# Evaluating the Effectiveness of a Tutorial Dialogue System for Self-Explanation

Vincent Aleven, Amy Ogan, Octav Popescu, Cristen Torrey, Kenneth Koedinger

Human Computer Interaction Institute, Carnegie Mellon University
5000 Forbes Ave, Pittsburgh, PA 15213, USA
+1 412 268 5475
aleven@cs.cmu.edu, {octav,koedinger}@cmu.edu
{aeo,ctorrey}@andrew.cmu.edu,

**Abstract.** Previous research has shown that self-explanation can be supported effectively in an intelligent tutoring system by simple means such as menus. We now focus on the hypothesis that natural language dialogue is an even more effective way to support self-explanation. We have developed the Geometry Explanation Tutor, which helps students to state explanations of their problem-solving steps in their own words. In a classroom study involving 71 advanced students, we found that students who explained problem-solving steps in a dialogue with the tutor did not learn better overall than students who explained by means of a menu, but did learn better to state explanations. Second, examining a subset of 700 student explanations, students who received higher-quality feedback from the system made greater progress in their dialogues and learned more, providing some measure of confidence that progress is a useful intermediate variable to guide further system development. Finally, students who tended to reference specific problem elements in their explanations, rather than state a general problem-solving principle, had lower learning gains than other students. Such explanations may be indicative of an earlier developmental level.

## 1 Introduction

A self-explanation strategy of learning has been shown to improve student learning [1]. It has been employed successfully in intelligent tutoring systems [2, 3]. One approach to supporting self-explanation in such a system is to have students provide explanations by means of menus or templates. Although simple, that approach has been shown to improve students' learning [2]. It is likely, however, that students learn even better when they explain their steps in their own words, aided by a system capable of providing feedback on their explanations. When students explain in their own words, they are likely to pay more attention to the crucial features of the problem, causing them to learn knowledge at the right level of generality. They are also more likely to reveal what they know and what they do not know, making it easier for the system to provide detailed, targeted feedback. On the other hand, in comparison to

J.C. Lester et al. (Eds.): ITS 2004, LNCS 3220, pp. 443–454, 2004.

explaining by means of a menu or by filling out templates, free-form explanations require more time and effort by the students. Free-form explanations require that students formulate grammatical responses and type them in. Further, templates or menus may provide extra scaffolding that is helpful for novices but missing in a natural language dialogue. Indeed, menus have been shown to be surprisingly effective in supporting explanation tasks [2, 4, 5], although it is not clear whether menus help in getting students to learn to *generate* better explanations. Whether on balance the advantages of dialogue pay off in terms of improved learning is thus an empirical question.

To answer that question, we have developed a tutorial dialogue system, the Geometry Explanation Tutor, that engages students in a natural language dialogue to help them state good explanations [6, 7]. Tutorial dialogue systems have recently come to the forefront in AI and Education research [6, 8, 9, 10, 11]. The Geometry Explanation Tutor appears to be unique among tutorial dialogue systems in that it focuses on having students explain and provides detailed (but undirected) feedback on students' explanations. A number of dialogue systems have been evaluated with real students, some in real classrooms (e.g., [12]). Some success has been achieved, but it is fair to say that tutorial dialogue systems have not yet been shown to be definitively better than the more challenging alternatives to which they have been compared.

The current paper reports on the results of a classroom study of the Geometry Explanation Tutor, which involved advanced students in a suburban junior high school. As reported previously, there was little difference in the learning outcomes of students who explained in their own words and those who explained by means of a menu [12], as measured by a test that involved problem-solving items, explanation items, and various transfer items. Yet even if there were no significant differences in students' overall learning gains, it is still possible that students who explained in a dialogue with the system may have acquired better geometry communication skills. Further, the result does not explain why there was no overall difference between the conditions, how well the system's natural language and dialogue components functioned, and whether one might reasonably expect that improvements in these components would lead to better learning on the part of the students. Finally, the result does not illuminate how different students may have employed different strategies to construct explanations in a dialogue with the system and how those strategies might correlate to their learning outcomes. Answers to those questions will help to obtain a better understanding of the factors that determine the effectiveness of a tutorial dialogue system that supports self-explanation. We address each in turn.

