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

Informing Teachers Live about 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 [7] 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. This reporting system is robost enough to support the 800 students currntly using our system.

Introduction

MCAS (Massachusetts Comprehensive Assessment System) is a graduation requirement in which all students educated with public funds in the tested grades are required to participate. 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, we are integrating assistance and assessment by utilizing "Assistment" system [7] supported by the U.S. Department of Education. The Assistments system offers instructions to students while providing a more detailed evaluation of their abilities to the teacher than is possible under current approaches. Each assistment consists of an *original item* and a list of *scaffolding questions*[1] which only show up to the students who have given wrong answers to original questions. Our supporting website "www.assistment.org" has been running for around 7 months, providing more than 100 assistments built using our online authoring tools [8] and is being used by 9 teachers and about 800 students.

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 [3] 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 item 19 from the 2003 MCAS 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).

[1] We use the term scaffolding question because they are like scaffolding that will help students solve the problem (and can "faded" later) so the scaffolds are meant to scaffold their learning. [2]

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've 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."

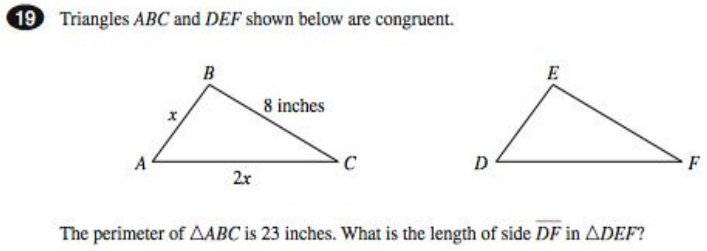


Figure 1: Item 19 from 2003 MCAS

There are several reasons for this poor MCAS reporting: 1) the reasonable desire to give problems tap-multiple knowledge components, 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.

1. 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 [5] 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 for about 7 months. Currently log records in our database show that about 50,000 MCAS items have been done and more than 600,000 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.

2. Transfer Model

A transfer model [4] 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 (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 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 hypothesize 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.

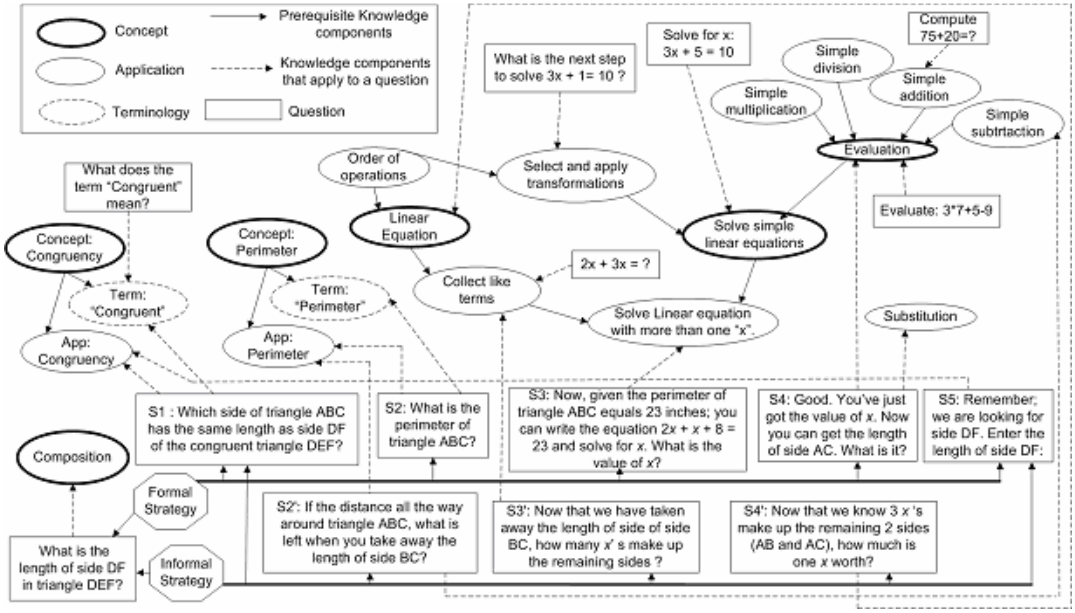


Figure2:A small piece of the WPI300 transfer model showing both how 14 questions (out of 245 in the WPI300) tap 19 knowledge components (out of 174 in the WPI300).

