Aleven V., Koedinger, K. R., & Popescu, O. (2003). A Tutorial Dialog System to Support Self-Explanation: Evaluation and Open Questions. In U. Hoppe, F. Verdejo, & J. Kay (Eds.), *Proceedings of the 11th International Conference on Artificial Intelligence in Education, AI-ED 2003* (pp. 39-46). Amsterdam: IOS Press. **One of five finalists for Best Paper Award AI-ED 2003.**

A Tutorial Dialog System to Support Self-Explanation: Evaluation and Open Questions

Vincent ALEVEN
Kenneth R. KOEDINGER
Octav POPESCU
Human Computer Interaction Institute
Carnegie Mellon University
5000 Forbes Ave, Pittsburgh, PA 15213, USA
aleven@cs.cmu.edu, koedinger@cmu.edu, octav@cs.cmu.edu

Abstract. We conducted an experiment to test the hypothesis that self-explanation can be supported more effectively by means of natural language dialog than by means of menu selection of explanations. The study was carried out in a school in the Pittsburgh area and involved 78 students, 42 of whom satisfied all requirements. There was some evidence that students whose self-explanations were supported by dialog acquired better problem-solving skills than students who selected explanations from a menu. The evidence however was not sufficient to decisively confirm the hypothesis. Lessons learned include: it is more difficult for students to produce accurate free-form explanations than we expected and it may be better to delay strict requirements on explanation quality for students early in learning. It might be valuable to increase explanation strictness gradually. Also, knowledge-based natural language understanding capabilities push on hardware capabilities and have potentially negative consequences for software useability.

A number of cognitive science studies have shown that self-explanation leads to better learning (Aleven & Koedinger, 2002; Bielaczyc, Pirolli, & Brown, 1995; Chi, 2000; Renkl, 1999). Students who explain worked-out examples, expository text, or their own problem solving steps to themselves come away with deeper understanding. However, many studies have also highlighted the fact that it is hard for many students to generate good explanations (Renkl, 1999). Therefore, research is needed to find out how to support self-explanation most effectively in practical educational settings.

Prompting students to self-explain and to clarify vague explanations helps but may not be practical in actual classrooms. Further, even when prompted, many students do not produce good explanations (Renkl, 1999; Wong, Lawson, & Keeves, 2002). Explicit instruction in self-explanation strategies (e.g., Bielaczyc, Pirolli, & Brown, 1995) is likely to have some effect, but more continued guidance may be needed than can be offered by such an instructional program. Interactive learning environments (ILEs) may offer such guidance and scaffolding. Aleven and Koedinger (2002) showed that adding a simple form of support for self-explanation to an existing Cognitive Tutor for geometry problem solving helps students learn with greater understanding. Conati and VanLehn (2000) investigated the effectiveness of more extensive support for self-explanation in the context of example studying.

The role of natural language in learning with ILEs is an area of active research. Some researchers commenting on the remarkable effectiveness of human one-on-one tutors have hypothesized that natural language dialog is the key to this effectiveness. A few systems have been built that interact with students in natural language or to some degree mimic the discourse

of human tutors. Although some attempts are being made to compare these systems to ITSs without natural language capabilities, no one to our knowledge has yet shown a clear instructional benefit for such capabilities.

It seems likely that students will learn more effectively if they self-explain in their own words, rather than by means of a structured user interface with menus and templates, such as those employed in the systems described above. First, natural language is natural. There is no need to learn a new user interface. Second, when students explain in their own words, this makes it easier to build on their partial knowledge. Students can express what they know and the tutor may help them fill in what they do not know, thus providing more targeted scaffolding. Third, articulation may force attention to relevant features to a greater degree than explaining by means of a structured interface. Finally, when students explain in their own words, they are less likely to have problems with jargon or unfamiliar terminology.

