A composite AL score (i.e., the average of the nine standardized test scores) was included in the analysis as another independent variable along with the interaction between AL and environment.

TABLE 2.2
Summary Statistics of Posttest Scores, Time-on-Tutor, and Pretest Scores by Environment

Variable	*Mean*	*SD*	*Minimum*	*Maximum*
Rule-Application (N - 139)				
Posttest 1	73.7	12.1	44.3	96.7
Posttest 2	43.4	27.1	0.0	100.0
Posttest 3	60.8	28.0	0.0	100.0
Posttest 4	19.9	28.2	0.0	100.0
Time-on tutor (h)	11.25	4.2	5.2	25.6
Pretest 1	65.3	11.1	42.9	92.0
Pretest 2	22.7	22.1	0.0	88.3
Rule-Induction (N - 142)				
Posttest 1	74.1	13.5	40.9	98.1
Posttest 2	41.1	26.9	0.0	100.0
Posttest 3	57.2	22.6	0.0	100.0
Posttest 4	14.5	19.4	0.0	100.0
Time-on-tutor (h)	11.29	3.5	5.8	20.32
Pretest 1	64.7	11.1	42.3	92.3
Pretest 2	23.6	21.6	0.0	75.0

Results showed the following. First, there was no main effect of learning environment on learning outcome: $F_{(4,271)} = 1.58$. As can be seen in Table 2.2, the pretest and posttest data were remarkably similar between the two learning environments. Next, there was a significant main effect of AL on outcome: $F_{(4,271)} = 16.06$, $p < .001$. Individuals with high AL scores performed better than low AL subjects, overall, on the posttests. Furthermore, each of the four univariate F ratios was significant beyond the .001 level.

The nonsignificant effect of environment on learning outcome was unanticipated, suggesting that the feedback manipulations were simply too subtle to result in learning outcome differences. It is interesting to note that if I had not analyzed the effects of environment in relation to aptitude levels, I would have erroneously concluded that there was *no difference* between the two environments in terms of their effects on learning outcome. This is not unlike the conclusions reached by Sleeman et al. (1989) and Anderson et al. (1989), discussed earlier. But, in fact, there *was* a significant interaction between AL and environment on learning outcome: $F_{(4,271)} = 5.62$, $p < .001$. So, given this significant (albeit, general) interaction, the next step was to determine more precisely its nature—that is, its pattern across the four posttests.

An interaction term was computed by multiplying the composite AL score by environment (coded 0, 1). Multiple regression analyses were then computed regressing the four posttest scores, individually, on the following variables: AL, environment, AL × environment, Pretest 1, and Pretest 2. The pretest data were included in the equation to statistically control for incoming related knowledge, designed to correspond to the full MANOVA.

[7] Two MANOVAs were actually computed—with and without the pretest data as covariates. In both analyses, the F ratios and significance levels were the same.

TABLE 2.3
Multiple Regression Solution Predicting Posttest 1 Scores (Multiple R = .75)

Variable	*Sum of Squares*	*df*	*Unique R^2*	*F*	*Significance*
AL	582.70	1	1.3%	7.97	.005
Environment	2.54	1	0.0%	0.03	.852
AL x Environment	298.73	1	0.7%	4.09	.044
Pretest 1	12,565.70	1	27.5%	171.95	.000
Pretest 2	257.57	1	0.6%	3.52	.062
Model	25,743.05	5	56.3%	70.45	.000
Residual	2,003.69	274	43.7%		

Results from these multiple regression analyses were as follows. Predicting Posttest 1 scores[8] (declarative knowledge acquisition), there were significant main effects of AL and pretest data on outcome, and no main effect of environment. However, of much more interest, a significant interaction appeared involving AL and environment for this outcome measure. These data may be seen in Table 2.3.

To illustrate this interaction, expected values were computed from the regression equation for all four groups of subjects: individuals one standard deviation above and below the mean AL in each of the two learning environments.[9] The results can be seen in the upper left part of Fig. 2.8. The subjects with higher associative learning skills performed better on this declarative knowledge test if they were in the rule-induction environment. However, subjects showing lesser associative learning skills performed better on Posttest 1 if they were in the rule-application environment.

Results from the regression analysis predicting Posttest 2 data (qualitative understanding) showed that the only significant independent variables were the pretest data. There was no main effect due to AL, environment, or the interaction between AL and environment. These data are summarized in Table 2.4. Although nonsignificant, a graph of the interaction data (expected values) is included in Fig. 2.8, upper right quadrant, to illustrate the trend of the interaction, AL by environment, across the four posttests.

Next, the regression analysis computed with Posttest 3 data (procedural skill acquisition) as the dependent variable yielded findings similar to Posttest 1 results, with an interesting twist. Similar to the regression solution predicting Posttest 1 scores, this analysis of Posttest 3 data produced significant main effects due to AL and pretest data, but not to environment. In addition, the interaction involving AL and environment predicting Posttest 3 data was also significant. This solution may be seen in Table 2.5. The "twist" was that for this outcome measure, high-AL subjects performed better in the *rule-application* environment than in the *rule-induction* environment. In addition, environment did not affect outcome performance for low-AL subjects. This finding can be compared to Posttest 1 results where high-AL subjects performed better in the *rule-induction* environment than in the *rule-application* environment and low AL subjects performed better in the rule-application environment than in the rule-induction environment. This interaction may be seen in Fig. 2.8, lower left quadrant.

