Eliminating the Gap between the High and Low Students through Meta-Cognitive Strategy Instruction

Min Chi and Kurt VanLehn
Learning Research and Development Center & Intelligent System Program
University of Pittsburgh, PA, 15260
{mic31, vanlehn+}@pitt.edu

Abstract. One important goal of Intelligent Tutoring Systems (ITSs) is to bring students up to the same level of mastery. We showed that an 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, named the BC-strategy, and the other is drawing students' attention on the characteristics of each individual domain principle, named 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 former learned just as effectively as the latter in physics. We concluded that the effective element of the taught strategy seemed not to be the BC-Strategy, but the principle-emphasis skill.

Keywords: Intelligent Tutoring Systems, meta-cognitive skills, domain-independent Problem-Solving Strategies.

1 Introduction

Bloom [2] argued that human tutors not only raised the mean of 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 on the low end of the distribution begin to catch up with those on the high end. Another way to measure the same phenomenon is to split students into High and Low groups based on their incoming competence. One then measures 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 the High ones' that their performance in the post-test ties with the High ones. That is, one benefit of tutoring is to narrow or even eliminate the gap between High and Low.

Previously, we found that Pyrenees [11], an ITS that explicitly taught a problem-solving strategy, was more effective than Andes [12], an ITS that did not explicitly teach any strategy not only in the domain where it was used, but in a second domain where it was not used [3]. The strategy seemed to have lived up to our expectations and transferred from one domain to another. In this paper, we investigated whether explicit strategy instruction exhibited an aptitude-treatment interaction, that is, whether it narrows or even eliminates the gap between High and Low; moreover, whether both High and Low indeed transferred the strategy to the second domain.

2 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. Deductive domains are common parts of math and science courses. Two common problem-solving strategies in deductive domains are forward chaining (FC) and backward chaining (BC) [7]. In FC, reasoning proceeds from givens toward goals; while in BC, it works backward from goals to givens. FC and BC have been widely used in computer science; however, they are rarely observed in a pure form in natural human problem solving. Early studies suggested that novices used BC and experts used FC [5], but later studies showed that both used fairly similar mixtures [6]. It appears that most human solvers use a mixture of strategies, heuristics, and analogies with past solutions as well as other general knowledge. Although human solvers don't seem to use FC and BC in their pure form, the strategies' success in guiding computer problem solvers suggests that teaching human solvers to use FC or BC might improve their problem-solving performance. Several ITS-based studies were conducted to test this hypothesis.

Trafton and Reiser [10] tested the benefits of explicit strategy instruction on an ITS called Graphical Instruction in Lisp. Three forms of instruction were compared: FC-only, BC-only or freely. After 13 training problems were completed in less than one hour, all three groups achieved the same learning gains. Scheines and Sieg [8] gave students over 100 training problems in sentential logic and they found students who were taught and required to use FC or BC learned just as effective as those who were not taught any strategy. VanLehn et al. [10] compared two ITSs that teach introductory college physics. One system explicitly taught students a version of BC; while the other did not teach or require students to follow any explicit strategy. Although some outcome measures differed between groups, overall performance on the post-test was quite poor, suggesting a floor effect.

In summary, most previous studies were conducted in a single domain and contrasted students who were taught a strategy and those who were not. In this paper, we investigated 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) [11], a domain-independent BC strategy, and the two selected domains were probability and physics. Probability covered 10 major principles in Axiom of Probability and Conditional Probability; and physics covered 10 principles in Work and Energy. During probability instruction, the Experimental students were trained on an ITS, Pyrenees, that explicitly taught the TVS; while the Control students 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:

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.

3 Methods

3.1 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.

3.2 Three ITSs

The three ITSs involved in this study were Pyrenees, Andes-probability, and Andes-physics respectively. The first two taught probability while the third taught physics. Apart from domain knowledge, Andes-probability and Andes-physics were the same and we use 'Andes' to refer to both. Pyrenees required students to follow the TVS while Andes did not require students to follow any explicit problem-solving strategy. Next, we will compare Pyrenees and Andes from the perspectives of both the user interface and students' behaviors.

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. The tutor-student interactions are quite different for each system.

