Further Results from the Evaluation of an Intelligent Computer Tutor to Coach Self-Explanation

Cristina Conati[1], Kurt VanLehn[2]

[1]Department of Computer Science, University of British Columbia
2366 Main Mall, Vancouver, B.C. Canada V6T 1Z4
email: conati@cs.ubc.ca
[2]Department of Computer Science, University of Pittsburgh,
Pittsburgh, PA, 15260, U.S.A
vanlehn@cs.pitt.edu

Abstract

We present further results on the educational effectiveness of an intelligent computer tutor that helps students learn effectively from examples by coaching self-explanation – the process of explaining to oneself an example worked-out solution. An earlier analysis of the results from a formative evaluation of the system provided suggestive evidence that it could improve students' learning. In this paper, we present additional results derived from a more comprehensive analysis of the experimental data. They provide a stronger indication of the system's effectiveness and suggest general guidelines for effective support of self-explanation during example studying.

1 Introduction

The research presented in this paper represents a step toward exploring innovative ways in which computers can enhance education and learning. While most intelligent tutoring systems support students during problem solving and teach domain specific skills, we have devised a computational framework that supports learning from examples and that coaches the general learning skill known as self-explanation - generating explanations and justifications to oneself to clarify an example solution. Several studies show that self-explanation greatly improves learning from examples (for overviews of these studies see [4] and [10]) and that coaching self-explanation can extend these benefits ([3], [4]). Our framework, known as the SE-Coach, aims to provide the individualized monitoring and guidance to self-explanation that has been proven so beneficial when administered by human tutors. It has been implemented and tested within Andes [11], a tutoring system that helps students learn Newtonian physics through both example studying and problem solving.

Other tutoring systems rely on examples as instructional means, but they use them to support students as they solve problems, not as a specific learning phase prior to and complementary to problem solving. These systems present students with relevant examples as they are solving problems and help students understand the connection between the example and the problems [12], [7], [1]. However, none of these systems

monitor how students study and understand the presented examples. Moreover, the systems themselves, rather than the students, generate explanations to help the students understand the examples. The Geometry Tutor [2] explicitly encourages students to explain the solution steps they have used to build geometry proofs, in terms of geometry axioms. However, the explanations are generated during problem solving and consist simply of selecting an item from a list of geometry axioms. The student does not have to explain the content of the axiom. Furthermore, the tutor makes the student explain each solution step, instead of trying to assess if some explanations may be more beneficial for the student than others.

Unlike the systems above, the SE-Coach includes an interface designed to encourage spontaneous, constructive self-explanation of examples. It also includes a help module that explicitly elicits further self-explanation tailored to a student's needs, as assessed by the SE-Coach probabilistic student model, when the interface scaffolding is not sufficient to overcome the natural reticence to self-explain that many students show [4], [10].

Self-explanation is a learning process whose underlying mechanisms are still unclear and under investigation. Since the SE-Coach is built on existing hypotheses about the features that make self-explanation effective for learning, an accurate evaluation of its effectiveness may allow us to shed light on the validity of the hypotheses and possibly suggest new ones. In [6], we presented initial results of a formal evaluation that we performed to test the usability and effectiveness of the system. These results indicated that the SE-Coach's interface is easy to use and generally effective in stimulating self-explanation. They also provided initial support on the SE-Coach's educational effectiveness. In this paper, we present a more detailed analysis of the experimental data that reveals a significant interaction between experimental condition and the learning stage in which students used the system, and provides insights on how the SE-Coach can more effectively bring students to constructively learn from examples.

2 Overview of the System

The SE-Coach's interface includes three different levels of scaffolding for self-explanation, to accommodate the varied propensity to self-explain that different students have, so as to provide each student with the minimum intervention sufficient to trigger constructive self-explanation.

The first level of scaffolding is given by a masking interface that presents different parts of the example covered by grey boxes (see Figure 1). In order to read the text or graphics hidden under a box, the student must move the mouse pointer over it. The fact that not all the example parts are visible at once helps students focus attention and reflect on individual example parts, and allows the SE-Coach to track student's attention [6]. The second level of scaffolding is provided by explicit prompts to self-explain. These prompts go from a generic reminder to self-explain, that appears when a student uncovers an example part, to more specific prompts for self-explanations that have been shown to correlate with learning in the self-explanation studies: (a) justify solution steps in terms of domain principles; (b) relate solution steps to goals in the underlying solution plan.