The last finding from this regression analysis involved Posttest 4 data (generalization of skills) as the dependent variable. These results were comparable to those discussed with Posttest 3 data. That is, there were significant main effects due to AL and pretest data, and there was no main effect due to environment. The interaction between AL and environment on Posttest 4 data was also significant (see Table 2.6). Again, high-AL subjects performed significantly better on this more difficult test in the rule-application environment than the rule-induction environment. For lower-AL subjects, environment did not affect performance. This interaction can be seen in Fig. 2.8, lower right quadrant.

[8] The posttest data used in all analyses were the raw scores, as recommended in Cronbach & Snow (1977, pp. 514–515).

[9] Error bars are included in each of the four graphs in Fig. 2-8. These represent the standard error measures per group (i.e., square root of mean-square error divided by N).

TABLE 2.4
Multiple Regression Solution Predicting Posttest 2 Scores (Multiple R = .35)

Variable	*Sum of Squares*	*df*	*Unique R^2*	*F*	*Significance*
AL	362.24	1	0.2%	0.55	.457
Environment	651.44	1	0.3%	1.00	.319
AL x Environment	1,004.70	1	0.5%	1.54	.216
Pretest 1	6,537.73	1	3.2%	10.01	.002
Pretest 2	4,233.84	1	2.1%	6.48	.012
Model	24,518.45	5	12.1%	7.50	.000
Residual	179,101.19	274	97.9%		

TABLE 2.5
Multiple Regression Solution Predicting Posttest 3 Scores (Multiple R = .69)

Variable	*Sum of Squares*	*df*	*Unique R^2*	*F*	*Significance*
AL	13,328.79	1	7.4%	38.89	.000
Environment	627.42	1	0.0%	1.83	.177
AL x Environment	1,415.65	1	0.8%	4.13	.043
Pretest 1	26,693.90	1	14.8%	77.88	.000
Pretest 2	2,517.41	1	1.4%	7.35	.007
Model	86,448.52	5	48.0%	50.44	.000
Residual	93,911.80	274	52.0%		

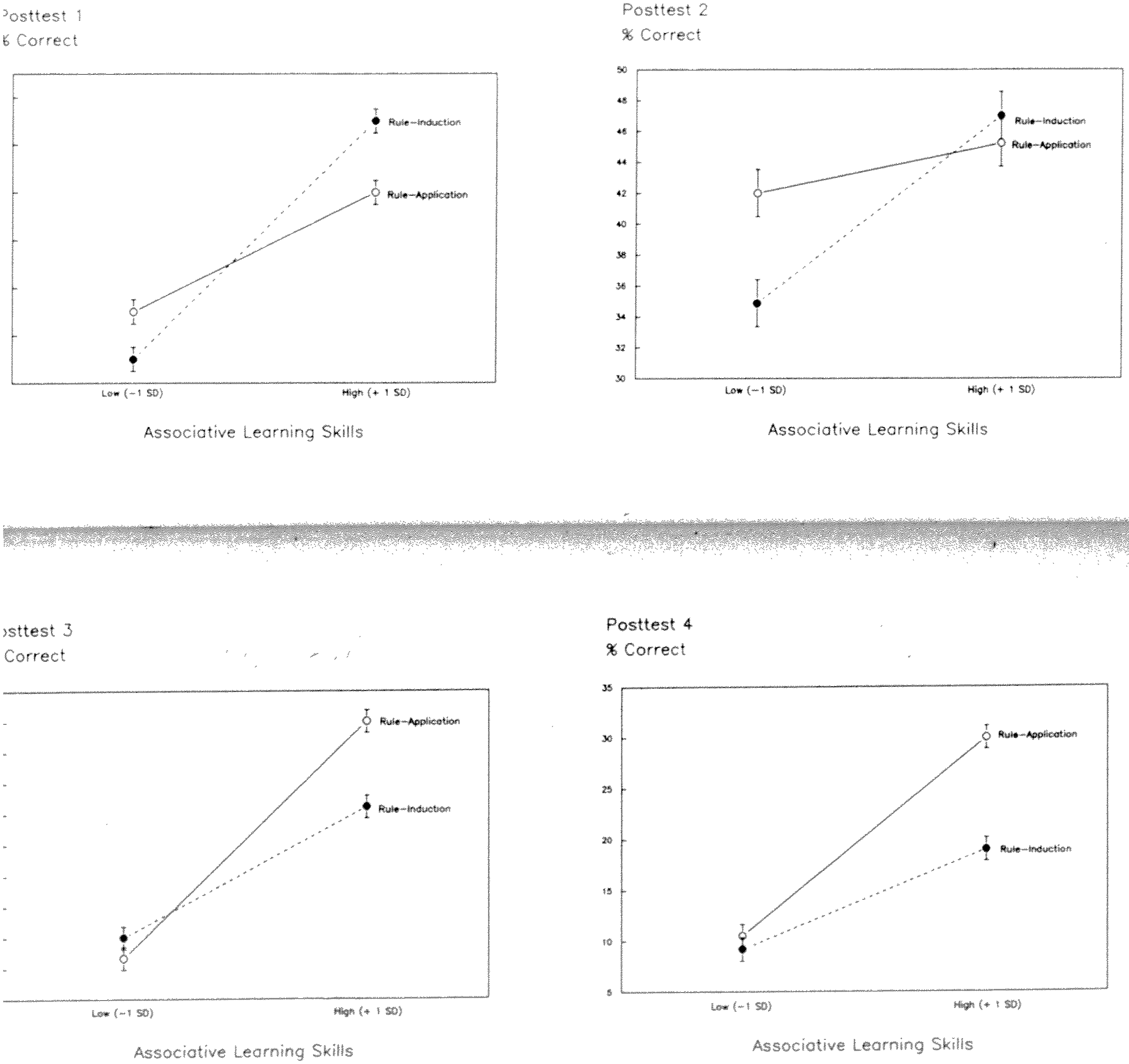


FIG. 2.8. **Associative learning ability by environment interactions in relation to four outcome measures.**

