Scaffolding vs. Hints in the Assistment System

Leena Razzaq, Neil T. Heffernan

Computer Science Department
Worcester Polytechnic Institute
100 Institute Road
Worcester, MA 01602
508 831-5006
{leenar, nth}@cs.wpi.edu

Abstract. Razzaq et al, 2005 reported that the Assistment system was causing students to learn at the computer but we were not sure if that was simply due to students getting practice or more due to the "intelligent tutoring" that we created and force students to do if they get an item wrong. Our survey indicated that some students found being forced to do scaffolding sometimes frustrating. We were not sure if all of the time we invested into these "fancy" scaffolding questions was worth it. We conducted a simple experiment to see if students learned on a set of 4 items, if they were given the scaffolds compared with just being given hints that tried to TELL them the same information that the scaffolding questions tried to ASK from them. Our results show that students that were given the scaffolds performed better although the results were not always statistically significant.

1 Introduction

Early evidence that the Assistment system was causing students to learn was reported by Razzaq et al, 2005 [9]. The Assistment system, a web-based system that aims to blend assisting students and assessing their knowledge, was causing students to learn 8th grade math at the computer, but we were uncertain if that was simply due to students getting more practice on math problems or more due to the "intelligent tutoring" that we created and force students to participate in if they got an item wrong. Our survey indicated that some students found being forced to do scaffolding sometimes frustrating. We were not sure if all of the time we invested into these "fancy" scaffolding questions was worth it. We conducted a simple experiment to see if students learned on a set of 4 items if they were forced to do the scaffolding questions, which would ASK them to complete each step required to solve a problem, compared with being given hints, which would TELL them the same information without expecting an answer to each step. In our study, our "scaffolding" condition represents a more interactive learning experience than the "hints" condition. Several studies in the literature have argued that more interactivity will lead to better learning.

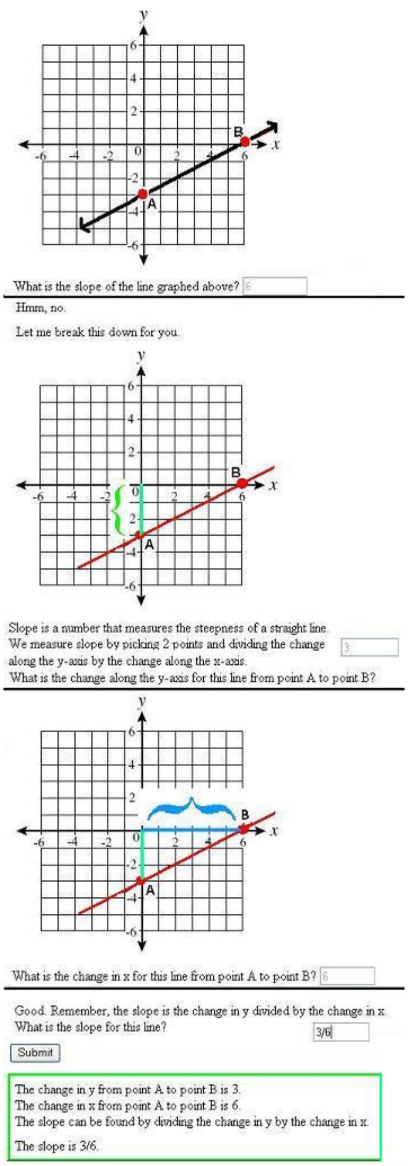

Fig. 1. An Assistment in progress.

Studies indicate that experienced human tutors provide the most effective form of instruction known [2]. They raise the mean performance about two standard deviations compared to students taught in classrooms. Intelligent tutoring systems can offer excellent instruction, but not as good as human tutors. The best ones raise performance about one standard deviation above classroom instruction [6].

In studying what makes a tutoring session successful, VanLehn, Siler and Murray (1998) [11] identified principles for effective teaching. One important principle was that tutors should not offer strong hints or apply rules to problems themselves when students make mistakes. Students miss the opportunity to learn how to solve a problem when they are given an answer and are not allowed to reason for themselves.