## 2 The Geometry Explanation Tutor

The Geometry Explanation Tutor was developed by adding dialogue capabilities to the Geometry Cognitive Tutor, which is part of a geometry curriculum currently being taught in approximately 350 high schools across the country. The combination of tutor and curriculum has been shown to improve on traditional classroom instruction

| | |
|---|---|
| **Student:** | the two angles add up to 180 degrees |
| **Tutor:** | Yes, the sum of the angle measures is 180. But you have to be more specific. In general, the sum of what kind of angles is 180? |
| **Student:** | The two angles add up to 180 degrees so the angles are supplimentary |
| **Tutor:** | [accepts explanation by means of implicit feedback] |

**Fig. 1.** A student dialog with the tutor, attempting to explain the Separate Supplementary Angles rule

[2]. The Geometry Cognitive Tutor focuses on geometry problem solving: students are presented with a diagram and a set of known angle measures and are asked to find certain unknown angles measures. Students are also required to explain their steps. We are investigating the effect of two different ways of supporting self-explanation: In the menu-based version of the system, students explain each step by typing in, or selecting from an on-line Glossary, the name of a geometry definition or theorem that justifies the step. By contrast, in the dialogue-based version of the system (i.e., the Geometry Explanation Tutor), students explain their quantitative answers in their own words. The system engages them in a dialogue designed to improve their explanations. It incorporates a knowledge-based natural language understanding unit that interprets students' explanations [7]. To provide feedback on student explanations, the system first parses the explanation to create a semantic representation [13]. Next, it classifies the representation according to a hierarchy of approximately 200 explanation categories that represent partial or incorrect statements of geometry rules that occur commonly as novices try to state explanation rules. After the tutor classifies the response, its dialogue management system determines what feedback to present to the student, based on the classification of the explanation. The feedback given by the tutor is detailed yet undirected, without giving away too much information. The student may be asked a question to elicit a more accurate explanation, but the tutor will not actually provide the correction. There are also facilities for addressing errors of commission that suggest that the student remove an unnecessary part of an explanation.

An example of a student-tutor interaction is shown in Fig. 1. The student is focusing on the correct rule, but does not provide a complete explanation on the first attempt. The tutor feedback helps the student in fixing his explanation.

## 3 Effect of Dialogue on Learning Results

A classroom study was performed with a control group of 39 students using the menu-based version of the tutor, and an experimental group of 32 students using the dialogue version (for more details, see [12]). The results reported here focus on 46 students in three class sections, 25 in the Menu condition and 21 in Dialogue condition, who had spent at least 80 minutes on the tutor and were present for the pre-test and post-test. All student-tutor interactions were recorded for further evaluation. The

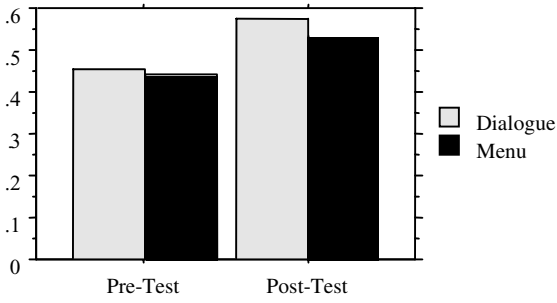


**Fig. 2.** Overall Pre/Post Test Score (proportion correct)

students completed a pre-test to measure prior knowledge and a post-test after the study.

