Chi, M., & VanLehn, K. (2010). Meta-Cognitive Strategy Instruction in Intelligent Tutoring Systems: How, When, and Why. *Educational Technology & Society*, *13* (1), 25–39.

Meta-Cognitive Strategy Instruction in Intelligent Tutoring Systems: How, When, and Why

Min Chi and Kurt VanLehn*
Learning Research and Development Center & Intelligent System Program, University of Pittsburgh, PA, USA // mic31@cs.pitt.edu
*Department of Computer Science & Engineering, Arizona State University, AZ, USA // Kurt.Vanlehn@asu.edu

ABSTRACT

Certain learners are less sensitive to learning environments and can always learn, while others are more sensitive to variations in learning environments and may fail to learn (Cronbach & Snow, 1977). We refer to the former as high learners and the latter as low learners. One important goal of any learning environment is to bring students up to the same level of mastery. We showed that an intelligent tutoring system (ITS) teaching a domain-independent problem-solving strategy indeed closed the gap between high and low learners, not only in the domain where it was taught (probability) but also in a second domain where it was not taught (physics). The strategy includes two main components: one is solving problems via backward chaining (BC) from goals to givens, called the BC strategy, and the other is drawing students' attention to the characteristics of each individual domain principle, called the principle-emphasis skill. Evidence suggests that the low learners transferred the principle-emphasis skill to physics while the high learners seemingly already had such skill and thus mainly transferred the other skill, the BC strategy. Surprisingly, the low learners learned just as effectively as the high learners in physics. We concluded that the effective element of transfer seemed not to be the BC strategy, but the principle-emphasis skill.

Keywords

Intelligent tutoring systems, Meta-cognitive skills, Domain-independent problem-solving strategies

Introduction

Certain learners are less sensitive to learning environments and can always learn; while others are more sensitive to variations in learning environments and may fail to learn (Cronbach & Snow, 1977). We refer to the former as high learners and the latter as low learners. Bloom (1984) argued that human tutors not only raised the mean of test scores, but also decrease the standard deviation of scores. That is, students generally start with a wide distribution in test scores but as they are tutored, the distribution becomes narrower: the students at the low end of the distribution begin to catch up with those at the high end. Another way to measure the same phenomenon is to split students into high and low groups based on their incoming competence then measure the learning gains of both groups. According to Bloom, a good tutor should exhibit an aptitude-treatment interaction: both groups should learn, and yet the learning gains of the low students should be so much greater than those of the high ones that their performance in the post-test ties with that of the high ones. That is, one benefit of tutoring is to narrow or even eliminate the gap between high and low. In order to fully honor the promises of learning environments, an effective system should narrow the gap as much as possible without pulling the high learners down. Many preexisting systems can decrease such differences but not eliminate them. This is due in part to the fact that we do not fully understand why such differences exist.

One of many hypotheses is that low learners lack certain specific skills about how to think, including general problem-solving strategies and meta-cognitive skills. If this hypothesis is true, we expect that teaching students an effective problem-solving strategy would not only improve students' learning gains but also decrease the gap between the low and the high learners. Furthermore, if such problem-solving strategy is domain independent, we expect that learners would learn how to apply the strategy and seek to transfer it to new learning environments. Past research has indicated that these skills can be transferred across domains (Lehman, Lempert, & Nisbett, 1988; Lehman & Nisbett, 1990). However, few studies have investigated transfer of problem-solving strategy across domains.

In this paper, we investigate these questions in a special class of learning environments, intelligent tutoring systems (ITSs) (VanLehn, 2006). We present a study in which two groups of college students studied probability first and then physics. The experimental group studied probability with Pyrenees, an ITS that explicitly taught and required students to employ a general problem-solving strategy (VanLehn et al. 2004); while the control group studied

probability with Andes, an ITS that did not teach or require any particular strategy (VanLehn et al. 2005). During subsequent physics instruction, both groups used Andes, which also did not teach or require students to employ any particular strategy.

As reported earlier (Chi & VanLehn, 2007), we found that the experimental group out-performed the control group not only in probability, where the strategy was taught and forced upon the participants, but also in physics where it was not forced upon the participants. Furthermore, the strategy seemed to have lived up to our expectations and transferred from probability to physics. In this paper, we determine whether explicit strategy instruction exhibits an aptitude-treatment interaction, that is, whether it narrows or even eliminates the gap between high and low and, moreover, whether both high and low indeed transfer the strategy to the second domain.

Background

A task domain is deductive if solving a problem requires producing an argument, proof, or derivation consisting of one or more inference steps, and each step is the result of applying a domain principle, operator, or rule. For instance, solving algebraic equations is a deductive domain, and in particular, $2x + 5 = 21$ can be done via two steps: 1) subtract the same term 5 from both sides of the equation; and 2) divide both sides by the non-zero term 2. Proving a geometry theorem is deductive, as is solving quantitative physics problems. Deductive task domains are common parts of mathematical and scientific courses such as probability and physics. Two common problem-solving strategies in deductive domains are forward chaining (FC) and backward chaining (BC) (Russell & Norvig, 2003).

In FC, the solver starts with the set of given propositions, applies a principle to some subset of them (which produces a new proposition), adds the new proposition to the known propositions, and repeats this process until the problem's goal is met or no new proposition can be produced. BC is goal-directed in that the goal is progressively broken down into sub-goals and sub-sub-goals, etc. This constructs a partial plan in the form of a goal stack, which the solver uses to guide its forward application of principles. This is easier to explain with the aid of an example, so we will combine that explanation with an introduction to one of the task domains, probability. A portion of the probability task domain is described in first row in Table 1.

Table 1. A subset of the probability principle and an example that can be solved by these rules

Rules:	R1: For any event E, $P(E) + P(\sim E) = 1$ R2: If events A and B are independent, $P(A \cap B) = P(A)*P(B)$. R3: If events A and B are independent, then A and ~B, ~A and B, and ~A and ~B are all independent events.
Problem:	Events A and B are independent, and $P(\sim A) = 0.9$, $P(\sim B) = 0.8$. Compute P(A ∩ B).

