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. (2005). 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*. 555-562. Amsterdam: ISO Press

Blending Assessment and Instructional Assisting

Leena RAZZAQ*, Mingyu FENG, Goss NUZZO-JONES, Neil T. HEFFERNAN, Kenneth KOEDINGER+, Brian JUNKER+, Steven RITTER°, Andrea KNIGHT+, Edwin MERCADO*, Terrence E. TURNER, Ruta UPALEKAR, Jason A. WALONOSKI, Michael A. MACASEK, Christopher ANISZCZYK, Sanket CHOKSEY, Tom LIVAK, Kai RASMUSSEN

Department of Computer Science
Worcester Polytechnic Institute, Worcester, MA, USA
assistments@wpi.edu

+*Human-Computer Interaction Institute*
Carnegie Mellon University, Pittsburgh, PA, USA

°*Carnegie Learning, Inc.*

Abstract. Middle school mathematics teachers are often forced to choose between assisting students' development and assessing students' abilities because of limited classroom time available. To help teachers make better use of their time, we are integrating assistance and assessment by utilizing a web-based system ("Assistment") that will offer instruction to students while providing a more detailed evaluation of their abilities to the teacher than is possible under current approaches. An initial version of the Assistment system was created and used last May with about 200 students and 800 students are using it this year once every two weeks. The hypothesis is that Assistments both assist students while also assessing them. This paper describes the Assistment system and some preliminary results.

Introduction

Limited classroom time available in middle school mathematics classes compel teachers to choose between time spent assisting students' development and time spent assessing students' abilities. To help resolve this dilemma, assistance and assessment are integrated in a web-based system ("Assistment"[1]) that will offer instruction to students while providing a more detailed evaluation of their abilities to the teacher than is possible under current approaches. The plan is for students to work on the Assistment website for about 20 minutes per week. The Assistment system is an Artificial Intelligence program. 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

* 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. Authors Razzaq and Mercado were 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.

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

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 was created and tested last May. That version of the system included 40 Assistment items. There are now approximately 150 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 testing approaches by using a "dynamic assessment" approach. In particular, Assistments use the amount and nature of the assistance that students receive as way to judge the extent of student knowledge limitations. Initial first year efforts to test this hypothesis of improved prediction of the Assistment's dynamic assessment approach are discussed below.

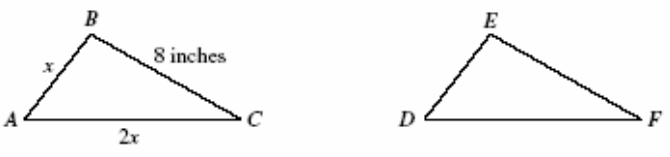
In preparation for fall of 2004, 75 Assistment items were created and 9 teachers and about 1000 students are currently using them in 3 schools. Currently, there are approximately 150 Assistments.

1. Assistment System and website development

In December of 2003, one of the authors met with the Superintendent of the Worcester Public Schools in Massachusetts, and was subsequently introduced to the three math department heads of 3 out of 4 Worcester middle schools. The goal was to get these teachers involved in the design process of the Assistment System at an early stage. The main activity done with these teachers was meeting about one hour a week to do "knowledge elicitation" interviews, whereby the teachers helped design the pedagogical content of the Assistment System.

The procedure for knowledge elicitation interviews went as follows. A teacher was shown a Massachusetts Comprehensive Assessment System (MCAS) test item and asked how she would tutor a student in solving the problem. What kinds of questions would she ask the student? What hints would she give? What kinds of errors did she expect and what would she say when a student made an expected error? These interviews were videotaped and the interviewer took the videotape and filled out an "Assistment design form" from the knowledge gleaned from the teacher. The Assistment was then implemented using the design form. The first draft of the Assistment was shown to the teacher to get her opinion and she was asked to edit it. Review sessions with the teachers were also videotaped and the design form revised as needed. When the teacher was satisfied, the Assistment was released for use by students.

Triangles ABC and DEF shown below are congruent.



The perimeter of $\triangle ABC$ is 23 inches. What is the length of side $\overline{DF}$ in $\triangle DEF$?

Figure 1: Item 19 from the 2003 MCAS

For instance, a teacher was shown a MCAS item on which her students did poorly, such as item #19 from the year 2003, which is shown in Figure 1. About 15 hours of knowledge elicitation interviews were used to help guide the design of Assistments.

