Tutoring Answer Explanation Fosters Learning with Understanding

Vincent Aleven, Kenneth R. Koedinger, Karen Cross
Human-Computer Interaction Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
{aleven/koedinger/kcross}@cs.cmu.edu

Abstract

In a previous formative evaluation of the PACT Geometry tutor, we found significant learning gains, but also some evidence of shallow learning. We hypothesized that a cognitive tutor may be even more effective if it requires students to provide explanations for their solution steps. We hypothesized that this would help students to learn with deeper understanding, as reflected by an ability to explain answers, transfer to related but different test items, and an ability to deal with unfamiliar questions.

We conducted two empirical experiments to test the effectiveness of learning at the explanation level. In each experiment, we compared two versions of the PACT Geometry Tutor. In the "answer-only" version of the tutor, students were required only to calculate unknown quantities in each geometry problem. In the "reason" version, students also needed to provide correct explanations for their solution steps, by citing geometry theorems and definitions. They did so by selecting from the tutor's Glossary of geometry knowledge, presented on the screen.

The results provide evidence that explaining answers leads to greater understanding. Students who explained their solution steps had better post-test scores and were better at providing reasons for their answers. They also did better at a type of transfer problem in which they were asked to judge if there was sufficient information to find unknown quantities. Also, students who explained solution steps were significantly better at steps where quantities sought were difficult to guess (in other words, require deeper knowledge) while answer-only students did better on easy-to-guess items indicating superficial understanding. Finally, students who explained answers did fewer problems during training, which suggests that the training on reason-giving transfers to answer-giving.

Introduction

Cognitive scientists and educators have long been interested in the question: How can we get our students to learn with understanding and acquire robust knowledge? Just exactly what it means to learn with understanding is somewhat difficult to pinpoint. Transfer to new and unfamiliar types of problems is often seen as a key criterion of understanding [Simon, 1987], as is removing misconceptions [Chi, *et al.,* 1989]. An ability to explain or justify problem-solving steps is also an important indicator.

The reality in many American high schools may be that few students achieve a robust understanding of the more difficult subjects such as algebra, physics, or geometry. They learn well enough to get a passing grade on tests, but have difficulty when dealing with even slightly unfamiliar problems and are unable to explain their answers. Shallow learning occurs in many forms of instruction, in many domains. In physics, for example, there are many common misconceptions, such as confusing mass and weight, or the direction of velocity and acceleration. In geometry, students often rely on the fact that angles look the same in the diagram, even if they cannot be shown to be so by reasoning logically from theorems and definitions. Such heuristics are often successful, but may fail in more complex problems, as is illustrated in Figure 1. Asked to find the unknown quantities in the diagram, a student makes the creative but unwarranted

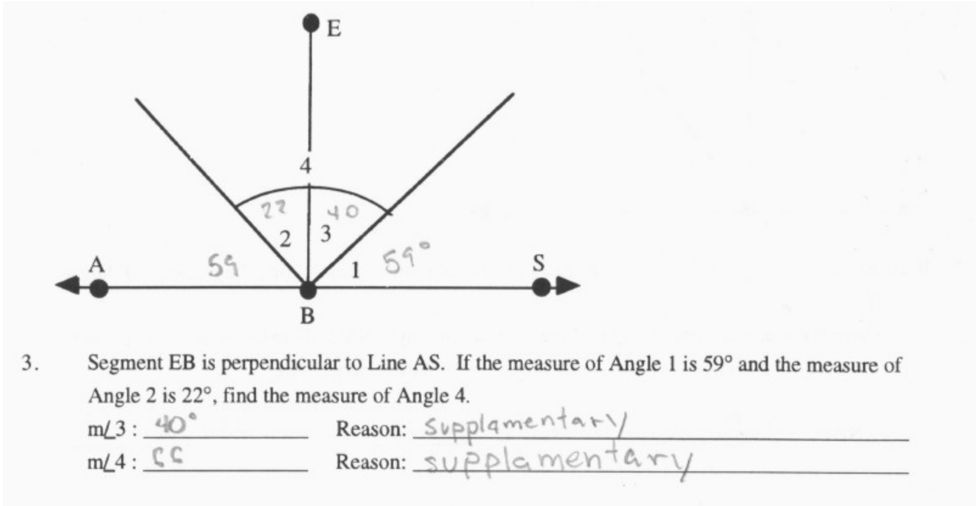


Figure 1: Example of a student's shallow reasoning

inference that the angle on the left has the same measure as angle 1 on the right. These angles look the same, sure enough, but this inference is not justified and leads to a wrong answer. The student is also not able to provide valid reasons for the answers, in terms of geometry theorems. The correct reasons would be complementary angles for m∠3 and angle addition for m∠4.

