Razzaq, L., Heffernan, N. T. (2007). What level of tutor feedback is best? In Luckin & Koedinger (Eds) Proceedings of the 13th Conference on Artificial Intelligence in Education. Amsterdam, Netherland: IOS Press. Pp222-229.

What Level of Tutor Feedback is Best?

Leena RAZZAQ and Neil T. HEFFERNAN
Computer Science Department, Worcester Polytechnic Institute
100 Institute Road, Worcester, Massachusetts, USA

Abstract. The results of a Spring 2006 experiment [11] showed that scaffolding led to higher averages on a post-test, although it was not statistically significant. However, when we looked at scores on particular post-test items that students had seen as pre-test items, we saw significant differences. For a 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 on performance on the post-test. This item on finding the y-intercept from an equation proved to be a difficult problem for all of the students and scaffolding helped significantly. We speculated that the scaffolding had a greater positive effect on learning for this item because it was much more difficult for the students than the other items. We thought that this result warranted a closer look at the link between the difficulty of an item and the effectiveness of scaffolding. In this paper, we report on an experiment that examines the effect of math proficiency and the level of feedback on learning.

Introduction

The Assistment system is a web-based system that blends assisting students with assessing their knowledge. The system tutors students on 7th, 8th and 10th grade content that is based on Massachusetts Comprehensive Assessment System (MCAS) released test items. There are currently over 1000 students using the Assistment system as part of their mathematics class.

Our results show that students are learning 8th grade math at the computer by using the system, but we were not certain if this is 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 a problem wrong. In Spring 2006, we conducted a simple experiment [11] to see if students learned on a set of 4 problems 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 on demand, which would TELL them the same information without expecting an answer to each step. In the study, the “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. In studying what makes a tutoring session successful, VanLehn, Siler and Murray (1998) [12] 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) [12] 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. [12] 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.

The results of the Spring 2006 [11] experiment described above showed that scaffolding led to higher averages on a post-test, although it was not statistically significant. We also looked at scores on particular post-test items that students had seen as pre-test 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. 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 between 0.5 and 1.2, implying that the effect size seems to be at least greater than 0.5. This item on finding the y-intercept from an equation proved to be a difficult problem for all of the students and scaffolding helped significantly. We speculated that 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 thought that this result warranted a closer look at the link between the difficulty of an item and the effectiveness of scaffolding.

Since teachers using our system have started to assign the Assistments for homework we also wanted to determine if Assistments can lead to better learning than traditional paper-and-pencil homework. The question is whether immediate feedback and coaching through math problems leads to better learning than the delayed feedback that is customary with traditional homework (i.e. after the teacher has graded it). We simulated the delayed feedback condition with Assistments that accepted whatever answer the student gave and did not provide any feedback until the end of the assignment.

The purpose of this experiment was to determine which level of feedback worked best for students learning math: scaffolding, hints on demand or delayed feedback, and how the difficulty of an item influenced the effectiveness of the feedback provided.

1. The Assistment System

Limited classroom time available in middle school mathematics classes requires teachers to choose between time spent assisting students' development and time spent assessing their abilities. To help resolve this dilemma, assistance and assessment are integrated in a web-based system called the Assistment[1] System that offers instruction to students while providing a more detailed evaluation of their abilities to the teacher than is available under most current approaches. Many teachers use the system by requiring their students to work on the Assistment website for about 20 minutes per week in their schools' computer labs. Each week when students work on the website, the system "learns" more about the students' abilities and thus, it can hypothetically provide increasingly accurate predictions of how they will do on a standardized mathematics test. The Assistment System is being built to identify the difficulties individual students - and the class as a whole – are having. It is intended that teachers will be able to use this detailed feedback to tailor their instruction to focus on the particular difficulties identified by the system. Unlike other assessment systems, the Assistment technology also provides students with intelligent tutoring assistance while the assessment information is being collected.

An initial version of the Assistment system was created and tested in May, 2004. That version of the system included 40 Assistment items. There are now over 1000 Assistment items. The key feature of Assistments is that they provide instructional assistance in the process of assessing students. The hypothesis is that Assistments can do a better job of assessing student knowledge limitations than practice tests or other on-line amount and nature of the assistance that students receive as a way to judge the extent of student knowledge limitations.

It is easy to carry out randomized controlled experiments in the Assistment System. Items are arranged in modules in the system. The module can be conceptually subdivided into two main pieces: the module itself, and sections. The module is composed of one or more sections, with each section containing items 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 module 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).