A 2x2 repeated-measures ANOVA on the test scores, with test time (Pre/Post) as an independent factor (as in Fig. 2), revealed no significant difference between the conditions ($F(1,44) = .578$, $p > .4$), consistent with the result reported in [12]. The high pre-test scores (Dialogue .47, Menu .43) may explain in part why no significant differences were found in learning gains. However, a significant difference emerged when focusing on the Explanation items, that is, items that ask for an explanation of a geometry rule used to find the angle measure in the previous step (see Fig. 3). These items were graded with a scheme of .33 points for giving the correct name of the rule to justify their answer, .67 points for attempting to provide a statement of the correct rule but falling short of a complete and correct statement, and a full point for a complete statement of the correct rule[1]. A repeated-measures ANOVA revealed a significant difference in learning gains between the conditions. Even with an initial advantage in Explanation score for the students in the Dialogue condition ($F(1,44) = 4.7$, $p < .05$), they had significantly greater learning gains on Explanation items compared to the Menu condition ($F(1,44) = 18.8$, $p < .001$). It may appear that the grading scheme used for Explanation items favors students in the Dialogue condition, since only complete and correct rule statements were given full credit and only students in the Dialogue condition were required to provide such explanations in their work with the tutor. However, even with a scheme that awards full credit for any attempt at explaining that references the right rule, regardless of whether it is a complete statement, there is no significant advantage for the Menu group. No significant difference was found between the two conditions on the other item types.

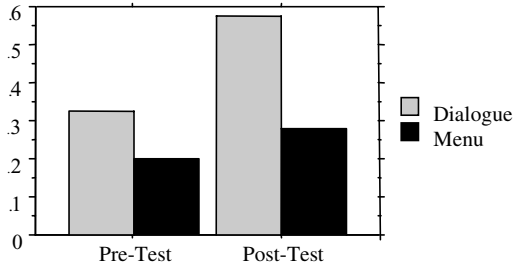


**Fig. 3.** Score on explanation items (proportion correct).

[1] In a previously-published analysis of these data [3], a slightly different grading scheme was used for Explanation items: half credit was given both for providing the name of a correct rule name and for providing an incomplete statement of a rule. The current scheme better reflects both standards of math communication and the effort required to provide an explanation.

A closer look at the Explanation items shows distinct differences in the type and quality of explanations given by students in each condition (see Fig. 4). In spite of written directions on the test to give full statements of geometry rules, students in the Menu condition only attempted to give a statement of a rule 29% of the time, as opposed for example to merely providing the name of a rule or not providing any explanation. The Dialogue condition, however, gave a rule statement in 75% of their Explanation items. When either group did attempt to explain a rule, the Dialogue condition focused on the correct rule more than twice as often as the Menu group (Dialogue .51 ± .27, Menu .21 ±.24; $F(1,44) = 16.2$, $p < .001$), and gave a complete and correct statement of that rule almost seven times as often (Dialogue .44 ± .27 Menu .06 ± .14; $F(1,44) = 37.1$, $p < .001$). A selection effect in which poorer students follow instructions better cannot be ruled out but seems unlikely. The results show no difference for correctness in answering with rule names (Dialogue .58, Menu .61), but the number of explanations classified as rule names for the Dialogue group (a total of 12) is too small for this result to be meaningful.

To summarize, in a student population with high prior knowledge, we found that students who explained in a dialogue learned better to state high-quality explanations than students who explained by means of a menu, at no expense for overall learning. Apparently, for students with high prior knowledge, the explanation format affects communication skills more than that it affects students' problem-solving skill or understanding, as evidenced by the fact that there was no reliable difference on problem-solving or transfer items.

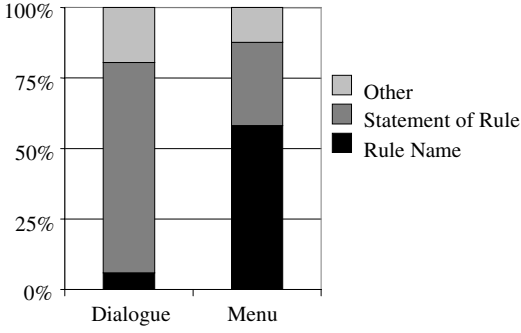


**Fig. 4.** Relative frequency of different explanation types at the post-test

## 4 Performance and Learning Outcomes

In order to better understand how the quality of the dialogues may have influenced the learning results, and where the best opportunities for improving the system might be, we analyzed student-tutor dialogues collected during the study. A secondary goal of the analysis was to identify a measure of dialogue quality that correlates well with learning so that it could be used to guide further development efforts.