But how to design instruction that fosters understanding? Some take the point of view that instruction should be anchored in a rich context [CTGV, 1993]. Others have found that self-explanation is an important key to learning with understanding [Chi, *et al*, 1989].

Cognitive tutors have been shown to be effective in raising students' test scores, both in laboratory studies and in actual classroom use [Koedinger, *et al.*, 1997]. In spite of this success, they may not be immune from the shallow learning problem. In this paper, we focus on the question of whether and how a cognitive tutor can lead to better understanding and thus be even more effective. We focus on the PACT Geometry Tutor, a cognitive tutor for teaching geometry problem-solving skills developed by our research group This tutor is currently in use in four schools in the Pittsburgh area as an adjunct to a full-year course of high school (plane) geometry. During a previous formative evaluation, we found significant learning gains attributable to the tutor, combined with classroom instruction. But we also found some evidence of shallow learning [Aleven, *et al.*, 1998]. On the post-test, students were better at finding answers (i.e., calculating unknown quantities in geometry problems) than they were at justifying their answers in terms of geometry theorems and definitions. Although an inability to explain is not necessarily indicative of shallow learning, it often is, as illustrated perhaps in the example shown above.

We hypothesized that students would learn with greater understanding if we modified the tutor so that students had to explain their reasoning steps during problem solving. We hypothesized that this would not only help students to learn better to explain reasoning steps, but also that they would learn better to solve problems, and that there would be transfer to extra-challenging problems in which they are asked to judge whether enough information is given to make certain inferences (e.g., to calculate certain unknowns in a geometry diagram).

If students have to explain their answers, this may force them to apply the requisite knowledge more consciously, and to rely less on superficial strategies. Teaching students to explain their answers may have a similar effect as self-explanation, which leads to better learning [Chi, *et al.*, 1989; 1994]. Explaining steps while solving problems (with a computer tutor) is not unlike explaining worked-out examples, except for the fact that students explain their own reasoning steps, perhaps checked by the tutor, not those provided in the examples, and that they can receive feedback from the tutor on whether their explanations are correct. (In the self-explanation studies mentioned above, students typically do not receive such feedback.)

In fact, an argument can be made that the combination of problem solving and explanation with feedback may be more effective than unguided self-explanation. One study reports that merely prompting students to provide self-explanations helps improve understanding [Chi, *et al.*, 1994]. A computer tutor can prompt students very consistently. Also, the Chi *et al.* studies report that self-explanation requires self-regulatory skills and reports rather large individual differences with respect to students' ability to self-explain. Our target population may not contain a large percentage of good self-explainers. It may well be that feedback on reasons is needed for students without advanced self-explanation skills. Although there is evidence that self-explanation skills can be taught [Bielaczyc *et al.*, 1995], it remains difficult to achieve in classrooms. (One project attempting to design a computer tutor for self-explanation is [Conati, *et al.*, 1997].) On the other hand, having to enter explanations into a computer interface may provide an obstacle. Also, selecting reasons from a Glossary on the computer screen may not guarantee or require the amount of deep processing on the part of the students as is ideal.

In this paper, we present the PACT Geometry tutor, redesigned so that it requires students to explain their solutions. We also present results from two empirical experiments that we undertook to test our hypothesis that tutoring at the explanation level improves students' understanding.