2. Experimental Design

Problems in this experiment addressed interpreting linear equations. Figure 1 shows an item used in the experiment. The item shows the different feedback that students can receive once they have answered a question incorrectly. (We call this top-level question the original question.) A student in the scaffolding condition is immediately presented with the first scaffolding question. Students must answer a scaffolding question correctly

[1] The term Assistment was coined by Kenneth Koedinger and blends Assisting and Assessment.

to proceed and receive the next scaffolding question (or finish the problem). They cannot go back and answer the original question, but rather are forced to work through the problem. Students in the hints condition receive a buggy message, outlined in red, of "Sorry, that is not correct". The hint shown outlined in green appears when the student requests a hint by pressing the Hint button. Students do not see the hints unless they ask for them. Students in the delayed feedback condition did not receive any feedback on the problems that they did until they had finished all of the problems. At that time, the students were presented with the answers and explanations of how to solve the problems. Figure 2 shows the explanation that students in the delayed feedback condition received for the item shown in Figure 1.

Based on the results of the Spring 2006 experiment, we hypothesized that less proficient students would need more interaction and benefit more from the scaffolding than more proficient students.

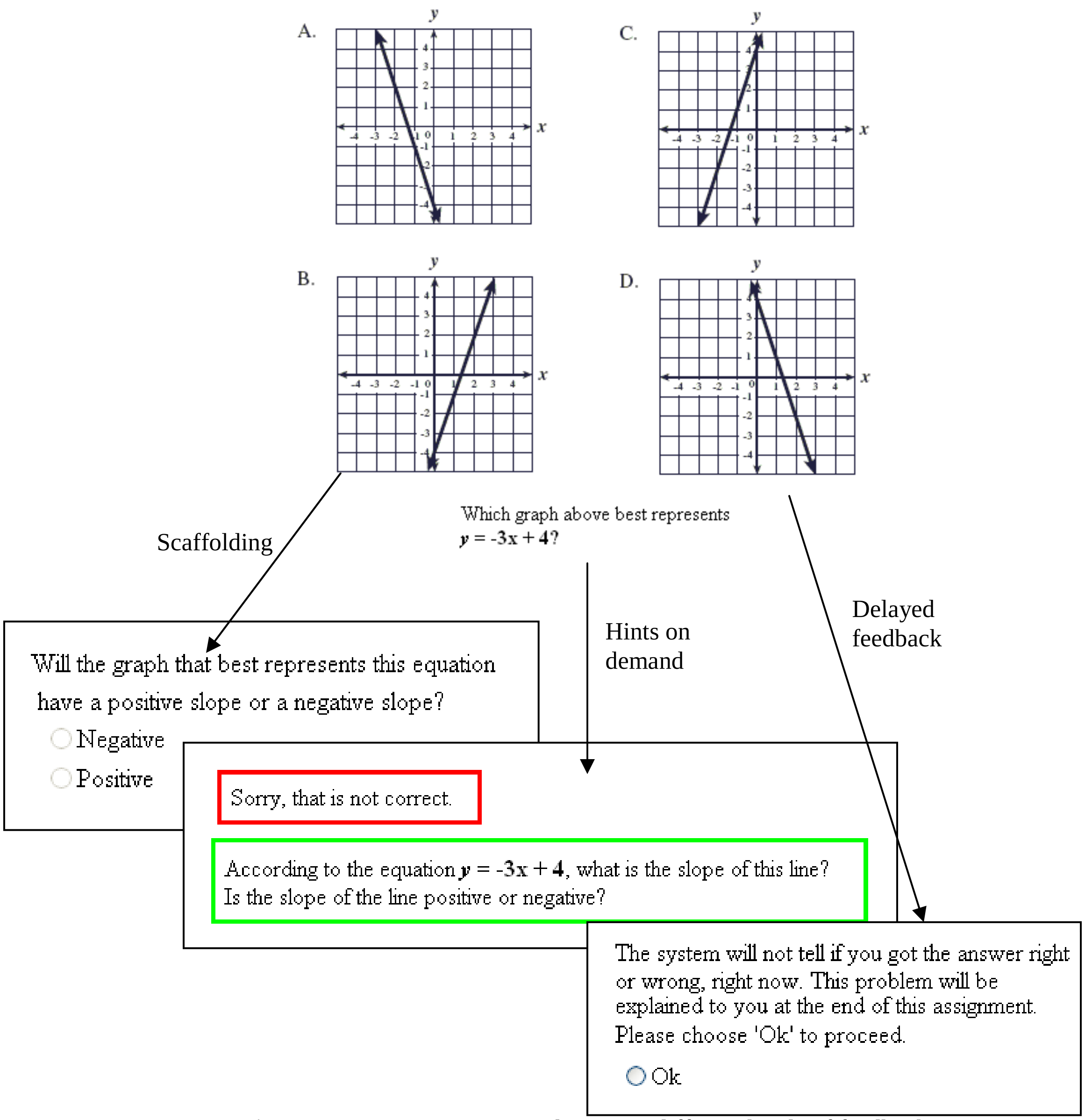


Figure 1. An Assistment item showing 3 different levels of feedback.

For this experiment, the number of problems was held constant, but students took as much time as they needed to finish all of the problems. Students were presented with 2 pre-test problems, 4 experiment problems and 4 post-test problems that addressed the topic of interpreting linear equations. There were 3 versions of the experiment problems, one for each condition. Two of the pre-test problems were the same problems as two of the post-test problems.

The Assistment system randomly assigned students to the hints, scaffold or delayed feedback condition with equal probability. 392 8th grade students in the Worcester Public Schools in Worcester, Massachusetts participated in the experiment, 138 students were in honors level classes and 254 were in regular math classes.

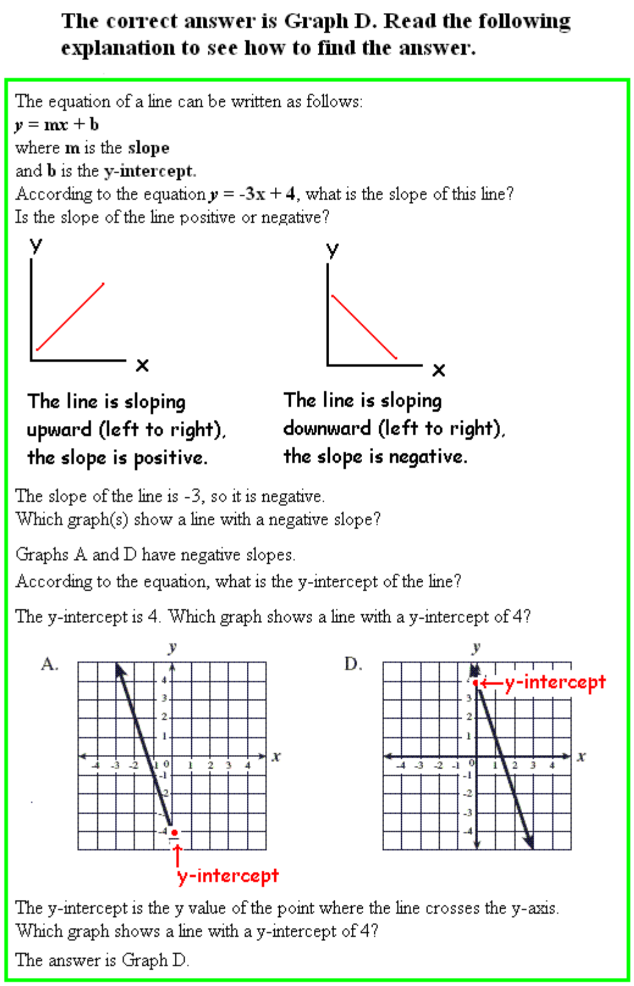
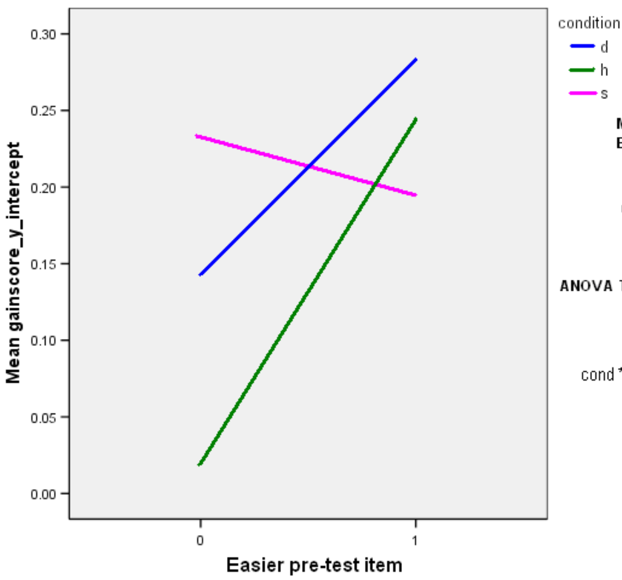


Figure 2. The delayed feedback condition received explanations at the end of the assignment.