The analysis focused on testing a series of hypothesized relations between the system's performance, the quality of the student/system dialogues, and ultimately the students' learning outcomes. First, it is hypothesized that students who tend to make progress at each step of their dialogues with the system, with each attempt closer to a complete and correct explanation than the previous, will have better learning results than students who do not. Concisely, *greater progress → deeper learning.* Second,

we hypothesize that students who receive better feedback from the tutor will make greater progress in their dialogues with the system, or *better feedback → greater progress → deeper learning*. Finally, before this feedback is given, the system's natural language understanding (NLU) unit must provide an accurate classification of the student's explanation. With a good classification, the tutor is likely to provide better, more helpful feedback to the student. The complete model we explore is whether *better NLU → better feedback → greater progress → deeper learning*.

To test the hypothesized relations in this model, several measures were calculated from a randomly-selected subset of 700 explanations (each a single student explanation attempt-tutor feedback pair) out of 3013 total explanations. Three students who did not have at least 10% of their total number of explanations included in the 700 were removed because the explanations included might not represent an accurate picture of their performance.

First, the quality of the system's performance in classifying student explanations was measured as the extent to which two human raters agreed with the classification provided by the NLU. Each rater classified the 700 explanations by hand with respect to the system's explanation hierarchy and then their classifications were compared to each other and to the system's classification. Since each explanation could be assigned a set of labels, a partial credit system was developed to measure the similarity between sets of labels. A formula to compute the distance between the categories within the explanation hierarchy was used to establish a weighted measure of agreement between the humans and the NLU. The closer the categories in the hierarchy, the higher the agreement was rated (for more details, see [7]). The agreement between the two human raters was 94% with a weighted kappa measurement [14] of .92. The average agreement between the humans and the NLU was 87% with a weighted kappa of .81.

Second, the feedback given by the tutor was graded independently by two human raters. On a one-to-five scale, the quality of feedback was evaluated with respect to the student's response and the correct geometry rule. Feedback to partial explanations was placed on the scale based on its appropriateness in assisting the student with correcting his explanation, with 1 being totally unhelpful and 5 being entirely apropos. Explanations that were complete yet were not accepted by the tutor, as well as explanations that were not correct yet were accepted as such, were given a rating of one. Responses where the tutor correctly acknowledged a complete and correct explanation were given a five. The two raters had a weighted agreement kappa of .75, with 89% agreement.

Finally, the progress made by the student within a dialogue was assessed. Each of the 700 explanations was paired with its subsequent student explanation attempt in the dialogue and two human raters independently evaluated whether the second explanation in each pair represented progress towards the correct explanation, compared to the first. The raters were blind with respect to the tutor's feedback that occurred in between the two explanations. (That is, the feedback was not shown and thus could not have influenced the ratings.) Responses were designated "Progress" if the student advanced in the right direction (i.e., improved the explanation). "Progress & Regression" applied if the student made progress, but also removed a crucial aspect of the

geometry rule or added something incorrect. If the explanation remained identical in meaning, it was designated "Same". The final category was "Regression," which meant that the second explanation was worse than the first. The two raters had an agreement of 94% in their assessment, with a kappa of .55.

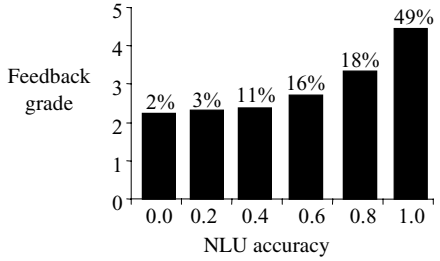


**Fig. 5.** Average feedback grade as a function of NLU accuracy. The percentages shown above the bars indicate the frequency of the accuracy scores.