Figure 2 shows an Assistment that was built for the item 19 shown above. Each Assistment consists of an *original item* and a list of *scaffolding questions* (in this case, 5 scaffolding questions). The first scaffolding question appears only if the student gets the item wrong. Figure 2 shows that the student typed "23" (which happened to be the most common wrong answer for this item from the data collected). After an error, students are not allowed to try the item further, but instead must then answer a sequence of scaffolding questions (or "scaffolds") presented one at a time[2]. Students work though the scaffolding questions, possibly with hints, until they eventually get the problem correct. If the student presses the hint button while on the first scaffold, the first hint is displayed, which is the definition of congruence in this example. If the student hits the hint button again, the hint that is shown in Figure 2 appears, which describes how to apply congruence to this problem. If the student asks for another hint, the answer is given. Once the student gets the first scaffolding question correct (by typing AC), the second scaffolding question appears.

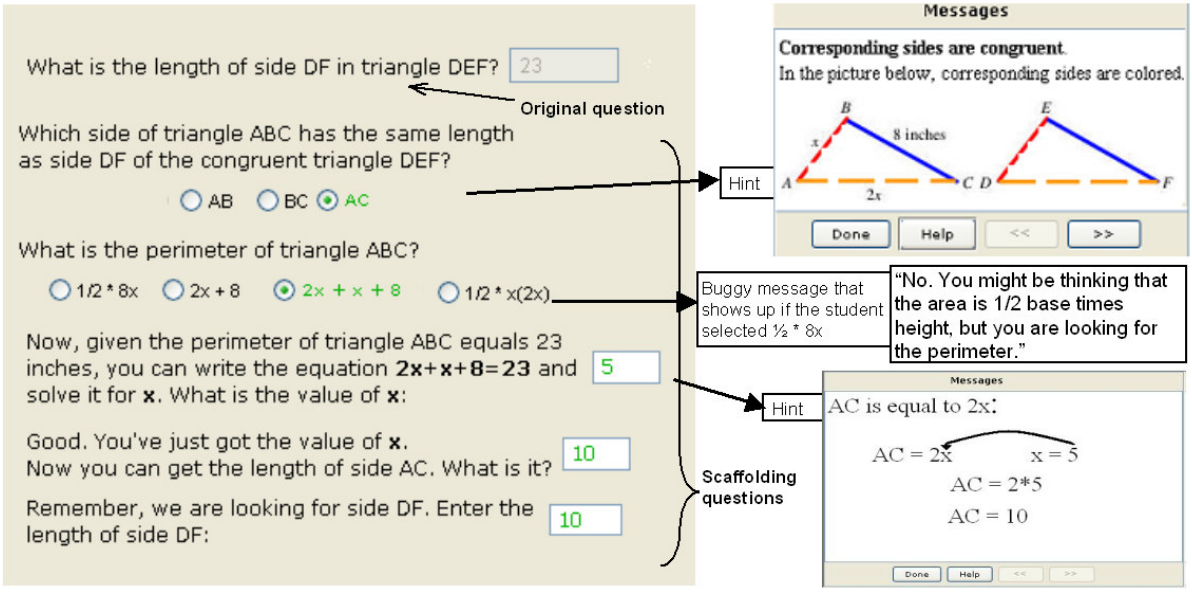


Figure 2: An Assistment shown just before the student hits the "done" bottom, showing two different hints and one buggy message that can occur at different points.

If the student selects ½ * 8x, the *buggy message* shown would appear suggesting that it is not necessary to calculate area. (*Hints* appear on demand, while *buggy messages* are responses to a particular student error). Once the student gets the second question correct, the third appears, and so on. Figure 2 shows the state of the interface when the student is done with the problem as well as a hint for the 4th scaffolding question.

About 200 students used the system in May 2004 in three different schools from about 13 different classrooms. The average length of time was one class period per student. The teachers seemed to think highly of the system and, in particular, liked that real MCAS items were used and that students received instructional assistance in the form of scaffolding questions. Teachers also like that they can get online reports on students'

[2] As future work, once a predictive model has been built and is able to reliably detect students trying to "game the system" (e.g., just clicking on answer) students may be allowed to re-try a question if they do not seem to be "gaming". Thus, studious students may be given more flexibility.

progress from the Assistment web site and can even do so while students are using the Assistment System in their classrooms. The system has separate reports to answer the following questions about **items**, **student, skills** and student **actions**: Which **items** are my students finding difficult? Which **items** are my students doing worse on compared to the state average? Which **students** are 1) doing the best, 2) spending the most time, 3) asking for the most hints etc.? Which of the approximately 80 **skills** that we are tracking are students doing the best/worst on? What are the exact **actions** that a given student took?