TABLE 2.6
Multiple Regression Solution Predicting Posttest 4 Scores (Multiple R = .62)

Variable	Sum of Squares	df	Unique R^2	F	Significance
AL	9,962.94	1	6.1%	27.05	.000
Environment	1,310.86	1	0.9%	3.75	.062
AL x Environment	1,412.56	1	1.0%	3.85	.050
Pretest 1	16,136.37	1	9.8%	43.81	.000
Pretest 2	5,201.54	1	3.2%	14.12	.000
Model	63,447.44	5	38.6%	34.45	.000
Residual	100,925.84	274	61.4%		

Post hoc comparisons were computed in order to establish some basis for the differential relations found between the aptitude-treatment interaction. These consisted of an overall posttest score and three orthogonal contrasts. Average was simply the sum of the four standardized outcome measures: $(Y_1 + Y_2 + Y_3 + Y_4)$. This represented a general outcome factor, independent of type of learning requirements inherent in the individual tests. The first contrast examined performance on Posttests 1 and 2 relative to 3 and 4: $(Y_3 + Y_4) - (Y_1 + Y_2)$. This new variable, *DecPro*, represented a declarative versus procedural distinction because Tests 1 and 2 required conceptual (declarative) understanding of the subject matter, whereas Tests 3 and 4 required procedural skills. The next orthogonal contrast, *MentMod*, compared outcome Measures 2 and 4 against 1 and 3: $(Y_2 + Y_4) - (Y_1 + Y_3)$. Posttests 2 and 4 required subjects to solve problems qualitatively; that is, they had to develop a mental model of how current, voltage and resistance interacted in the solution of a circuit problem. On the other hand, Tests 1 and 3 required a specific response to problems—a word/concept or a number (i.e., retrieval of facts or procedures in the solution process, not the more abstract creation of mental models). Finally, *Last* was defined as the remaining orthogonal contrast: $(Y_1 + Y_4) - (Y_2 + Y_3)$. It was not interpreted in terms of psychological meaning.

The orthogonal contrasts were analyzed with the same MANOVA design as described earlier, except that the *contrasts* (rather than the actual posttests) served as the dependent variables. Results from this MANOVA were as follows. The analysis was first computed for *Average*, and the interaction (AL by environment) was not significant ($F_{(1,274)} = 0.05$; $p = .82$). This was not surprising because combining the posttests obscured differentiating information. Results from the orthogonal contrasts, on the other hand, did reveal a significant aptitude by treatment interaction: $F_{(3,272)} = 7.47, p < .001$. The univariate F tests show clearly the basis for this finding. The only significant contrast was *DecPro*. The other contrasts were not significant: *DecPro* ($F_{(1,272)} = 20.07$, $p < .001$); *MentMod* ($F_{(1,272)} = 0.03$, $p = .87$); and *Last* ($F_{(1,272)} = 0.01$, $p = .93$).

Figure 2.9 depicts the AL by environment interaction in relation to the *DecPro* contrast.[10] Large positive *DecPro* values reflect higher scores on the more difficult, procedural tests in relation to the simpler declarative knowledge tests. Large negative values imply just the opposite (i.e., higher scores on the declarative than procedural tests). A *DecPro* value of zero indicates no difference between test scores. This figure shows that high-AL subjects acquired new procedural skills better (in relation to declarative knowledge acquisition) if they learned from the rule-application environment. High-AL subjects in the rule-induction environment were greatly impaired with regard to procedural skill acquisition. For low-AL subjects, the contrast between environments was not so great. Large negative *DecPro* values were associated with the rule-application environment, and there were no large positive *DecPro* values for the low-AL subjects. So, in regard to learning outcome, a significant ATI was found. Furthermore, the effects of the interaction differed by type of outcome (e.g., declarative vs. procedural skills).

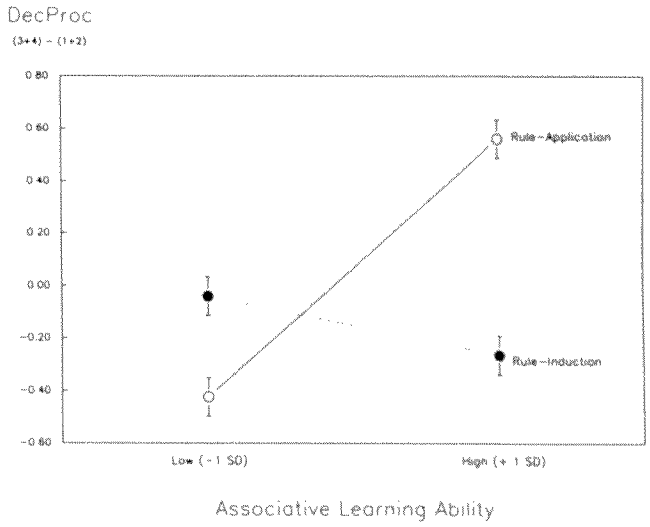


FIG. 2.9. Associative learning ability by environment interactions in relation to declarative/procedural outcome measure.

Was learning efficiency influenced by ATIs? An ANOVA was computed on subjects' time to complete the tutor (dependent variable) by learning environment, aptitude, and the interaction between AL and environment (independent variables). The findings were similar to those reported above with outcome as the criterion. That is, there was no main effect of learning environment on this

[10] Expected values (+1, −1 standard deviation for high/low AL) were computed from the regression equation and plotted separately by environment. Standard error bars are included for each group.