However, a number studies with ILEs that prompt for self-explanations in natural language without providing students with feedback on their explanations have yielded mixed results on the key variable of increased learning. Aleven and Koedinger (2000) found that prompts for self-explanations are not very effective in the absence of feedback on explanations. Students often ignored the prompts and produced very few good explanations. Hausmann and Chi (2002) found that a simple prompting system that does not provide feedback can help increase the number of self-explanations generated by students, as compared to the number of self-explanations typed spontaneously into a text editor, but they found also that students produce considerably fewer self explanations when they type than when they explain orally. Trafton and Trickett (2001) found that adding an on-line notepad to an interface for on-line problem solving helps performance and learning in an abstracted scientific reasoning task, relative to a system without a notepad, even in the absence of feedback,. The studies by Renkl (2002) and Schworm and Renkl (2002) showed that providing either instructional explanations or prompts is useful in the context of example studying (although providing both is detrimental), again in the absence of feedback. But Renkl also reported that there were significant individual differences and that many students were not effective explainers. We are not aware of any studies that have contrasted self-explanations in menus with self-explanations in natural language.

In the current research, we explore the hypothesis that self-explanation with an ILE is most effective when students explain in their own words, rather than through menu selections, and the system guides them in constructing good explanations by means of dialog. We have developed a tutorial dialog systems that supports self-explanation, by means of a restricted form of tutorial dialog, in the context of geometry problem solving. The paper reports on an evaluation study of the system.

The Geometry Explanation Tutor

The Geometry Explanation Tutor was built on top of an existing Cognitive Tutor (Anderson, Corbett, Koedinger, & Pelletier, 1995) for geometry problem solving, the Geometry Cognitive Tutor™. This tutor is an integrated part of a full-year high-school curriculum for geometry. The combination of the Geometry Cognitive Tutor and curriculum has been shown to be better than traditional geometry classroom instruction (Koedinger, Corbett, Ritter, & Shapiro, 2000). The tutor and the curriculum are being marketed commercially and are in use in about 100 schools in the United States (see http://www.carnegielearning.com).

Like all Cognitive Tutors, the Geometry Explanation Tutor assists students in solving problems: it follows students in their own individual approaches to problems, providing hints

and feedback on students' solution steps and selecting problems on an individual basis (Anderson, et al., 1995). In addition, the Geometry Explanation Tutor provides support for self-explanation. It requires that students provide general explanations, in their own words, entered via the keyboard, for their problem-solving steps. The tutor helps them, through a restricted form of dialog, to improve their explanations and to arrive at explanations that are (close to) being mathematically precise. For example, when a student types "angles in a triangle are 180", the tutor replies with "is each angle in a triangle 180?" The student might reply "angles in a triangle *sum to* 180," which would be accepted by the system as a correct explanation (Aleven, Popescu, & Koedinger, 2001; 2002). Dialog capabilities have been implemented for one of the six units that make up the curriculum of the Geometry Cognitive Tutor, namely, the Angles unit, which deals with the geometric properties of angles.

The architecture of the Geometry Explanation Tutor has been described elsewhere (Aleven, Popescu, & Koedinger, 2001), so here we provide only a brief overview. The architecture has two main components, a Cognitive Tutor component, which is largely the same as the existing Geometry Cognitive Tutor on which the Explanation Tutor was built, and a knowledge-based natural language understanding (NLU) component. The NLU component evaluates students' explanations and provides feedback on their quality. Each student utterance is assumed to be an attempt to state a geometry rule and is processed in three steps: First, the system's natural language understanding (NLU) component parses the utterance and builds a representation of its semantic content, using the LCFLEX left-corner chart parser (Rosé & Lavie, 1999) and the Loom description logic system (MacGregor, 1991). The semantic representation is then classified, by Loom's classifier, with respect to a fine-grained set of categories of student explanations, including correct, incorrect, and incomplete explanations. For example, category ANGLES-OF-TRIANGLE-180 represents all statements that mean "the angles in a triangle are 180," which is an incomplete statement of the triangle sum theorem. Each category is defined by a statement in Loom's terminological language. Finally, the system provides detailed feedback, based on the set of categories under which the explanation is classified. Each category is associated with a number of increasingly more directed feedback messages. The system selects the category that is closest to being a correct explanation and gives the next message associated with the selected category.

An Evaluation Study

We conducted a study to evaluate how natural language self-explanation supported by dialog compares to a simpler form of computer-supported self-explanation, namely, self-explanation "by reference". In this approach, students explain their problem-solving steps by giving the name of a geometry rule (theorem or definition) that justifies the step. They can type the name or select it from an on-line Glossary of geometry knowledge. Although simple, this control condition is no strawman: it was shown to be better than problem solving with a Cognitive Tutor that does not support self explanation (Aleven & Koedinger, 2002), which in turn had been shown to be better than classroom instruction (Koedinger, et al., 2000).