WPI300, which actually contains only 174 knowledge components so far, is the first model we have created. In the model, knowledge components are arranged in a hierarchy based on prerequisite structure. So far, 102 knowledge components in this transfer model have been used to tag 92 assistments (including 853 questions) in our system. Figure 2 shows 19 of the 174 knowledge components that we used to explain both a "formal" and "informal" problem solving strategy related to the item shown in Figure 1. We added a few other questions (like "What does the word 'congruent' mean?") to help define what a knowledge component means. Each of the scaffolding questions (S1 to S5) are mapped to one or two knowledge components. Tagging the scaffolding questions enable us to assess individual knowledge components instead of only overall performance. Each knowledge component might have prerequisite knowledge so that for a student to know "What does the word 'congruent' mean?" the student first needs to have mastered the "Concept of congruency" as shown by there being an arc between them.

Currently, we have been able to generate reports based on Massachusetts Curriculum Framework, as well as the WPI300 transfer model which reveals more detailed information about students' knowledge learning and knowledge components contained in problems. And we hope to be able to show that WPI300, as a finer grained cognitive model, will be more predictive. This is one subject of our current research.

3. Reporting System

3.1.1 Student Grade Book Report

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

The "Grade Book", shown in Figure 3.1, 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

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.	
Tom	34	0	15	3	20%	200	Failing	30	16	53%	15	N.1.8-understanding-number-representations (Error times: 5/6)
Dick	32	0	38	26	68%	242	Proficient	81	56	69%	4	P.1.8-understanding-patterns (Error times: 2/6)
Harry	33	0	20	9	45%	220	Needs improv.	63	28	44%	63	P.1.8-understanding-patterns (Error times: 8/10)

Figure 3.1: Grade Book on real student data

his MCAS score and his performance level[2,3] . Besides presenting information on the item level, it also summarizes the student's actions in an "Assistment metric": 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 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" [1].

In Figure 3.1, we see that these 3 students have used the system for about 30 minutes. (Many students have used it for about 250 minutes). "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 is 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

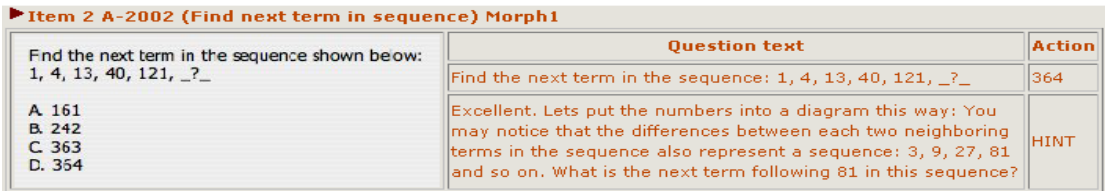


Figure 3.2. Items tagged with difficult knowledge component

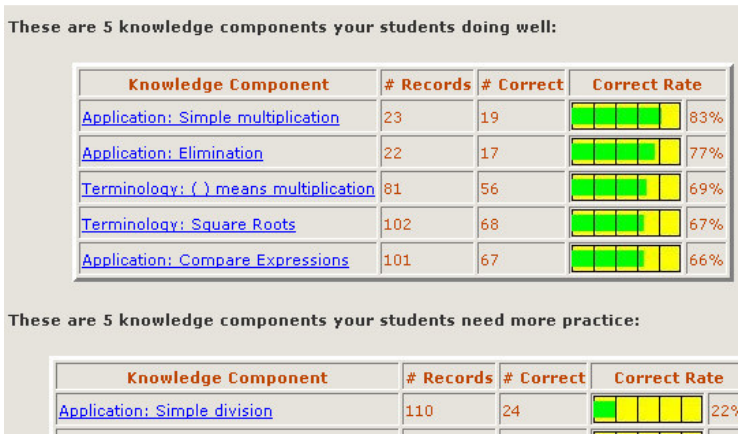
[2] Our "prediction" of a student MCAS score is at this point primitive. The column is currently simply a function of percent correct. We might even remove these two columns related to MCAS score prediction until we feel more confident in our prediction, in another word, "rough and ready".

[3] In our recent 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) [7]. But since that is a rather small group of students compared to the number of students now (68 vs. 8000), we'll continually refine our prediction function based on this year's data.

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 3.2) 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.

3.1.2 Class Summary Report



These are 5 knowledge components your students doing well:

Knowledge Component	# Records	# Correct	Correct Rate
Application: Simple multiplication	23	19	83%
Application: Elimination	22	17	77%
Terminology: ( ) means multiplication	81	56	69%
Terminology: Square Roots	102	68	67%
Application: Compare Expressions	101	67	66%

These are 5 knowledge components your students need more practice:

Knowledge Component	# Records	# Correct	Correct Rate
Application: Simple division	110	24	22%

Figure 3.3 Class summary report for a teacher's classes

"Class Summary" is a report we developed to inform teachers about the knowledge status of classes. 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 3.3 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 3.3), 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.

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.