The three teachers from this first use of the Assistment System were impressed enough to request that all the teachers in their schools be able to use the system the following year. Currently that means that about 1,000 students are using the system for about 20 minutes per week for the 2004-2005 school year. Two schools have been using the Assistment System since September. A key feature of the strategy for both teacher recruitment and training is to get teachers involved early in helping design Assistments through knowledge elicitation and feedback on items that are used by their students.

Assistments are based on Intelligent Tutoring System technology that is deployed with an internet-savvy solution (for more technical details on the runtime see [6]). In the first year's solution, when students started an Assistment item, a Java Web Start application was downloaded and reported each students' actions (other than their mouse movements) to a database at WPI, thus enabling completely live database reporting to teachers. Database reporting for the Assistment Project is covered extensively in [3]. In the second year, the application has been delivered via the web and requires no installation or maintenance. We have spent considerable time observing its use in classrooms; for instance, one of the authors has logged over 50 days, and was present at over 300 classroom periods. This time is used to work with teachers to try to improve content and to work with students to note any misunderstandings they sometimes bring to the items. For instance, if it is noted that several students are making similar errors that were not anticipated, the "Assistment Builder" [4] web-based application can be logged into and a buggy message added that addresses the students' misconception. The application is being prepared for its statewide release in May 2005.

The current Assistment System web site is at www.assistment.org, which can be explored for more examples.

2. Analysis of data to determine whether the system reliably predicts MCAS performance

One objective the project had was to analyze data to determine whether and how the Assistment System can predict students' MCAS performance. In Bryant, Brown and Campione [2], they compared traditional testing paradigms against a dynamic testing paradigm. In the dynamic testing paradigm a student would be presented with an item and when the student appeared to not be making progress, would be given a prewritten hint. If the student was still not making progress, another prewritten hint was presented and the process was repeated. In this study they wanted to predict learning gains between pretest and posttest. They found that static testing was not as well correlated ($R = 0.45$) as with their "dynamic testing" ($R = 0.60$).

Given the short use of the system in May, there was an opportunity to make a first pass at collecting such data. The goal was to evaluate how well on-line use of the Assistment System, in this case for only about 45 minutes, could predict students' scores on a 10-item post-test of selected MCAS items. There were 39 students who had taken the posttest. The paper and pencil posttest correlated the most with MCAS scores with an R-value of 0.75.

A number of different metrics were compared for measuring student knowledge during Assistment use. The key contrast of interest is between a static metric that mimics paper practice tests by scoring students as either correct or incorrect on each item, with a dynamic assessment metric that measures the amount of assistance students need before they get an item correct. MCAS scores for 64 of the students who had log files in the system were available. In this data set, the static measure does correlate with the MCAS, with an R-value of 0.71 and the dynamic assistance measure correlates with an R-value of -0.6. Thus, there is some preliminary evidence that the Assistment System may predict student performance on paper-based MCAS items.

It is suspected that a better job of predicting MCAS scores could be done if students could be encouraged to take the system seriously and reduce "gaming behavior". One way to reduce gaming is to detect it [1] and then to notify the teacher's reporting session with evidence that the teacher can use to approach the student. It is assumed that teacher intervention will lead to reduced gaming behavior, and thereby more accurate assessment, and higher learning.

The project team has also been exploring metrics that make more specific use of the coding of items and scaffolding questions into knowledge components that indicate the concept or skill needed to perform the item or scaffold correctly. So far, this coding process has found to be challenging, for instance, one early attempt showed low inter-rater reliability. Better and more efficient ways to use student data to help in the coding process are being sought out. It is believed that as more data is collected on a greater variety of Assistment items, with explicit item difficulty designs embedded, more data-driven coding of Assistments into knowledge components will be possible.

Tracking student learning over time is of interest, and assessment of students using the Assistment system was examined. Given that there are approximately 650 students using the system, with each student coming to the computer lab about 7 times, there was a table with 4550 rows, one row for each student for each day, with an average percent correct which itself is averaged over about 15 MCAS items done on a given day. In Figure 3, average student performance is plotted versus time. The y-axis is the average percent correct on the original item (student performance on the scaffolding questions is ignored in this analysis) in a given class. The x-axis represents time, where data is bunched together into month, so some students who came to the lab twice in a month will have their numbers averaged. The fact that most of the class trajectories are generally rising suggests that most classes are learning between months.

