Towards Live Informing and Automatic Analyzing of Student Learning: Reporting in Assistment System

Mingyu FENG, Neil T. HEFFERNAN

Dept. of Computer Science, Worcester Polytechnic Institute, Worcester, MA 01609

{mfeng|nth}@cs.wpi.edu

Abstract. Limited classroom time available in middle school mathematics classes forces teachers to choose between assisting students' development and assessing students' abilities. 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 (refer to (Razzaq et al., 2005) for more details about the ASSISTment system). In this paper we describe the types of reports that we have designed and implemented that provide real time reporting to teachers in their classrooms. And experiment analysis tools are available to facilitate researchers to carry out randomized controlled learning experiments so that they are able to compare different tutoring strategies. Additionally, reports to principals are in progress. This reporting system is robost enough to support the 2000 students currntly using our system.

1. Introduction

Given the limited classroom time available in mathematics classes, teachers are compelled 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 called the "ASSISTment" system (Razzaq et al., 2005) that will offer instruction to students while providing a more detailed evaluation of their abilities to the teachers than is possible under current approaches. The plan is for students to work on the ASSISTment website for about 20 minutes per week. Every time when students work in the system, the system "learns" more about their abilities. Students' performance is tracked by the reporting system which will provide live online reports to inform teachers about students' learning results.

1.1 The Massachusetts Comprehensive Assessment System (MCAS)

MCAS is a high-stakes testing system required by the *No Child Left Behind Act*. In Massachusetts, MCAS is the graduation requirement in which all students in state educated with public funds in the tested grades are required to participate. It is administered as standardized test that produces rigorous tests in English, math, science and social studies for grades 3 to 10 every year. Students need to pass the math and English portions of the 10th grade versions in order to get a high school diploma. Because students are more likely to fail the mathematics portion of the test, the state is focusing efforts on mathematics. The state of Massachusetts has singled out student performance on the 8th grade math test as an area of highest need for improvement (see http://www.doe.mass.edu/mcas/2002/results/ summary.pdf). Therefore, 8th grade math became where the ASSISTment project started to help students get better prepared for the MCAS. In Massachusetts, the state department of education has released 8 years worth of 8th grade MCAS test items on math, totalling over 300

items, which we have turned into assistments by adding "tutoring". An example of the MCAS test item can be seen in Figure 1 (without the break-down questions).

1.2 Background on the ASSISTment system

The ASSISTment system is an e-learning and e-assessing system that is about 1.5 years old. In the 2004-2005 school year some 600+ students used the system about every two weeks. Eight math teachers from two schools would bring their students to the computer lab, at which time students would be presented with randomly selected MCAS test items. If students got the item correct they were given a new one. If they got it wrong, they were provided with a small "tutoring" session where they were forced to answer a few questions that broke the problem

Figure 1 An ASSISTment shown just before a student hits the "Submit" bottom, showing 2 scaffolding questions, one buggy message and a hint message that can occur at different points.

down into steps. The key feature of assistments is that they provide instructional assistance in the process of assessing students. Razzaq et al. (2005) addressed the learning outcome of the system and some evidence was shown that the students were learning due to the instructional assistance within the system. Though *learning* has been one of the focus points of our research, detailed discussion of the learning effect is beyond the scope of this paper.
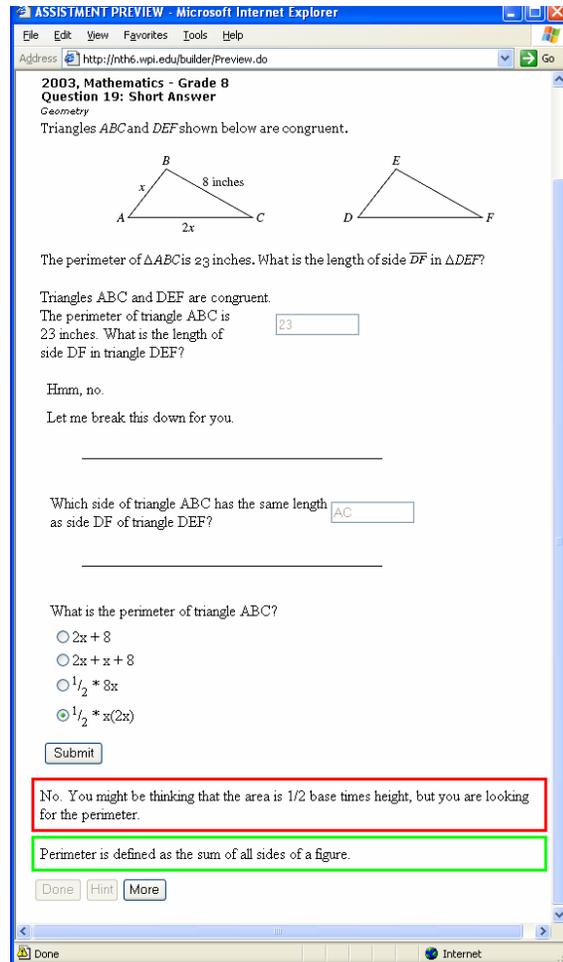
Each assistment consists of an original question and a list of scaffolding questions. An assistment that was built for item 19 of the 2003 MCAS is shown in Figure 1. In particular, Figure 1 shows the state of the interface when the student is partly done with the problem. The first scaffolding question appears only if the student gets the item wrong. We see 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. Students work through 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 would be the definition of congruence in this example. If the student hits the hint button again, the second hint 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. Buggy messages will show up if the student types in a wrong answer as expected by the author. Figure 1 shows a buggy messages that appeared after the student clicked on "½*x(2x)" suggesting he might be thinking about area. Once the student gets this question correct he will be asked to solve 2x+x+8=23 for 5, which is a scaffolding question that is focused on equation-solving. So if a student got the original item wrong, what skills should be blamed? This example is meant to show that the ASSISTment system has a better chance of showing the utility of fine-grained skill modeling due to the fact that we can ask scaffolding questions that will be able to tell if the student got the item wrong because they did not know congruence versus not knowing perimeter, versus not being able to set up and solve the equation. Most questions' answer fields have been converted to text entry style from the multiple choice style they originally appear as in the MCAS tests. As a matter of logging if the student got an original question right or wrong, the student is only marked as getting the item correct if they answered the questions correctly before asking for any hints or encountering scaffolding.