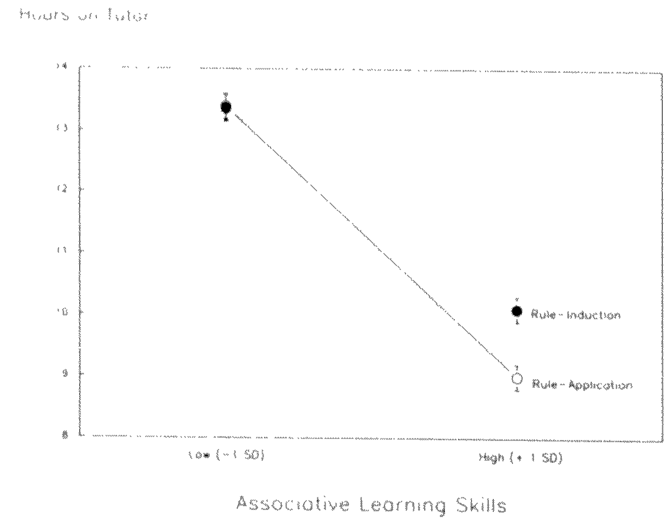


FIG. 2.10. Associative learning ability by environment interactions in relation to time on tutor.

learning efficiency measure [$F_{(1,277)} = 0.01$]. There was a significant main effect due to AL [$F_{(1,277)}$) = 67.41, $p < .001$). Finally, there was a marginally significant interaction between AL and environment $F_{(1,277)} = 3.73$, $p < .057$).

The same plotting convention was used as with the outcome data: expected values were plotted from the regression equation (regressing hours on AL, environment, and AL × environment) using plus and minus one standard deviation to represent high- and low-AL groups. Error bars are included for each group in the graph (see Fig. 2.10). First note that individuals in both learning environments took, on average, about the same amount of time to complete the curriculum (see Table 2.2). However, when AL was included in the analysis, the results showed that the rule-application environment was associated with more "efficient" behavior (i.e., it took less time to complete). This was true for the high-AL subjects but not for the low-AL subjects.

DISCUSSION

Three studies involving learning environment manipulations were discussed in this chapter. None of the three showed main effects of environment on outcome performance for different domains. First, Sleeman et al. (1989) reported no main effects on high school students' learning of algebra when they compared two types of mediation techniques (i.e., model-based remediation versus reteaching the subject matter). Second, Anderson et al. (1989) reported no main effects of student- versus tutor-controlled feedback on an outcome quiz measuring acquisition of LISP procedures. Third, in the study reported in this chapter, no main effects were found between the inductive versus more applied learning environments in terms of acquiring basic principles of electricity. However, when associative learning data were examined in relation to learning electricity principles, then learning environments did appear to be differentially effective.

In the focal study, I created two different learning environments (i.e., rule-application and rule-induction) to systematically examine possible aptitude-treatment interactions. Four outcome measures were developed to assess declarative knowledge (Tests 1 and 2) and procedural skill (Tests 3 and 4) acquisition. The four main results were as follows:

Declarative Knowledge Acquisition:

1. High-ability subjects learned more if they had been assigned to the rule-induction environment, and
2. Low-ability subjects learned more if they had been assigned to the rule-application environment.

Procedural Skill Acquisition:

3. High-ability subjects developed more skill if they had been assigned to the rule-application environment, and
4. Low-ability subjects performed poorly on the procedural skills tests, regardless of learning environment.

To understand these findings, consider the cognitive activities invoked by each environment in conjunction with the learning outcome being assessed and the cognitive abilities of the learner. For example, the rule-induction environment invoked declarative representations. Learners had to first understand the concepts involved in a given problem, then formulate a rule by connecting relevant concepts together in a meaningful way. To illustrate, an early, relatively simple principle to be learned was: The current is the same before and after a resistor. In the inductive environment, learners would receive a problem involving this principle. They had to determine the relevant variables embedded in the problem (i.e., current and resistance), then induce the functional relationships—what happens to current after it crosses a resistor (i.e., increases, decreases, or stays the same). Finally, they had to verify whether this relationship held up in related problems involving current and resistors. Cognitive resources would thus be wrapped up in elaborative processing and testing.

High-ability subjects in the inductive environment performed well on the declarative knowledge tests (Finding 1). A possible explanation is that there was a good match among learning environment, outcome measure, and cognitive ability: (a) the rule-induction environment supported declarative representations, (b) the outcome tests required accessing declarative representations, and (c) the high-AL subjects possessed relevant cognitive skills.

Another good match accounted for Finding 3. The rule-application environment simply informed learners of the appropriate rule underlying each problem. For related problems, learners promptly applied the rule during the solution process. The cognitive activity supported by this environment was the proceduralization of skills. Subjects with good associative learning skills performed well on the procedural skills tests in the applied environment because (a) the application environment supported proceduralization, (b) the outcome tests required the application of rules and procedures in the solution of problems, and (c) the high-AL subjects possessed good cognitive abilities.

Low-ability individuals acquired more declarative knowledge from the tutor if they were in the rule-application environment (Finding 2) as opposed to the induction environment. This was probably due to its straightforward instructional approach (i.e., the explication of rules). Furthermore, these low-ability subjects' deficient skills were not as burdened as they would have been in the induction environment. Because the computer provided the relevant rules explicitly (and repeatedly), this should have enabled memory for the associated principle, thus enhancing performance on the declarative knowledge tests. When the outcome being measured was procedural, however, neither learning environment enhanced outcome performance for these low-ability subjects. They scored equally poorly (Finding 4).