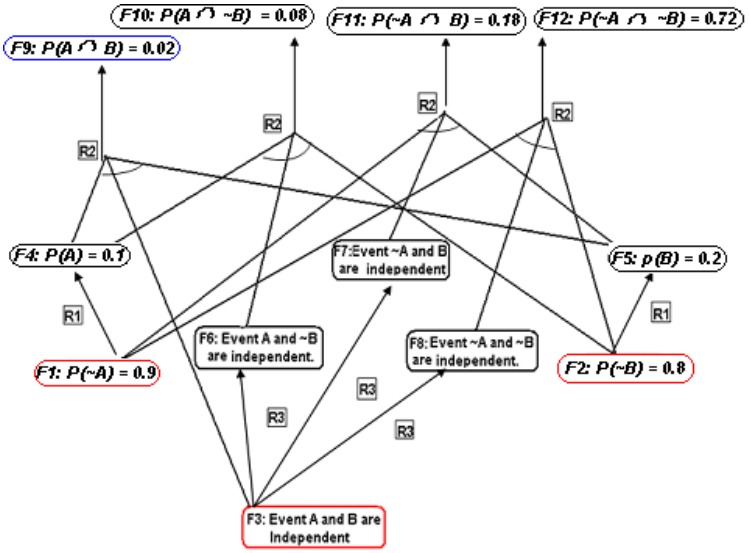

Figure 1. A solution graph using forward chaining

An example of solving a problem using forward chaining is shown in Figure 1 and the problem is listed in the last row of Table 1. One reads this graph starting at the bottom and working upward. The lowest propositions are given. Rules are applied to produce the propositions above. Although forward chaining stops once the goal of the problem, F9, is found, a few more propositions (F10, F11, and F12) are also shown in the graph because it is possible that they may be produced before F9.

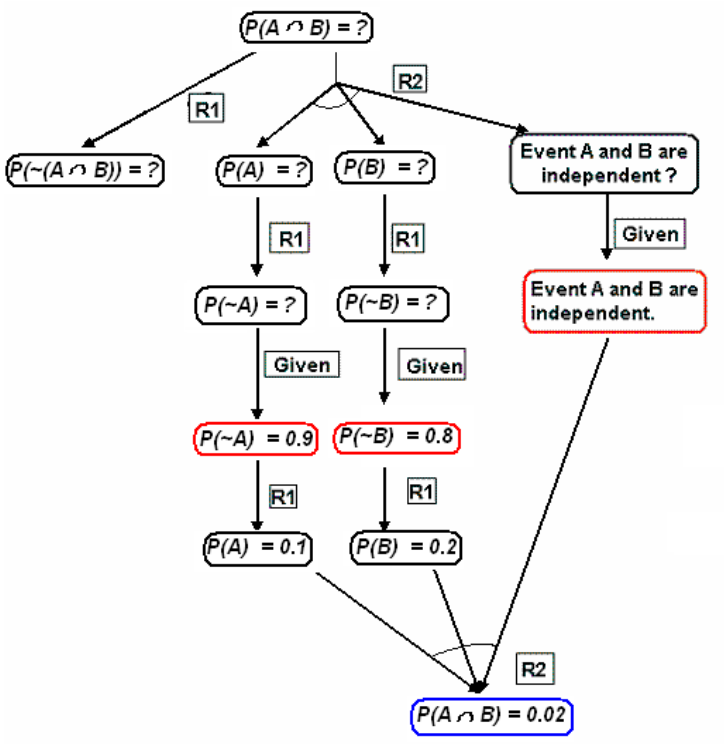

Figure 2. A solution graph using backward chaining

Figure 2 presents solving the same problem via BC. It is read from the top down. The problem's goal is decomposed by R1 into a sub-goal $P(\sim(A \cap B))$, but doesn't look promising because no other rules can be fired further. Thus, BC tries decomposing via R2. That yields three sub-goals, which look promising, so it continues decomposing each of them until they all eventuate in sub-goals that match given propositions. It then applies the rules in the forward direction, guided by the goal and sub-goal links that it has saved, doing computations as it goes and eventually calculating the problem's answer.

FC is complete (Russell & Norvig, 2003) but can be inefficient because its inference process is not directed toward solving the problem's goal. In Figure 1, FC did a lot of irrelevant work. Of the nine facts inferred, only three were needed for solving the problem. BC, on the other hand, is focused on achieving the problem's goals but can meet dead ends. For instance, $P(\sim(A \cap B))$ in Figure 2, is one such dead end because it triggers no more rules in Table 1.

Studies comparing strategy instruction with no-strategy instruction

Although FC and BC are widely used in computer science, they are seldom observed in a pure form in natural human problem solving. In Newell and Simon's (1972) seminal study of logic problem solving, *none* of the subjects used FC or BC in a pure form. Early studies of expert and novice physics problem solvers suggested that novices used BC and experts used FC (Larkin, McDermott, Simon, & Simon, 1980; Simon & Simon, 1978), but later studies showed that both used a mixture and, in fact, used fairly similar mixtures (Priest & Lindsay, 1992). Eventually, work in this area diminished, perhaps because it appeared that most human solvers used a mixture of strategies, analogies,

heuristics, and many other kinds of knowledge during their problem solving.

Although neither experts nor novices seem to use FC and BC in their pure form, the strategies' success in guiding computer problem solvers suggests that teaching students to use pure FC or BC might improve their problem solving. Several studies were conducted to test this hypothesis. Next, we will give a brief review.

Sweller and his colleagues conducted a series of studies comparing goal-free problem solving to ordinary problem solving (Owen & Sweller, 1985; Sweller, 1989; Tarmizi & Sweller, 1988). In the goal-free problem solving condition, students were not told a specific goal of a problem. Instead, they were asked to derive everything they could from the given facts. Students could only use FC on these goal-free problems, as BC requires a goal to start from. In the ordinary problem-solving condition, students were given problems with a specific goal, so they could use FC, BC, or a mixture. In all of these studies, the goal-free group learned more than the ordinary problem-solving group, thus suggesting that "teaching" a single problem-solving strategy in the form of pure FC improves learning. However, the number of inferences made to solve a goal-free problem is generally much larger than the number of inferences made to solve a goal-specific problem. Thus, the goal-free students could have benefited simply from having more practice in applying the domain principles. Indeed, when the experimenters modified the study so that students in both conditions applied the same number of domain principles, the difference between conditions disappeared (Owen & Sweller, 1985; Sweller, 1988). Thus, although these studies are consistent with the hypothesized benefits of explicit strategy instruction, there are other explanations for the results as well.

Trafton and Reiser (1991) tested the benefits of explicit strategy instruction in a sub-domain of computer programming, wherein students had to compose primitive functions in order to produce composite goal function. Three forms of instruction were compared based on the way in which the goal function could be assembled: forward-only, backward-only or freely. After completing 13 training problems in less than an hour, all three groups achieved the same learning gains. Although it is always hard to interpret a null result, it could be that the task domain was too simple to allow an explicit instruction on problem-solving strategies to demonstrate benefits.

Scheines and Sieg (1994) gave students over 100 training problems in sentential logic over a five-week period. One group of students was taught to use FC; a second group was taught to use BC; while a third group, the unconstrained group, was not taught any strategy and operated freely. After five weeks of instruction, no significant differences were found among the three groups on the mid-term exam (post-test). When the FC and BC groups were aggregated as a one-way strategy condition, there were still no significant differences between them and the unconstrained group on post-test scores. However, contrary to our hypothesis, the unconstrained students gained more than the one-way strategy students on difficult problems, where one would expect an explicit search strategy to be especially helpful. The experiment suggested that constraining students to use just one strategy may actually harm their performance.