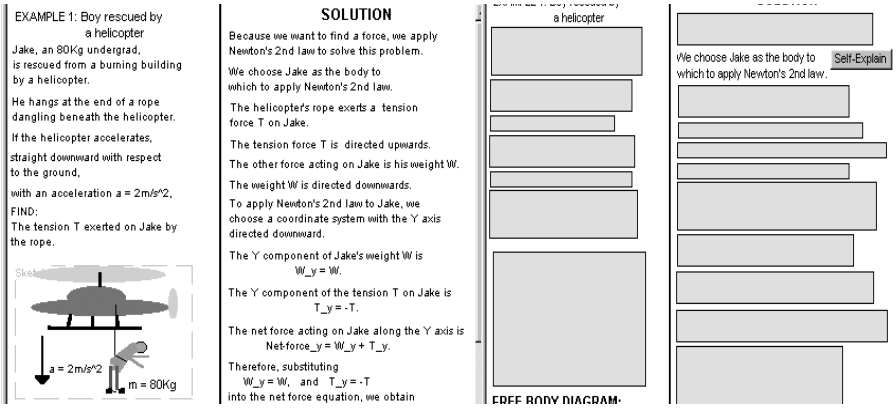

Figure 1: A physics example (left), as it is presented in the masking interface (right)

The third level of scaffolding consists of menu-based tools designed to provide constructive but controllable ways to generate the above self-explanations, to help those students that would be unable to properly self-explain if left to their own devices [10]. If a student selects the prompt to self-explain in terms of domain principles ("This is true because..."), a Rule Browser is displayed in the right half of the window (see Figure 2a), while if the student selects the prompt to self-explain in terms of the solution plan ("The purpose of this step is..."), a Plan Browser is activated instead.

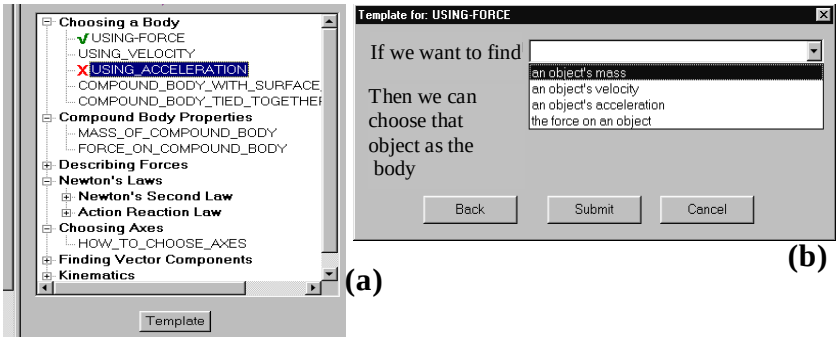

Figure 2: (a) Selections in the Rule Browser and (b) Template filling

The rule browser contains a hierarchy of physics rules, reflecting the content of the SE-Coach's knowledge base. The student can browse the rule hierarchy to find a rule that justifies the currently uncovered part. The SE-Coach will use a green check or a red cross to provide feedback on the correctness of the student's selection (see Figure 2a). To explain more about the actual content of a rule, the student can click on the "Template" button in the rule browser. A dialog box comes up (see Figure 2b) with a partial definition of the rule that the student can complete by selecting appropriate

fillers from available pull down menus. The SE-Coach gives immediate feedback on the student's selections.

The plan browser is similar to the rule browser, but it displays a hierarchical tree representing the solution plan for a particular example instead of the SE-Coach's physics rules. The student explains the role of the uncovered part by selecting in the plan hierarchy the step that most closely motivates the fact.

The SE-Coach includes a probabilistic student model based on a Bayesian network. The Bayesian network comprises a model of correct self-explanation for the current example, probabilities estimating the student's physics knowledge and nodes representing the student's reading and self-explanation actions. At any time during the interaction, probabilities in the Bayesian network assess how well the student understands the example solution and how the student's knowledge changes as a result of the interaction with the system [5]. Using this assessment, the SE-Coach prompts the student to generate further self-explanation to fix gaps in the student's example understanding.

Initially, self-explanation is voluntary. However, if a student tries to close the example when the student model indicates that there are still some lines left to self-explain, then the SE-Coach generates a warning and colors pink the corresponding masking interface boxes. It also provides more directive advice as of what interface tool should be used to better self-explain each line. The SE-Coach's tutorial interventions represent a fourth, stronger level of scaffolding for self-explanation, directed to help those students that do not self-explain because they tend to overestimate their understanding [4].