Merrill, Reiser, Ranney and Trafton (1992) [7] compared the effectiveness of human tutors and intelligent tutoring systems. They concluded that a major reason that human tutors are more effective is that they let the students do most of the work in overcoming impasses, while at the same time provided as much assistance as necessary. [5] argues that the main thing human tutors do is to keep students on track and prevent them from following "garden paths" of reasoning that are unproductive and unlikely to lead to learning. [5] pointed to the large number of remarks made by tutors that helped keep students on track while learning Lisp programming. Modeling, coaching, and scaffolding are described by Collins, Brown and Hollum (1991) [3] as the heart of cognitive apprenticeship, which they claim "help students acquire an integrated set of skills through processes of observation and guided practice." An important part of scaffolding is fading, which entails progressively removing the support of scaffolding as the student demonstrates proficiency [3].

VanLehn et al (2005) [10] reviews several studies that hypothesize that the relationship between interactivity and learning exists, as well as a few studies that failed to find evidence for this relationship. [10] found that when students found text to be too difficult, tutoring was more effective than having the students read an explanation of how to solve a problem. We believe that our results show that this was true for one of the problems in our experiment which proved to be very difficult for the students.

This experiment would show that it MIGHT be beneficial to have this scaffolding, but the experiment would consciously confound on time as students being forced to do the scaffolding questions would take longer. If this experiment worked we would follow up with an experiment that controlled for time on task. Our results showed that students that were given the scaffolds performed better with an effect size of 0.3. Our survey results seem in line with this result in that students that said they tried to get through difficult problems as quickly as possible were negatively correlated with learning during the course of the year according to Feng et al (2005) [4]. We now plan a follow up study to see if it is worth the extra time.

In this paper, we will present a brief introduction of the Assistment system, how an experiment is executed and our experimental design followed by our results and discussion.

2 The Assistment System

Two years ago, Heffernan and his colleague Ken Koedinger received funding[1] to develop a web-based assessment system, designed to collect formative assessment data on student math skills. Since the assessment is delivered online, students can be tutored on items that they get incorrect. We are currently working with teams of paid and volunteer Worcester Polytechnic Institute (WPI) and Carnegie Mellon students and teacher volunteers to create the Assistment website, which is reported on in [9].

2.1 What is an Assistment?

Once students log into the system they are presented with math items taken from one of the Massachusetts Comprehensive Assessment System (MCAS) tests for math given in previous years. The MCAS test is a state test given to all public school students in Massachusetts. Figure 1 shows a screenshot of an Assistment for an

[1] This research was made possible by the US Dept of Education, Institute of Education Science, "Effective Mathematics Education Research" program grant #R305K03140, the Office of Naval Research grant # N00014-03-1-0221, NSF CAREER award to Neil Heffernan, and the Spencer Foundation. Razzaq was funded by the National Science Foundation under Grant No. 0231773. All the opinions in this article are those of the authors, and not those of any of the funders.

MCAS problem from 2003. If the student had answered correctly, she would have moved on to a new item. The screen shot shows that she incorrectly typed 6 and that the system responded with, “Hmm, no. Let me break this down for you” and followed that up with a question isolating the first step for finding slope, finding the rise. Once she answered that question correctly, she was asked a question focusing on the second step, finding the run. After successfully identifying rise and run, the student was asked to divide these two values and find the slope, repeating the *original question* (we use this term to distinguish it from the other questions we call *scaffolding questions* that help break the problem into pieces). We see that the student then asked for a hint and was told, “The change in y from point A to point B is 3. The change in x from point A to point B is 6. The slope can be found by dividing the change in y by the change in x.” This student asked for a second hint and received “The slope is 3/6.”

2.2 Reporting in the Assistment System

Teachers think highly of the Assistment system not only because their students can get instructional assistance in the form of scaffolding questions and hint messages while working on real MCAS items, but also because they can get online, live reports on students’ progress while students are using the system in the classroom.

Student Name	Total time before (min)	Time spent today (min)	Original Items					Scaffolding + Orig. Items				Most Difficult MA. Stan
			# Done	# Correct	% Corr.	MCAS Score*	Perf. Level	# Done	# Correct	% Correct	# Hint Req.	
Tom	34	0	15	3	20%	200	Failing	30	16	53%	15	N.1.8-understanding-number representations (Error times:
Dick	32	0	38	26	68%	242	Proficient	81	56	69%	4	P.1.8-understanding-pattern times: 2/6)
Harry	33	0	20	9	45%	220	Needs improv	63	28	44%	63	P.1.8-understanding-pattern times: 8/10)

Fig. 2. The Grade book

The “Grade Book”, shown in Figure 2, is the most frequently used report by teachers. Each row in the report represents information for one student, including how many minutes the student has worked on the Assistments, how many minutes he has worked on the Assistments today, how many problems he has done and his percent correct, our prediction of his MCAS score and his performance level.