VanLehn et al. (2004) compared an explicitly taught version of backward chaining to unconstrained problem-solving. Students who had not taken college physics were taught elementary mechanics over several multiple-hour training sessions. On some post-test measures, the students who were explicitly taught a strategy scored higher than those who were not taught a strategy and could solve problems in any order. However, on other measures, the two groups did not differ. Overall, performance on the post-test was quite poor, suggesting a floor effect — the post-test was too difficult for both groups.

In summary, most studies above were conducted in a single domain and contrasted students who were taught a strategy and those who were not. In this paper, we investigate the impact of explicit strategy instruction on eliminating the gap between high and low across two unrelated domains and two different ITSs. The problem-solving strategy chosen is the target variable strategy (TVS), a domain-independent BC strategy (VanLehn et al., 2004), and the two selected domains were probability and physics. During probability instruction, students in the experimental group were trained on an ITS, Pyrenees, that explicitly taught the TVS; while students in the control group were trained on another ITS, Andes, without explicit strategy instruction. During subsequent physics instruction, both groups were trained on the same ITS, which did not teach any strategy. On both probability and physics post-tests, we expect the following: high-experimental = low-experimental = high-control > low-control. That is, for both task domains, the low students should catch up with the high students, but only if they were taught the TVS.

Methods

Participants

Participants were 44 college students who received payment for their participation. They were required to have a basic understanding of high-school algebra, but not to have taken college-level statistics or physics courses. Students were randomly assigned to the two conditions. Two students were eliminated: one for a perfect score on the probability pre-test and one for deliberately wasting time.

Target variable strategy (TVS)

The TVS consists of three phases: 1) translating the problem statement, 2) applying principles and generating equations, and 3) solving equations. The central part of TVS happens in the second phase: applying principles and generating equations, in which students follow:

(1) Choose one of the *sought* (unknown) variables as the target variable
(2) Select a principle application that will generate an equation containing the target variable
(3) Define new variables as necessary to ensure that every quantity in the equation has a variable
(4) Write the equation in terms of the defined variables
(5) Mark the target variable *known*
(6) Mark the unknown variables in the equation *sought*

This procedure is repeated until there is no variable marked sought anymore, and then students go to the final phrase, solving equations. More details on TVS are described in (VanLehn et al., 2004). To illustrate, we will compare two example solutions for a probability problem. Table *2* contains a TVS solution by following the TVS while Table 3 contains a non-TVS solution. Note that $P(A)$, $P(A \cap B)$, $P(\sim A \cup \sim B)$, etc. are algebraic variables even though their names make them look like functions.

Table 2. Solving a problem by following the TVS

Problem: **Given $P(A) = 1/3$, $P(B) = 1/4$, $P(A \cap B) = 1/6$, find the probability: $P(\sim A \cup \sim B)$.**		
Step	Proposition	Justification
Phase 1: Translating the problem statement		
1	$P(A) = 1/3$	Given
2	$P(B) = 1/4$	Given
3	$P(A \cap B) = 1/6$	Given
4	$P(\sim A \cup \sim B)$	Sought
Phase 2: Applying principles and generating equations		
5	$P(\sim A \cup \sim B) = P(\sim(A \cap B))$	To find $P(\sim A \cup \sim B)$, apply De Morgan's theorem. Delete *sought* from $P(\sim A \cup \sim B)$ and mark $P(\sim(A \cap B))$ as sought
6	$P(A \cap B) + P(\sim(A \cap B)) = 1$	To find $P(\sim(A \cap B))$, apply the complement theorem. Delete *sought* from $P(\sim(A \cap B))$
Phase 3: Solving equations.		
7	$P(\sim(A \cap B)) = 5/6$	Solve 6.
8	$P(\sim A \cup \sim B) = 5/6$	Solve 5

Table 2 shows that in the first phase the student defines four variables, gives values to three of them, and marks $P(\sim A \cup \sim B)$ the sought variable. Then the student moves to the second phase. During each cycle of the second phase, the student must make two decisions. One is which sought variables to select as the target variable if there is more than one variable marked sought. Since all sought variables must eventually be selected as target variables, the order in which they are chosen does not affect the solvability of the problem. The other decision is which principle application to use if there happen to be several that contain the target variable. If the student makes an unlucky selection, the problem will be unsolvable. If so, the student must back up and make a different selection. In this example, during the first cycle, only $P(\sim A \cup \sim B)$ was marked as sought, so it is selected as the target variable. Then

the student chooses De Morgan's theorem as the principle to solve for $P(\sim A \cup \sim B)$. To do so, the student defines a variable $P(\sim(A \cap B))$ and then enters the equation $P(\sim A \cup \sim B) = P(\sim(A \cap B))$. Then the student removes the sought mark from $P(\sim A \cup \sim B)$, and marks the other variable in the equation, $P(\sim(A \cap B))$, as sought. This ends the first cycle. On the next cycle, again, only one variable is marked as sought, $P(\sim(A \cap B))$, so the student selects it as the target variable, applies the complement theorem to it, and removes the sought mark from it. At this point, no variables are marked sought, so the second phase ends. During the third and final phase, the student solves the equations in reverse chronological order.

Table **3. *A non-TVS solution of the same problem***

Step	Proposition	Justification
1	$P(A \cup B) = P(A)+P(B)- P(A \cap B)$	Addition theorem for two events: A and B
2	$P(\sim B)+p(B) = 1$	Complement Theorem
3	$P(A \cap B)+ P(\sim A \cup \sim B) = 1$	De Morgan's Law and Complement Theorem
4	$P(A \cap B) = 1/6$	Given
5	$P(\sim(\sim A \cup \sim B)) = 1/6$	Solve 3
6	$P(\sim A \cup \sim B) = 5/6$	Solve 4

Table 3 presents a solution to the same problem derived without the TVS. The solution is neither FC nor BC. It includes an equation (Step 2) that is not necessary for solving the problem and another equation (Step 3: $P(A \cap B) + P(\sim A \cup \sim B) = 1$) that is a combination of two principle applications: $P(\sim(A \cap B)) = P(\sim A \cup \sim B)$ and $P(A \cap B) + P(\sim(A \cap B)) = 1$. This is typical of the solutions by students who were not taught the TVS.

Students can use the TVS to solve problems in probability, physics, and many other tasks domains (VanLehn et al., 2004). In general, the TVS applies in task domains where solving a problem consists of generating a solvable set of equations. The TVS can be proved complete in that it will generate a set of equations that solves the problem if such a set exists. So far, we have described only the tedious procedural aspects of the TVS. There is only one key aspect of the TVS that has not yet been described, and we will describe it next.