A Tutor that Teaches at the Explanation Level

The PACT Geometry Tutor, is one of the cognitive tutors developed by our research group. This tutor was designed to be an integrated part of a new high-school geometry course, developed in tandem. The tutor curriculum consists of five lessons covering the topics of Area, Pythagorean Theorem, Angles, Similar Triangles, and Circles. A lesson on Quadrilaterals is under development. Following new guidelines from the National Council of Mathematics Teachers [NCTM, 1989], the curriculum emphasizes the use of geometry as a problem solving tool (unlike the "old" Geometry tutors [Koedinger & Anderson, 1993], which focused on proof skills). The problems in the tutor curriculum therefore often involve a real-world problem situation that calls for geometric reasoning. For example, in the problem shown in Figure 2, geometric reasoning is the tool of choice to help an archeological research team with a vexing identification problem.

The tutor was designed to get students to explain their answers and to reason explicitly with the rules and definitions of the domain. In order to complete a problem, students must calculate the unknown quantities, and must state a reason for each solution step, citing a geometry definition or rule that was used. (In other words, they must complete the answer sheet shown on the left in Figure 2. The quantity labels are provided by the tutor.) When students enter values or reasons in the answer sheet, the tutor tells them if they are right or not. At any point during their work with the tutor, students can consult a Glossary, which lists relevant definitions and theorems (Figure 2). When students click on a Glossary item, they get to see a short description of the rule and an example, illustrated with a diagram. They can enter a reason into the answer sheet by selecting the relevant rule in the Glossary, then clicking on the "Select" button. Students who are stuck on a problem can search the Glossary for a rule that applies. In addition, the tutor provides hints on request. The tutor's hints encourage students to search the Glossary for applicable rules. As part of its hinting strategy, the tutor may highlight a small set of rules in the Glossary, in order to narrow down the Glossary search. The tutor provides two additional tools, not shown in Figure 2. The DiagramTool displays an abstract diagram representing the problem and lets students record information in the diagram. The second tool, the Equation Solver, lets students enter an algebraic equation that relates the quantities in the problem and helps in solving it step by step.

The PACT Geometry tutor is based on the technology of cognitive tutors, which draws heavily on the ACT-R theory of cognition and learning [Anderson, 1993]. A crucial component of each tutor is its cognitive model, which captures the skills of an ideal student, expressed as a set of production rules. The tutor uses the model to assess the student's solution steps and to provide feedback and hints (through model tracing). Also, the cognitive model is the basis for student modeling. The tutor maintains estimates of the probability that the student masters each skill in the model (knowledge tracing). The tutor uses this student model to assign remedial problems, targeting skills for which the student has not yet reached mastery, and to decide when a student may move on to the next section or lesson (namely, upon mastery of all skills).

First Experiment to Evaluate Teaching at the Explanation Level

We conducted two experiments to test whether students learn more effectively when they explain their solutions during training. We focused on the third lesson of the tutor curriculum, which