What are the implications of mismatching conditions? One mismatch between environment, outcome, and subject abilities included high-AL subjects assigned to the rule-application environment and tested on their declarative knowledge acquisition. These subjects performed poorly on declarative knowledge tests compared to high-AL subjects in the inductive environment. This may be explained by ACT* (see Anderson, 1983). That is, when learning a new cognitive skill, initial learning is declarative. With practice, the skill can be executed progressively faster. The cost of this speed-up is the gradual inability to describe the underlying procedures. In other words, as a skill becomes more automatic, the ability to talk about constituent procedures decreases. Because the applied environment fostered proceduralization, and the outcome measured in the mismatch condition was declarative, high-AL subjects may have lost access to the original declarative representation during the process of practicing and proceduralizing new skills.

One other mismatch with negative consequences involved high-AL subjects in the rule-induction environment being tested on procedural skill acquisition. Their procedural skills were significantly worse compared to high-AL individuals from the rule-application environment. The disadvantage of the rule-induction environment in relation to the complex procedural tests was that it did not provide time for necessary practice. Instead, it continued to demand and use cognitive resources in estimating variable relationships. If cognitive resource demands are continually kept high and learners never have the opportunity to practice certain skills, they will inevitably fail on the complex tasks that require high levels of proficiency (see Ackerman, 1988). So the more demanding rule-induction environment simply does not "pay off," except if the outcome measures declarative knowledge acquisition. The rule-application environment, in contrast, does not "waste" cognitive resources in the induction of variable relationships. By providing these relationships to subjects explicitly in the form of specific feedback, learners can proceed immediately to apply them across various circuits. In order to solve the more complex procedural skills tests, a learner must have had sufficient and consistent practice across a variety of circuit types (for more on practice effects, see Regian & Schneider, 1990; Schneider & Shiffrin, 1977).

Relating these findings back to cognitive diagnosis, the question posed earlier concerned which aptitudes should be diagnosed, and when. In the present study, a composite measure of associative learning skill was found to be an informative predictor of various learning outcome measures. Any or all of these tests could be administered prior to ITS instruction. But the decision about what aptitudes to measure should depend on the subject matter being instructed as well as the desired knowledge outcome. For example, suppose you wanted to teach 12th-century English history, and you wanted your students to walk away with declarative (propositional) knowledge. An individual's verbal skills, associative learning skills, and/or general knowledge represent reasonable and relevant aptitudes to assess prior to instruction. As different ITS are tested across a variety of domains, more precise information about important cognitive correlates should be forthcoming. At the Armstrong Laboratory, we have begun this mapping process between cognitive factors and knowledge and skill acquisition in a variety of areas, including logic gates, microeconomics, Pascal programming, principles of electricity, and flight engineering.

In terms of macroadaptation, the pertinent question becomes: Does preliminary aptitude testing increase tutor effectiveness? Although the range of applicability remains an empirical question, results from this study suggest that it potentially can increase performance (i.e., instruction and consequently learning). For instance, consider the following decision rules determining the optimal environment for persons based on their AL score (see Fig. 2.9).

> When outcome = *declarative knowledge* (posttests 1 and 2)—If high AL (greater than or equal to the mean AL), then rule-induction environment, else rule-application environment (for low AL).
>
> When outcome = *proceduralization of skills* (posttests 3 and 4)—If high or low AL, then rule-application environment since that environment is better for the high AL subjects and does not effect outcome for low-AL subjects.

Also related to tutor and learner improvements, a pragmatic concern involves the payoff of this proposed approach. There are two parts to this question. The first part concerns the cost of employing the macroadaptive method (i.e., preliminary aptitude testing and global adjustment of environment based on the results). The second part addresses the impact on learning.

The cost of prior testing is minimal. The average time to finish the associative learning tests ranges from 2 to 10 min (mean completion time per test = 4.8 min). In addition, the odd–even reliabilities of these tests are high, ranging from .82 to .98 (mean reliability of all nine tests > .90). The cost associated with altering the tutor's learning environment is also very small. To illustrate, in the electricity tutor discussed in this chapter, simple modifications were made to the feedback, with all else equal. At most, it required 2 hr to rewrite the feedback from the explicated form (e.g., "The principle involved in this kind of problem is that current before a resistor is equal to the current after a resistor in a parallel net") to the more inductive form (e.g., "What you need to know to solve this type of problem is how current behaves, both before and after a resistor, in a parallel net"). The variables remained the same, but the structure of the sentences was altered. This resulted in a single computer program with a yes/no "environment flag" denoting learning environment. In this study, when the program was initialized, the flag was set to either rule-application (environment flag = yes) or rule-induction (environment flag = no). Given the exploratory nature of this study, assignment to learning environment was random. However, the computer could just as easily set the flag itself based on the results from preliminary aptitude testing (i.e., the action taken in response to the evaluation of a decision rule).

To answer the second part of the question concerning the benefit(s) of such an approach, consider the amount of variance explained by these independent variables: AL, environment, AL × environment, Pretest 1 and Pretest 2. These variables accounted for 56% of Posttest 1 outcome variance (declarative knowledge acquisition), 12% of Posttest 2 variance (qualitative understanding), and 48% and 39% of the outcome variance, respectively, for Posttests 3 and 4 (proceduralization and generalization of skills). Also, the AL by environment interaction accounted for unique, significant variance underlying Posttests 1, 3, and 4 (see Tables 2.3, 2.5, and 2.6). These findings suggest that tutor improvements are possible using the appropriate decision rules. Additional analyses are planned that will investigate the relationship(s) among other cognitive process measures (e.g., working-memory capacity, information-processing speed), learning environment, and learning outcome and efficiency measures. This may lead to even more complex and informative decision rules.