A key detail and its implications for learning

In the second phase of the TVS, applying principles and generating equations, students select a principle application whose equation will contain the target variable. To do this, the tutoring system has students first pick a principle by name from a list of the domain principles that have been taught. It then has students specify how to apply that principle, again by making menu selections. For example, in step 6 of Table 2, after students pick $P(\sim(A \cap B))$ as the target variable and select the complement theorem to apply, but before they input the correct equation, $P(A \cap B) + P(\sim(A \cap B)) = 1$, the tutoring system would say:

> *You have chosen the complement theorem to apply for the target variable. To apply the principle, you must have noticed that there is a set of events that are mutually exclusive and collectively exhaustive. What are these events?*

Students should input the two events, $\sim(A \cap B)$ and $(A \cap B)$. In AI terms, the students are being asked to supply values for the arguments of the principle, thus establishing which of many possible instances of the principle should be applied. Only when an instantiation has been selected is the equation generated by the principle application completely determined and the tutoring system able to ask the student to enter it.

Therefore, the TVS is not simply a BC strategy, like the one used by Bhaskar and Simon (1977), which has students simply enter an equation. It instead has them specify both a principle and its arguments. Even if students know what equation they want, they have to figure out which principle and which arguments will generate it. Thus, they have to learn the principles deeply instead of simply learn a syntactic version of the available equations. As for the example above, students following the TVS are more likely to learn that the complement theorem should only apply in the events that are mutually exclusive and collectively exhaustive instead of in normal events. The TVS taught the students to focus their attention on acquiring a deep understanding of the principles, as that was all they needed in

order to run the TVS. To summarize, the TVS includes two main components: one is to solve problems via BC from goals to givens, called the BC-strategy, and the other is to focus attention to the domain principles, called the principle-emphasis skill.

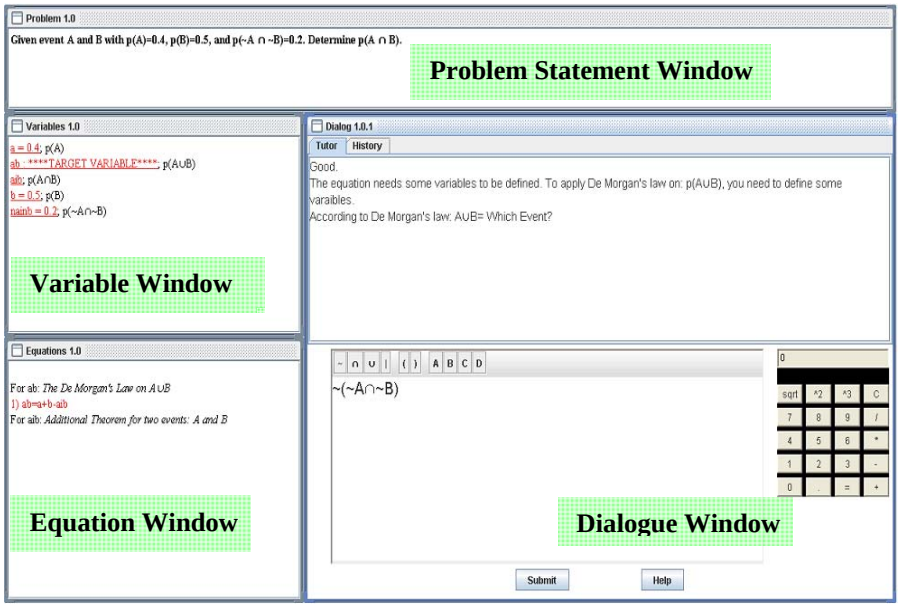

Figure 3. Pyrenees's interface

Three ITSs

The three ITSs involved in this study were Pyrenees, Andes probability, and Andes physics. Their corresponding screen shots are shown in Figure 3, *Figure 4*, and *Figure 5*, respectively. The first two taught probability, whereas the third taught physics. Apart from their domain knowledge, Andes probability and Andes physics were identical, so we use "Andes" to refer to both. Pyrenees required students to follow the TVS, while Andes did not require students to follow any problem-solving strategy. In this study, students in the experimental group learned probability in Pyrenees, and then learned physics in Andes; while student in the Control group learned both probability and physics in Andes. Next, we will compare Pyrenees and Andes from the perspectives of both the user interface and students' behaviors.