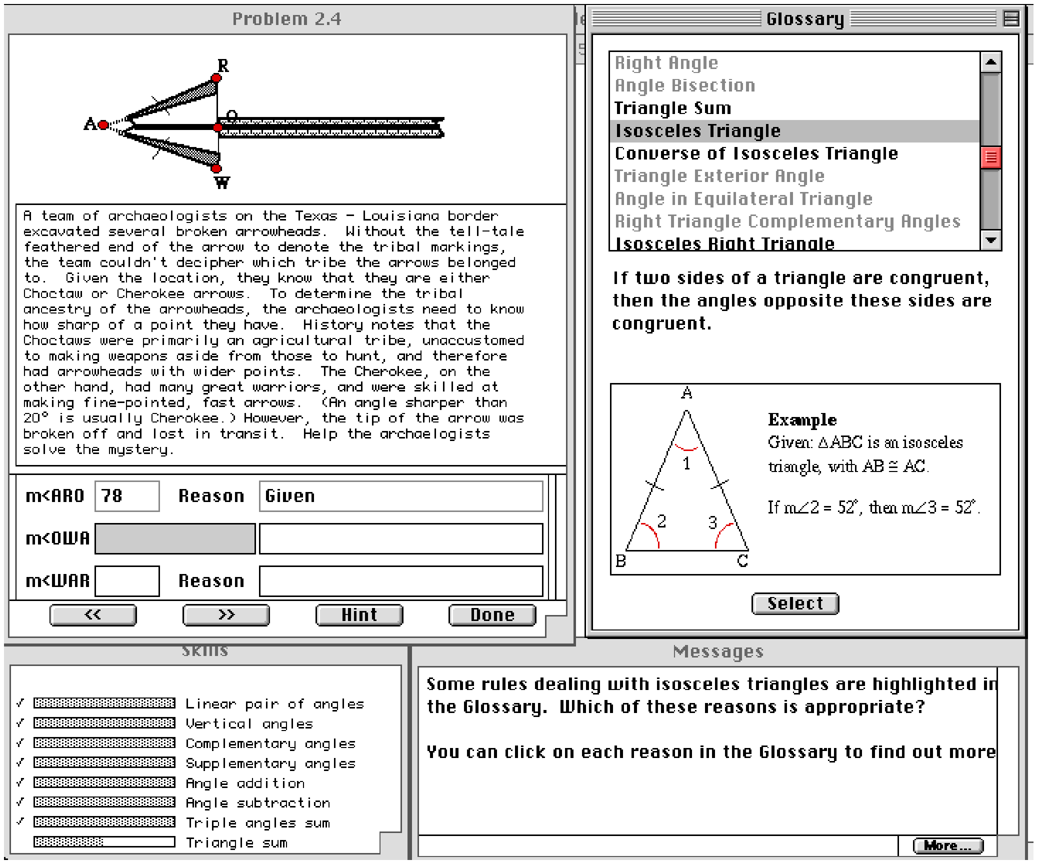


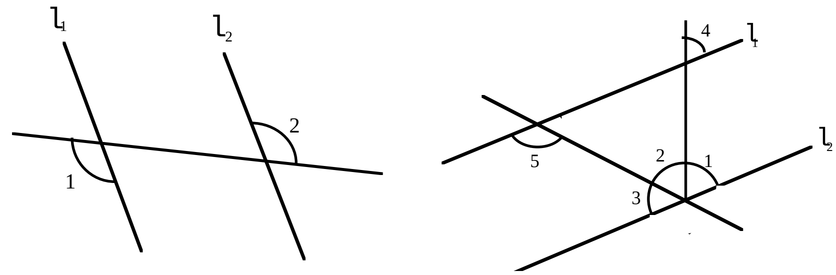
Figure 2: The PACT Geometry Reason Tutor

deals with geometric properties of angles. Both experiments took place in a suburban high school near Pittsburgh, which uses the PACT Geometry tutor as part of its regular geometry instruction. In both experiments, we compared learning with the PACT Geometry Reason Tutor, described above, against a control condition that involved an "answer only" version of the same tutor. The answer-only condition was the same as the "reason version", except that students were not required to state reasons for their answers. There was no column for entering reasons in the tutor's answer sheet, as there is in the reason condition. The main difference between the two experiments was the way in which we tried to control the time that students spent with the tutor. This is discussed further below.

The first experiment involved 41 students taking the geometry course (two periods). The students were assigned to the reason condition or the answer-only condition on the basis of their scores (on quizzes, tests, and homework assignments) in the course prior to the experiment, so as to create groups that were balanced in terms of prior ability. 23 of the 41 students completed both the tutor curriculum and the post-test. The results reported here pertain to those 23 students.

All students took a pre-test before they started working on the tutor's Angles lesson and completed a post-test afterwards. Students started and finished the tutor at different times. In order to complete the tutor, the students had to reach mastery level for all skills targeted in each of the three sections. (That is, the tutor assigned problems until mastery level was reached for all skills.) The students also received classroom instruction on the topics covered in the Angles lesson. Some of that took place before the pre-test, some of it in between pre-test and post-test.

The pre-test and post-test involved 6 problems, designed to assess students' ability to solve geometry problems and to explain their answers. In these problems, students were asked to compute unknown quantities in a diagram and for each had to provide a reason why their answer was correct, in terms of geometric theorems. (Subsequently we refer to these items as "Answer" and "Reason" items. They are illustrated in Figure 1.) Students were provided with a sheet listing acceptable reasons and definitions and were told that they could freely reference it.