At present, we are focused on 8th grade mathematics and certain amount of content (about 50 assistments, 2 hours' work) for 10th grade mathematics has been released. Though, we believe the system is flexible enough to be used to build tutors for other subjects, such as English, and physics. Our supporting website "www.assistment.org" has been running for 2 and a half years, providing more than 400 assistments built using our online authoring tools (Turner, Macasek, Nuzzo-Jones, Heffernan, & Koedinger, 2005; Heffernan et al., 2006) and over 2000+ students from 20+ teachers from 5 schools were using the system every two weeks during the school year of 2005 to 2006.

1.3 Why we need a new reporting system beyond MCAS reports?

Schools seek to use the yearly MCAS assessments in a data-driven manner to provide regular and ongoing feedback to teachers and students on progress towards instructional objectives. But teachers do not want to wait six months for the state to grade the exams. Teachers and parents also want better feedback than they currently receive. While the number of mathematics skills and concepts that a student needs to acquire is on the order of hundreds, the feedback on the MCAS is broken down into only 5 mathematical categories, known as "Strands". However, a detailed analysis of state tests in Texas (Confrey, Valenzuela & Ortiz, 2002) concluded that such topic reporting is not reliable because items are not equated for difficulty within these areas. To get some intuition on why this is the case, the reader is encouraged to try the item shown in Figure 1. Then ask yourself "What is the most important thing that makes this item difficult?" Clearly, this item includes elements from four of the 5 "strands" Algebra, Geometry (congruence), Number Sense (arithmetic operations) and Measurement (perimeter). Ignoring this obvious overlap, the state chose just one strand, Geometry, to classify the item, which might also be the first feeling of most people. However, as we will show below, we have found evidence there is more to this problem. The question of tagging items to learning standards is very important because teachers, principals and

superintendents are all being told to be "data-driven" and use the MCAS reports to adjust their instruction. As a teacher has said "It does affect reports... because then the state sends reports that say that your kids got this problem wrong so they're bad in geometry-and you have no idea, well you don't know what it really is- whether it's algebra, measurement, or geometry."

There are several reasons for this poor MCAS reporting: 1) the reasonable desire to give problems tap-multiple knowledge components (knowledge component is the way we refer to "strand" or "skill" in our system), 2) the fact that paper and pencil tests cannot figure out, given a student's response, what knowledge components to credit or blame, 3) there are knowledge components that deal with decomposing and recomposing multi-step problems, yet are currently poorly understood by cognitive science. So a teacher cannot trust that putting more effort on a low scoring area will indeed pay off in the next round of testing.

The reporting in the ASSISTment system was 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 the detailed feedback to tailor their instruction to focus on the particular difficulties identified by the system. Compared to the MCAS reports, reports provided by the reporting in the ASSISTment system is live so that teachers do not need to wait. We have built more multi-mapping models that allow one problem to be tagged with multiple knowledge components and finer grained models that break down the 5 strands into about 100 knowledge components and code the problems (also the scaffolding questions) with the new knowledge components. Moreover, the reporting system provides more performance analyzing tools for teachers and school principals to make comparison among different groups and to run learning experiments.

The remainder of the paper is organized as follows. Section 2 discusses related work. Section 3 discusses the data source we used in our reporting system. Then in Section 4, we introduce the transfer models we have built and related work that has been done using the new transfer models. Different reports for teachers will be shown in Section 5 and we provide

teachers' feedback at the end of Section 5 too. Reports for principals are discussed in Section 6. And the experiment analysing tools are discussed in Section 7. We conclude in Section 8.

2. Related Work

Many researchers have been interested in constructing assessment/tutoring systems on different subjects, many of which provide the similar tutoring functionality as the ASSISTment system and various reports to teachers to help instructing student learning.

Measures of Academic Progress (MAP - http://www.nwea.org) are state-aligned computerized adaptive tests provided by the Northwest Evaluation Association (NWEA) and it is also the most commonly used assessment system by Worcester Public Schools. MAP covers subjects other than math and gives similar online reports such as class rosters, student progress report, class by subject report to educators to guide their instructions. Unlike the ASSISTment system, as an assessment system, MAP provides no tutoring to help student learning and it sticks to the strands and categorization given by the state. Therefore, it lacks the ability to analyze a problem in further detail. The Online Learning Initiative (OLI - http://www.cmu.edu/oli/) from Carnegie Mellon University provides a collection of online tutors directed at many subject areas. While the OLI provides a wide range of online tutors, the tutors lack extensibility to other tutor types and domains, resulting in a high cost for creating content. Cognitive Tutors (Koedinger et al., 2004), created by LearnLab (http://www.learnlab.org/), also provide tutoring in addition to being extendable to other domain or content. They have been successful in raising students' math test scores in high school and middle-school classrooms. Authoring tools, named, CTAT, are provided to make content creation easier for experts and possible for novices in cognitive science. However, the cognitive tutors lacks the administrative tools necessary for non-experts to effectively manage the system and they are not web-based and do not provide comprehensive reports about students' progress. The National Center for Research on Evaluation, Standards and Student

Testing (CRESST) (Vendlinski et al., 2005) provides an online system (not limited to math) and has a collection of tools to support the creation and distribution of content. However the CRESST system does not offer tutoring, instead it allows for open ended questions that are then evaluated by a human teacher, nor does the CRESST system provides reports for teachers. Effective Educational Technologies (EET) developed a series of online assessment and tutoring programs (MasteringPhysics - http://www.masteringphysics.com/, MasteringGeneralChemistry, and MasteringAstronomy) together with the authoring tools for content creation. Most like in the ASSISTment system, with the mastering program, students receive feedback based on common wrong answers and misconceptions. By capturing the step-by-step difficulties of individual students, the Mastering platform responds to each student with individualized hints and instructions. The program provides tools to find problems of the wanted type, topic coverage, and level (functioned as problem difficulty report as described in Section 4) and to monitor class/student performance via a "gradebook"; tracks students' work on the sub-problems (similar to the scaffolding questions in assistments) and awards partially credit when evaluating students' performance. MasteringPhysics has been widely used as homework system while the ASSISTment project just started its first step into the picture. LON-CAPA (http://www.lon-capa.org/) is a special assessment system because of its distributed learning content management that allows the sharing of assessment materials across institutions and disciplines. It provides assessment analysis gives an overview of how students are performing in the courses. The report shows all the attempts made by a student on each problem and it can also analyse one problem across all students, which is rather simple, comparing reports in the MAP, MasteringPhysics, or the ASSISTment system. Although many of above systems provide reports for teachers, none of them offer reports for principals and tools for researchers to conduct learning experiments and analyze learning effects, as described in Section 6 and 7.