2.3 Experiments in the Assistment System

The Assistment System allows randomized controlled experiments to be carried out [8] fairly easily. Problems are arranged in curriculums in the system. The curriculum can be conceptually subdivided into two main pieces: the *curriculum* itself, and *sections*. The *curriculum* is composed of one or more *sections*, with each *section* containing *problems* or other *sections.* This recursive structure allows for a rich hierarchy of different types of *sections* and *problems*.

The *section* component is an abstraction for a particular listing of problems. This abstraction has been extended to implement our current *section* types, and allows for future expansion of the *curriculum* unit. Currently existing *section* types include "Linear" (*problems* or sub-*sections* are presented in linear order), "Random" (*problems* or sub-*sections* are presented in a pseudo-random order), and "Experiment" (a single *problem* or sub-*section* is selected pseudo-randomly from a list, the others are ignored).

When an experiment has been carried out, the Experiment Analysis tool can be used to extract the data from the experiment. This tool, developed by Shane Gibbons and Emilia Holban at WPI, allows a researcher to enter a curriculum number, which is a unique identifier, and returns a list for every section in the curriculum. The list contains students who completed problems in the section and whether they got the item correct or incorrect and how much time they spent on each problem. The Experiment Analysis tool is also able to automatically compare performance on particular items or sections.

3 Experimental Design

An experiment carried out in 2004 tested to see whether scaffolding in the Assistment system get students to learn more than hints. In that experiment, 11 MCAS items on probability were presented to 8th grade students in Worcester, Massachusetts. We will refer to this as the Probability Experiment. Some students received the scaffold version of the item while others received the hint version. In the scaffold condition, the computer broke each item down into 2-4 steps (or scaffolds) if a student got the original item wrong. In the hints condition, if students made an error they simply got hints upon demand. The number of items was controlled for. When students completed all 11 items, they saw a few items that were very similar to test if they could do "close"-transfer problems.

The results of the statistical analysis showed a large gain for those students that got the scaffolding questions, but it was discovered that there was a selection-bias. There were about 20% less students in the scaffolding condition that finished the curriculum, and those students that finished were probably the better students, thus invalidating the results. This selection bias was possible due to a peculiarity of the system that presents a list of assignments to students. The students are asked to do the assignments in order, but many students choose not to, thus introducing this bias. This will be easy to correct by forcing students to finish a curriculum once they have started it. Another reason for this bias could be due to fact that students in the hint condition can finish problems faster than students in the scaffold condition. We tried to address both of these issues in the new experiment.

For the new experiment, we chose to focus on items that involved slope and intercept, which according to data from within the Assistment system, students found difficult. We will refer to this experiment as the Slope Experiment. Four MCAS items were chosen for the experiment and four more were chosen as transfer items to test whether the students had learned how to do slope problems. Two of the transfer items were also presented at the beginning of the experiment to serve as pre-test items. Students who got both pretest items correct did not participate in the experiment as they probably had already mastered the material. Students who got a pre-test item wrong were not told the answer or given any tutoring on the item. They were shown a message that told them that they would come back to this problem at the end of class.

To make sure that all of the students had the opportunity to complete the transfer items, we timed the students during the Slope Experiment. The students were given 20 minutes to work on a curriculum containing the two pretest items and four experiment items. They were then given 15 minutes to complete another curriculum containing the 4 transfer items. Unlike the Probability experiment, students had to complete the curriculums before proceeding to any other assignment. This procedure also ensured that students would work on the transfer items regardless of which condition they were in.