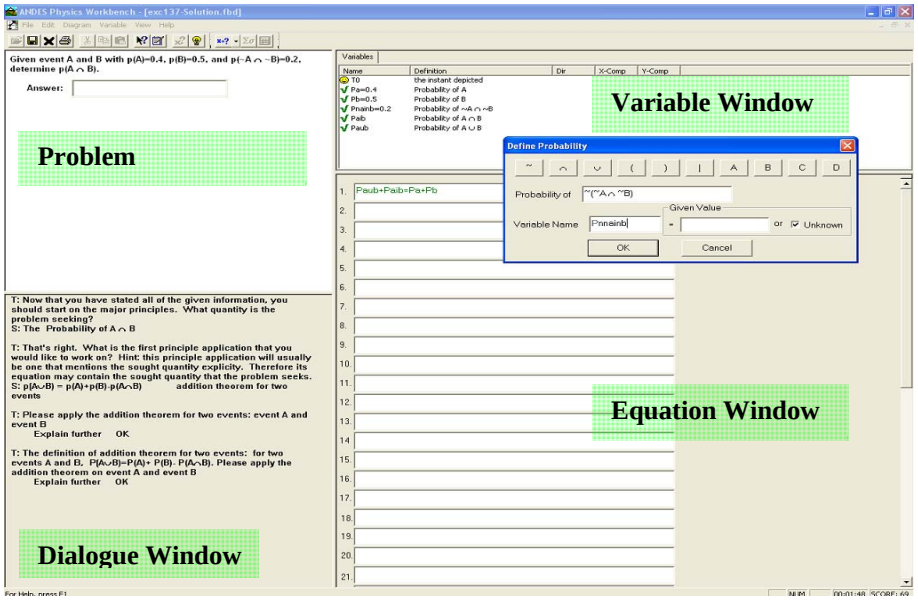

Figure 4. Andes probability's interface

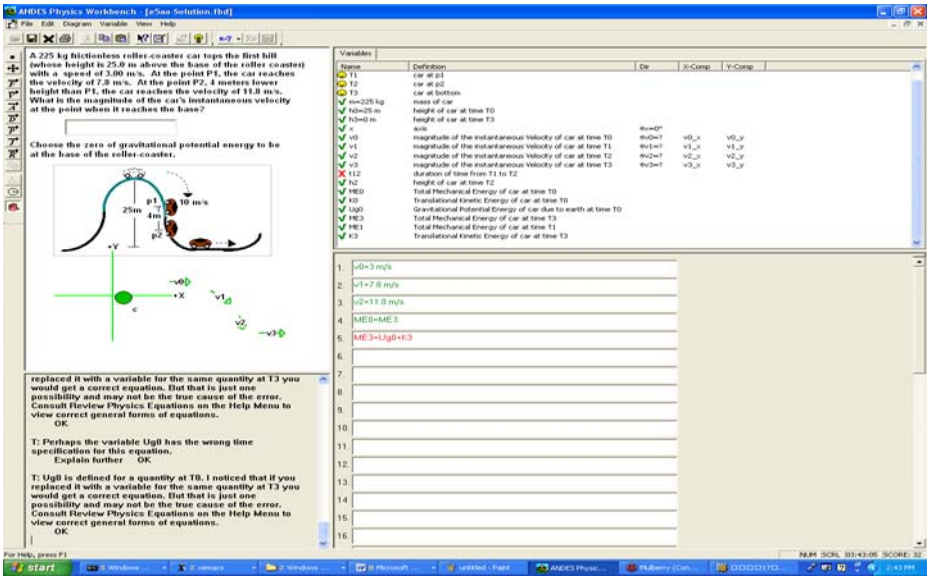

Figure 5. Andes physics' interface

User Interfaces Perspectives

Both Pyrenees and Andes provide a multi-paned screen that consists of a problem-statement window, a variable window for listing defined variables, an equation window, and a dialog window (see Figures 3–5). However, the computer-student interactions were quite different for each system. Pyrenees guided students in applying the TVS by prompting them to take steps dictated by the TVS. For example, when the TVS determined that it was time to define a variable, Pyrenees popped up a tool for that purpose (see Figure 3). Thus the interaction with Pyrenees was a turn-taking dialogue, where the tutor's turns always ended with a question to which the student must reply. All interaction with Pyrenees took place in the dialogue window. In Andes, on the other hand, students used GUI tools to construct and manipulate a solution. Thus the interaction with Andes was open-ended and event-driven. Students could edit or interact with any of the four windows by drawing vectors in the top left window, writing or editing equations in any row of the equation window, and so on. Once an entry or edit was made successfully, Andes provided no further prompting for the next step. If students didn't know what to do next, they could ask for a hint by clicking on the next-step help button.

Interactive behaviors perspectives

Both Andes and Pyrenees provide immediate feedback. However, their standard of correctness differs. Andes considers an entry correct if it is true, regardless of whether it is useful for solving the problem. On Pyrenees, however, an entry is considered correct if it is true and strategically acceptable to the TVS. Moreover, students can enter an equation that is the algebraic combination of several principle applications on Andes but not on Pyrenees because the TVS requires students to apply one principle at a time.

Both systems provide hints when students ask. When an entry is incorrect, students can either fix it independently or ask for what's-wrong help. When they do not know what to do next, they can ask for next-step help. Both next-step help and what's-wrong help are provided via a sequence of hints that gradually increase in specificity. The last hint in the sequence, called the bottom-out hint, tells the student exactly what to do. Pyrenees and Andes give the same what's-wrong help for any given entry, but their next-step help differs. Because Pyrenees requires students to follow the TVS, it knows what step they should be doing next so it gives specific hints. In Andes, however, students can always enter any correct step, so Andes does not attempt to determine their problem-solving plans. Instead, it asks students what principle they are working on. If students indicate a principle that is part of a solution to the problem, Andes provides as a hint an uncompleted step from the principle application. If no acceptable principle is chosen, Andes picks an unapplied principle from the solution that they are most likely to be working on.

Two domains

Two deductive domains, probability and physics, were involved in this study as the initial and transfer domain, respectively. Each domain contained ten major principles. Probability included the complement theorem, Bayes rule, and so on; while physics included the definition of kinetic energy, conservation of total mechanical energy, and so on.

Procedure

The procedure in this study had four main parts: background survey, probability instruction, Andes interface training, and physics instruction (shown in the left column of Table 4). All materials were online. The background survey asked for high school GPA, SAT scores, experience with algebra, and other information.

Table 4**.** Experiment procedure

Part	Experimental	Control
Survey	Background survey	
Probability instruction	Pre-training	
	Pre-test	
	Training on Pyrenees	Training on Andes probability
	Post-test	
Andes interface training	Solve a probability problem on Andes probability	
Physics instruction	Pre-training	
	Pre-test	
	Training on Andes physics	
	Post-test	

The probability and physics instruction each consisted of the same four phases: 1) pre-training, 2) pre-test, 3) training on the tutoring system, and 4) post-test. We describe each phase in turn, pointing out relevant differences, if any, between the two task domains.

Pre-training

During pre-training all students studied the domain principles. For each principle, they read a general description, reviewed some examples, and solved a series of single-principle and multi-principle problems. After solving a problem, the answer was marked correct or incorrect, and the correct solution was displayed. If the answer was incorrect, the students were asked to solve another problem isomorphic to the one that they had just failed to solve; this repeated until they either succeeded in solving a problem or failed three times. On multiple-principle problems, students had only one chance to solve the problem and were not asked to solve an isomorphic problem if their answer was incorrect.

Pre-tests

During the pre-tests, after an answer was submitted, students automatically proceeded to the next question without any feedback on the correctness of the answer. Students were not allowed to go back to earlier questions. This was the procedure for the post-tests as well. All students took the same pre- and post-tests. All test problems were open-ended and required students to derive an answer by writing and solving one or more equations.

Training on ITSs

In Phase 3, students first watched a video that demonstrated problem solving in the corresponding ITS. During probability instruction, the strategy students also read a text description of the TVS. Then, all students solved the

same twelve probability problems or eight physics problems in the same order. More specifically, students in the experimental group solved all twelve probability problems in Pyrenees and students in the control group solved them in Andes probability. Both conditions solved the eight physics problems on Andes physics. Students could also access the domain textbook at any time during training. During the probability training, students in the experimental group were able to access a description of the TVS. Each main domain principle was applied at least twice in both trainings.

Post-tests

Finally, all students took a *post-test.* Five problems on both post-tests were isomorphic to training problems in Phase 3. In addition, there were five non-isomorphic, novel, multiple-principle problems in the probability post-test and eight in the physics post-test. Table 5 shows the distribution of single-principle and multiple-principle problems in the experiment.