3. Data Source

The ASSISTment system is deployed with a completely internet savvy solution whereby students can simply open a web browser and login in to work on the problems. Our Java-based runtime system (Nuzzo-Jones, Walonoski, Heffernan, & Livak, 2005) will post each student's actions (other then their mouse movements) to a message server as an xml message that includes action timestamp, student ID, problem ID, student's action type (did they attempt or just ask for help), student's input and response, etc. The messages will be stored in the database server at WPI. As mentioned above, about 800 students of 9 teachers have been using the ASSISTment system every other week during the school year of 2004 - 2005. Currently log records in our database show that about 120,000 MCAS items have been done and more than 1,500,000 actions made by these students. Since students are arranged to use our system regularly, our database will continually receive new data for the students. This allows our reporting system to assess students' performance incrementally and give more reliable assessment as time goes on. These large amounts of student data also offer valuable material for further learning analysis using data mining or statistical techniques.

4. Transfer Model

A transfer model (Croteau, Heffernan, & Koedinger, 2004) is a cognitive model that contains a group of knowledge components and maps existing questions (original items and scaffolding questions) to one, or more of the knowledge components. It also indicates the number of times a particular knowledge component has been applied for a given question. It is called a "transfer model" since we hope to use the model to predict when learning and knowledge transfer will happen. Also as a predictive tool, transfer models are useful in selecting the next problem to work on. In the next section, we will show that transfer models are quite important for quality reporting.

Massachusetts Curriculum Frameworks breaks the 5 strands (will be referred to as the MCAS-5) (Patterns, Relations and Algebra; Geometry; Data Analysis, Statistics and Probability; Measurement; Number Sense and Operations) into 39 "learning standards" for 8th grade math and tags each item with one of the 39 standards. As we have shown in Figure 1, item 19 from Year 2003 has been tagged with "G.2.8 Congruence and similarity", the 2nd learning standard in the **G**eometry strand.

We have made several attempts of using the 39 MCAS learning standards (will be referred to as the MCAS-39) to "code up" items, first using the state's mapping with one standard per question, and then with our own coding which allows each question to be tagged with multiple standards. However, we could not get statistically reliable coefficients on the learning standards. So we hypothesise that a finer grained model would help. Additionally, we need a more detailed level of analysis for reporting to teachers and for predicting students' responses on questions. WPI-106 is a much finer-grained transfer model we have created in WPI with 106 knowledge components. In the model, knowledge components are arranged in a hierachy based on prerequisite structure. So far, 78 knowledge components in this transfer model have been used to tag the assessments, together with all the scaffolding questions, in our system. Tagging the scaffolding questions enable us to assess individual knowledge components instead of only overall performance. Mappings between WPI-106 and the Massachusetts Curriculum Frameworks have been constructed by nesting a group of fine-grained knowledge components into a single category in a coarse model. Table 1 shows the hierarchal nature of the relationship between WPI-106 and the models in Massachusetts Curriculum Frameworks.

Table 1 – Knowledge Components Transfer Table

WPI-106	MCAS-39	MCAS-5
Inequality-solving Equation-Solving Equation-concept	setting-up-and-solving-equations	Patterns-Relations-Algebra
…	…	
Plot Graph	modeling-covariation	
X-Y-Graph Slope	understanding-line-slope-concept	
…	…	…
Congruence Similar Triangles	understanding-and-applying-congruence-and-similarity	Geometry
…		
…	…	
Perimeter Circumference Area	using-measurement-formulas-and-techniques	Perimeter
…		
…	…	

Consider the item in Figure 1, which had the first scaffolding question tagged with "congruence", the second tagged with perimeter, the third tagged with equation-solving. In the MCAS-39, the item was therefore tagged with "setting-up-and-solving-equations", "understanding-and-applying-congruence-and-similarity" and "using-measurement-formulas-and-techniques". The item was tagged with three skills at the level of the MCAS-5.

At present, we are able to generate reports based on Massachusetts Curriculum Framework, as well as the WPI-106 transfer model which reveals more detailed information about students' knowledge learning and knowledge components contained in problems. Our most recent research work (Feng, Heffernan, Mani, & Heffernan, 2006, in press; Pardos, Heffernan, Anderson & Heffernan, 2006, in press) shows that WPI-106, as a finer-grained cognitive model, can produce better tracking of student performance than MCAS-5 as measured by ability to predict student performance on MCAS test.

5. Reporting System for Teachers

5.1 Student Grade Book Report

Right now, we only have anecdotal information from our teachers that they find the reporting helpful. Teachers seem to 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. Standard
			# Done	# Correct	% Corr.	MCAS Score*	Perf. Level	# Done	# Correct	% Correct	# Hint Req.	
	34	0	15	3	20%	200	Failing	30	16	53%	15	N.1.8-understanding-number-representations (Error times: 5/6)
	32	0	38	26	68%	242	Proficient	81	56	69%	4	P.1.8-understanding-patterns (Error times: 2/6)
	33	0	20	9	45%	220	Needs improv.	63	28	44%	63	P.1.8-understanding-patterns (Error times: 8/10)

