Feng, M., Heffernan, N.T., & Koedinger, K.R. (2009). Addressing the assessment challenge in an Online System that tutors as it assesses. *User Modeling and User-Adapted Interaction: The Journal of Personalization Research (UMUAI journal).* 19(3), 243-266, August, 2009.

Addressing the Assessment Challenge with an Online System That Tutors as it Assesses

Mingyu Feng[1], Neil Heffernan[1], Kenneth Koedinger[2]
[1] Department of Computer Science, Worcester Polytechnic Institute
[2] Human Computer Interaction Institute, Carnegie Mellon University

Abstract

Secondary teachers across the United States are being asked to use formative assessment data (Black & Wiliam, 1998a, 1998b; Roediger & Karpicke, 2006) to inform their classroom instruction. At the same time, critics of US government's No Child Left Behind legislation are calling the bill "No Child Left Untested". Among other things, critics point out that every hour spent assessing students is an hour lost from instruction. But, does it have to be? What if we better integrated assessment into classroom instruction and allowed students to learn during the test? We developed an approach that provides immediate tutoring on practice assessment items that students cannot solve on their own. Our hypothesis is that we can achieve more accurate assessment by not only using data on whether students get test items right or wrong, but by also using data on the effort required for students to solve a test item with instructional assistance. We have integrated assistance and assessment in the ASSISTment system. The system helps teachers make better use of their time by offering instruction to students while providing a more detailed evaluation of student abilities to the teachers, which is impossible under current approaches. Our approach for assessing student math proficiency is to use data that our system collects through its interactions with students to estimate their performance on an end-of-year high stakes state test. Our results show that we can do a reliably better job predicting student end-of-year exam scores by leveraging the interaction data, and the model based on only the interaction information makes better predictions than the traditional assessment model that uses only information about correctness on the test items.

Keywords: Intelligent Tutoring System, ASSISTments, dynamic assessment, assistance metrics, interactive tutoring

1 Introduction

In many US states there are concerns about poor student performance on new high-stakes standards-based tests that are required by the No Child Left Behind Act (NCLB) legislation of the US government. For instance, the high-stakes Massachusetts Comprehensive Assessment System (MCAS) is the graduation requirement in which all students in the state educated with public funds are required to participate. It administers rigorous standardized tests in English, math, history and science in grades 3–10 every year. Students need to pass the math and English portions of the 10th grade versions in order to get a high school diploma. In 2003, a full 10% of high school seniors were predicted to be denied a high school diploma due to having failed to pass the test on their fourth try. Moreover, the state of Massachusetts singled out student performance on the 8th grade math test as an area of highest need for improvement[1]. Partly in response to this pressure, and partly because teachers, parents, and other stakeholders want and need more immediate feedback about how students are doing, there has recently been intense interest in predicting student performance on end-of-year tests (Olson, 2005). There is a large interest in "Formative Assessment" (Boston, 2002; Black & Wiliam, 1998a, 1998b; Roediger & Karpicke, 2006) in K-12 Education (Olson, 2004) with many companies[2] providing such services. Some teachers make extensive use of practice tests and released test items to help

[1] http://www.doe.mass.edu/mcas/2002/results/summary.pdf
[2] Including nwea.org/assessments/, measuredprogress.org, Pearson and ww.cddre.org/Services/4Sight.cfm

identify learning deficits for individual students and the class as a whole. However, such formative assessments not only require great effort and dedication, but they also take valuable time away from instruction. Some online testing systems (such as Renaissance Learning, www.renlearn.com) automatically grade students and provide reports but they may not be informative as they do not maintain sufficiently rich data records for students and therefore cannot report on a fine-grained model of student knowledge. They also do not provide instruction while students engage in formative assessment.

The limited classroom time available in middle school mathematics classes compels teachers to choose between time spent assisting students' development and time spent assessing students' abilities. A possible solution might involve a way whereby students can take an assessment, but also learn as they are being assessed; unfortunately statisticians have not done a great deal of work to enable assessment of students while they are learning during the test.[3] This solution allows teachers to get the benefit of data-driven instruction, but at the same time, make sure that their students' time is spent primarily on learning. To help resolve this dilemma, the U.S. Dept of Education funded us to build a web-based tutoring system that would also do assessment at the same time. Assistance and assessment are integrated in the system ("ASSISTment"[4]) that offers instruction to students while providing a more detailed evaluation of their abilities to the teacher than is possible under current approaches.