Having established the three measures, we tested whether there was (correlational) evidence for the steps in the model. First, we looked at the relation *better NLU → better feedback.* A chi-square test shows that the correlation is significant ($\chi^2(12) = 262.8$, $p < .0001$). Fig. 5 refers to the average feedback grade for a particular range of agreement with the NLU. In the figure, frequency of each accuracy score is listed above the column. A higher NLU rating is indicative of a higher feedback grade.

**Table 1.** Progress toward a correct and complete explanation as a function of the quality of the system feedback received

| Feedback Grade | Regres-sion | Same | Progress & Regression | Progress | N |
|---|---|---|---|---|---|
| 1 | 20.5% | 67.1% | 2.3% | 10.2% | 88 |
| 2 | 16.0% | 41.3% | 9.3% | 33.3% | 75 |
| 3 | 19.1% | 29.2% | 9.0% | 42.7% | 89 |
| 4 | 13.5% | 18.3% | 19.2% | 49.0% | 104 |
| 5 | 3.9% | 26.9% | 7.7% | 61.5% | 78 |

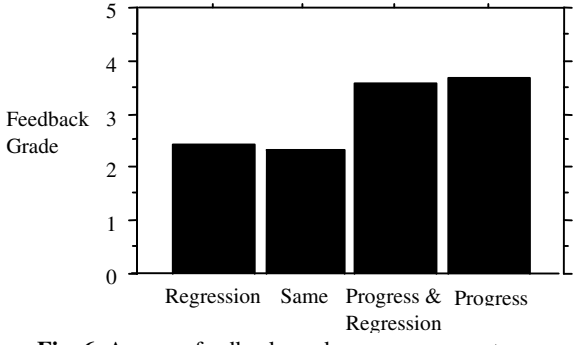


**Fig. 6.** Average feedback grade per progress category

We tested the relation *better feedback → greater progress* by looking at the relative frequency of the progress categories following feedback of any given grade (1 through 5). As shown in Table 1, the higher the feedback rating, the more likely the student is to make progress (i.e., provide an improved explanation). The lower the feedback grade, the more likely it is that the student regresses. A chi-square test shows that the correlation is significant ($\chi^2(12) = 92.7$, $p < .0001$). Fig. 6 is a condensed view that shows the average feed-

back grade for each category, again illustrating that better feedback was followed by greater progress.

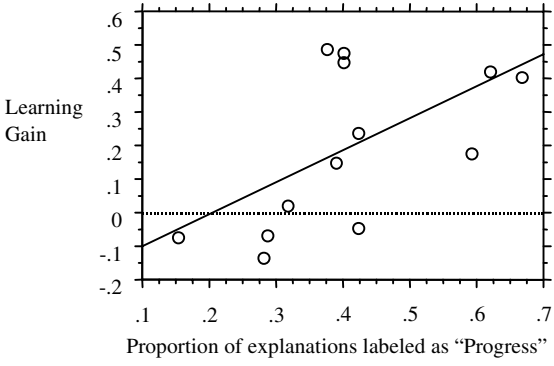


**Fig. 7.** Best Fit Progress vs. Learning Gain

Finally, we looked at the last step in our model, *greater progress → deeper learning.* Each student was given a single progress score by computing the percentage of explanations labeled as "Progress." Learning gain was computed as the commonly-used measure *(post – pre) / (1 – pre)*. While the relation between learning gain and progress was not significant ($r = .253$, $p > .1$), we hypothesized that this may in part be a result of greater progress by students with high pre-test scores, who may have had lower learning gains because their scores were high to begin with. This hypothesis was confirmed by doing a median split that divided the students at a pre-test score of .46. This correlation was significant within the low pre-test group ($r = .588$, $p < .05$) as seen in Fig. 7, but not within the high pre-test group ($r = .031$, $p > .9$). We also examined the relation *better feedback → deeper learning,* which is a concatenation of the last two steps in the model. The relation between learning gain and feedback grade was statistically significant ($r = .588$, $p < .01$).