The study took place in the context of a regular Cognitive Tutor Geometry course in an urban school. The participants came from among 88 students taking this course, three class periods of two different teachers. The students in one of these class periods were honors students, meaning that they were among the best of their year in the given school in terms of academic merit and diligence. At the start of the experiment, 10 students had already started to work on the Angles unit as part of their regular instruction and were therefore excluded from the experiment. The remaining students were assigned to two conditions, "Dialog" and "Rule

Reference". We wanted to make sure that about equal numbers of students of each teacher were assigned to each condition. Further, to the extent possible, we wanted to avoid assigning students in the same class period to different conditions, in order to avoid complaints from students that they had to do more work than their peers (which in fact was not the case). Thus, for the teacher who taught two of the three participating class periods, each period was assigned to one condition. The students in the remaining class period, which was taught by the other teacher and involved the honors students, were assigned randomly to the conditions.

All students took an in-class pre-test. They then worked on the Angles unit, the students in the Dialog condition using a tutor version with the dialog capabilities described above, the students in the Rule Reference condition a version that supports Explanation by Reference. These tutor versions were the same in almost all other respects. Both tutor versions employed a time limit of 7 hours for the Angles unit, with idle time factored out. Since the work on the tutor was self-paced, each student started and finished their work on the Explanation Tutor at a different time. A considerable number of students in both conditions worked on the Angles unit a second time, prior to the post-test. They did so using the regular Geometry Cognitive Tutor, which was used throughout the year as part of the regular geometry instruction. This tutor is practically identical to the Explanation by Reference version used by Rule Reference condition. This extra work on the Angles unit was not planned and may have been due to a miscommunication between the teachers and the experimenter, or other reasons. Finally, the students completed a post-test, again administered in class.

The pre-test and post-test included a number of different types of items. The "Answer" and "Reason" items were similar to the items students encountered during their work with the tutor: They were asked to compute unknown angle measures in a diagram and to explain their answers by giving a statement of an applicable geometry theorem or definition. In addition, the tests included transfer items of different types, intended to measure students' level of understanding as well as their skill in dealing with mathematical text. In some test problems, the students were asked to judge whether there was enough information to find unknown angle measures. At the post-test, if the measure could not be determined, the students were asked also to indicate what additional information would have enabled them to find it. These items test students' understanding, since superficial strategies such as "if angles look the same, their measures are the same," which may achieve some level of success on Answer items, are likely to lead students to wrong answers when insufficient information is available. Items for which there was not enough information, as well as the explanations of such items, we call "Not Enough Info" items. Items for which there was sufficient information were grouped with the Answer and Reason items. Finally, as a test of students' ability to interpret mathematical text, the post-test included "Verbal" items in which students were presented with a general statement about geometry and were asked to indicate if the statement was true or false and, if false, either to correct the statement or to draw a diagram explaining why it was false.

Given the hypothesis that tutored self-explanation in students' own words leads to greater understanding and to greater proficiency in expressing mathematical ideas, one expects to see better performance of students in the Dialog condition on the transfer items (i.e., the Not Enough Info and the Verbal items) as well as the Reason items. The effect on Answer items is more difficult to predict. Greater understanding due to self-explanation in one's own words is likely to manifest itself in better performance on these items. But explaining items in one's own words takes more time than selecting references from a menu, resulting in less practice on Answer items. We predict (without a complete theoretical basis) that "less is more" and that students who explain in their own words will do better on Answer items.

Table 1: On-line measures of the interactions that students had with the Geometry Explanation Tutor

Condition	Time (mins)	#. of Pro-blems	# of Answer Steps	# of Reason Steps	Answer Time (mins)	Reason Time (mins)	Answer %Cor-rect	Reason %Cor-rect
Dialog	398	34	77	76	130	220	59	75
Rule Reference	396	54	174	174	206	120	57	65

Results

Of the 78 students who participated in the experiment, 51 completed both the pre-test and the post-test. Of these students, 42 worked on the experimental tutor for at least two-thirds of the required 7 hours (i.e., 280 minutes), 21 in each condition. In the remainder of the paper, we present the results of these 42 students. As shown in Table 1, the students in both conditions worked on the experimental tutor for about 400 minutes (idle time factored out). The students in the Dialog condition worked an additional 107 minutes on the Angles unit the second time around, the students in the Rule Reference condition 122 minutes. As expected, the students in the Dialog condition spent more time explaining steps (Reason Time) and less time finding unknown angle measures (Answer Time) than their peers in the Rule Reference group. They completed fewer problems and fewer Answer and Reason steps, yet in spite of the lesser amount of practice, they performed slightly better on the Answer steps and considerably better on the Reason steps than the students in the Rule Reference condition. (In the Rule Reference tutor, many problems had multiple questions and therefore had more steps. Therefore, there were more steps per problem in the Rule Reference condition.)