The implications of these findings for ITS design are as follows. To teach procedural skills, use a more structured, rule-application environment, allowing for sufficient practice on the skills being instructed. In the present study, high-AL subjects were shown to perform better in this environment, whereas low-AL subjects were not affected by environment. But to teach declarative or conceptual knowledge, assign high-AL subjects to a more self-directed, inductive environment and low-AL subjects to a more tutor-directed, applied environment. The present findings thus go beyond what Anderson et al. (1989) found, that "well-designed feedback can minimize the time and pain of learning but has no effect on final instructional outcome" (p. 498). In fact, feedback may affect final outcome performance, but it depends on learner traits as well as the outcome measures.

A number of empirical research questions remain. What are the characteristics of learners who perform better in what types of learning environments? Are certain domains better suited for specific instructional methods? At what point should feedback be provided, what should it say, how is it best presented, and what is the relationship of feedback to learner characteristics? How much learner control should be allowed? What other learner attributes influence outcome performance (e.g., motivation, interests, activity level, independence). Do the same aptitudes predict learning outcome and efficiency across various subject matters? What treatment effects should be manipulated and how?

ATI research, conducted with ITS, can help answer some of these questions. Furthermore, the learning skills taxonomy (Kyllonen & Shute, 1989) provides a framework for conducting systematic and controlled ATI studies that was not possible prior to the arrival of ITSs. Instead of continuing to add to the growing pool of ITSs, the field can profit from controlled research altering, systematically, the design of existing ones and evaluating the results of those changes in accordance with a principled approach. Results from this study showed that different learning environments are differentially effective for learners; however, treatment conditions mostly affected the performance of high ability subjects. Research is needed to determine what kinds of environments promote learning for low-ability persons, precisely the population that needs help the most.

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 Press.

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

Bunderson, V. C., & Olsen, J. B. (1983). *Mental errors in arithmetic skills: Their diagnosis in precollege students* (Final project report, NSF SED 80-125000). Provo, UT: WICAT Education Institution.

Campbell, V. N. (1964). Self-direction and programmed instruction for five different types of learning objectives. *Psychology in the Schools, 1*, 348–359.

Chase, W. G., & Simon, H. A. (1973). The mind's eye in chess. In W. G. Chase (Ed.), *Visual information processing* (pp. 215–281). New York: Academic Press.

Chi, M. T. H., Glaser, R., & Rees, E. (1982). Expertise in problem solving. In R. Sternberg (Ed.), *Advances in the psychology of human intelligence* (Vol. 1, pp. 7–75). Hillsdale, NJ: Lawrence Erlbaum Associates.

Clancey, W. J. (1986). *Intelligent tutoring systems: A tutorial survey* (Report No. KSL-86-58). Stanford, CA: Stanford University.

Cronbach, L. J., & Snow, R. E. (1977). *Aptitudes and instructional methods: A handbook for research on interactions*. New York: Irvington.

Kyllonen, P. C., & Christal, R. E. (1990). Cognitive modeling of learning abilities: A status report of LAMP. In R. Dillon & J. W. Pellegrino (Eds.), *Testing: Theoretical and applied issues* (pp. 112–137). San Francisco: Freeman.

Kyllonen, P. C., & Shute, V. J. (1989). A taxonomy of learning skills. In P. L. Ackerman, R. J. Sternberg, & R. Glaser (Eds.), *Learning and individual differences* (pp. 117–163). New York: Freeman.

Kyllonen, P. C., & Tirre, W. C. (1988). Individual differences in associative learning and forgetting. *Intelligence, 12*, 393–421.

Kyllonen, P. C., Woltz, D. J., Christal, R. E., Shute, V. J., Tirre, W. C., & Chaiken, S. (1990). *CAM-4: Computerized battery of cognitive ability tests*. Unpublished computer program, Brooks Air Force Base, TX.

Lesgold, A. M., Bonar, J., Ivill, J., & Bowen, A. (1989). An intelligent tutoring system for electronics troubleshooting: DC-circuit understanding. In L. Resnick (Ed.), *Knowing and learning: Issues for the cognitive psychology of instruction* (pp. 29–53). Hillsdale, NJ: Lawrence Erlbaum Associates.

Malmi, R. A., Underwood, B. J., & Carroll, J. B. (1979). The interrelationships among some associative learning tasks. *Bulletin of the Psychonomic Society, 13*, 121–123.

Regian, J. W., & Schneider, W. (1990). Assessment procedures for predicting and optimizing skill acquisition. In N. Fredericksen, R. Glasser, A. Lesgold, & M. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 297–323). Hillsdale, NJ: Lawrence Erlbaum Associates.

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: Detection, search, and attention. *Psychological Review, 84*, 1–66.

Shute, V. J. (1990). *A comparison of rule-induction and rule-application learning environments: Which is better for whom and why?* Paper presented at the American Educational Research Association (AERA), Boston, MA.

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

Shute, V. J., & Glaser, R. (1990). A large-scale evaluation of an intelligent discovery world: Smithtown. *Interactive Learning Environments, 1*(1), 51–77.

Shute, V. J., Glaser, R., & Raghavan, K. (1989). Inference and discovery in an exploratory laboratory. In P. L. Ackerman, R. J. Sternberg, & R. Glaser (Eds.), *Learning and individual differences* (pp. 279–326). San Francisco: Freeman.

Siegler, R. S., & Richards, D. D. (1982). The development of intelligence. In R. J. Sternberg (Ed.), *Handbook of human intelligence* (pp. 897–971). Cambridge, England: Cambridge University Press.

Sleeman, D. (1987). PIXIE: A shell for developing intelligent tutoring systems. In R. W. Lawler & M. Yazdani (Eds.), *Artificial intelligence and education* (vol. 1, pp. 239–265). Norwood, NJ: Ablex.