3 Empirical Evaluation of the SE-Coach

To test the system's effectiveness for learning, we performed a formal study with 56 college students. The SE-Coach does not provide any introductory physics instruction, because it is meant to complement regular classroom activities. Therefore, an evaluation of the SE-Coach requires subjects who have the right level of domain knowledge for using the system. Students generally benefit more from examples when they are studying a new topic, whereas as the students' knowledge improves, problem solving becomes more effective for learning [8]. Hence, to evaluate the SE-Coach adequately, subjects need to have enough knowledge to understand the topic of the examples, but not so much knowledge to find the examples not worthy of attention. The ideal evaluation setting for the SE-Coach would be in the context of an introductory physics course, where it is possible to control when students are ready to study examples on a new topic. Unfortunately, we could not coordinate the SE-Coach's evaluation with a specific physics course. Instead, we conducted the study in our laboratory, with students who were taking introductory physics classes at four different colleges: the University of Pittsburgh (20 students), Carnegie Mellon University (14 students), Community College of Allegheny County (5 students) and U.S. Naval Academy (17 students). The best we could do to get subjects at comparable learning stages was to run the subjects after their first class on Newton's Second Law and before they took a class test on the topic.

The one-session study comprised: 1) solving four pre-test problems on Newton's Second Law; 2) studying examples on Newton's Second Law with the system; 3)

solving post-test problems equivalent but not identical to the pre-test ones; The study had two conditions. In the *experimental (SE)* condition, 29 students studied examples with the complete SE-Coach. In the *control* condition, 27 students studied examples with the masking interface and Plan Browser only[1]. They had no access to the Rule Browser and Templates, nor feedback or coaching.

3.1. Effectiveness of the SE-Coach

As we reported in [6], the analysis of the log data file from the study shows that the SE-Coach's interface is easy to use and is quite successful at stimulating self-explanation. The gains scores between post-test and pretest were higher for the SE condition, although the difference between gain scores of the two conditions was not statistically significant. Since then, we have sought to better understand the reason behind the above outcome by restricting the analysis to the subgroups of subjects coming from different colleges. We found that the SE condition of CMU (Carnegie Mellon) and CCAC (Community College of Allegheny County) students performed better than the control condition (see Figure 3). The performance difference, as measured by an Analysis of Covariance with post-test as dependent variable, pre-test as covariate and condition as main effect, was statistically significant for CMU students ($p < 0.04$) and nearly significant ($p = 0.0576$) for CCAC students. In contrast, in the Pitt (Univ. of Pittsburgh) and USNA (U.S. Naval Academy) subgroups, students in the control condition performed slightly better than students in the SE condition (see Figure 3), although the difference was not statistically significant

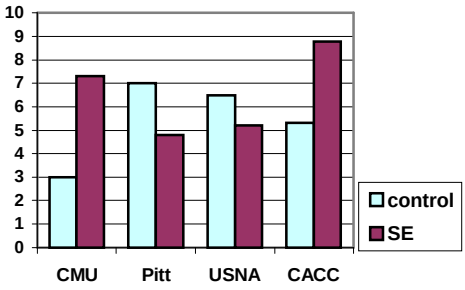

Figure 3: Gains scores for the four subgroups

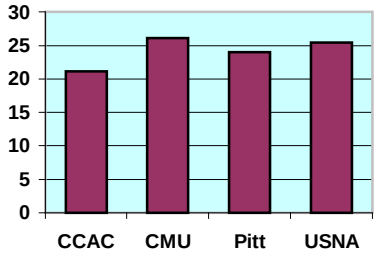

Figure 4: pretest scores for the four subgroups

The commonality of behavior between CMU and CACC students is quite surprising, because CMU and CCAC are supposed to be, respectively, the best and the worst among the four colleges in the study. This ranking is confirmed by the pretest scores shown in Figure 4. The difference in pretest performance between CMU and CCAC is the only one that approaches significance ($p = 0.0561$), among the pretest performances of the four groups.

[1] We let the control students access the Plan Browser because introductory physics courses usually do not address solution planning, therefore control students would have had too much of a disadvantage if they had not been able to see what a solution plan is through the Plan Browser.

To understand what may have caused this different learning behavior, we collapsed and analyzed the data in two subgroups with the same learning outcome, CMU-CCAC and Pitt-USNA. Within the CMU-CCAC group, students in the SE condition performed significantly better than students in the control condition, after covarying out the pretest ($p = 0.021$). Pitt-USNA students in the control condition performed slightly better than those in the SE condition, but the difference is not statistically significant ($p > 0.2$).