To compare the learning gains between the two conditions, we ran an ANOVA on the test scores, shown in Figure 1, with two levels of repeated measures, Test Time (pre v. post), and Item Type (Answer, Not Enough Info, and Reason), and with Condition as independent factor. The Verbal items are not included in the current analysis, since we do not have repeated measures for these items. There was a significant main effect of test time ($F(1,40) = 27.3$, $p < .0001$), with students' test scores increasing from pre-test to post-test. There was a significant main effect of Item Type ($F(2,80) = 22.7$, $p < .0001$) and a significant interaction between Item Type and Test Time ($F(2,80) = 26.9$, $p < .0001$). At the post-test, the students did best on the Answer items, while at the pre-test the students did best on the Not Enough Info items. There was a significant interaction between Condition and Test Time $F(2,80) = 6.9$, $p = .012$), with students in the Rule Reference condition having higher pre-test scores, and students in the Dialog condition doing better at the post-test. There was also a significant interaction between Condition and Item Type ($F(2,80) = 11.3$, $p < .0001$), with the Dialog condition doing better on Answer and Reason items and the Rule Reference Condition doing better on Not Enough Information items. The Rule Reference condition also did better on the Verbal items.

Discussion

It is tempting to interpret the significant Condition by Test Time interaction as strong evidence that students who self-explained in their own words learned more than students who explained by selecting references from a menu. However, there is reason to be cautious in interpreting these test results. First, the interaction was observed with a subset of the test items. The Verbal items were not included, since we did not have repeated measures for these items. If we consider all items, including the Verbal items, the advantage of the Dialog condition may be smaller than suggested by the 2x2x3 ANOVA, since the Rule Reference group did better on the Verbal items than the Dialog group. Second, the pattern of results