Merging the results of these separate analyses, we see that each step in the hypothesized chain of relations, *better NLU → better feedback → greater progress → deeper learning,* is supported by means of a statistically significant correlation. We must stress, however, that the results are correlational, not causal. While it is tempting to conclude that better NLU and better feedback *cause* greater learning, we cannot rule out an alternative interpretation of the data, namely, that the better students somehow were better able to stay away from situations in which the tutor gives poor feedback. They might more quickly figure out how to use the tutor, facilitated perhaps by better understanding of the geometry knowledge. Nonetheless, the results are of significant practical value, as discussed further below.

## 5 Students' Explanation Strategies and Relation with Learning

In order to get a better sense of the type of dialogue that expands geometric knowledge, we investigated whether there were any individual differences in students' dialogues with the tutor and how such differences relate to students' learning outcomes. First we conducted a detailed study of the dialogues of four students in the Dialogue condition. Two students were randomly selected from the quarter of students with the

highest learning gains, two from the quarter with the lowest learning gains. In reviewing these case studies, we observed that the low-improvement students often referred to specific angles or specific angle measures in their explanations. For example, one student's first attempt at explaining the Triangle Sum rule is as follows: "I added 154 to 26 and got 180 and that's how many degrees are in a triangle." In contrast, both high-improvement students often began their dialogue by referring to a single problem feature such as "isosceles triangle." In doing so, students first confirmed the correct feature using the feedback from the tutor, before attempting to express the complete rule.

Motivated by the case-study review, the dialogues of all students in the Dialogue condition were coded for the occurrence of these phenomena. An explanation which referred to the name of a specific angle or a specific angle measure was labeled "problem-specific" and an explanation which named only a problem feature was labeled "incremental." The sample of students was ordered by relative frequency of problem-specific instances and split at the median to create a "problem-specific" group and a "no-strategy" group. The same procedure was done again, on the basis of the frequency of incremental instances, to create an "incremental" group and a "no-strategy" group.

The effect of each strategy on learning gain was assessed using a 2X2 repeated-measures ANOVA with the pre- and post-test scores as repeated measure and strategy frequency (high/low) as independent factor (see Fig. 8). The effect of the incremental strategy was not significant. However, the effect of the problem-specific strategy on learning gain was significant ($F(2,23) = 4.77$, $p < .05$). Although the problem-specific group had slightly higher pre-test scores than the no-strategy group, the no-strategy group had significantly higher learning gains.

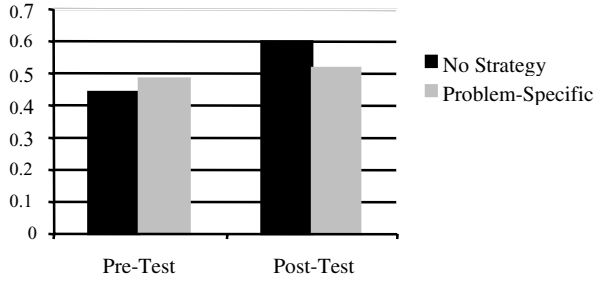


**Fig. 8.** Overall test scores (proportion correct) for frequent and infrequent users of the problem-specific strategy

It was surprising that the incremental strategy, which was used relatively frequently by the two high-improving students in the case studies, was not related with learning gain in the overall sample. Apparently, incremental explanations are not as closely tied to a deep understanding of geometry as expected. Perhaps some students use this strategy to "game" the system, guessing at keywords until they receive positive feedback, but this cannot be confirmed from the present analysis.

On the other hand, students who used the problem-specific strategy frequently ended up with lower learning gains. One explanation of this phenomenon may be that the dialogues that involved problem-specific explanations tended to be longer, as il-

**Student:** angles lgh and tgh are supplementary

**Tutor:** Yes, the angles are supplementary. You need to state a geometry rule that explains why the angles are supplementary.

**Student:** angles lgh and tgh are supplementary, because when you have 180 and subtraction the one angle, you get the second angle

**Tutor:** Can you state a rule that says that certain types of angles are supplementary?

**Student:** angles lgh and tgh are supplementary, the sum of their angles is 180

**Tutor:** In this problem, you have two adjacent angles. Can you state a general rule that says that those kinds of angles are supplementary?