Table 5. Number of various problems during pre-training, pre-test, training, and post-test

		Single-principle	Multiple-principle	Total
Probability	Pre-training	14	5	19
	Pre-test	10	4	14
	Training	−	12	12
	Post-test	10	10	20
Physics	Pre-training	11	3	14
	Pre-test	9	5	14
	Training	−	8	8
	Post-test	5	13	18

Only the students in the experimental group took the third part, Andes interface training. Its purpose was to familiarize them with the Andes GUI without introducing any new domain knowledge. The problem used was one of the 12 probability training problems that they had previously solved on Pyrenees. Pilot studies showed that one problem was sufficient for most students to become familiar with Andes GUI.

To summarize, the procedural difference between the two conditions were: 1) during probability instruction, students in the experimental group trained on Pyrenees while students in the control group trained on Andes probability; 2) students in the experimental group learned how to use the Andes user interface before they received physics instruction.

Grading criteria

We used two scoring rubrics: binary and partial credit. Under the binary rubric, a solution is worth 1 point if it is completely correct or 0 if not. Under the partial credit rubric, each problem score is a proportion of correct principle applications evident in the solution. If they correctly applied four of five possible principles they would get a score of 0.8. Solutions were scored by a single grader blind to conditions.

Results

In order to measure aptitude-treatment interaction, we needed to define high and low groups based on some measure of incoming competence. We chose to use MSAT scores because probability and physics are both math-like domains. Our split point was 640, which divide into high (n = 20) and low (n = 22). Except for the MSAT scores and high-school GPA, no significant difference was found between high and low on other background information such as age, gender, VSAT scores, and so on. As expected, the high group out-performed the low group during the probability pre-training and the probability pre-test under the binary scoring rubric: $t(40) = 3.15$, $p = 0.003$, $d = 0.96$, $t(40) = 2.15$, $p = 0.038$, $d = 0.66$, and $t(40) = 2.27$, $p < 0.03$, $d = 0.70$ on single-principle, multiple-principle problems during probability pre-training, and overall in probability pre-test, respectively. The same pattern was found under partial

rubric in the probability pretest. Thus, the MSAT score successfully predicted the incoming competence of the students, which justifies using it to define our high vs. low split.

Incoming competence combined with conditions partitioned the students into four groups: high-experimental (n = 10), low-experimental (n = 10), high-control (n = 10), and low-control (n = 12). Fortunately, random assignment balanced the experimental vs. control conditions for ability, and this balance persisted even with the groups subdivided into high and low via MSAT score. On every measure of incoming competence, no significant difference was found between the experimental and control groups, the low-experimental and low-control ones, or the high-experimental and high-control ones. These measures were the background survey, the probability pre-test, probability pre-training scores, the time spent reading the probability textbook, and the time spent solving the pre-training problems. Averaged over all students, the total time for each training phase were 2.4 hrs and 2.7 hrs for probability pre-training and training and 1.5 hrs and 3.0 hrs for physics pre-training and training, respectively. No significant differences were found among the four groups on any of these times.

Test scores

Error! Reference source not found. shows that the test score results are consistent with our hypothesis. After training on Pyrenees, the low-experimental students scored significantly higher than their low-control peers on all three assessments: probability post-test, physics pre-test and physics post-tests: $t(20) = 4.43$, $p < 0.0005$, $d = 1.90$; $t(20) = 3.23$, $p < 0.005$, $d = 1.34$; and $t(20) = 4.15$, $p < 0.0005$, $d = 1.84$, respectively. More importantly, the low-experimental students even seemed to catch up with the high ones: no significant difference was found among the high-experimental, low-experimental, and high-control on all three assessments, even though the two experimental groups seemed to out-perform the high-control group in Figure 6.

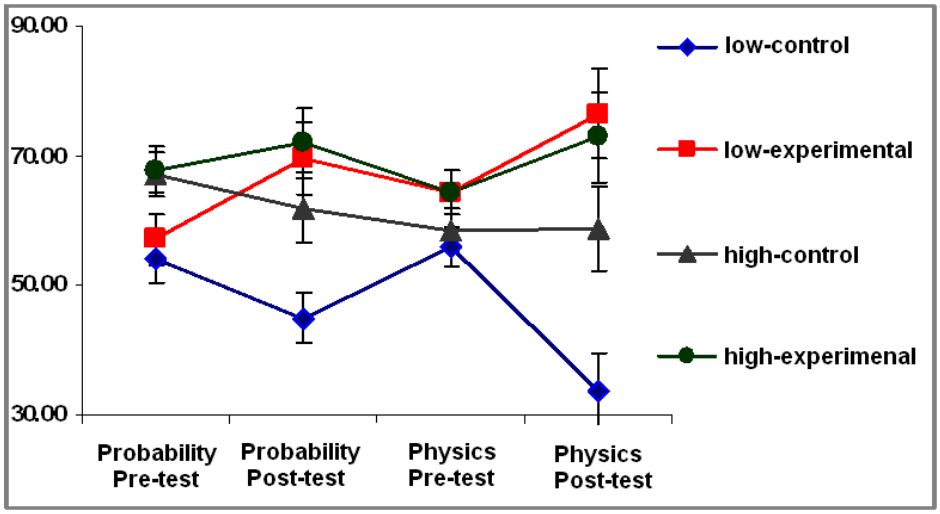

Figure 6. Comparison of four groups on four tests (maximum score = 100)

Thus, the Pyrenees instruction in probability caused the low-experimental group to learn more effectively than those in the low-control group during probability training, physics training, and even physics pre-training. They seemed to have caught up to the high ones while the low-control ones did not. Moreover, while the high-experimental group didn't benefit much from the TVS, they were not harmed either.

Dynamic assessments

While test results are the most common assessment of learning performance, one can also compare students' behaviors as they learn. Such comparisons are called dynamic assessments (Haywood & Tzuriel 2002). In performing dynamic assessments, we can identify students who are effective learners even though their test scores

may be equal to or even lower than those of others. Here we investigated students' interactive behaviors on Andes during physics training, as all students received the identical procedure during that period.

Frequency of help requests

Andes physics logs every user's interface action performed, including help requests, tool usage, and equation entries. We first tried to characterize the overall difference in students' solutions via the amount of help they requested. On each of eight physics training problems, the low-experimental students made significantly fewer next-step help requests than the low-control ones. No significant difference was found among the low-experimental, high-experimental and high-control groups. This suggests that the low-experimental students may have transferred the TVS. However, there are other possible explanations, so we conducted several other analyses.

Triage of logs

Solution logs were grouped into three categories: smooth, help-abuse, and rocky. Smooth solutions included no help requests, except on problems that required more than eight principle applications. Students were permitted up to two what's-wrong help requests. Help-abuse solutions were produced when every entry was derived from one or more next-step helps. Otherwise, the solution was categorized as rocky because students appeared capable of solving part of the problem on their own, but needed help on the rest.