Pyrenees is a restrictive tutor-initiative ITS. It guides students in applying the TVS by prompting them to take each step as dictated by the strategy. For example, when the TVS determines that it is time to define a variable, Pyrenees will pop up a tool for that purpose. Thus the tutor-student interactions in Pyrenees take the form of a turn-taking dialogue, where the tutor's turns end with a prompt or question to which the student must reply and all interactions only takes place in the dialogue window. Andes, on the other hand, is a nonrestrictive mixed-initiative ITS. Students use GUI tools to construct and manipulate a solution, so the interaction is event-driven. Students may edit or interact with any of the four windows: by drawing vectors in vector window, writing or editing equations in the equation window, and so on. Once an entry or edit has been made successfully, Andes provides no further prompt to make the next step.

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 asked. 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 hints 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.

3.3 Procedure

The study had 4 main parts: background survey, probability instruction, Andes Interface training, and physics instruction (shown in the left column of Table 1). All materials were online. The background survey asked for High school GPA, SAT scores, experience with algebra and other information.

Table 1. 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 four phases: 1) Pre-training, 2) Pre-test,3) Training on the ITS, and 4) Post-test. During pre-training, students studied domain principles. For each principle, they read a text description, reviewed some worked examples, and solved some single-principle and multiple-principle problems. After solving a problem, their answer was marked correct or incorrect, and the expert's solution was also displayed. The students then took the pretests. All students took the same pre- and post-tests. All test problems were open-ended and required students to derive answers by writing and solving one or more equations. In phase 3, students in both conditions solved the same problems in the same order, albeit on different ITSs. Each of the domain principles was applied at least twice in both trainings. The Experimental group learned probability in Pyrenees and physics in Andes-physics while the Control group learned both domains in Andes. Students could access the domain textbook at any time during the training. Finally, students took the post-tests. On each post-test, 5 problems were isomorphic to a training problem in phase 3. There were also 5 novel, non-isomorphic multiple-principle problems on the probability post-test and 8 on the physics post-test.

Only the Experimental students 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 twelve 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 differences between the two conditions were: 1) during the probability training, the Experimental condition trained on Pyrenees while the Control condition trained on Andes-probability; 2) the Experimental students learned how to use Andes' GUI before physics instruction.

3.4 Grading Criteria

We used two scoring rubrics: binary and partial credit. Under binary, a solution is worth 1 point if it was completely correct or 0 if not. Under partial credit, each problem score is a proportion of correct principle applications evident in the solution. If they correctly apply 4 of 5 possible principles they would get a score of 0.8. Solutions were scored by a single grader blind to condition.

4 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 were 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 times for each training phase were: 2.4 hrs and 2.7 hrs for probability pre-training and training; 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.

4.1 Test Scores

Figure 1 shows that the test score results are consistent with our hypothesis: after trained 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 significantly 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 in Figure 1.

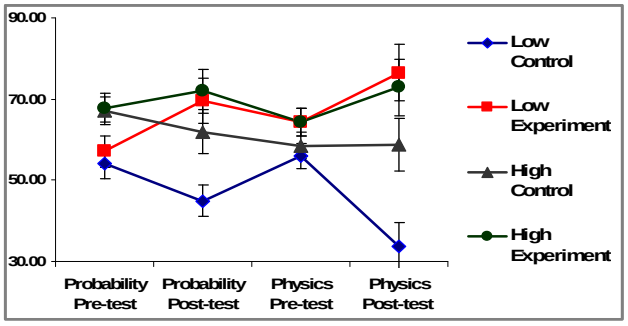

Figure 1. Compare four groups on four tests (maximum score = 1)

Thus, explicit strategy instruction in probability caused the Low-Experimental group to learn more effectively than 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.

4.2 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* [2]. In so doing, we can identify students who are effective learners even though their test scores may be equal to or even lower than 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 8 physics training problems, the Low-Experimental students made *significantly fewer* next-steps help requests than the Low-Control ones. No significant difference was found among the Low-Experimental, the High-Experimental and High-Control groups. This suggests that the Low-Experimental 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 3 categories: smooth, help-abuse, and rocky:

Smooth solutions included no help requests, except on problems that required more than eight principle applications. There students were permitted up to two *what's-wrong help* requests.