If the measure of Angle 2 is 101°, *do you have enough information* to find the measure of Angle 1?

Line l1 is parallel to line l2. If the measure of Angle 1 is 65°, ... *do you have enough information* to find the measures of the other angles?

Figure 3: Problems designed to test students' understanding

In order to investigate issues of deep learning, we also included extra-challenging problems, designed to measure students' understanding, illustrated in Figure 3. In these problems, students were asked to judge whether there was enough information to compute certain quantities, and if so, to compute them and state a reason. In some problems, angle measures could not be calculated because geometric constraints were missing, such as the constraint that lines were parallel (as in the item on the left in Figure 3). In other problems, the values of needed premises were not given (right item, Figure 3). Items with missing information are subsequently referred to as "Not Enough Info" items. On these items, superficial visual reasoning such as that illustrated in Figure 1 is likely to lead a student astray. We predicted that students in the reason condition would perform better on the Not Enough Info items. The tutor curriculum does not include problems of this type. Thus, these problems provide a measure of how well skills learned with the tutor transfer to unfamiliar but related problems.

Reasons were graded as correct if they described an accurate geometric relationship between the answer and other angles in the diagram, from which the measure could be determined. We had 6 test forms, assigned randomly to students at pre-test and post-test in order to counterbalance for test difficulty.

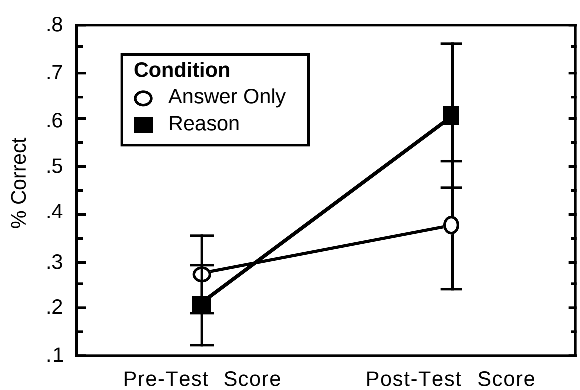


Figure 4: Test Scores in Experiment 1

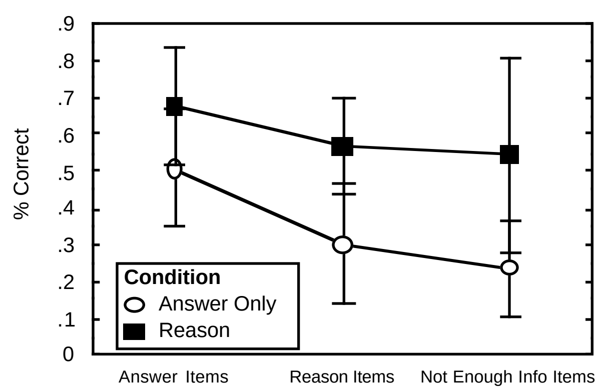


Figure 5: Post-Test Scores in Experiment 1

We found that students' test scores (shown in Figure 4) increased from pre-test to post-test ($F(1,21) = 27.2$, $p<.0001$). Further, there was a significant interaction between condition (answer-only vs. reason) and test time (pre vs. post) ($F(1,21) = 9.19$, $p<.01$), indicating that the students in the reason condition improved significantly more than their counterparts in the answer-only condition. Students in the reason condition also performed significantly better on the post-tests than students in the answer-only condition ($F(1,21) = 6.25$, $p<.05$).

Students in the reason condition did better on each of the different types of test items (Answer, Reason, and Not Enough Information) than students in the answer condition. These results are presented in Figure 5. The analysis shows significant effects of item type ($F(2,21)=8.4$, $p<.001$) and of condition with test item as a covariant ($F(1,21)=7.3$, $p<.01$), but no significant interaction ($F(1,21)=0.97$, $p>.35$).