**Student:** angles lgh and tgh are supplementary, because they are on a straight line the sum of their angles is 180

**Tutor:** Yes. And what do you know about the angle measures? If angles form a line, their measures must...

**Student:** the angles on a straight line sum to 180

**Tutor:** You are very very close. But not ALL angles on a line add up to 180.

**Student:** the angles on a straight line sum to 180

**Tutor:** The sum of angles on a line is 180 only if the angles are adjacent.

**Student:** The adjacent angles on a straight line sum to 180

**Fig. 9.** Example of Problem-Specific Dialogue

lustrated in Figure 9. The extended length of these dialogues may be resulting in this group's weaker learning gains. The problem-specific group averaged only 52.5 problems, compared to the no-strategy group's average of 71 problems in the same amount of time. An alternative explanation is that the problem-specific group could be less capable, in general, than the no-strategy group, although the pre-test scores revealed no difference. Problem-specific explanations might reveal an important aspect of student understanding. Their reliance on superficial features might indicate a weakness in their understanding of geometric structures, in their ability to abstract. Possibly, they illustrate the fact that students at different levels of geometric understanding "speak different languages" [15]. While the implications for the design of the Geometry Explanation Tutor are not fully clear, it is interesting to observe that students' explanations reveal more than their pre-test scores.

## 6 Conclusion

The results of a classroom study show an advantage for supporting self-explanation by means of dialogue, as compared to explaining by means of a menu: Students who explain in a dialogue learn better to provide general explanations for problem-solving steps, in terms of geometry theorems and definitions. However, there was no overall difference between the learning outcomes of the students in the two conditions, possi-

bly because the students in the sample were advanced students, as evidenced by high pre-test scores, and thus there was not much room for improvement. It is possible also that the hypothesized advantages of explaining in one's own words did not materialize simply because it takes much time to explain.

Investigating relations between system functioning and student learning, we found correlational evidence for the hypothesized chain of relations, *better NLU → better feedback → greater progress → deeper learning,* Even though these results do not show that the relations are causal, it is reasonable to concentrate further system development efforts on the variables that correlate with student learning, such as progress in dialogues with the system. Essentially, progress is a performance measure and is easier to assess than students' learning gains (no need for pre-test and post-test and repeated exposure to the same geometry rules).

Good feedback correlates with students' progress through the dialogues and with learning. This finding suggests that students do utilize the system's feedback and can extract the information they need to improve their explanation. On the other hand, students who received bad feedback regressed more often. From observation of the explanation corpus, other students recognized that bad feedback was not helpful and tended to enter the same explanation a second time. Generally, students who (on average) received feedback of lesser quality had longer dialogues than students who received feedback of higher quality ($r = .49$, $p < .05$). A study of the 10% longest dialogues in the corpus revealed a recurrent pattern: stagnation (i.e., the repeated turns in a dialogue in which the student did not make progress) followed either by a "sudden jump" to the correct and complete explanation or by the teacher's indicating to the system that the explanation was acceptable (using a system feature added especially for this purpose). This analysis suggests that the tutor should be able to recover better from periods of extended stagnation. Clearly, the system must detect stagnation – relatively straightforward to do using its explanation hierarchy [6] – and provide very directed feedback to help students recover.

The results indicate that accurate classification by the tutor's NLU component (and here we are justified in making a causal conclusion) is crucial to achieving good, precise feedback, although it is not sufficient– the system's dialogue manager must also keep up its end of the bargain. Efforts to improve the system focus on areas where the NLU is not accurate and areas where the NLU is accurate but the feedback is not very good, as detailed in [7, 12].

Finally, an analysis of the differences between students with better/worse learning results found strategy differences between these two groups of students. Two specific strategies were identified, an incremental strategy that focused on using system feedback first to get "in the right ballpark" with minimal effort, and then to expand the explanation. A second strategy was a problem-specific strategic in which students referred to specific problem elements. Students who used the problem-specific explanation strategy more frequently had lower learning gains. Further investigations are needed to find out whether the use of the problem-specific strategy provides additional information about the student that is not apparent from their numeric answers to problems and if so, how a tutorial dialogue system might take advantage of that information.