Figure 7 shows a significant difference among the four groups on the distribution of the three types of solutions. While no significant difference was found between the high-experimental and low-experimental groups, there was a significant difference between the low-control and the high-control groups: $\chi^2(2) = 11.33$, $p(\chi^2) = 0.003$. Furthermore, a significant difference was found between the low-experimental and high-control groups: $\chi^2(2) = 15.322$, $p(\chi^2) < 0.001$, and between the high-experimental and high-control groups: $\chi^2(2) = 11.585$, $p(\chi^2) < 0.005$. Qualitatively, the results appear to be as follows: high-experimental = low-experimental > high-control > low-control.

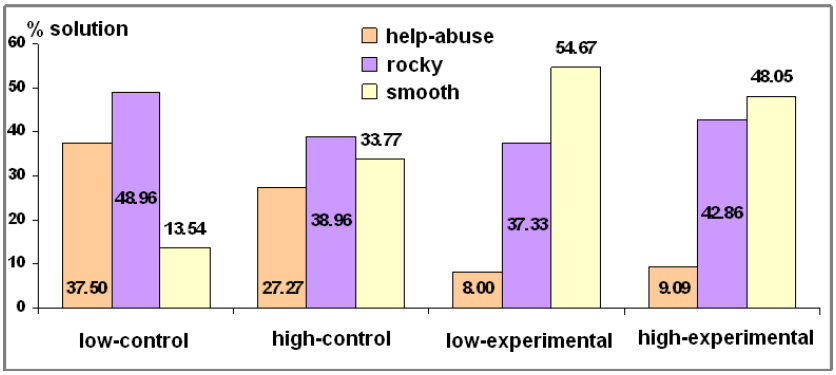

Figure 7. Solution percentage by type

For a more quantitative measure, we used a smaller unit of analysis, individual equations. We coded each correct equation entry in the solution logs with 3 features:

- *Relevance*: The equation was labeled relevant or irrelevant based on whether it contributed to the problem solution.
- *Help:* The equation was labeled "Help" if it was entered after the student asked for help from Andes physics. Otherwise, it was labeled "No-help."
- *Content*: The equation's content was coded as either "a correct equation with new physics content" or "others."

We sought to find out how frequently students made progress toward solving a problem without asking for any help from Andes. In terms of the three-feature coding mentioned above, such a "desirable" equation would be coded as "relevant," "no-help," or "correct equation with new physics content." We called these desirable equations *desirable*

steps and defined the desirable steps ratio (DSR):

$$DSR = \frac{Desirable\ \ steps\ \ in\ the\ solution}{All\ \ steps\ \ in\ the\ solution}$$

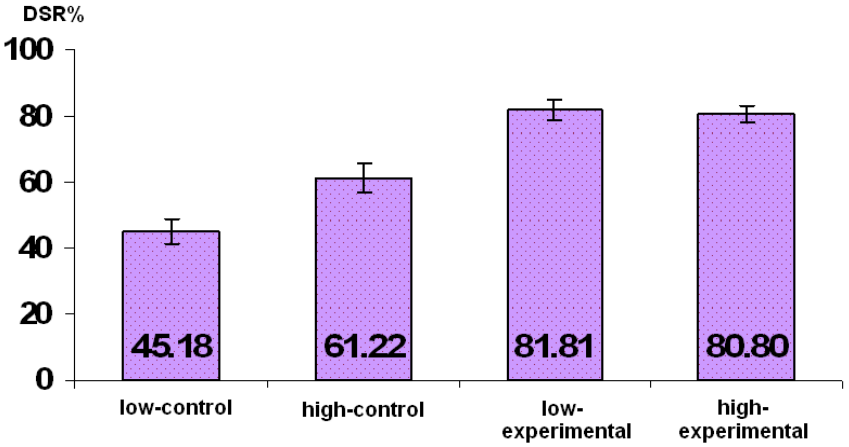

Figure 8. DSR on overall solutions

As shown in Figure 8, the low-experimental students had significantly higher DSR than the low-control ones: $t(169) = 7.50$, $p < 0.0001$. In fact, the former even made significantly more progress than the high-control group: $t(150) = 3.84$, $p < 0.001$. While there is a significant difference between the low-control and high-control groups: ($t(171) = 2.83$, $p < 0.01$), there is no such difference between the two experimental groups. In short, this dynamic assessment showed that the following: high experimental = low experimental > high-control > low-control.

To summarize, both test scores and dynamic assessments show that the low students catch up with the high ones in the experimental condition but not in the control condition. On some measures, the low-experimental students even surpass the high-control ones. Next, we'll investigate what was transferred from probability to physics that made the low experimental students so successful.

Transferring the two cognitive skills of the TVS

As we described above, the TVS includes two main components: solving problems via backward-chaining (BC) from goals to givens, called the BC-strategy, and drawing students' attention to the characteristics of each individual domain principle, called the principle-emphasis skill. In the following, we will investigate whether either or both skills were transferred by the two experimental groups to physics. In order to determine the BC-strategy usage, we analyzed students' logs to see whether the order of equations in their solutions followed the BC strategy. For the principle-emphasis skill, we used the single-principle problems as our litmus test. Students who had applied the BC-strategy would have no particular advantage because solving these single-principle problems need to apply only one principle. On the other hand, students who had learned the idea of focusing on domain principles should be at an advantage.

Transferring the BC strategy

If students engaged in the BC-strategy, we expect they would apply the BC strategy when they had difficulties, that is, on rocky solutions. On smooth solutions, students don't have any difficulties since they may solve problems mainly based on existing schemas (Sweller, 1989). Thus, we subcategorized each desirable step in the logs as BC or non-BC, where non-BC included FC, combined equations, and so on. We then defined BC% as the proportion of desirable steps that were coded as BC. Figure 9 showed that on Rocky solutions the high-experimental group applied BC significantly more frequently than the other three groups: $t(40) = 2.25$, $p = 0.03$ while the low-experimental group used the BC as frequently as the two control groups. Thus, apparently it was the high-experimental group alone who transferred the BC-Strategy to physics.