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

Swan, M. B. (1983). *Teaching decimal place value. A comparative study of conflict and positively-only approaches* (Research Report No. 31). Nottingham, England: University of Nottingham, Sheel Center for Mathematical Education.

Underwood, B. J. (1975). Individual differences as a crucible in theory construction. *American Psychologist, 30*, 128–134.

Underwood, B. J., Boruch, R. F., & Malmi, R. A. (1978). Composition of episodic memory. *Journal of Experimental Psychology: General, 107*, 393–419.

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

TRANSCRIPTION OF DISCUSSION

John Anderson: That was very nice data. In our study that you mentioned, we didn't find aptitude-treatment interactions, but our students all have Math SAT scores around 600–800 (so no aptitude variance). Although we have not been able to identify such interactions, for a long time we've had a suspicion that there was a trade-off going on with respect to learning. Some situations call for students to work out things themselves. And there is a whole lot of research suggesting that better memory results when you generate answers yourself rather than being told things. On the other hand, if you can't generate the answers yourself, then obviously you have to be told. We have been struggling with ways to do this. We think this happens on an item-by-item basis. Even within a particular student, there are going to be problems that the student can't actually solve, and other points of difficulty that they can't control. In that direction, we've been trying to adapt instruction on a problem-by-problem basis. This might be a reasonable way of organizing your results. That is, when it's the case that students are having difficulty, then a more direct approach helps. But, in general, leaving problem solution more in the students' hands is positive. In the first study, you analyzed different learning outcome measures. In the case where the questions were easy (the declarative questions), high-aptitude students were doing better in a more discovery-based environment. But then, when the questions were getting presumably more difficult, even for the high-ability students, that was the point in which more directed instruction was needed.

Val Shute: That's a pretty good summary. What I envision is a kind of *interplay* between macro- and microadaptation. I think that the problem-by-problem adaptation of instruction that you mentioned is a good idea. This takes place during the learning process where the computer deals with each problem-solving episode individually. That's what I think of as "microadaptation." But the nature of these low-level computer responses is a function of "macroadaption." Suppose you had information about a student, like she had a high aptitude. And you also had a specific learning goal in mind, like the student should be able to effortlessly apply Ohm's law in solving circuit problems. The results from my research suggest that she should be placed in a more applied learning environment to achieve the goal state. This environment would provide her with the principles and the time to practice solving circuit problems rather than making her spend time inducing principles first. Although these were exploratory studies, it seems that both micro- and macroadaption may be important in terms of optimizing learning.

Dan Fisk: Can you tell me a little bit more about what the associative learning tasks are and how they relate to a measure I might be familiar with?

Val Shute: These computerized tests measure how quickly a person can form associations between different things, like words, numbers, or geometric shapes. For example, one of the verbal associative learning tests showed subjects eight pairs of words at the top of the screen that they had to remember. The pairings between the words always stayed the same and always remained at the top of the screen, but for each new test item, the pairs appeared in random order. Subjects had to answer true/false questions presented at the bottom of the screen, like: lawyer/bed? For subjects to answer the questions quickly and accurately, they needed to memorize the top pairs so they wouldn't have to keep looking up and searching through the list.

Dan Fisk: And how they relate to other measures?

Val Shute: The associative learning tests were all pretty highly correlated with one another (from .3 to .7) and the odd–even reliabilities of these tests were all above .95. The relation to the other cognitive factors showed these tests correlated most highly with working memory (.6), general knowledge (.5), and inductive reasoning skills (.5).

Jim Pellegrino: Were your primary dimensions operating in your analyses of cognitive abilities the abilities themselves or the content domain?

Val Shute: I've done some analyses that show the data cluster more on the process than the content dimension. For example, when I computed a factor analysis on all of the test data, three factors emerged. But the factors weren't verbal, quantitative, and spatial. Instead, they were (a) working memory and associative learning skills, (b) general knowledge and inductive reasoning skills, and (c) processing speed. Within each factor, all content domains were mixed up. This factor analysis accounted for a whole bunch of the variance (75%).

Bill Johnson: As I got my presentation material together, I asked myself what a human tutor would do at any given point in an instructional scenario. When I think about aptitude, a human tutor knows the general characteristics about the population that he or she is training. We know that they have X number of years of school, it is an environment where they have 2 or 3 years of on-the-job experience, and they understand some of the prerequisites coming in. Therefore, the human tutor can predict some of the aptitude. Now you use a pretest battery of tests to understand what some of the aptitudes are. I guess my question is: What does an intelligent tutor system do at the very beginning? I know that Doug Towne's system asks a student what his or her skill level is. That's a good idea, but what are some other ways early on what would help assess what the student's aptitude is, make adjustments, and then decide what kind of tutoring to provide?

Val Shute: That's a really good question. I think that the answer depends on what you want to teach and what you want your students to walk away with. That information would constrain the decision about what to assess. For example, if you wanted to teach conceptual understanding of Ohm's law, then the computer could administer a 20- to 30-minute test measuring verbal aptitude, like word knowledge or associative learning skills. Results from that testing would inform the tutor about which environment would be best. By using tutors that teach different domains by different instructional methods, I'd like to be able to figure out the best combinations among domain, learning environment, and desired knowledge outcome. I suspect that some learning environments may be better for certain domains with specific knowledge outcomes in mind. So more empirical studies can begin filling in the missing pieces of this puzzle, and we might eventually be able to really constrain decisions about how to teach a particular person. Right now I'm just limited to talking about teaching electricity and Pascal programming and have only tested two contrasting environments (inductive and applied).

John Anderson: In the university environment, most people actually walk in with lots of good measures. Is that true in other environments too?