3.2. Possible Differences in the Student Populations

One possible explanation for the above results could be a difference in physics and background knowledge between the two subgroups of CMU-CCAC and Pitt-USNA students. However, an ANCOVA with post-test as dependent variable and subgroup and condition as main effects, shows that there is still a significant interaction ($p < 0.01$) of subgroup with condition after covarying out pretest only and both pretest and SAT scores. Although 10 subjects are excluded from the latter ANCOVA (we did not have these subjects' SAT scores), these data still provide a strong indication that physics and background knowledge do not explain the different performance of the two subgroups.

A second explanation for the different learning behavior of the CMU-CCAC and Pitt-USNA subgroups could be that subjects in the two subgroups used the system differently. The one thing that CMU and CCAC have in common, and that distinguishes them from Pitt and USNA students, is that they start the semester more than a week later. Therefore, although all the subjects participated to the experiment after they had their lectures on Newton's laws and before they took a class test on the topic, Pitt and USNA subjects were ahead in the course schedule and had likely spent more time on Newton's laws than CMU and CCAC subjects when they participated to the study. Our data show that this did not significantly influence the pretest performance of the two subgroups. However, it may have caused the students in the two subgroups to have a different attitude toward the example study task we made them perform.

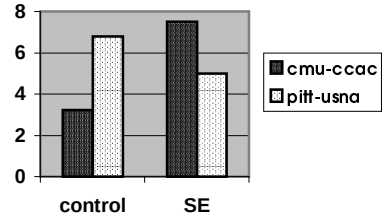

Figure 5: gains scores of the two subgroup in each condition

If we analyse the learning patterns of the two subgroups within each condition, we find that in the SE condition, CMU-CCAC students learned more than Pitt-USNA students (see figure 5), although the difference is not statistically significant ($p > 0.1$). In the Control condition, Pitt-USNA students learned significantly more than CMU-CCAC students ($p < 0.03$). These outcomes could be due to two reasons:

- In the SE condition, Pitt-USNA students did not use the SE-Coach as extensively and effectively as the CMU-CCAC students did.
- In the Control condition, Pitt-USNA students self-explained spontaneously more that the CMU-CCAC students did.

We will now verify these two hypotheses by comparing the log data of the two subgroups within the SE and the control condition.

Log data analysis of the two subgroups within the SE condition

To test whether CMU-CCAC students used the SE-Coach better than the Pitt-USNA students in the SE condition, we compared time on task, statistics describing how subjects used the interface self-explanation tools (Rule Browser, Plan Browser and Templates) and how they reacted to the SE-Coach's advice to further self-explain.

Rule Browser	**CMU-CCAC (12)**	**Pitt-USNA (17)**	**p**
Initiated	63.6%	61.4%	0.8
Correct	88%	86%	0.6
Attempts before correct	1.1	1.3	0.35
Max # attempts	7.8	10.2	0.45
Attempts before abandon	4.3	3.7	0.7

Template	**CMU-CCAC (12)**	**Pitt-USNA (17)**	**p**
Initiated	57.6%	53.8%	0.7
Correct	97.2%	96.8%	0.8
Attempts before correct	0.47	0.51	0.8
Max # attempts	2.2	2.7	0.3
Attempts before abandon	3	0.15	**0.011**

Plan Browser	**CMU-CCAC (12)**	**Pitt-USNA (17)**	**p**
Initiated	36.2%	45%	0.55
Correct	92%	81%	0.15
Attempts before correct	1	1	0.9
Max # attempts	3.9	3.8	0.96
Attempts before abandon	1.4	1.1	0.77

Table 1: Statistics on interface tools usage for CMU-CCAC and Pitt-USNA students

	CMU-CCAC (12)	**Pitt-USNA (17)**	**p**
Rule prompts followed	41%	37%	0.8
Plan prompts followed	50%	36%	0.36
Read prompts followed	31%	35%	0.88