Unlike other assessment systems, the ASSISTment technology also provides students with tutoring assistance while the assessment information is being collected. Each week when students work on the website, the system "learns" more about the students' abilities and thus, it can hypothetically provide increasingly accurate predictions of how they will do on a standardized mathematics test (Anozie & Junker, 2006). It helps students to work through tough problems by breaking them into sub-steps; meanwhile, it collects data related to different aspects of student performance such as accuracy, speed, help-seeking behavior and attempts as students interact with the system. Based on the rich source of data, various reports, for individual students and for groups, have been developed to help teachers and other stakeholders to better understand students' performance and progress (e.g. Feng & Heffernan, 2007). Our work in this paper is to see if it is possible to do assessment better if we take advantage of the student-system interaction information that is normally not available in traditional practice tests. These measures include response efficiency that accounts for the time it takes students to come up with an answer to a problem, the time they take to correct an answer if it is wrong, help-seeking behavior (e.g. the number of hints they request), and their performance on the sub-steps (called scaffolding questions).

Much of the power of a computer tutoring system comes from its ability to assess students. Yet, assessing students automatically, continuously and accurately without interfering with student learning is an appealing but also a challenging task. Corbett & Bhatnagar (1997) describe an early and successful effort to increase the predictive validity of student modeling in the ACT

[3] Standard psychometric theory requires a fixed target for measurement (e.g. van der Linden & Hambleton, 1997), which requires that learning during testing be limited. Some attempts to combine standard psychometric models with Markov learning models have been attempted (as far back as Jannarone, 1986), and some work has been done on psychometric measurement of growth and change (e.g. Tan, Imbos & Dos, 1994; Embretson, 1992; Fischer & Seliger, 1997). However, making predictions from testing data in which students are actively learning material has only been pursued energetically in the realm of computer-based intelligent tutoring (e.g., Corbett, Anderson & O'Brien, 1995).

[4] The term "ASSISTment" was coined by Kenneth Koedinger and blends **assist**ing and assess**ment**.

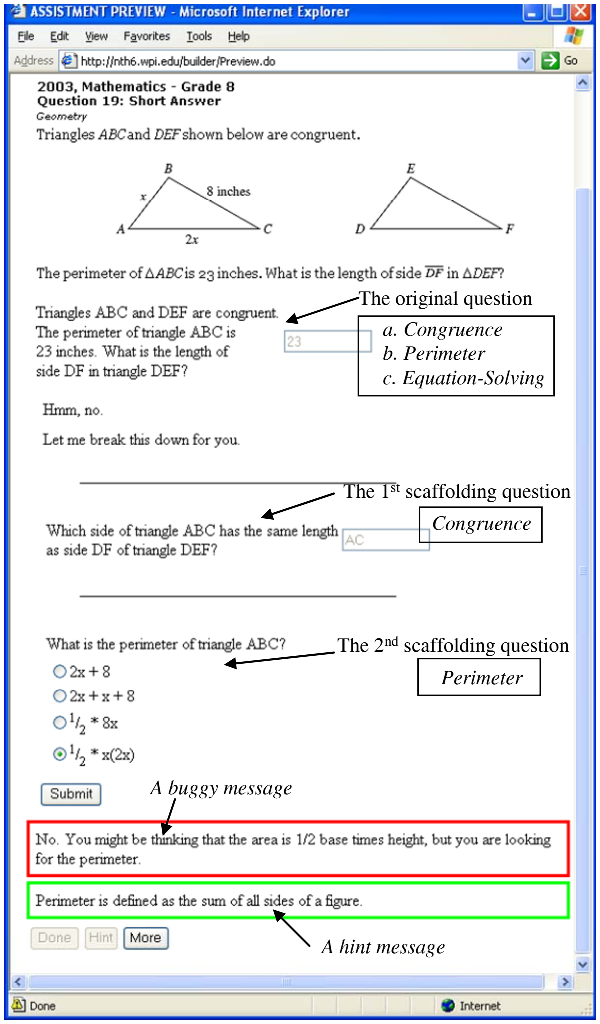
Programming Tutor (APT). They used assessments from a series of short tests to adjust the knowledge-tracing process in the tutor and more accurately predict individual differences among students in the post test. Beck & Sison (2006) used knowledge-tracing to construct a student model that can predict student performance at both coarse-grained (overall proficiency) and fine-grained (for a particular word in the reading tutor) sizes. Feng, Heffernan & Koedinger (2006) developed a group of dynamic metrics (which will be discussed in more detail later) to measure the amount of assistance students need to solve a problem, and showed that a student's test score can be better predicted using these metrics. Anozie & Junker (2006) pursued a rather different approach, looking at the changing influence of online ASSISTment metrics on MCAS performance over time. They computed monthly summaries of online metrics similar to those used by Feng, Heffernan & Koedinger, and built several linear prediction models, predicting end-of-year raw MCAS scores for each month.