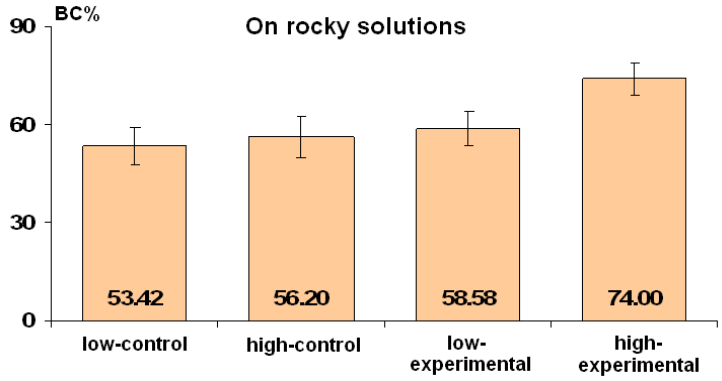

Figure 9. BC usage on rocky solutions

Transfer of the principle-emphasis skill

The low-experimental students scored just as high as the high-experimental ones even though they used BC no more frequently than the students in the two control groups. Our hypothesis is that they transferred the principle-emphasis skill. We divided both post-tests into single-principle and multiple-principle problems. Furthermore, we divided the multiple-principle problems into those that were isomorphic to a training problem and those that were not. If students in the low-experimental group applied the principle-emphasis skill, we expected them to out-perform students in the low-control group on all three types of problems in both post-tests. This turned out to be the case (see Table 6). In Table 6, the third and fourth columns list the means of test scores of the low-experimental and low-control groups. The low experimental group had reliably higher means than the low-control group in both probability and physics post-tests across three types of problems: simple-principle, isomorphic multiple-principle and non-isomorphic multiple-principle. This suggests that a main effect of teaching the TVS to the low students was to get them to focus on the domain principles. Further analysis showed no significant difference among the students in high-control, low-experimental, and high-experimental groups on any types of problems, which indicates that high students may already have such skill.

Table 6. Scores on three types of problems in both probability and physics post-tests

Test	Problem type	Mean (low-experimental)	Mean (low-control)	Statics
Probability post-test	Single	0.93	0.70	$t(20) = 3.62$, $p = 0.002$, $d = 1.58$
	Multiple, isomorphic	0.48	0.23	$t(20) = 3.71$, $p = 0.001$, $d = 1.55$
	Multiple, non-isomorphic	0.44	0.17	$t(20) = 3.734$, $p = 0.013$, $d = 1.15$
Physics post-test	Single	0.93	0.8	$t(20) = 4.33$, $p < 0.001$, $d = 1.85$
	Multiple, isomorphic	0.60	0.18	$t(20) = 4.55$, $p < 0.001$, $d = 1.93$
	Multiple, non-isomorphic	0.70	0.19	$t(20) = 3.734$, $p < 0.001$, $d = 2.10$

Conclusions

Overall, our instructional manipulation indeed exhibited an aptitude-treatment interaction: the gap between high-experimental and low-experimental students seemed to be eliminated in both probability and physics, whereas it remained between the high-control and low-control groups. More detailed analyses of the training behavior and post-test results suggest that students in the low-experimental group transferred the principle-emphasis skill to physics while those in the high-experimental group apparently already possessed it. On the other hand, students in the high-experimental group transferred the BC strategy.

These results suggest that it is not the BC strategy that is most important to teach low learners. Instead, one should teach the meta-cognitive skill of focusing on individual principle applications. It could be that low and high learners differed initially in that low students lacked this "how to learn" meta-cognitive knowledge for a principle-based domain like probability or physics. Such results suggest building an ITS that does not teach the TVS explicitly, but instead just teaches to focus on principle applications in deductive domains. Perhaps it would be just as effective as Pyrenees. Indeed, because its students need not learn all the complicated bookkeeping of the BC strategy, which may cause cognitive overload (Sweller, 1989), it might even be more effective than Pyrenees not only for an initial domain where the ITS was used but also for subsequent domains where it is not used.

References

Bhaskar, R., & Simon, H. A. (1977). Problem solving in semantically rich domains: An example from engineering thermodynamics. *Cognitive Science, 1,* 193–215.

Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher, 13,* 4–16.

Chi, M. & VanLehn, K. (2007). Accelerated future learning via explicit instruction of a problem solving strategy. In R. Luckin, K. R. Koedinger, & J. Greer (Eds.) *The13th International Conference on Artificial Intelligence in Education* (pp. 409–416). Amsterdam, Netherlands: IOS Press.

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

Haywood, H.C. & Tzuriel, D. (2002). Applications and challenges in dynamic assessment. *Peabody Journal of Education, 77(2),* 40–63.

Larkin, J., McDermott, J., Simon, D. P., & Simon, H. A. (1980). Expert and novice performance in solving physics problems. *Science, 208,* 1335–1342.

Lehman, D. R., Lempert, R. O., & Nisbett, R. E. (1988). The effects of graduate training on reasoning: Formal discipline and thinking about everyday-life events. *American Psychologist, 43,* 431–442.

Lehman, D. R., & Nisbett, R. E. (1990). A longitudinal study of the effects of undergraduate training on reasoning. *Developmental Psychology, 26,* 431–442

Newell, A., & Simon, H. A. (1972). *Human problem-solving.* Engelwood Cliff, NJ: Prentice-Hall.

Owen, E. & Sweller, J., (1985) What do students learn while solving mathematics problems? *Journal of Educational Psychology 77(3),* June 1985, 272–284.

Priest, A. G., & Lindsay, R. O. (1992). New light on novice–expert differences in physics problem-solving. *British Journal of Psychology, 83,* 389–405.

Russell, Stuart J. & Norvig, Peter (2003), *Artificial Intelligence: A Modern Approach* (2nd Ed.), Upper Saddle River, NJ: Prentice Hall.

Simon, D. P., & Simon, H. A. (1978). Individual differences in solving physics problems. In R. S. Siegler (Ed.), *Children's thinking: What develops?* Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Sweller, J. (1989). Cognitive technology: Some procedures for facilitating learning and problem-solving in mathematics and science. *Journal of Educational Psychology, 8 (4),* 457–466.

Tarmizi R. A., & Sweller, J. (1988). Guidance during mathematical problem-solving, *Journal of Educational Psychology 80(4),* 424–436.

Trafton, J. G., & Reiser, B. J. (1991). Providing natural representations to facilitate novices' understanding in a new domain: Forward and backward reasoning in programming. *Proceedings of the Thirteenth Annual Conference of the Cognitive Science Society* (pp. 923–927). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

VanLehn, K. (2006) The behavior of tutoring systems. *International Journal of Artificial Intelligence in Education. 16* (3) 227–265.

VanLehn, K., Bhembe, D., Chi, M., Lynch, C., Schulze, K., Shelby, R., Taylor, L., Treacy, D., Weinstein, A., & Wintersgill, M. (2004). Implicit versus explicit learning of strategies in a non-procedural cognitive skill. In J. C. Lester, R. M. Vicari, & F. Paraguacu (Eds.), *Intelligent Tutoring Systems: 7th International Conference* (pp. 521-530). Berlin: Springer-Verlag.

VanLehn, K., Lynch, C., Schulze, K. Shapiro, J. A., Shelby, R., Taylor, L., Treacy, D., Weinstein, A., & Wintersgill, M. (2005). The Andes physics tutoring system: Lessons learned. *International Journal of Artificial Intelligence and Education*, *15(3),* 1–47.