Figure 2: Grade Book on real student data

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 . Our "prediction" of a student MCAS score is at this point primitive. The column is currently simply a function of percent correct. We might even remove these two columns related to MCAS score prediction until we feel more confident in our prediction, in another word, "rough and ready". In our past research, we have found a strong correlation between our prediction for the 68 students who have used our system May 2004 and their real MCAS raw score (r = .7) (Razzaq et al., 2005). And we were continually refining our prediction function based on new data (See Feng, Heffernan, & Koedinger, 2006a, 2006b). In these works, we showed that we were able to predict students' MCAS score pretty well with Mean Absolute Difference being 5.533 out of a full score of 54 points. Besides presenting information on the item level, it also summarizes the student's actions in "ASSISTment

metrics": how many scaffolding questions have been done, student's performance on scaffolding questions and how many times the student asked for a hint. The "ASSISTment metrics" are good measurements of the amount of assistance a student needs to finish a problem. Feng, Heffernan, & Koedinger (2006a, 2006b) found evidence showing that the ASSISTment system, as an online assessment system, can do a better job of predicting student knowledge by being able to take into consideration how much tutoring assistance was needed. In addition, the "ASSISTment metric" tells more about students' actions besides their performance. For example, it exposes students' unusual behaviour like making far more attempts and requesting more hints than other students in the class, which might be evidence that students did not take the assistments seriously or was "gaming the system" (Baker, Corbett, & Koedinger, 2004; Walonoski, & Heffernan, 2006).

In Figure 2, we see that these 3 students have used the system for about 30 minutes. (Many

Figure 3. Items tagged with difficult knowledge component

students have used it for about 250 minutes during the school year of 2004 - 2005). "Dick" has finished 38 original items and only asked for 4 hints. Most of the items he got correct and thus our prediction of his MCAS score was high. We can also see that he has made the greatest number of errors on questions that have been tagged with the standard "P.1.8 understanding patterns". The student had done 6 problems tagged with "P.1.8" and made errors on 2 of those problems. Teachers can also see "Harry" has asked for too many hints (63 compared to 4 and 15). Noticing this, a teacher could go and confront the student with evidence of gaming or give him a pep-talk. By clicking the student's name shown as a link in our report, teachers can even see each action a student has made, his inputs and the tutor's response and how much time he has spent on a given problem (which we will not present here for lack of space). The "Grade

Book" is so detailed that a student commented: "It's spooky", "He's watching everything we do" when her teacher brought students to his workstation to review their progress.

By clicking the link of the most difficult knowledge component, the teacher can see what those questions were and what kind of errors the student made (See Figure 3). Knowing students' reactions to questions helps teachers to improve their instruction and enable them to correct students' misunderstandings in a straightforward way. Finding out students' difficult knowledge components also offers a chance to improving our item selection strategy. Currently, random and linear are the only two problem selection strategies supported by our runtime system. Another option could be added if we can reliably detect difficult knowledge components of each individual student, which requires the runtime system to preferentially pick items tagged with those hard knowledge components for the students so that students would have more opportunity to practise on their weak point.

5.2 Reports by Knowledge Component

Tagging questions with knowledge components in different transfer models enables us to develop reports to inform teachers about the knowledge status of classes and of individual student. "Class Summary" report and "Student-Level Knowledge Component" report were developed for this purpose. As shown in Figure 4, teachers can select their favourite transfer model, specify the number of knowledge components to be shown in the report. Knowledge components are ranked according to their correct rate which is students' correct rate (demonstrated in Figure 4 as green bars together with percent correct as values) at the items tagged with those knowledge components. By clicking the name of a knowledge component (shown as a hyperlink in Figure 4), teachers are redirected to another page showing the items tagged with the knowledge components. In the new page, teachers are able to see the question text of each item and continue to preview or analyze the item if they want to know more about the item.
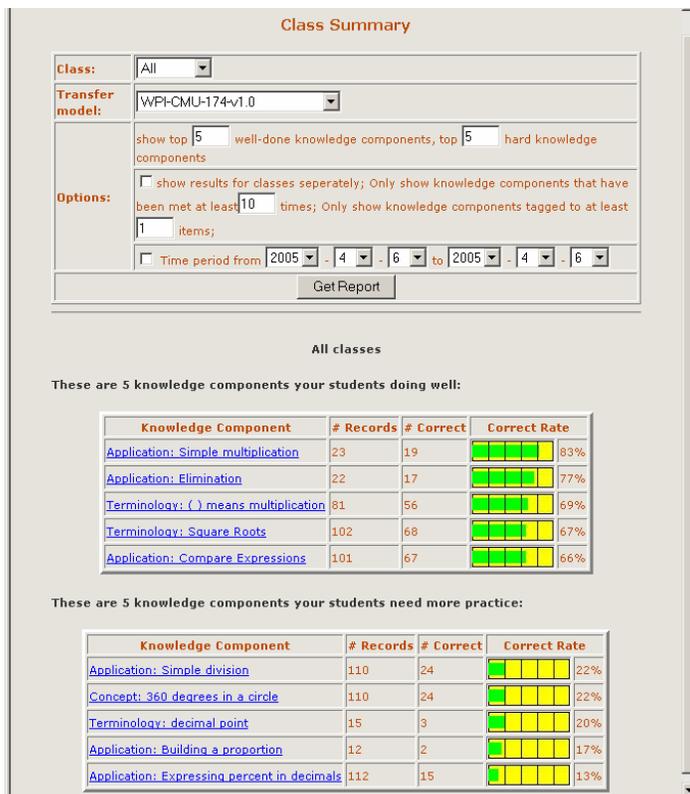
Figure 4 Class summary report for a teacher's classes

By presenting such a report, we hope we can help teachers to decide which knowledge components and items should be focused on to maximize the gain of students' scores at a class level when instructional time is limited. We would like to evaluate the effectiveness of the report by comparing the learning gain in a limited time of the classes for which the teachers have been exposed to this report to the control groups for which this report is not accessible.

In addition to the class level knowledge component report, we present a student level report (developed by Quyen Do Nguyen at WPI) to teachers which shows the knowledge status of individual students. Similar to the class level report, strong and weak knowledge components are listed, but only for the particular student specified by the teacher. The student level knowledge component report comes after the class level report and quickly becomes a favourite report of our cooperating teachers. Teachers love the fact that they can see in this report the weak points of a particular student in their classes so that they will pay more attention to those knowledge components when giving instructions to the student. Since both original items and scaffolding steps have been tagged in different grain sized transfer models in the ASSISTment system, we claim that we can more accurately detect what are the real obstacle knowledge components for each student.