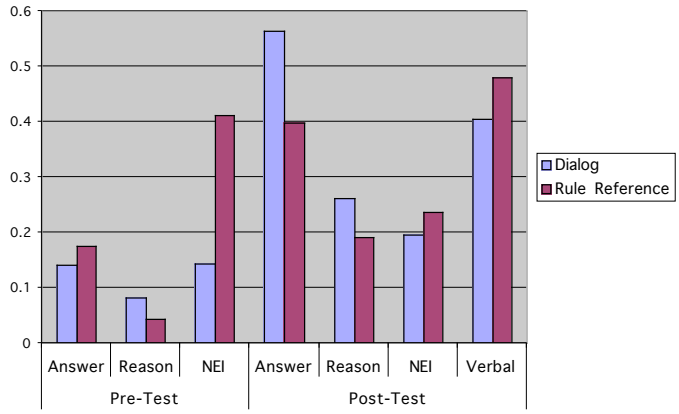

Figure 1: Test scores by Item Type for the 42 students in the sample

related to the different types of test items is different from that which we expected to see. It is difficult to explain this pattern in terms of the underlying knowledge components that students may have acquired. The greatest advantage of the Dialog condition was observed on the items for which we expected to see the smallest advantage, namely, the Answer items. Possibly, this advantage indicates that the students in the Dialog condition acquired better problem-solving knowledge, for example, stronger declarative knowledge whose visual and verbal components were better integrated (see also Aleven and Koedinger, 2002). However, that same knowledge would also give an advantage on the other types of test items, which was not observed in the data. In particular, contrary to our expectations, the Dialog condition did not have a clear advantage with respect to the Reason items and Verbal items. Possibly this lack of advantage is due to that the students in the Dialog condition may have tried, but only partially successfully, to provide correct and complete statements of geometry rules, as they had done during their training. The students in the Rule Reference condition, on the other hand, may have ignored the instructions given on the test form and, also following what they did during training, provided only the name of the geometry rule. This strategy may well be an easier way to get partial credit. The data provide some support for this interpretation: At the post-test, there were more references to rules in the Rule Reference condition (2.9 v. 1.4 per student), whereas in the Dialog condition there were more attempts at stating a geometry rule (0.67 v. 0.33), although the number of attempts was low in both conditions. Further, there were very few complete and correct explanations at the post-test (Dialog: 6.2%, Rule Reference: 4.3%), confirming that explaining is hard and perhaps reflecting the limitations of the copying strategy employed by students during their work with the tutor. A third way in which the pattern of test scores deviates from the expectations is that the students in the Rule Reference condition did better on the Not Enough Info items and that their pre-test scores for these items were unusually high—higher for example than the corresponding post-test scores. It is difficult to explain why this might be so, except perhaps if students consistently wrote "No" for all items of which they were asked to judge whether there is enough information. The data provide some evidence to support this explanation: Students in the Rule Reference condition were 3 times more likely to write "No" on items where there actually was enough information. Be that as it may, if the high pre-test scores in the Rule Reference condition on the Not Enough Info items are somewhat of an anomaly, this means that the difference between the conditions may appear larger than it should have. Thus, we feel compelled to conclude that the data do not fully confirm the hypothesis that natural language self-explanation is superior to explanation by reference, although they certainly do not disconfirm it either.

Although it is rather remarkable that the system was used in an actual classroom over an extended period of time, there was a sense that the students were struggling somewhat as they were using the Geometry Explanation Tutor, perhaps not always in productive ways. Providing mathematical explanations is hard—we are beginning to appreciate exactly how hard. The students very frequently did not try to provide an explanation in their own words, but rather, looked up the relevant rule in the system's on-line Glossary of geometry knowledge and copied the text found there. This strategy no doubt explains the high correctness rate on the Reason steps in the tutor, shown in Table 1. We had not observed this strategy in a previous pilot study in a different school, perhaps due to better-prepared students (Aleven, Popescu, & Koedinger, 2002), or perhaps to the teachers' encouraging students to use the Glossary to look up geometry rules in the context of Answer steps. It is likely however that copying is not as conducive to learning as trying to state a rule in one's own words. Further, the system sometimes appeared overly picky to the students and teachers. Its feedback, aimed at communicating gaps in students' geometry knowledge, was not always helpful in getting students to fix typos in explanations that they copied from the Glossary. Further, the system's response times were sometimes rather long, which was frustrating to some students. The long response times may have had the positive side effect of forcing students to be deliberate when typing explanations, which may be an additional reason for the high correctness rate on explanations.

Some of these difficulties relate to aspects of human-computer interaction (HCI). At least some of the explanations currently required by the system are difficult and may not be within the students' zone of proximal development, or if they are, may require too much cognitive effort. For example, it is difficult for many people to state the angle addition theorem in words: "the measure of an angle formed by adjacent angles is equal to the sum of the measures." One of the teachers commented that the explanations expected from the students are too far removed from the way in which they usually talk about geometry, although somewhat ironically, the language that the system requires is the same language that the students are supposed to understand, such as that found in textbooks. The system requires explanations that are fairly close to being mathematically precise, because precision of expression is an instructional objective in mathematics. However, perhaps one should not expect too much too soon from students. Perhaps the system should initially accept explanations that identify critical features of the problem, even if they are not mathematically precise and then gradually tighten its criteria for correctness. For example, initially the system might accept explanations of the angle addition theorem such as: "you can sum angles that lie next to each other" or even "sum the smaller angles to get the bigger angle". It might help students improve these explanations (e.g., "What is the term for angles that lie next to each other?") without insisting that students immediately go all the way to the statement of the angle addition shown above. We plan to work with teachers to explore this possibility further.

Other difficulties seem related to the system's dialog and natural language understanding capabilities. The system's feedback was not always helpful in fixing copying errors, as it was designed to give feedback at the level of geometry understanding. The system corrects spelling errors and is quite robust in the face of sloppy grammatical language, but it is not able to provide feedback when its analysis fails due to grammar errors or typos that it cannot correct/ignore (e.g., duplicated words, wrong propositions such as "on" instead of "of", etc.). Further, the system did not always respond quickly to explanations by entered by students. We are working to address these problems, by changing the parser somewhat and by adding a "semantic repair mechanism."