**Acknowledgements.** The research reported in this paper has been supported by NSF grants 9720359 and 0113864. We thank Jay Raspat of North Hills JHS for his inspired collaboration.

## References

1. Chi, M. T. H. (2000). Self-Explaining Expository Texts: The Dual Processes of Generating Inferences and Repairing Mental Models. In R. Glaser (Ed.), *Advances in Instructional Psychology*, (pp. 161-237). Mahwah, NJ: Erlbaum.
2. Aleven V., Koedinger, K. R. (2002). An Effective Meta-cognitive Strategy: Learning by Doing and explaining with a Computer-Based Cognitive Tutor. *Cog Sci,* 26(2), 147-179.
3. Conati C., VanLehn K. (2000). Toward Computer-Based Support of Meta-Cognitive Skills: a Computational Framework to Coach Self-Explanation. *Int J Artificial Intelligence in Education,* 11, 398-415.
4. Atkinson, R. K., Renkl, A., Merrill, M. M. (2003). Transitioning from studying examples to solving problems: Combining fading with prompting fosters learning. *J Educational Psychology*, 95, 774-783.
5. Corbett, A., Wagner, A., Raspat, J. (2003). The Impact of Analysing Example Solutions on Problem Solving in a Pre-Algebra Tutor. In U. Hoppe et al. (Eds.), *Proc 11th Int Conf on Artificial Intelligence in Education* (pp. 133-140). Amsterdam: IOS Press.
6. Aleven V., Koedinger, K. R., Popescu, O. (2003). A Tutorial Dialog System to Support Self-Explanation: Evaluation and Open Questions. In U. Hoppe et al. (Eds.), *Proc 11th Int Conf on Artificial Intelligence in Education* (pp. 39-46). Amsterdam: IOS Press.
7. Popescu, O., Aleven, V., & Koedinger, K. R. (2003). A Knowledge-Based Approach to Understanding Students' Explanations. In V. Aleven, et al. (Eds.), *Suppl Proc 11th Int Conf on Artificial Intelligence in Education, Vol. VI* (pp. 345-355). School of Information Technologies, University of Sydney.
8. Evens, M. W. et al. (2001). CIRCSIM-Tutor: An intelligent tutoring system using natural language dialogue. *Twelfth Midwest AI and Cog. Sci. Conf*, MAICS 2001 (pp. 16-23).
9. Graesser, A. C., VanLehn, K., Rosé, C. P., Jordan, P. W., Harter, D. (2001). Intelligent tutoring systems with conversational dialogue. *AI Magazine, 22*(4), 39-51.
10. Rose C. P., Siler, S., VanLehn, K. (submitted). Exploring the Effectiveness of Knowledge Construction Dialogues to Support Conceptual Understanding.
11. Rose C. P., VanLehn, K. (submitted). An Evaluation of a Hybrid Language Understanding Approach for Robust Selection of Tutoring Goals.
12. Aleven V., Popescu, O., Ogan, A., Koedinger, K. R. (2003). A Formative Classroom Evaluation of a Tutorial Dialog System that Supports Self-Explanation. In V. Aleven et al. (Eds.), *Suppl Proc 11th Int Conf on Artificial Intelligence in Education, Vol. VI* (pp. 345-355). School of Information Technologies, University of Sydney.
13. Rosé, C. P., Lavie, A. (1999). LCFlex: An Efficient Robust Left-Corner Parser. User's Guide, Carnegie Mellon University.
14. Carletta, J. (1996). Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics*, 22(2):249-254.
15. Schoenfeld, Alan H. "On Having and Using Geometric Knowledge." In *Conceptual and Procedural Knowledge: The Case of Mathematics,* J. Hiebert (Ed.), 225-64. Hillsdale, N.J: Lawrence Erlbaum Associates, 1986.