Class	Date	# Correct	# Total	# Student	Avg. Score	Std. Dev.
Period 3	2004-09-21	153	382	23	18	9.95
Period 3	2004-10-27	427	773	23	25	11.18
Period 3	2004-11-10	630	1119	24	26	11.03
Period 3	2004-12-01	879	1437	22	29	10.20
Period 3	2004-12-15	1167	1790	21	32	8.24
Period 3	2005-02-02	1341	2029	20	33	7.96
Period 3	2005-02-16	1702	2576	23	33	6.67
Period 3	2005-03-02	1972	3065	24	33	6.61
Period 3	2005-03-16	2106	3288	23	33	6.58

Figure 5. Preliminary progress report for a class

5.3 Class Progress Report

Since our teachers let their students using the ASSISTment system every two weeks, we thought it would be helpful for teachers to track the change of students' performance if we can show to teachers students' progress by looking at their performance at each time they worked on the assistments.

Figure 5 shows our preliminary progress report for a teacher's class. In this report, we can see this class has been using our system since September 21st, 2004 and has used it as a class 9 times. The average of students' predicted MCAS raw score increased from 18 to 33, and kept being 33 for a while. [Note, we are being conservative in calculating these predicted MCAS scores, in that we calculate for each students their predict scores using every items them have even done in our system, instead of 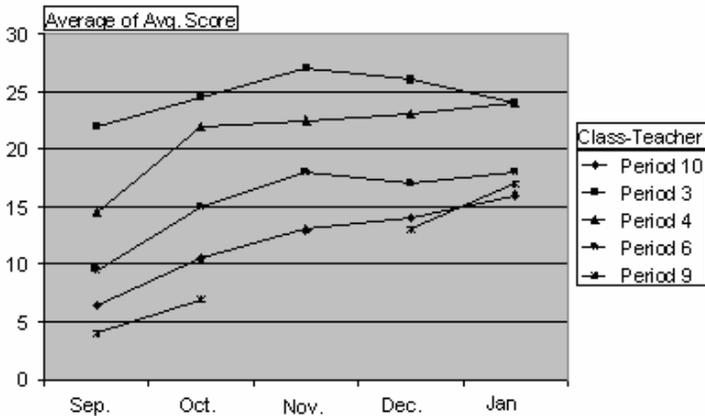using only the items done on day they came to the lab.] Standard deviation of scores is also displayed as a column to help teachers see performance variance in the class.

Figure 6: Predicted MCAS Score over months

The progress of students' predicted MCAS raw score over months is more clearly shown in Figure 6. Those students of the five different classes (all from school A) have

been using our system for more than 5 month starting from Sep., 2004. We can see in this graph students' predicted MCAS score on average increase steadily with month passing (even for class "Period 9" which "left" us for two months). In our recent work, the nicely "time-tagged" progress data (at student level) has been used to construct longitudinal models and thus track students' learning over time (See Feng, Heffernan, & Koedinger, 2006a).

5.4 Analysis of Items

Item 20 N-2003 Morph (3/4 of 1 2/3)	24%
Item 20 N-2003 (2/3 of 1 1/2) Morph2	26%
Item 18 G-1998 (Angle in isosceles triangle)	27%
Item 35 G-2001 (Angle between clock hands)	27%
Item 13 D-1998 (Eiffel Tower model)	29%

Figure 7: Problems order by correct rate

A report is built to show difficulty each problem in our system. (See Figure 7: 5 lines of the 200+ lines that are in the report). By breaking original items into scaffolding questions and tagging scaffolding questions with knowledge components, we are able to analyze individual steps of a problem. Figure 8 is what we call a scaffolding report because it reports statistics on each of the scaffolding questions that are associated with a particular original item.

Figure 8: A scaffolding report generated by Assistment reporting system

On the first line of Figure 8, we see this problem is hard since only 12% of the students got it correct on their first attempt. Of the 180 students having done this item so far, 154 students could not get the correct answer to the original question, thus forced by the system to go through scaffolding questions to eventually solve the problem. One may notice that 154 is less than 88% of 180, which should be about 158. And the number of attempts on later scaffolding questions went down more. That's because students could log out and log back in to redo the original question to avoid going through all scaffolding questions. This problem has been solved. 56% of students asked for a

hint, telling you something about students' confidence when confronted with this item. (It is useful to compare such numbers across problems to learn which items students think they need help on but don't, and vice versa). Remember that the state classified the item according to its "congruence" (G.2.8) shown in bold. The other MA learning standards (M.3.8, P.7.8) are the learning standards we added in our first attempt to code using the MCAS 39 standards. We see that only 23% of students that got the original item incorrect can correctly answer the first scaffolding question lending support to the idea that congruence is tough. But we see a as low percent correct 25% on the 3rd question that asks students to solve for x. The statistics result gives us a good reason to tag "P.7.8-setting-up-and-solving-equations" to the problem. Teachers want to know particular skills or knowledge components that cause trouble to students while solving problems. Unfortunately the MCAS is not designed to be cognitively diagnostic. Given the scaffolding report can provide lower level of cognitive diagnosis, our cooperating teachers have carefully designed scaffolding questions for those tough problems to find out the answer. For example, one teacher designed an assistment for ("What's ¾ of 1 ½?"), item 20 of year 2003 8th grade MCAS. The first scaffolding question for the assistment is "what mathematical operation does the word 'of' represent in the problem". This teacher said, "Want to see an item that 97% of my students got wrong? Here it is… and it is because they don't know 'of' means they should multiply." The report has confirmed the hypothesis. 40% of students could not select "multiplication" with 11 of them selecting "division".

Table 2: "Red & Green" distribution matrix

Q0								154								22
Q1			119						35								
Q2		85			34			12			23				na		
Q3	72		13		21		13		8	4		18		5			
Q4	45	8	5	7	15	6	3	10	6	2	1	3	15	3	1	4

The scaffolding report has helped us to develop our tutors in an iterative way. For each question, the report shows top common errors and corresponding "buggy" messages. When building the Assistments, we have tried to "catch" common errors students could make and

give them instructive directions based on that specific error, such as correcting students' misunderstanding of question texts or knowledge concepts. But given that students may have different understandings of concepts, assistments may give no messages for some errors, which means our tutor lost chances to tutor students. Also, students may feel frustrated if they are continually being told "You are wrong" but get nothing instructive or encouraging. As shown in Figure 8, the wrong answer "15" to the third question has been given 13 times, but the assistment gave no instructive messages. Noticing this, the assistment builders can improve their tutor online by adding a proper "buggy" message for this error.