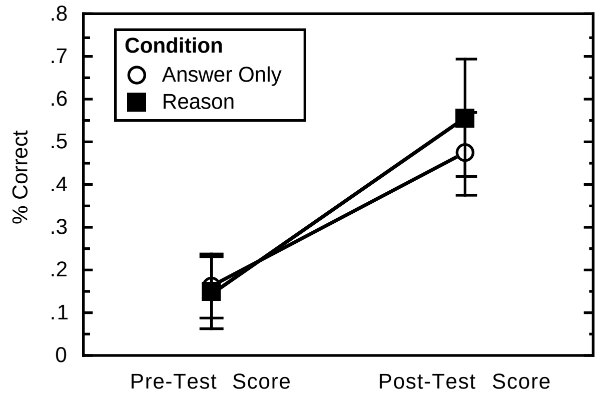


Figure 6: Test scores in Experiment 2

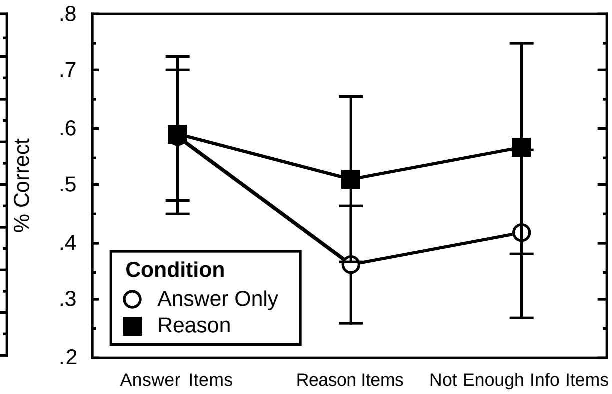


Figure 7: Post-Test scores in Experiment 2

The students in the reason condition spent about 14% more time working on the tutor than did the students in the answer-only condition (436 ± 169 minutes for the reason condition, 383 ± 158 for the answer-only condition). The difference is not statistically significant ($F(1,21) = .61$, $p = .45$). The main reason for this difference is that students in the reason condition have more work to do per problem, since they have to provide reasons for their solution steps. There was some evidence that students in the reason condition got more out of each problem: They needed fewer problems to reach the tutor's mastery level criterion: 102 v. 135 problems on the average, but the difference was not statistically significant $F(1,21) = 2.74$, $p = .11$).

At the outset both conditions were balanced in terms of their prior grades in the course. But of the students (23 in total) who completed the experiment, the reason group students had slightly better prior scores (87.3 vs. 82.3), although that difference is not statistically significant ($F(1,21) = 2.19$, $p = .15$). The difference may be due to the fact that the reason condition involved slightly more work, so that the better students were more likely to finish before the semester was over.

In conclusion, the results suggest that there are considerable advantages to teaching students to explain answers. It seems to lead to better performance in providing these reasons. Also, it transfers to better overall performance in both providing answers and making judgments about whether there is enough information to give an answer. On the other hand, we cannot rule out an alternative interpretation, namely, that the students in the reason condition performed better because they spent more time on the tutor or were a better than average sample. While it seems unlikely that this would account for all post-test differences between the two conditions, concern about these issues motivated us to do a second experiment.

Second Experiment

The second experiment had the same goal as the first: to test whether teaching students to give reasons improves their understanding. As in the previous experiment, we compared a reason condition and an answer-only condition. To make sure that both groups spent the same amount of time on the tutor, we changed the criterion for finishing the tutor: Instead of a mastery level criterion, we used a time limit of 7 hours. This was the average amount of time spent by students in the reason condition in the first experiment. The experiment involved 53 students in two periods of a geometry course. 43 of these students completed the tutor and the post-test. We excluded from the analysis the data from two students: One who left 75% of the post-test blank, one who spent less than 7 hours on the tutor. The teacher may have advanced this student early.

Both groups spent an equal amount of time on the tutor (reason group: 513±74 minutes, answer only group: 501±62). (These numbers are higher than the tutor's time limit of 7 hours because they include students' idle time, whereas the tutor factored out idle time in keeping track of the amount of time spent by each student.) The reason group did significantly fewer problems than the answer-only group: 76±14 vs. 111±23 ($F(1,39) = 32.83$, $p < .0001$).