Table 2: SE-Coach prompts statistics for CMU-CCAC and Pitt-USNA students

For each interface tool, we computed the following data summaries (see Table 1): *Initiated:* percentage of the explanations that students initiated out of all the explanations that could be generated with that tool for the available examples. *Correct*: percentage of the initiated explanations that were generated correctly. *Attempts before correct:* average number of attempts the students made before achieving a correct self-explanation. An attempt is the submission of an incorrect self-explanation. *Max # attempts:* average maximum number of attempts needed to achieve a correct self-explanation. *Attempts before abandon:* average number of attempts before abandoning a self-explanation. We also computed how many of the different prompts generated during the SE-Coach tutorial interventions (prompts to

self-explain using the Rule Browser, the Plan Browser or by reading more carefully) the students actually followed (see Table 2). There is no statistically significant difference in the average time on task for the two subgroups ($p > 0.1$). The only significant difference in the way CMU-CCAC and Pitt-USNA students used the system in the SE condition is that CMU-CCAC students performed a significantly higher number of attempts before giving up on a Template explanation (see Table 1, Template data). This suggests that the CMU-CCAC students had a higher level of motivation to learn from the SE-Coach self-explanation tools, consistently with the fact that students in the CMU-CCAC group had started studying Newton's Laws later than Pitt-USNA students and thus they were likely more willing to put substantial effort in learning from examples on the topic.

The CMU-CCAC students' higher level of motivation can explain why they learned more from the SE-Coach than the Pitt-USNA students did, although in general they did not use the system more easily and extensively (as Table 1 and Table 2 show). Selecting items in the browsers and filling templates does not necessarily trigger constructive learning if students do not reflect on what they are doing. Indeed, if students are not motivated to put substantial effort in studying examples, the actions of browsing and Template filling may act as distracters from learning. Students may concentrate their attention on selecting items to get positive feedback on their interface actions, but not actually reflect on the physics behind the actions and behind the worked out solution. Thus, we argue that CMU-CCAC students in the SE condition learned more from the same self-explanation actions than Pitt-USNA students because, being more motivated, they reasoned more constructively on their self-explanation actions and on the physics underlying them.

This argument is supported by the correlation between post-test scores and the number of rules that reached high probability in the student model. The correlation is very low ($r < 0.1$) for Pitt-USNA students and it is higher ($r = 0.33$) for the CMU-CCAC students. Since the probabilities in the student model are driven upward by correct self-explanations conducted on the SE-Coach's interface, the high correlation of the CMU-CCAC group suggests that their self-explanations drove their understanding upward just as they drove the model's probabilities upward, whereas the low correlation of the Pitt-USNA group suggests that their learning was independent of their use of the SE-Coach's self-explanation tools.

Log data analysis of the two subgroups within the control condition

The hypothesis that the learning of Pitt-USNA students in the control condition is due to spontaneous self-explanation is not easy to verify, because in this condition students could not express their self-explanation through the SE-Coach. The only log data file that could indirectly indicate self-explanation in the control condition are: (1) average number of multiple accesses to example lines; (2) standard deviation of the above measure; (3) average time spent on each example line; (4) standard deviation of the above; (5) time on task; (6) number of accesses to the Plan Browser; (7) number of selections in the Plan Browser.

We ran a regression analysis of post-test on the above variables for the Pitt-USNA control group and we found a marginally significant correlation of post-test scores with average and standard deviation of line accesses ($p = 0.083$ and $p = 0.057$ respectively). We found no significant correlations in the same regression analysis

for the CMU-CCAC control group. These results support the hypothesis that Pitt-USNA control students were selectively reviewing example lines because they were self-explaining specific example parts, while the CMU-CCAC control students' reviewing actions were not accompanied by effective self-explanation. The hypothesis that Pitt-USNA students self-explained more in the control condition is consistent with the fact that Pitt-USNA students had started studying Newton's Laws earlier and had probably gained more knowledge on the topic. This knowledge was not strong enough to make Pitt-USNA students perform better in problem solving tasks (their pretest performance was comparable to the CMU-CCAC students' one). However, it was sufficient to enable Pitt-USNA control subjects to generate effective self-explanations under the minimal scaffolding provided by the masking interface. We argue that it is indeed the minimality of the scaffolding that allowed Pitt-CMU control students to bring to bear their knowledge at best. Because of their more advanced learning stage, spontaneous self-explanation triggered by the masking interface likely came quite effortlessly to Pitt-USNA control students and therefore was not suffocated by the lower level of motivation that prevented Pitt-USNA students in the SE condition to learn effectively from the SE-Coach self-explanation tools.

4 Conclusions and Future Work

In this paper, we discussed the results of a formal study to evaluate an intelligent computer tutor that coaches the meta-cognitive skill known as self-explanation – generating explanations to oneself to clarify an example worked out solution. The tutor provides different levels of tailored scaffolding for self-explanation, to provide each student with the minimum intervention sufficient to trigger self-explanation while maintaining the spontaneous, constructive nature of this learning strategy.