3. Analysis

We first checked to make sure that the groups were not significantly different at pre-test by doing an ANOVA on pre-test averages by condition. There was no significant difference between groups at pre-test ($p = 0.556$). Students learned overall from pre-test to post-test ($p = .005$). When we look at performance on the post-test by condition, the difference is not significant; however there is an interesting trend when we separate students by math proficiency. The regular students seem to benefit more from the scaffolding condition, while honors students seem to benefit more from the delayed feedback condition. We decided to take a closer look at the item that proved most difficult for students. The problem concerned finding the y-intercept from an equation and was presented to students in the pre-test and again in the post-test. We did a one-way ANCOVA with performance on the easier pre-test item as a covariate. The main effect of condition ($F(2, 349) = 3.8$, $p = 0.02$) implies that the delayed feedback works best and the hints are least effective. The statistical significance of the covariate's coefficient implies that students with higher scores on the easier pretest item are more likely to learn more averages over the three conditions. The statistical significance of the interaction term (Covariate*Condition) implies that the main effects of pre-test and condition are different.

We interpret the low p-values on the interaction term to mean that there are different rates of learning on the single items based upon the interaction between the covariate and condition. Students who come in with less knowledge benefit more from the scaffolding than students who come in with more knowledge. Students who come in with more knowledge benefit from the delayed feedback more than the other groups.



Means Table for gain
Effect: cond

	Count	Mean	Std. Dev.	Std. Err.
h	116	.164	.372	.035
d	121	.231	.423	.038
s	118	.220	.416	.038

ANOVA Table for gain

	DF	Sum of Squares	Mean Square	F-Value	P-Value
cond	2	1.239	.619	3.869	.0218
easier pre-test item	1	.721	.721	4.502	.0346
cond * easier pre-test item	2	1.130	.565	3.528	.0304
Residual	349	55.874	.160		

Figure 3. Condition is significant on performance

4. Discussion

The results of this experiment were surprising. We did not expect students in the delayed feedback condition to learn more than in other groups, however the honors students did better in this condition than in the scaffolding or hints conditions. The regular students performed better in the scaffolding condition. One possible explanation is that less proficient students benefit from more interaction and coaching through each step to solve a problem while more proficient students benefit from seeing problems worked out and seeing the big picture. Another possible explanation, put forth by one of the eighth grade teachers, is that honors students are often more competitive and like to know how they do on their work. The delayed feedback group had to wait till the end of the assignment to see how they did and perhaps the honors students ended up reading through the explanations more carefully than they would have read the scaffolding questions or hints.

The students in the hints condition did not perform as well as the delayed feedback groups for both proficient and less proficient students. One possible explanation is the more proactive nature of the delayed feedback explanations. Murray and VanLehn [8] found that proactive help was more effective for some students. “Proactive help when a student would otherwise flounder can save time, prevent confusion, provide valuable information at a time when the student is prepared and motivated to learn it, and avoid the negative affective consequences of frustration and failure.” In the Assistment system, students only see hints if they ask for them and they are less likely to ask for hints on multiple choice questions when they can guess more easily.

We believe the results of this experiment present a good case for tailoring tutor feedback to types of students to maximize their learning.

Acknowledgements

We would like to thank all of the people associated with creating the Assistment system listed at www.assistment.org including the investigators Kenneth Koedinger, and Brian Junker at Carnegie Mellon. We would also like to acknowledge funding from the US Department of Education, the National Science Foundation, the Office of Naval Research and the Spencer Foundation. All of the opinions expressed in this paper are those solely of the authors and not those of our funding organizations.

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] Murray, C., VanLehn K. (2006) A Comparison of Decision-Theoretic, Fixed-Policy and Random Tutorial Action Selection. In Ikeda, Ashley & Chan (Eds.). Proceedings of the 8th International Conference on Intelligent Tutoring Systems. Springer-Verlag: Berlin. pp. 116-123. 2006.

[9] 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

[10] 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

[11] Razzaq, L., Heffernan, N.T. (2006). Scaffolding vs. hints in the Assistment System. In Ikeda, Ashley & Chan (Eds.). *Proceedings of the 8th International Conference on Intelligent Tutoring Systems*. Springer-Verlag: Berlin. pp. 635-644. 2006.

[12] 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

[13] 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).