As in the first experiment, we found significant performance gains across all students ($F(1, 39) = 19.56$, $p < .0001$). (See Figure 6.) The pre-test scores of both groups were about equal: 15% correct in the reason condition, 16% correct in the answer condition. Students in the reason

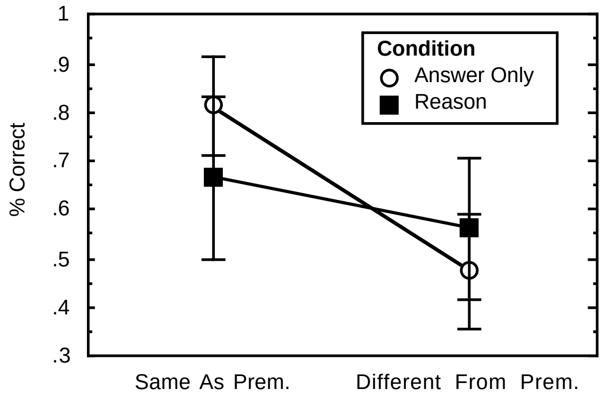


Figure 8: Post-Test scores in Experiment 2, for items where the unknown quantity was the same as a premise and for items where the unknown was different from all premises

condition did better on the post-test than students in the answer-only condition (56% vs. 48%) and had better gain scores (49% vs. 37%)[1]. Looking at the different item types (Figure 7), there is no difference between the conditions on the answer items (59% vs. 59%). In contrast, on the items testing deeper learning, the reason condition does substantially better, both on the reason items (51% vs. 36%) and on the not enough info items (57% vs. 42%). Comparing the performance of the conditions on the answer items vs. these deep items (which we averaged together), we found a statistically significant interaction of condition and item type ($F(1,39) = 5.4$, $p < .05$). As predicted reason students are doing significantly better on problems requiring deeper reasoning.

In order to further investigate issues of deep learning, we divided the post-test items into easier-to-guess items and more-difficult-to-guess items. A quantity sought is easy to guess when it is equal to a quantity on which it depends ("Same As Prem."), and that it is difficult to guess when it is different from all premises ("Different From Prem."). This follows from our observation that students often use a shallow rule "if it looks the same, it is the same" as illustrated in Figure 1. Such a rule is successful on "Same As Prem." items despite the lack of understanding it reflects. As shown in Figure 8, the reason condition students performed better on the more demanding "Different From Prem." items while the answer-only students performed better on the "Same As Prem." items ($F(1, 39) = 6.10$, $p<.001$). This suggests that the students in the reason condition had more robust knowledge and relied less on guessing.

An analysis of students' errors reveals that students in the answer-only condition make more errors of commission (as a percentage of their total number of errors), whereas students in the reason condition have a larger proportion of errors of omission. In other words, students in the answer-only condition are more likely to guess, whereas students in the reason condition are more likely to either know they know the answer or else leave it blank. This is true for "Same as Prem." items (reason group: 54% commission errors; answer group: 78% commission errors) and "Different from Prem." items (69% v. 85%). This confirms that students in the reason condition may have learned better to avoid shallow heuristics.

Discussion and conclusion

We conducted two experiments to evaluate if a cognitive tutor can be improved if it requires students to provide reasons for their problem-solving steps. We hypothesized that this would lead to better understanding on the part of students. Specifically, that it would help students to learn to provide better reasons and also that there would be transfer to finding solutions and to extra challenging problems where superficial strategies are likely to fail. We compared two versions of the PACT Geometry Tutor: an "answer-only" version and a "reason" version.

In both experiments, we found significant learning gains attributable to the tutor, combined with classroom instruction. We found strong evidence for our hypothesis stated above. In the first experiment, the students who had explained their problem-solving steps during training did better on all relevant post-test measures than students who had not explained steps. They were better at finding answers, giving reasons, and handling not enough information items. However, we cannot attribute these differences entirely to the training in reason giving, because we also found that the reason group had a slight advantage in terms of time on tutor and prior ability.

[1] The gain score is defined as (post-test score – pre-test score) / (1 – pre-test score).