We also display a table that we call "Red & Green" distribution matrix as shown in Table 2 in the scaffolding report. Numbers in the cells show how many students got correct (indicted by green number in un-shaded cells) or wrong (indicated by red in shaded cells) on a question. We split the number as the questions' sequence number grows so that it also represents how those students have done on previous questions. In this example, we see that 4 students who have answered the original question wrong went through all of the scaffolding questions correctly. Given that, we tend to believe those students have mastered the knowledge components required by each step and but need instruction on how to "compose" those steps. It's also worth pointing out that there are 8 students who answered original question wrong but answered correctly to the last question, which asks the same question as the original one. Since the assistment breaks the whole problem into scaffolding steps and gives hints and "buggy" messages, we would like to believe those students learned from working on the previous steps of this assistment.

5.5 Performance evaluation

Our reporting system was first used in May, 2004. In the early stage, it worked well and most reports at the class level could be generated in less than 10 seconds. And it took 10 to 20 seconds to generate a scaffolding report at "system" level, depending on the number of

students who have worked on the item and the number of scaffolding questions the item has. The performance went down when the number of recorded student actions increased past 1 million. In particular, we have seen the "Grade Book" report took more than 2 minutes, which we consider unacceptable as a live report. We then switched to Oracle database which provides mechanisms, such as view, stored procedure, to improve query performance. We also updated the approaches we used to generate the reports. Now we can generate the "Grade Book" report in about 7 seconds on average. The time required to generate the system level scaffolding report for Item 19 (See Figure 8) is about 5 seconds.

5.6 Teachers' attitude towards the system

Nice things have been mentioned about the system by our cooperating teachers. To collect usage feedback from teachers, we created an online survey of teacher attitude about the ASSISTment system and how do they use the data from the system during the school year 2005 to 2006. The responses are positive. Teachers in general liked the feature that the assistments lead students step by step when they incorrectly answered a question and "it was great to have the 'hints' that are tailored to their individual needs". They also consider using the system as a good MCAS practice and loved the fact that they can receive scores at the end of the class. Among the 11 teachers who responded, 8 teachers strongly agreed that they thought their students learned by using the system and 3 agreed somewhat. And nine of the teachers would consider assigning assistment problems as homework for students with computers at home.

We noticed a discrepancy that although eight of the 11 teachers thought the data provided the system was helpful, only 3 teachers said that they did use the data to change what they did instruction in class while 7 others mentioned that they only did this somewhat. We hypothesize that one reason for this difference can be the availability of the reports. Most teachers are not accustomed to frequently log in to the system to access the reports on their

own. To some teachers, doing this also adds extra work. Actually when being asked the opinion on receiving automatic email reports, 9 teachers thought that would be great since it would be "a much easier and faster way of obtaining the information" and it would eliminate work for them so that allow them "more time to focus on certain strategies or concepts in class". Developers at WPI are now working on automatically generating and emailing certain reports, as described below in Section 5.

Another thing we care about is how the teachers use the reports. In the survey, most teachers only mentioned that they reviewed common mistake problems with the whole class, which indicated that many functions provided the reporting system have been ignored. Again, availability of the reports might be one explanation. Another reason, we speculate, can lie in the fact that different reports in the system are not quite well organized and there is no demonstration or function specification on the website to help people get started. One teacher did say that she/he was not able to using the data until she/he was shown (by the second author) step by step on how to retrieve the information and then how to make use of it. We are now seeking better communication approach to help teachers discover real values of the reports.

6. Reporting for Principals and Related Results

Most of the reports described in Section 5 were for normal teachers. As a supplementary to those reports, we have been working on new reports for principals and administrators which will allow them to see which groups of students need most attention on a wider scope across teachers/classes, based on their gender, special education status, if they get free-lunch and if they are underrepresented. Given these reports, users can also compare teachers and see that which schools/teachers produced more learning than others. The reports are made possible by the fact that we have trained up longitudinal data analysis models (Feng, Heffernan, & Koedinger, 2006a). Though the reports themselves are still under development, we describe the

supporting longitudinal data analysis approach for the reports and show the results we got based on the data collected during the school year of 2004 – 2005.

"Singer and Willett" (Singer & Willett, 2003) style longitudinal data analysis is an approach for investigating change over time; in this case, the change of students' performance over the course year. It allows us to learn a slope that represents a student's learning rate and intercept that represents the estimate of incoming knowledge for each individual student. This is achieved by fitting a multilevel statistical model (also referred to as mixed-effects model) that simultaneously builds two sub-models, in which level-1 sub-model fits *within-person* change and describes how individuals change over time and level-2 sub-model tracks *between-person* change and describes how these changes vary across individuals.

We applied the longitudinal analysis approach on the log data of 324 students coming from 8 different teachers' classes of two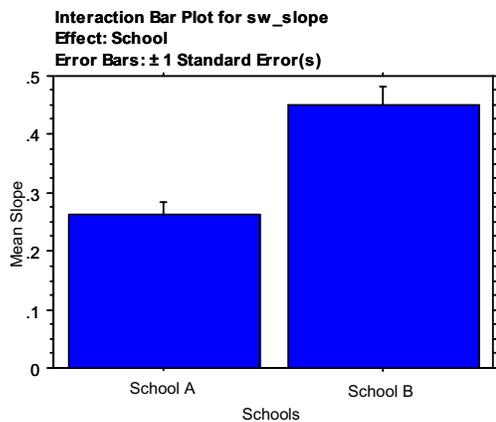 schools and obtained the slope (i.e. the learning rate) and intercept (i.e. the incoming knowledge) for each student. For all these students, we record certain characteristics such as their gender, special education status, if they get free-lunch and if they belong to traditionally underrepresented groups.

Figure 9: Compare learning rates of schools

The first thing we want to test is whether the students from the two schools differ by their learning rate. Before doing this, we noticed that for schools there was a clear difference in incoming students' scores, which makes sense with regard to the fact that one school draws students from the more affluent side of town. We then ran an ANOVA on the slope introducing *school* as the effect. The result showed a p-value of smaller than 0.0001 with an effect size of 0.595 (See Figure 9), which suggested that one school has caused more learning in students than the other.