Figure 3 shows a slope item used in the experiment. The item on the left, in the scaffolding condition, shows that a student has answered incorrectly and is immediately presented with a scaffolding question. The item on the right, in the hints condition, shows that a student got the item wrong and received the buggy message, outlined in red, of “That is incorrect”. The hint shown outlined in green appears when the student requests a hint by pressing the Hint button. We tried to make the hints in the hints condition similar to the scaffolding questions so that the scaffolding condition did not have an unfair advantage. However, the hints tended to be shorter than scaffolding questions. The difference is that the students in the scaffolding condition were forced to give answers to the individual steps in the problem. We hypothesize that if there is a difference between scaffolding and hints in this

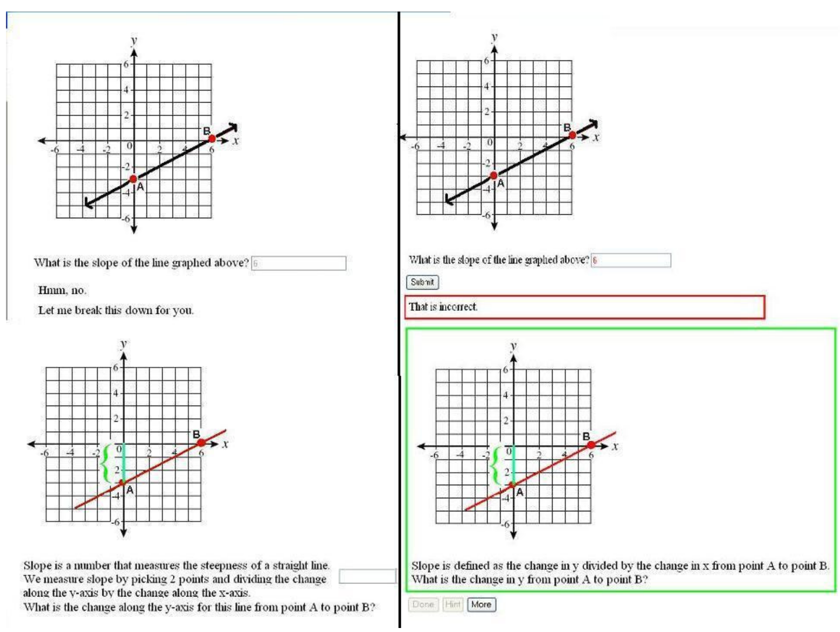

Fig. 3. A scaffold item in the experiment is shown on the left. A hint item is shown on the right.

experiment it will be due to forcing students to work actively to solve a problem, i.e. learning by doing, rather than allowing them to be passive.

174 students from 3 middle schools in Worcester participated in the Slope Experiment. 25 students were excluded for getting both pretest items correct, 11 in the scaffold condition and 14 in the hints condition. Another 5 students were excluded because they had not completed any transfer items, 2 in the scaffold condition and 3 in the hints condition. After these exclusions, there were 75 students in the scaffold condition and 69 students in the hints condition.

4 Results

We first ran an ANOVA to test whether the two conditions differed by pre-test. The result was not statistically significant so we were able to conclude that the groups were fairly balanced in incoming knowledge. We did know that of the two pretest items given, one of them was much harder than the other; 18% of the students got the first pretest item correct as opposed to 45% who got the second pretest item correct. The first pretest item concerned finding the y-intercept from an equation (What is the y-intercept in this equation: $y = 3/4x - 2$?). The second pretest item presented the student with 3 points and asked them to choose the graph that contained the points.
We report two different ways of analyzing our data, as we did not know ahead of time which method would be more likely to detect an effect. The first method, Analysis #1, takes into account 4 items on the posttest, while the second method, Analysis #2, only uses a single item, but has the advantage of being able to use performance on the pretest. Is it more important to have more items on your test, or is it more important to use information from the pretest? We did not know. In Analysis #1, we compared

Inclusion criteria: not1aveonpre from onerowdatawith0insteadofnulls

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
condition	1	.213	.213	2.483	.1173	2.483	.329
Residual	146	12.517	.086				

Means Table for avetrans
Effect: condition
Inclusion criteria: not1aveonpre from onerowdatawith0insteadofnulls