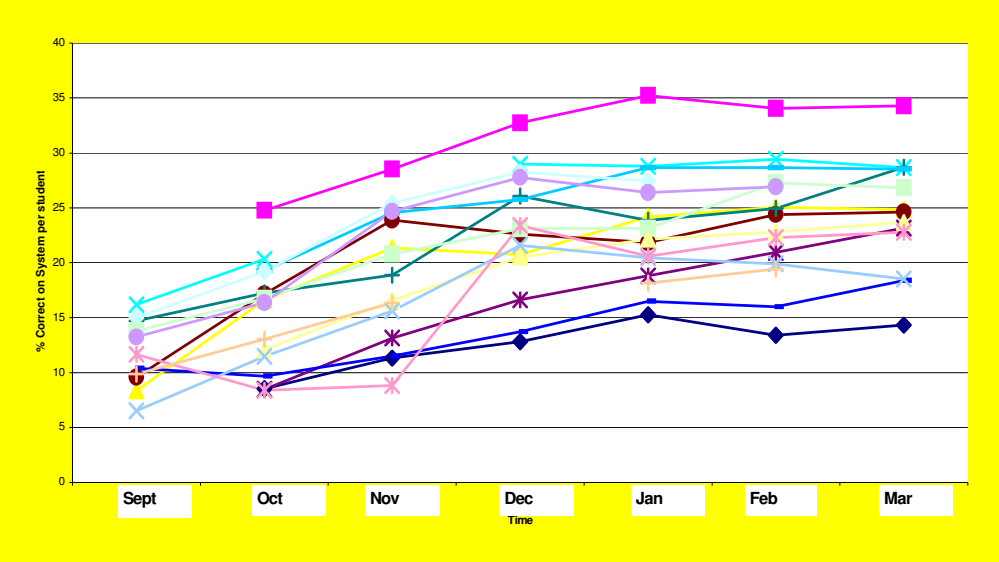


Figure 3: Average student performance is plotted versus time.

Given that this is the first year of the Assistment project, new content is created each month, which introduces a potential confounder of item difficulty. It could be that some very hard items were selected to give to students in September, and students are not really learning but are being tested on easier items. Next year, this confound will be eliminated by sampling items randomly. Adding automated applied longitudinal data analysis [7] is currently being pursued.

3. Analysis of data to determine whether the system effectively teaches.

The second form of data comes from within Assistment use. Students potentially saw 33 different problem pairs in random order. Each pair of Assistments included one based on an original MCAS item and a second "morph" intended to have different surface features, like different numbers, and the same deep features or knowledge requirements, like approximating square roots. Learning was assessed by comparing students' performance the first time they were given one of a pair with their performance when they were given the second of a pair. If students tend to perform better on the second of the pair, it indicates that they may have learned from the instructional assistance provided by the first of the pair.

To see that learning happened and generalized across students and items, both a student level analysis and an item level analysis were done. The hypothesis was that students were learning on pairs or triplets of items that tapped similar skills. The pairs or triplet of items that were chosen had been completed by at least 20 students.

For the student level analysis there were 742 students that fit the criteria to compare how students did on the first opportunity versus the second opportunity on a similar skill. A gain score per item was calculated for each student by subtracting the students' score (0 if they got the item wrong on their first attempt, and 1 if they got it correct) on their 1st opportunities from their scores on the 2^{nd} opportunities. Then an average gain score for all of the sets of similar skills that they participated in was calculated. A student analysis was done on learning opportunity pairs seen on the same day by a student and the t-test showed statistically significant learning ($p = 0.0244$). It should be noted that there may be a selection effect in this experiment in that better students are more likely to do more problems in a day and therefore more likely to contribute to this analysis.

An item analysis was also done. There were 33 different sets of skills that met the criteria for this analysis. The 5 sets of skills that involved the most students were: Approximating Square Roots (6.8% gain), Pythagorean Theorem (3.03% gain), Supplementary Angles and Traversals of Parallel Lines (1.5% gain), Perimeter and Area (4.3% gain) and Probability (3.5% gain). A t-test was done to see if the average gain scores per item were significantly different than zero, and the result ($p = 0.3$) was not significant. However, it was noticed that there was a large number of negative average gains for items that had fewer students so the average gain scores were weighted by the number of students, and the t-test was redone. A statistically significant result ($p = 0.04$) suggested that learning should generalize across problems. The average gain score over all of the learning opportunity pairs is approximately 2%. These results should be interpreted with some caution as some of the learning opportunity pairs included items that had tutoring that may have been less effective. In fact, a few of the pairs had no scaffolding at all but just hints.

4. Experiments

The Assistment System allows randomized controlled experiments to be carried out. At present, there is control for the number of items presented to a student, but soon the system will be able to control for time, as well. Next, two different uses of this ability are described.