Val Shute: If by "good measures" you mean "high aptitudes," then I guess other environments are comparable. Sometimes I test Air Force recruits. These people are selected on the basis of their ASVAB scores, so there is a selection criterion, but not as high as, say, Carnegie-Mellon students. The subject populations I use in my large-scale studies are not university students or Air Force recruits. They come from temporary employment agencies. So my subjects are students, housewives, unemployed carpenters, and "others." These people are very heterogeneous, showing a range on all the aptitude measures. And there are lots of individual differences on the outcome measures as well.

Alan Lesgold: I have two questions. It seems to me that calling this an "aptitude issue" is a peculiar convenience. I think, although I could be mistaken, the model you are heading toward is sort of a Swellard-like model, that says that people are lacking certain capabilities. There is too much stuff to do in the *middle* of problem solving, and if you give them one more thing to do, we simply overburden them. We might be better off figuring out for them what information they need. And then giving them that information rather than giving them the extra burden of trying to figure out what information to go after, and when to try to figure something out on their own. If that's the case, then the question arises: "What is the utility of the aptitude approach?" The aptitude approach says that, for some reason, we think we would like to measure enduring characteristics of these folks and then use an instructional strategy that is tailored to those enduring characteristics. The alternative would be to try to find ways to assess information about people while they are in the middle of doing complex activities, when they're being swamped by the burdens that are being placed on them by the instructional approach, and essentially microadapting. You made a strong case for adapting and I'm inclined to believe it. The reason I would raise the possibility of microadapting is that there is just a chance that this ability to, this sort of "learning to learn" capability that is involved in inductive learning, might be part of (somewhere in the back of your mind) what you think you are teaching. In particular, what about teaching Air Force jobs, where you hope that if they suddenly need to do a different job, they can pick it up fairly quickly without too much formal instruction. When you choose to go the route of *macroadapting*, saying that we are going to give you the rules and be as efficient as possible, are you costing people anything? I don't know the answer to that. You showed us that adaptation is important, but have you showed us that macroadaptation is preferable to some more micro-oriented or some more domain-centered adaptation?

Doug Towne: Of course, they don't have to be mutually exclusive either.

Val Shute: I view micro- and macroadaptation as complementary approaches, hand-in-glove, rather than being mutually exclusive. Now, with macroadaptation, I have seen a lot of instances where people placed in an inductive environment (like Smithtown) either thrive and do well, or just flounder around and do poorly. For some people, garden paths are fruitful, more fruitful than, say, the straight and narrow. For example, when I get new software, I rarely open the manual, but I learn a lot just by trial and error. But other people do better in a directed environment. If we just default to using inductive kinds of environments, then that would only benefit some folks, and impede learning for others. Also, by using this macro- and microadaptive approach, if we select a learning environment at the outset based on aptitude data but the learning-in-progress data indicate that it's not being effective, well, nothing is written in stone. In other words, there's no reason why the learning environment couldn't be switched midstream, why the computer couldn't make another "macro" decision rule later on based on new information. So microadaptation is always going on, and macroadaptation can occur at the outset of learning, and possibly during the course of learning. So there's a constant interplay between micro- and macroadaptive responses.

Walt Schneider: I'm kind of concerned that after 30 years of searching for aptitude-treatment interactions, so many studies were below the level of statistical chance. So going after it now, anew, should be at least somewhat concerning. In the cases that you've illustrated, for example, one out of the four cases looks like an interaction. You have not just one test, there were a number of aptitudes you compared to a number of treatments, so how do you do the post hoc verification, whether there is anything there, is iffy at best. In order to have a chance at being able to impact an instructional domain, first you have to be able to have an a priori specification of an aptitude-treatment interaction. Then you have to be able to design a number of tutor systems that do all of those treatments so you have now increased your costs, perhaps significantly. The trick is coming up with an a priori interaction. I totally agree that, particularly in the military environment, you have lots of aptitude measures that just come in for free. Recruits come in and the first thing you do is test them. If there is something there, you need to be able to show when that treatment will be true and when it will be generalizable, so that it can impact the next person's tutor. In each case that you deal with, you need to figure out how to specify the decision rules. What is your reason for optimism?

Val Shute: My reasons for optimism are because the older ATI studies, hundreds of which are reviewed in Cronbach and Snow's 1977 book, are filled with confounded, noisy data. Those studies used different classrooms, different teachers with different personalities, different instructional materials that weren't controlled, and so on. A whole lot of noise entered into the equation so I'm not surprised that there were so few significant ATIs reported. But with ITSs, you can control these variables. I guess my optimism is partly innate, but also springs from having been successful in finding several significant aptitude-treatment interactions. And actually my tests were very conservative. I may have given the wrong impression that I just tested a bunch of interactions until I found some that were significant. That's not what I did. What I *did* was first compute a MANOVA and found that the overall aptitude by environment interaction was significant for all outcome measures considered simultaneously. That permitted me to then zoom in on the data, and I chose to start with a really basic ability: associative learning skill. So the MANOVA told me that, yes, there is something there, in general. This macro/microadaptive approach does provide a systematic way of fitting learning environments to individuals to optimize learning, which is the name of the game in ITS design.

Dan Fisk: But your optimism for generalizing these findings still has to be somewhat task specific.

Val Shute: Yes, absolutely. I am limited to only speaking about teaching different aspects of electricity, like conceptual knowledge or procedural skills. But the approach provides a framework for conducting additional studies, those that can systematically permute various things like learning environment and domain. Then I can collect more data and start making generalizations. I've just completed another large-scale study with a tutor that teaches flight engineering skills. I developed alternative learning environments, and have found some tantalizing ATIs involving working-memory capacity and general knowledge by different learning environments. But that's a story for another time.