Then we switched to compare teachers. We did an ANOVA using *teacher* as a factor and got a p-value that was statistically significant ($p < 0.0001$). This result led us to conclude that some teachers have done a better job helping student learning than others.

The next thing is to investigate which groups of students have shown more knowledge gain over the same period of time as measured by their learning rate. We are especially interested in questions such as "Which group is better at learning math, boys or girls?", "Does students from under/over-represented groups show different rates of learning on math?" To answer these questions, we tried different factors in ANOVA, namely gender, under/over-represented, special education status, free-lunch or not. It turned out that for the selected data, none of these factors are statistically significant ($p > 0.05$). Among all these tests, the difference in slope parameter of free-lunch was near significance ($p = 0.08$) and suggested that the students who got free lunch showed more learning than those who did not. Given this, we went ahead to test if there are difference in the incoming scores of these two group of students and found out that students who did not get free lunch started with a significant higher score (3.33 points higher, $p < 0.0001$). This is consistent with a general fact that we found before, that is, groups with higher estimated initial scores showed lower rates of learning. Our preliminary speculation on this fact is that 1) this may be attributed to the "ceiling effect": it is hard for top students to make fast progress; 2) good students were assigned to Algebra class and learning content that won't be tested until 10th grade and won't appear in the ASSISTment system. Further investigation needs to be done to explain this phenomenon. Currently, we are working on automating all the above analyses and implementing the corresponding reports.

7. Reporting as Learning Experiment Tools for Researchers

The ASSISTment system allows randomized controlled experiments to be carried out (Razzaq et al., 2005) fairly easily. There is control for the number of items presented to a student, but the system will be able to support control for time soon. 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).

Researchers can select items to put into the experiment curriculums, and then assign them to classes. Below shows a real experiment (Razzaq et al., 2005) that was designed to compare two different tutoring strategies when dealing with proportional reasoning problems and investigated whether students would learn better if asked to set up proportions. The item is 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?" The author built two different assistments that differed only by one extra scaffolding question. One of the conditions involved coaching the students to solve the problem by first setting up the proportion, while the other one did not use the formal notion of proportion. The author made a second "morphed" version of each by changing the "cover story". Finally, the author selected two items to posttest for "far transfer" (See Figure 10). Students participating in the experiment will be randomly assigned
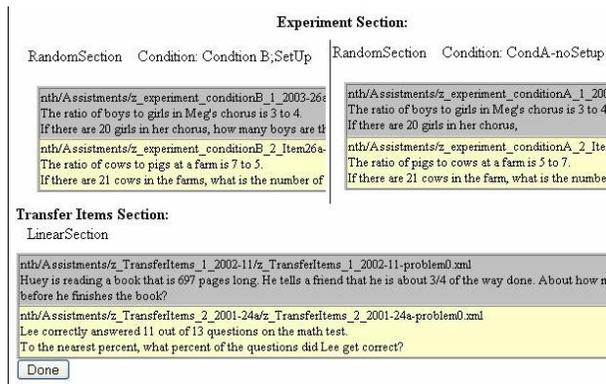

Figure 10: An experiment curriculum

Your Current Experiments Being Analyzed:				
Report Name	Last Updated			
Does Scaffolding Work?	Fri Jul 01 11:59:56 EDT 2005	Cancel This Report	Run Live Analysis	Edit
Practice Experiment	Thu Jun 30 13:37:46 EDT 2005	Cancel This Report	Run Live Analysis	Edit
for jay	Thu Jun 30 16:05:56 EDT 2005	Cancel This Report	Run Live Analysis	Edit

Other User's Active Reports:			
Report Name	Curriculum	Last Updated	Owner
dateTest	MA/curriculum_IgorExperimentv03.xml	Thu Jun 30 13:37:44 EDT 2005	1695
AnalysisThree	MA/curriculum_IgorExperimentv03.xml	Thu Jun 30 13:37:47 EDT 2005	1695
example	MA/curriculum_IgorExperimentv02.xml	Thu Jun 30 15:12:12 EDT 2005	1695
MyNextAnalysis	MA/curriculum_IgorExperimentv02.xml	Thu Jun 30 13:37:49 EDT 2005	1695

Figure 11: Online experiment analysis report

to either condition. After they finished the first two items in the random section, they all will encounter the "far transfer" items as posttest. Participants' performance on the posttest as well as on the second item in condition will be utilized to evaluate the effectiveness of different tutoring strategies.

The experiment set-up/analysis tools (implemented mainly by Shane Gibbons and Emilia Holban at WPI) were developed to facilitate the running of experiments. The set-up tool allows researchers to schedule when they want to be notified of the results of their experiments during/after the experiments have been carried out. They can get daily, weekly or monthly reports of the situation of their experiments and the notification can also be set up based on the statistically significance (the effect size, p-value, or the number of subjects who have participated in the experiments). If they like, users can type in their email address and have the reports ready in their mail box when the analysis is done.

After the experiments were set up and run, the system automatically does the analysis and presents the reports online (See Figure 11) or sends the results to users' mail box according the settings. There are two types of analyses the project is interested in full automating. The first is to run the appropriate ANOVA to see if there is a difference in performance on the transfer items by condition, and the second is to look for learning in the condition, and see if there is a disproportionate amount of learning by condition. Figure 12 shows the "SetupRatio" condition to have better learning within the condition as well as better learning on the posttest/transfer items (reported in Razzaq et al., 2005)
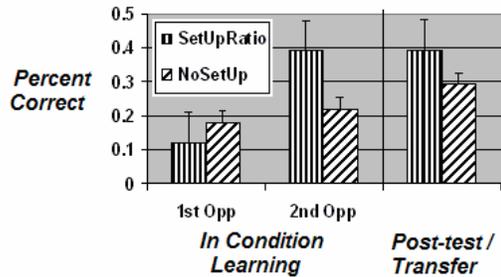
Figure 12: Learning results

Different kinds of experiments have been run in the ASSISTment system. In addition to the one as described above that investigates how different coaching strategies affect learning, experiments have been run to answer the question that "Are scaffolding questions useful compared to just hints on the original questions". The survey results 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. Thus, a simple experiment was conducted to find the answer, and the results showed that students that were given the scaffolds performed better although the results were not always statistically significant (Razzaq, & Heffernan, 2006).