	Count	Mean	Std. Dev.	Std. Err.
1	71	.359	.276	.033
2	77	.435	.307	.035

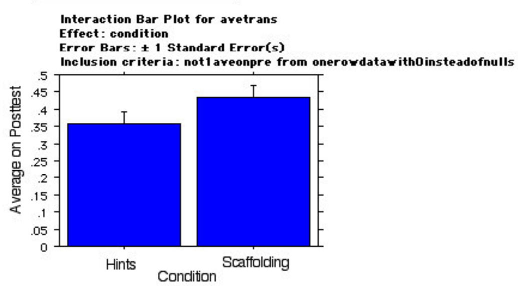

Fig. 4. Results for average on posttest items by condition.

the two groups' average posttest/transfer scores but ignored pretest scores, while in Analysis #2 we looked at differing performance on the harder of the two pretest items that was repeated in the posttest/transfer section. In both analyses, we report the p-values and the effect sizes. We also report the confidence intervals on the effect sizes.

4.1 Analysis #1

For Analysis #1, we ran an ANOVA on the average scores on the transfer items by condition. We remind the reader that there were 4 posttest/transfer items so the scores were either 0%, 25%, 50%, 75% or 100%.

The result showed a p-value of 0.117 with an effect size of 0.3 (See Figure 4). We also calculated the 95% confidence interval for this effect size of .3 and got [-0.03, 0.6]. Because zero is included in this interval, we do not have 95% confidence that the effect size is real. We wanted to get a sense of the significance of this effect size so we calculated the 90% confidence interval and found the range to be [0.01, 0.56]. This implied that the effect size was great than .01 with 90% confidence. We interpret this as somewhat weak evidence in support of the hypothesis that students learned more in the scaffolding condition.

4.2 Analysis #2

We also looked at scores on the transfer items that students had seen as pretest items. For the first pre-test item, which concerned finding the y-intercept from an equation, the ANOVA showed a statistically significant p-value of 0.005 with an effect size of 0.85 (See Figure 5). The 95% confidence interval of the effect size of 0.85 is [0.5, 1.2], meaning that we are 95% confident that the effect size is somewhere

ANOVA Table for trans1com
Inclusion criteria: not1onpre1 from onerowdatawith0insteadofnulls

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
condition	1	.748	.748	8.212	.0048	8.212	.827
Residual	139	12.657	.091				

Means Table for trans1com
Effect: condition
Inclusion criteria: not1onpre1 from onerowdatawith0insteadofnulls

	Count	Mean	Std. Dev.	Std. Err.
1	67	.030	.171	.021
2	74	.176	.383	.045

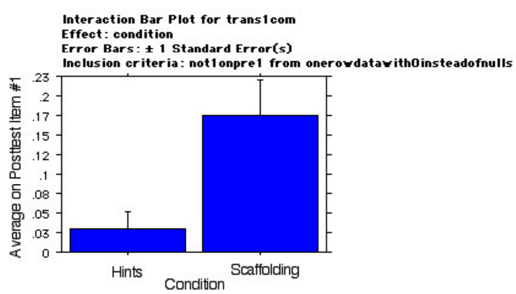

Fig. 5. Results on the transfer item for the first pre-test item by condition

between 0.5 and 1.2, implying that the effect size seems to be at least greater than 0.5, which is a very respectable effect size.

For the second pre-test item, the scaffold condition did better on the transfer item than the hint condition, but the result was not statistically significant.

5 Discussion

In the previous section, we did two different analyses to look for effects of learning. Before we did the analyses we were not sure which was the better way of detecting differences. The results seem to show that there is more learning with scaffolding than with hints, although the difference was not always significant between the two conditions.

The first pretest item on finding the y-intercept from an equation proved to be a difficult problem for all of the students and scaffolding helped significantly. Perhaps the scaffolding had a greater positive effect on learning for the first pretest item because it was much more difficult for the students than the second pretest item. We cannot prove that yet, but would like to study the link between the difficulty of an item and the effectiveness of scaffolding.

In May, 2004, we gave students who were using the Assistment system a survey. 324 students participated in the survey where they were asked their opinions on the Assistment system and math in general. Students who said they tried to get through difficult problems as quickly as possible were negatively correlated with learning [4] during the course of the year. We believe that this falls in line with the results to the Slope Experiment in that students who were in the hint condition could finish items faster. Students who were in the scaffolding condition were forced to spend more time doing scaffolding and ended up learning more. Students who thought that breaking a question down into smaller steps did not help them understand how to solve similar problems was negatively correlated with MCAS scores. Over 60% of the students surveyed thought that the Assistment system helped them prepare for the MCAS. Students who liked using the Assistment system better than normal classroom activity were positively correlated to MCAS scores.