Formal studies are fundamental to assess why and how a computer tutor does or does not support learning. Understanding how students use and learn from the SE-Coach is especially important, because the SE-Coach focuses on a learning process that no other tutoring system has tackled so far and whose underlying mechanisms are still unclear and under investigation. In particular, different studies have shown that both simple prompting [4] and more elaborate scaffolding [3] enhance self-explanation and learning, but no study has yet addressed the explicit comparison of these different kinds of intervention. The study that we performed provides initial insights on this issue. In this paper, we have presented data analysis results indicating that the stage of learning in which the students use the system influences how much they benefit from versions of the system that provide different amounts of scaffolding for self-explanation. The data suggest the following conclusions on the SE-Coach effectiveness and, in general, on the effectiveness of support for self-explanation during example studying.

- Rich scaffolding for self-explanation, like the one provided by the complete SE-Coach in the experimental condition, can improve students' performance at an early learning stage. At this stage, students are still unfamiliar with the subject matter. Hence, they benefit more from structured help in using domain knowledge to generate effective self-explanations and are more motivated to put substantial effort in exploiting this help.

- As students become more proficient in the subject matter, even minimal prompting, like the one provided by the masking interface in the control condition, can help improve their self-explanations. At this stage, more elaborate scaffolding can actually be less effective, if it requires students to put too much effort in studying examples, because they may lack the motivation to do so.

Of course, more data is necessary to confirm these conclusions. We plan to gather the data by running a study in the context of classroom instruction, where it is easier to control at what stage of learning the students use the system. If the study confirms the results presented in this paper, it may be beneficial to add to the SE-Coach the capability to automatically tailor the available levels of scaffolding depending upon the student's familiarity with the examples topic.

5 References

1. Aleven, V., & Ashley, K. D. (1997). Teaching case-based argumentation through a model and examples: Empirical evaluation of an intelligent learning environment. In *Proc. of AIED '97, 8th World Conference of Artificial Intelligence and Education*, Kobe, Japan.
2. Aleven, V., Koedinger, K. R., & Cross, K. (1999). Tutoring answer-explanation fosters learning with understanding. In *Proc. of AIED '99, 9th World Conference of Artificial Intelligence and Education*, Le Mans, France.
3. Bielaczyc, K., Pirolli, P., & Brown, A. L. (1995). Training in self-explanation and self-regulation strategies: Investigating the effects of knowledge acquisition activities on problem-solving. *Cognition and Instruction, 13*(2), 221-252.
4. Chi, M. T. H. (in press). Self-Explaining: The dual process of generating inferences and repairing mental models. *Advances in Instructional Psychology*.
5. Conati, C. (1999). An intelligent computer tutor to guide self-explanation while learning from examples. *Unpublished Ph.D. thesis*, University of Pittsburgh, Pittsburgh.
6. Conati, C., & VanLehn, K. (1999). Teaching meta-cognitive skills: implementation and evaluation of a tutoring system to guide self-explanation while learning from examples. In *Proc. of AIED'99, 9th World Conference of Artificial Intelligence and Education*, Le Mans, France.
7. Gott, S. P., Lesgold, A., & Kane, R. S. (1996). Tutoring for transfer of technical competence. In B. G. Wilson (Ed.), *Constructivist Learning Environments* (pp. 33-48). Englewood Cliffs, NJ: Educational Technology Publications.
8. Nguyen-Xuan, A., Bastide, A., & Nicaud, J.-F. (1999). Learning to solve polynomial factorization problems: by solving problems and by studying examples of problem solving, with an intelligent learning environment. In *Proc. of AIED '99, 9th World Conference of Artificial Intelligence and Education*, Le Mans, France.
9. Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, CA: Morgan-Kaufmann.
10. Renkl, A. (1997). Learning from worked-examples: A study on individual differences. *Cognitive Science, 21*(1), 1-30.
11. VanLehn, K. (1996). Conceptual and meta learning during coached problem solving. In C. Frasson, G. Gauthier, & A. Lesgold (Eds.), *ITS96: Proc. of the 3rd Int. Conference on Intelligent Tutoring Systems, Montreal, Canada*. New York: Springer-Verlag.
12. Weber, G., & Specht, M. (1997). User modelling and adaptive navigation support in WWW-based tutoring systems. In *Proc. of User Modeling '97*.