8. Conclusions

In conclusion we feel that we have developing some state-of-the-art online reporting tools that will help teachers and researchers be better informed about what their students know. Our implicit evaluation is that we have made it possible for all these reports to work live in the classroom. We feel we have a lot to do in automating yet further the statistical analysis of learning experiments. We have done some learning analysis with this year's data set environing over 800 students and 30 Learning Opportunity Groups. In particular we see students are about 5% on their second opportunity and this was statistically significant (Razzaq et al., 2005). Also since doing learning analysis by hand is both time consuming and fallible, another aim of our reporting system is to automat learning analysis process. We have done some preliminary work towards this direction: let teachers create content, and send them email automatically when we know that their content is better (or worse) than what we are currently using in the assistment systems.

References

Baker, R. S., Corbett, A. T., & Koedinger, K. R. (2004). Detecting student misuse of intelligent tutoring systems. *Proceedings of 7th International Conference on Intelligent Tutoring System*s. Maceio, Brazil.

Collins, A., Brown, J. S., & Newman, S. E. (1989). Cognitive apprenticeship: teaching the crafts of reading, writing, and mathematics. In L. B. Resnick (Ed.), *Knowing, Learning, and Instruction: Essays in Honor of Robert Glaser.* (pp. 453-494). Hillsdale, NJ: Lawrence Erlbaum Associates.

Confrey, J., Valenzuela, A., & Ortiz, A. (2002). Recommendations to the Texas State Board of Education on the setting of the TAKS standards: a call to responsible action. from http://www.syrce.org/State_Board.htm

Croteau, E., Heffernan, N. T., & Koedinger, K. R. (2004). Why are algebra word problems difficult? using tutorial log files and the power law of learning to select the best fitting cognitive model. *Proceedings of the 7th International Conference on Intelligent Tutoring System*. Maceio, Brazil.

Feng, M., Heffernan, N.T, Koedinger, K.R. (2006a). Addressing the testing challenge with a web-based e-assessment system that tutors as it assesses. *Proceedings of the Fifteenth International World Wide Web Conference.* (pp. 307-316). New York, NY: ACM Press.

Feng, M., Heffernan, N.T, Koedinger, K.R. (2006b). Predicting state test scores better with intelligent tutoring systems: developing metrics to measure assistance required. In Ikeda, Ashley & Chan (Eds.), Proceedings of the Eighth International Conference on Intelligent Tutoring Systems. (pp. 31-40). Heidelberg, Germany: Springer Berlin.

Feng, M., Heffernan, N. T., Mani M., & Heffernan C. (2006, in press). Using mixed-effects modelling to compare different grain-sized skill models. AAAI'06 Workshop on Educational Data Mining. Boston. 2006.

Heffernan N. T., Turner T. E., Lourenco A. L. N., Macasek M. A., Nuzzo-Jones G., & Koedinger K. R. (2006). The ASSISTment builder: towards an analysis of cost effectiveness of ITS creation. *Proceedings of the 19th International FLAIRS Conference*. Florida.

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 International Conference on Intelligent Tutoring Systems*. (pp.162-173). Maceio, Brazil.

Mostow J., Beck J. E., Chalasani R., Cuneo A., & Jia P. (2002). Viewing and analyzing multimodal human-computer tutorial dialogue: a database approach. *Proceedings of the Fourth IEEE International Conference on Multimodal Interfaces (ICMI 2002)*.

Nuzzo-Jones, G., Walonoski, J. A., Heffernan, N. T., Livak, T. (2005). The eXtensible tutor architecture: a new foundation for ITS. In C.K. Looi, G. McCalla, B. Bredeweg, & J. Breuker. (Eds.), *Proceedings of the 12th International Conference on Artificial Intelligence In Education*. (pp. 902-904). Amsterdam: ISO Press.

Pardos, Z. A., Heffernan, N. T., Anderson, B., & Heffernan C. (2006, in press). Using fine-grained skill models to fit student performance with Bayesian networks. Workshop on Educational Data Mining held at the 8th International Conference on Intelligent Tutoring Systems, Taiwan.

Razzaq, L, Feng, M., Nuzzo-Jones, G., Heffernan, N.T., Aniszczyk, C., Choksey, S., Livak, T., Mercado, E., Turner, T., Upalekar, R., Walonoski, J., Macasek, M., Rasmussen, K., Koedinger, K., Junker, B., Knight, A., & Ritter, S. (2005). The Assistment project: blending assessment and assisting. In C.K. Looi, G. McCalla, B. Bredeweg,

& J. Breuker (Eds.), *Proceedings of the 12th International Conference on Artificial Intelligence in Education.* (pp. 555-562). Amsterdam: ISO Press.

Razzaq L., & Heffernan, N. T. (2006). Scaffolding vs. hints in the Assistment system. In Ikeda, Ashley & Chan (Eds.), *Proceedings of the Eighth International Conference on Intelligent Tutoring Systems.* (pp. 635-644). Heidelberg, Germany: Springer Berlin.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and occurrence*. New York, NY: Oxford University Press.

Turner, T. E., Macasek, M. A., Nuzzo-Jones, G., Heffernan, N. T., Koedinger, K. R. (2005). The Assistment builder: a rapid development tool for ITS. In C.K. Looi, G. McCalla, B. Bredeweg, & J. Breuker (Eds.), *Proceedings of the 12th International Conference on Artificial Intelligence In Education*. (pp. 929-931). Amsterdam: ISO Press.

Vendlinski, T., Niemi, D., Wang, J., Monempour, S., & Lee, J. (2005). Improving formative assessment practice with educational information technology. From American Educational Research Association 2005 Annual Meeting.

Walonoski J., & Heffernan, N. T. (2006). Detection and analysis of off-task gaming behavior in intelligent tutoring systems. In Ikeda, Ashley & Chan (Eds.), *Proceedings of the Eighth International Conference on Intelligent Tutoring Systems.* (pp. 382-391). Heidelberg, Germany: Springer Berlin.