Help-abuse solutions are 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 2 shows there was a significant difference among four groups on the distribution of the three types of solutions. While no significant difference was found between the High-Experimental and Low-Experimental, there was a significant difference between the Low-Experimental and the High-Control: $\chi^2(2) = 11.74$, $p(\chi^2) < 0.003$; and between the High-Experimental and High-Control: $\chi^2(2) = 9.06$, $p(\chi^2) < 0.01$. Qualitatively, the results appear to be: High-Experimental = Low-Experimental > High-Control > Low-Control.

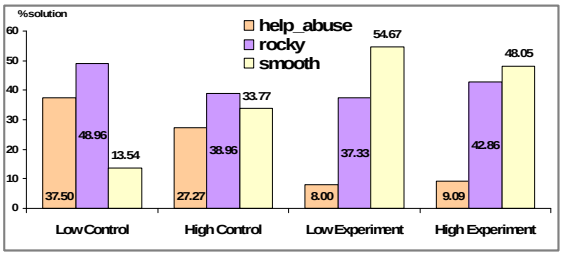

Figure 2. Solution Percentage by Type.

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", and "Correct equation with new physics content". We called them *desirable* steps and defined the desirable steps ratio DSR:

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

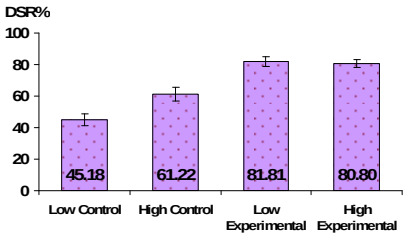

Figure 3. DSR on overall solutions.

As shown in Figure 3, the Low-Experimental had significantly Higher DSR than the Low-Control: $t(169) = 7.50$, $p<0.0001$. In fact, the former even made significantly more

progress than the High-Control: $t(150)= 3.84$, $p< 0.001$. 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: 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?

4.2 Transferring the Two Cognitive Skills of the TVS

The TVS is BC problem-solving strategy [11]. That is, it solves problems backwards from goals to givens. However, it differs from pure BC in that it requires students to explicitly identify principles before applying them. As an illustration, Table 2 presents the principle application part of a TVS solution.

Prior work on BC through equations required students to enter the equations alone [1]. Thus, they might only write the equations shown in the middle column of Table 2. Our TVS strategy, however, also requires them to attend to the application of individual domain principles, as shown in the right column of Table 2. For example, instead of simply entering an equation with one principle application each, students need to pick an unknown variable, select a principle that apply to the unknown variable, define all the variables appearing in its equation if undefined yet, write the equation, and finally remove the "sought" mark and mark new unknown variables. Students were also asked various questions on the characteristics of the principle. For example, in last row in Table 2, after students pick the complement theorem, Pyrenees would ask: "… *To apply the principle, you must have noticed that there are a set of events that are mutually exclusive and collectively exhaustive. What are these events?*" Students should answer: $\sim(A\cap B)$ and $(A\cap B)$. Therefore, the TVS is not only a BC strategy, but it draws students' attention to the characteristics of each individual domain principle, such as when it is applicable.

Table 2. Part of a TVS example solution.

Problem: **Given P(A)=1/3, P(B)=1/4, P(A ∩ B)=1/6, find: P(~A ∪ ~B).**		
Step	**Equations**	**Justification**
1	$P(\sim A \cup \sim B)=P(\sim(A\cap B))$	To find $P(\sim A \cup \sim B)$, apply De Morgan's theorem. Delete "sought" from $P(\sim A \cup \sim B)$ and add "sought" to $P(\sim(A\cap B))$
2	$P(A\cap B) + P(\sim(A\cap B))=1$	To find $P(\sim(A\cap B))$, apply the Complement theorem. Delete "sought" from $P(\sim(A\cap B))$

In short, we argue that the TVS includes two main components: one is to solve problems via BC from goals to givens, named the BC-strategy, and the other is to focus attention to the domain principles, named the principle-emphasis skill. In order to determine the BC-strategy usage, we analyzed students' logs to see whether the order of equations in their solutions follows the BC. For the principle-emphasis skill, we used the single-principle problems as our litmus test because students who had applied the BC-strategy would have no particular advantage on them because solving these single-

principle problems only need to apply one principle; while students who had learned the idea of focusing on domains principles should show an advantage on them.