In sum, our evaluation study produced limited evidence to support the hypothesis that natural language self-explanation is superior to explanation by reference, but does not conclusively confirm the hypothesis. The hypothesis however remains plausible and we plan to test it in further experiments. As we are improving the system, we find ourselves more and more involved in the HCI aspects of self-explanation, such as making sure that the explanation quality required by the system is appropriate for the learner's level of competence.

Acknowledgements

The research is supported by NSF grants 9720359 and 0113864. We thank Jeff Ziegler and Bradley Baker of Langley High School and Dara Weber of CMU for their assistance.

References

Aleven, V., & Koedinger, K. R. (2002). An Effective Meta-cognitive Strategy: Learning by Doing and Explaining with a Computer-Based Cognitive Tutor. *Cognitive Science,* 26(2), 147-179.

Aleven V., Popescu, O. , & Koedinger, K. R. (2002). Pilot-Testing a Tutorial Dialogue System that Supports Self-Explanation. In S. A. Cerri, G. Gouardères, & F. Paraguaçu (Eds.), *Proceedings of Sixth International Conference on Intelligent Tutoring Systems, ITS 2002* (pp. 344-354). Berlin: Springer Verlag.

Aleven V., Popescu, O. , & Koedinger, K. R. (2001). Towards Tutorial Dialog to Support Self-Explanation: Adding Natural Language Understanding to a Cognitive Tutor. In J. D. Moore, C. L. Redfield, & W. L. Johnson (Eds.), *Proceedings of AI-ED 2001* (pp. 246-255). Amsterdam, IOS Press.

Aleven, V., & Koedinger, K. R. (2000). The Need for Tutorial Dialog to Support Self-Explanation. In C. P. Rose & R. Freedman (Eds.), Building Dialogue Systems for Tutorial Applications, Papers of the 2000 AAAI Fall Symposium (pp. 65-73). Technical Report FS-00-01. Menlo Park, CA: AAAI Press.

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive Tutors: Lessons Learned. *The Journal of the Learning Sciences,* 4, 167-207.

Bielaczyc, K., Pirolli, P. L., & 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, 221-252.

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.

Conati C. & VanLehn K. (2000). Toward Computer-Based Support of Meta-Cognitive Skills: a Computational Framework to Coach Self-Explanation. *Intern. Journal of Artificial Intelligence in Education,* 11, 398-415.

Koedinger, K. R., Corbett, A. T., Ritter, S., & Shapiro, L. (2000). Carnegie Learning's Cognitive Tutor': Summary Research Results. White paper. Available from Carnegie Learning Inc., 1200 Penn Avenue, Suite 150, Pittsburgh, PA 15222, E-mail: info@carnegielearning.com, Web: http://www.carnegielearning.com.

Hausmann, R.G.M., & Chi, M.T.H. (2002). Can a computer interface support self-explaining? The *International Journal of Cognitive technology,* 7(1), 4-14.

MacGregor, R, (1991). The Evolving Technology of Classification-Based Knowledge Representation Systems. In J. Sowa (ed.), *Principles of Semantic Networks: Explorations in the Representation of Knowledge.* San Mateo, CA: Morgan Kaufmann.

Renkl, A. (2002). Learning from Worked-Out Examples: Instructional Explanations Supplement Self-Explanations. *Learning and Instruction.*

Renkl, A. (1999). Learning mathematics from worked-out examples: Analyzing and fostering self-explanations. *European Journal of Psychology in Education,* 16(4), 477-488.

Rosé, C. P. & Lavie, A. (1999). LCFlex: An Efficient Robust Left-Corner Parser. User's Guide, Carnegie Mellon University.

Schworm, S. & Renkl, A. (2002). Learning by Solved Example Problems: Instructional Explanations Reduce Self-Explanation Activity. *Proceeding of the 24nd Annual Conference of the Cognitive Science Society.* Mahwah, NJ: Erlbaum.

Trafton, J.G., & Trickett, S.B. (2001). Note-Taking for Self-Explanation and Problem Solving. *Human-Computer Interaction* 16, 1-38.

Wong, R.M.F., Lawson, M.J., & Keeves, J. (2002). The effects of self-explanation training on students' problem solving in high-school mathematics. *Learning and Instruction* 12, 233-262.