For future work, we plan a follow up study to see if scaffolding is worth the extra time where we will control for time.

Acknowledgements

This research was made possible by the US Dept of Education, Institute of Education Science, "Effective Mathematics Education Research" program grant #R305K03140, the Office of Naval Research grant #N00014-03-1-0221, NSF CAREER award to Neil Heffernan, and the Spencer Foundation. All of the opinions in this article are those of the authors, and not those of any of the funders.

This work would not have been possible without the assistance of the 2004-2005 WPI/CMU ASSISTment Team including Brian Junker at CMU, Andrea Knight, Abraao Lourenco, Michael Macasek, Goss Nuzzo-Jones, Kai Rasmussen, Steven Ritter at Carnegie Learning, Carolyn Rose at CMU, Terrence Turner, Ruta Upalekar, and Jason Walonoski.

References

1. Baker, R.S., Corbett, A.T., Koedinger, K.R. Detecting Student Misuse of Intelligent Tutoring Systems. Proceedings of the 7th International Conference on Intelligent Tutoring Systems, Maceio, Brazil (2004)
2. Bloom, B. S. The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-one Tutoring. Educational Researcher, 13, (1984) 4-16.
3. Collins, A., Brown, J. S., Holum, A. Cognitive Apprenticeship: Making Thinking Visible, American Educator, 6(11), (1991) 38-46.
4. Feng, M., Heffernan, N. T., Koedinger, K. Predicting State Test Scores Better with Intelligent Tutoring Systems: Developing Metrics to Measure Assistance Required. Submitted to the 8th International Conference on Intelligent Tutoring Systems (2005) http://nth.wpi.edu/pubs_and_grants/ITS2006/Submissons/Ming/feng.doc
5. Graesser, A. C., Person, N., Magliano, J. Collaborative Dialog Patterns in Naturalistic One-on-One Tutoring. Applied Cognitive Psychology, 9, (1995) 359-387
6. Koedinger, K. R., Anderson, J. R., Hadley, W. H. & Mark, M. A. Intelligent tutoring goes to school in the big city. In Proceedings of the 7th World Conference on Artificial Intelligence in Education, pp. 421-428. Charlottesville, VA: Association for the Advancement of Computing in Education. (1995)
7. Merrill, D., Reiser, B. Ranney, M., and Trafton, J. Effective Tutoring Techniques: A Comparison of Human Tutors and Intelligent Tutoring Systems. Journal of the Learning Sciences 2(3): 277-305. (1992)
8. Nuzzo-Jones, G., Walonoski, J.A., Heffernan, N.T., Livak, T. The eXtensible Tutor Architecture: A New Foundation for ITS. In C.K. Looi, G. McCalla, B. Bredeweg, & J. Breuker (Eds.) *Proceedings of the 12th Artificial Intelligence In Education*, Amsterdam: ISO Press (2005) 902-904
9. Razzaq, L., Feng, M., Nuzzo-Jones, G., Heffernan, N.T., Koedinger, K. R., Junker, B., Ritter, S., Knight, A., Aniszczyk, C., Choksey, S., Livak, T., Mercado, E., Turner, T.E., Upalekar. R, Walonoski, J.A., Macasek. M.A., Rasmussen, K.P. The Assistment Project: Blending Assessment and Assisting. In C.K. Looi, G. McCalla, B. Bredeweg, & J. Breuker (Eds.) *Proceedings of the 12th Artificial Intelligence In Education*, Amsterdam: ISO Press (2005) 555-562
10. VanLehn, K., Graesser, A. C., Jackson, G. T., Jordan, P., Olney, A., & Rose, C. P. When is reading just as effective as one-on-one interactive human tutoring? In Proceedings of the 27th Annual Meeting of the Cognitive Science Society (pp. 2259-2264). Mahwah, NJ: Erlbaum. (2005) 2259-2264
11. VanLehn, Kurt, Siler, Stephanie, Murray, Charles. What Makes a Tutorial Event Effective? Proceedings of the Twenty-first Annual Conference of the Cognitive Science Society (1998).