Transfer 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. Whereas on smooth solutions, students don't have any difficulties since they may solve problems mainly based on existing schemas [9]. 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 4 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.

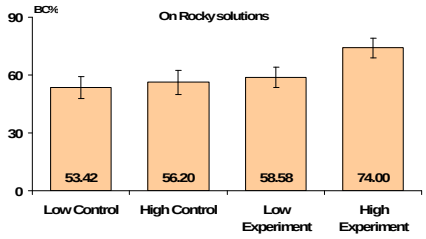

Figure 4. BC Usage on Rocky Solutions

Transfer of the Principle-Emphasis Skill: The Low-Experimental students scored just as high as the High-Experimental ones even though they used the BC no more frequently than two Control groups. Thus, they must have transferred something else of the TVS. 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 the Low-Experimental group applied the principle-emphasis skill, we expected them to out-perform the Low-Control group on all of them in both post-tests. This turned out to be the case (see Figure 5). It suggests that the 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 High-Control, the Low-Experimental, and High-Experimental on any types of problems, which indicates that High students may already have such skill.

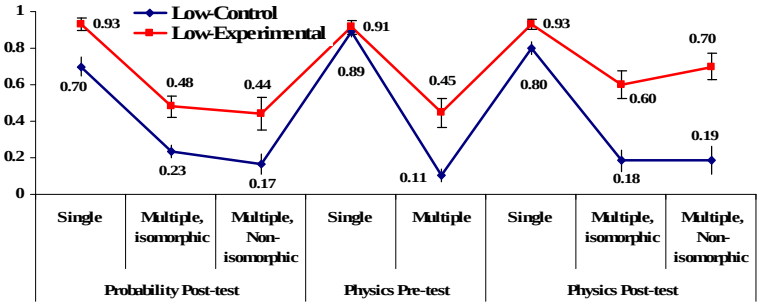

Figure 5. Scores on Three Types of Problems in Both Tests

5 Conclusions

Overall, we found teaching students the TVS indeed exhibited an aptitude-treatment interaction in deductive domains: the gap between High and Low students in the Experimental Condition seemed to be eliminated in both probability and physics. Although the two Experimental groups performed equally well in both physics pre- and post-tests, the Low-Experimental group transferred the principle-emphasis skill to physics while the High-Experimental apparently already possessed it and thus they mainly 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 may have 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 [9], it might even be more effective than Pyrenees not only for an initial domain where the ITS was used but subsequent domains where it is not used.

References

1. Bhaskar & Simon. (1977). Problem solving in semantically rich domains: An example from engineering thermodynamics. Cognitive Science, 1, 193-215.
2. Bloom (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. Educational Researcher, 13, 4-16.
3. Chi & VanLehn (2007). Accelerated future learning via explicit instruction of a problem solving strategy. 13th International Conference on AIED. pp. 409-416.
4. Haywood & Tzuriel (2002). Applications and challenges in dynamic assessment. Peabody Journal of Education, 77(2), 40-63.
5. Larkin, McDermott, Simon, and Simon (1980). Expert and novice performance in solving physics problems. Science, 208:1335-1342.
6. Priest and Lindsay (1992). "New Light On Novice-Expert Differences in Physics Problem-solving," British Journal of Psychology, 83, 389-405.
7. Russell, and Norvig (1995). Artificial Intelligence: A Modern Approach, Second Edition. Upper Saddle River, NY: Prentice-Hall.
8. Scheines, & Sieg. (1994). Computer environments for proof construction. *Interactive Learning Environments*, 4(2), 159-169.
9. Sweller. (1989). Cognitive Technology: Some Procedure for Facilitating learning and Problem-solving in mathematics and Science. Journal of EdPsych, 81-4, 457-466.
10. Trafton. and Reiser, (1991) Providing natural representations to facilitate novices' understanding in a new domain: Forward and backward reasoning in programming. Proceedings of the 13th Annual Conference of the Cognitive Science Society, 923-927.
11. VanLehn. et al. (2004). Implicit versus explicit learning of strategies in a non-procedural cognitive skill. 7th Conference on ITS, Maceio, Brazil.
12. VanLehn, et al. (2005). The Andes physics tutoring system: Lessons learned. International Journal of Artificial Intelligence and Education, 15(3), 147-204.