3.1.3 Class Progress Report

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 3.4 preliminary progress reports for a class

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

Figure 3.4 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.

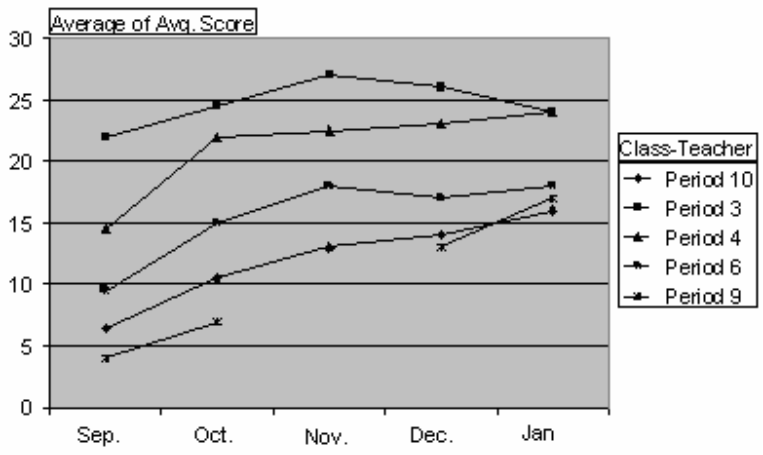


Figure 3.5 predicted MCAS Score over months

The progress of students' predicted MCAS raw score over months is more clearly shown in Figure 3.5. Those students (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).

3.2 Analysis of Items

A report is built to show difficulty each problem in our system. (See Figure 3.6: 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 3.7 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.

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 3.6: Problems order by correct rate

On the first line of Figure 3.7, 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[4] 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. 56% of students asked for a hint, telling you something about students'

ID	Question	Correct Answer	% Correct	Hint Req.	# Attempt	Common Errors			WPI's Use of MA. Standard	WPI's Knowledge Components
						Resp.	#	Buggy Message		
Triangles ABC and DEF are congruent. The perimeter of triangle ABC is 23 inches. What is the length of side DF in triangle DEF?		10	12%	56%	180	8	15	N/A	**G.2.8**, M.3.8, P.7.8	Composition, T.3, A.3, T.4, A.4, A.12, A.15, A.17
						16	13	N/A		
						23	8	N/A		
1	Which side of triangle ABC has the same length as side DF of triangle DEF?	ac	23%	50%	154	ab	13	Side AB corresponds to side DE of triangle DEF, not DF. Try again, please.	**G.2.8-congruence-and-similarity**	Term: "Congruency", Appl: Congruency
						DF	6	N/A		
2	What is the perimeter of triangle ABC?	2x+x+8	39%	20%	148	2x + 8	69	No. It looks like you have added just two of the sides of triangle ABC. Perimeter is the sum of all the sides.	M.3.8-using-measurement-formulas	Term: "Perimeter", Appl: Perimeter
3	Now, given the perimeter of triangle ABC equals 23 inches, you can write the equation **2x + x + 8 = 23** and solve it for **x**. What is the value of **x**:	5	25%	52%	147	15	13	N/A	P.7.8-setting-up-and-solving-equations	Appl: Solve linear equation
						13	10	N/A		
						8	10	N/A		
4	Remember, we are looking for side DF. Enter the length of side DF:	10	30%	43%	143	5	26	N/A	G.2.8-congruence-and-similarity	Appl: Congruency
						2x	2	N/A		
						8	3	N/A		

Figure 3.7: A scaffolding report generated by Assistment reporting system

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.

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".

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 3.7, 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 3.1 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.

Table 3.1: "Red & Green" distribution matrix

Original	154																22
Q1	119								35								N/A
Q2	85				34				12				23				
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	

3.3 Performance evaluation

Our reporting system was 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. 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 3.7) is about 5 seconds.

4. Conclusions

In conclusion we feel that we have developing some state-of-the-art online reporting tools that will help teachers 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 [7]. 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. Our long term vision is to 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. We feel we have taken some stops in that direction.

References

[1] 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 Systems*, 2004, Maceio, Brazil

[2] 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.

[3] 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. At http://www.syrce.org/State_Board.htm

[4] 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*, 2004, Maceio, Brazil

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

[6] J. Mostow, J.E. Beck, R. Chalasani, A. Cuneo, and P. Jia. Viewing and Analyzing Multimodal Human-computer Tutorial Dialogue: A Database Approach. *Fourth IEEE International Conference on Multimodal Interfaces* (ICMI 2002), October, 2002.

[7] 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. *Proceedings of the 12th Annual Conference on Artificial Intelligence in Education* 2005, Amsterdam, to appear.

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