In the second experiment, the two groups spent equal time on the tutor. We again found that teaching students to give reasons improves their understanding. The students who explained their answers during training had better post-test and gain scores than the students who did not. This better performance was detected in items assessing deeper understanding. Students in the reason group were better both in justifying their answers with reasons and on transfer problems designed to detect shallow reasoning. As a further indication of deeper learning in the reason condition, we found that reason students were better at harder-to-guess items, whereas the answer-only students were better at easier-to-guess items. In other words, students using the reason tutor learned with more understanding and learned better to avoid shallow reasoning.

Because the two groups spent equal time on the tutor, the results of the second experiment cannot be attributed to time on task. However, our results provide further support that time on task is a critical variable in instruction. This was indicated by the improvement in the answer condition performance from the first experiment to the second (from 38% to 48%). As a methodological point, experiments that do not control for time on task can be difficult to interpret. This issue creates a serious dilemma for any "less is more" instructional design where an effort is made to achieve deeper learning through fewer, but richer problems or activities. Our second experimental design provides a good model for addressing this dilemma.

In both experiments, we found strong evidence that explaining reasoning steps transfers to improved problem-solving skills (i.e., finding answers). At the post-test, students in the reason condition were as good as, or even better than, students in the answer condition at finding answers, in spite of the fact that they had significantly less practice on answer-giving during training (because they solved fewer problems). Moreover, the fact that the reason group students appeared to be better at explaining answers is important in its own right. Justification of reasoning steps is a cornerstone of mathematics. Curricular guidelines from the National Council of Mathematics Teachers emphasize communication of results and reasoning as an important teaching objective.

As a practical matter, our results suggest that having students explain their answers is a viable way of improving a cognitive tutor. Improving students' post-test scores and understanding is no easy matter, but we believe that experiments like these are the only scientifically reliable way to reach this important goal.

Acknowledgments

Chang-Hsin Chang, Colleen Sinclair, and Jaclyn Snyder contributed to this research. The research is sponsored by the Buhl Foundation, the Grable Foundation, the Howard Heinz Endowment, the Richard King Mellon Foundation, and the Pittsburgh Foundation. We gratefully acknowledge their contributions.

References

Aleven, V., K. R. Koedinger, H. C. Sinclair, and J. Snyder, 1998. Combatting Shallow Learning in a Tutor for Geometry Problem Solving. In *Proceedings ITS '98,* edited by B. P. Goettl, H. M. Halff, C. L. Redfield, and V. J. Shute, 364-373. Berlin: Springer.

Bielaczyc, K., P. L. Pirolli, and A. L. Brown, 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.

CTGV, 1993. (Cognition and Technology Group at Vanderbilt.) Anchored instruction and situated cognition revisited. *Educational Technology,* 33 (3), 52-70.

Conati, C., J. Larkin, And K. VanLehn, 1997. A Computer Framework to Support Self-Explanation. In *Proceedings of the AI-ED 97 World Conference,* edited by B. du Boulay and R. Mizoguchi, 279-286. Amsterdam: IOS Press.

Chi, M. T. H., M. Bassok, M. W. Lewis, P. Reimann, and R. Glaser, 1989. Self-Explanations: How Students Study and Use Examples in Learning to Solve Problems. *Cognitive Science,* 13, 145-182.

Chi, M. T. H., N. de Leeuw, M. Chiu, and C. Lavancher, 1994. Eliciting Self-Explanations Improves Understanding. *Cognitive Science,* 18, 439-477.

Koedinger, K. R., and J. R. Anderson, 1993. Reifying implicit planning in geometry. In S. Lajoie and S. Derry (eds.), *Computers as Cognitive Tools,* 15-45. Hillsdale, NJ: Erlbaum.

Koedinger, K. R., Anderson, J.R., Hadley, W.H., & Mark, M. A. (1997). Intelligent tutoring goes to school in the big city. International Journal of Artificial Intelligence in Education, 8, 30-43.

NCTM, 1989. *Curriculum and Evaluation Standards for School Mathematics*. National Council of Teachers of Mathematics. Reston, VA: The Council.

Simon, H. A., 1987. The Information-Processing Explanation of Gestalt Phenomena. In: *Models of Thought, Volume 2,* 481-493. New Haven, CT: Yale University Press.