4.1 Do different scaffolding strategies affect learning?

The first experiment was designed as a simple test to compare two different tutoring strategies when dealing with proportional reasoning problems like item 26 from the 2003 MCAS: "The ratio of boys to girls in Meg's chorus is 3 to 4. If there are 20 girls in her chorus, how many boys are there?" One of the conditions of the experiment involved a student solving two problems like this with scaffolding that first coached them to set up a proportion. The second strategy coached students through the problem but did not use the formal notation of a proportion. The experimental design included two items to test transfer. The two types of analyses the project is interested in fully automating is to 1) to run the appropriate ANOVA to see if there is a difference in performance on the transfer items by condition, and 2) to look for learning during the condition, and see if there is a disproportionate amount of learning by condition.

Two types of analyses were done. First, an analysis was done to see if there was learning during the conditions. 1st and 2nd opportunity was treated as a repeated measure and to look for a disproportionate rate of learning due to condition (SetupRatio vs. NoSetup). A main effect of learning between first and second opportunity ($p = 0.05$) overall was found, but the effect of condition was not statistically significant ($p = 0.34$). This might be due to the fact that the analysis also tries to predict the first opportunity when there is no reason to believe those should differ due to controlling condition assignment. Given that the data seems to suggest that the SetupRatio items showed learning a second analysis was done where a gain score (2nd opportunity minus 1st opportunity) was calculated for each student in the SetupRatio condition, and then a t-test was done to see if the gains were significantly different from zero and they were ($t = 2.5$, $p = 0.02$), but there was no such effect for NoSetup.

The second analysis done was to predict each student's average performance on the two transfer items, but the ANOVA found that even though the SetupRatio students had an average score of 40% vs. 30%, this was not a statistically significant effect.

In conclusion, evidence was found that these two different scaffolding strategies seem to have different rates of learning. However, the fact that setting up a proportion seems better is not the point. The point is that it is a future goal for the Assistment web site to do this sort of analysis automatically for teachers. If teachers think they have a better way to scaffold some content, the web site should send them an email as soon as it is known if their method is better or not. If it is, that method should be adopted as part of a "gold" standard.

4.2 Are scaffolding questions useful compared to just hints on the original question?

An experiment was set up where students were given 11 probability items. In the first condition, the computer broke each item down into 2-4 steps (or scaffolds) if a student got the original item wrong. In the other condition, if a student made an error they just got hints upon demand. The number of items was controlled for. When students completed all 11

items, they saw a few items that were morphs to test if they could do "close"-transfer problems.

The results of the statistical analysis were showing 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. New results are expected inside a month.

Conclusion

The Assistment System was launched and presently has 3 middle schools using the system with all of their 8th grade students. Some initial evidence was collected that the online system might do a better job of predicting student knowledge because items can be broken down into finer grained knowledge components. Promising evidence was also found that students were learning during their use of the Assistment System. In the near future, the Assistment project team is planning to release the system statewide in Massachusetts.

References

[1] Baker, R.S., Corbett, A.T., Koedinger, K.R. (2004) Detecting Student Misuse of Intelligent Tutoring Systems. *Proceedings of the 7th International Conference on Intelligent Tutoring Systems*, 531-540.

[2] Campione, J.C., Brown, A.L., & Bryant, N.R. (1985). Individual differences in learning and memory. *In R.J. Sternberg (Ed.). Human abilities: An information-processing approach,* 103-126. New York: W.H. Freeman.

[3] Feng, M., Heffernan, N.T., (2005). Informing Teachers Live about Student Learning: Reporting in the Assistment System. *Submitted to the Workshop on Usage Analysis in Learning Systems at 12th Annual Conference on Artificial Intelligence in Education 2005*, Amsterdam.

[4] Turner, T. E., Macasek, M. A., Nuzzo-Jones, G., Heffernan, N.T., (2005). The Assistment Builder: A Rapid Develoment Tool for ITS. *Submitted to the 12th Annual Conference on Artificial Intelligence in Education 2005,* Amsterdam

[5] Koedinger, K. R., Aleven, V., Heffernan. T., McLaren, B. & Hockenberry, M. (2004) Opening the Door to Non-Programmers: Authoring Intelligent Tutor Behavior by Demonstration. *Proceedings of 7th Annual Intelligent Tutoring Systems Conference*, Maceio, Brazil. page162-173

[6] Nuzzo-Jones, G., Walonoski, J. A., Heffernan, N.T., Livak, T.(2005). The eXtensible Tutor Architecture: A New Foundation for ITS. *Submitted to the 12th Annual Conference on Artificial Intelligence in Education 2005*, Amsterdam

[7] Singer, J. D. & Willett, J. B. (2003). Applied Longitudinal Data Analysis: Modeling Change and Occurrence. Oxford University Press, New York.