In this paper, we will focus on answering the research question: *Does the tutoring interaction provide valuable assessment information?* To answer this question we will compare the model built that considers only the *original question* response to models that take into consideration students' response data during the tutoring session and their help-seeking behavior. We have presented our prediction of students' "expected" MCAS test scores as a single column in one of our online teacher reports, the "Grade Book" report. The prediction was made based only upon student performance on the main questions (called original questions in ASSISTments, comparing to the *scaffolding questions* that break the main question into sub-steps). The report does not distinguish between two students who both got the original question wrong, but then needed very different levels of tutoring to get the problem correct eventually. A positive answer to the research question would help us to build a better predictive model and also improve our online teacher reporting.

The rest of this article is organized as follows. In section 2, we introduce the ASSISTment system that students and teachers interact with. Section 3 describes our reporting system that provides assessment on students' performance to teachers. In Section 4 we show that we can achieve more accurate assessment by not only using data on whether students get test items right or wrong, but by also using data on the effort required for students to learn how to solve a test item. And we conclude our work in Section 5.

2 Background on the ASSISTment System

The ASSISTment system is an online tutoring system that was first created in 2004. In Massachusetts, the state department of education has released 10 years (1998-2007) worth of 8th grade MCAS test items, almost 400 items, which we have turned into ASSISTments by adding "tutoring". If students working on ASSISTments get an item correct they are given a new one. If they get it wrong, they are provided with a small "tutoring" session where they must answer a few questions that break the problem down into steps. The key feature of ASSISTments is that they provide instructional assistance while assessing students. Razzaq & Heffernan (2006, 2007), Feng, Heffernan, Beck & Koedinger (accepted), addressed student learning due to the instructional assistance, while this paper is focused on skill model evaluation by assessing students' performance on a state test.


ASSISTMENT PREVIEW - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://nth6.wpi.edu/builder/Preview.do
Go
B
E
x
8 inches
A
C
D
F
2x
Submit
Done
Hint
More
Done
Internet

2003, Mathematics - Grade 8
Question 19: Short Answer
Geometry

Triangles *ABC* and *DEF* shown below are congruent.

The perimeter of ΔABC is 23 inches. What is the length of side $\overline{DF}$ in ΔDEF?

The original question

Triangles ABC and DEF are congruent. The perimeter of triangle ABC is 23 inches. What is the length of side DF in triangle DEF? 23

a. Congruence
b. Perimeter
c. Equation-Solving

Hmm, no.

Let me break this down for you.

The 1st scaffolding question

Which side of triangle ABC has the same length as side DF of triangle DEF? AC

Congruence

What is the perimeter of triangle ABC?

The 2nd scaffolding question

Perimeter

○ 2x + 8
○ 2x + x + 8
○ $^1/_2$ * 8x
● $^1/_2$ * x(2x)

A buggy message

No. You might be thinking that the area is 1/2 base times height, but you are looking for the perimeter.

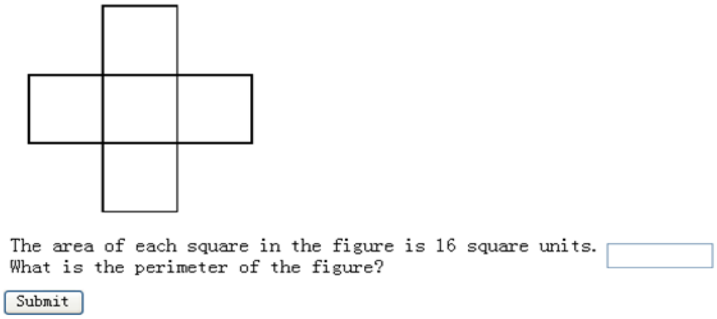
Perimeter is defined as the sum of all sides of a figure.

A hint message

Figure 1. An ASSISTment shown student working on an ASSISTment

Each ASSISTment consists of an *original question* and a list of *scaffolding questions*. The original question usually has the same text as in the MCAS test while the scaffolding questions were created by our content experts to coach students who fail to answer the original question. 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 that represents a common error. Figure 1 shows a buggy message 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 question 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 question wrong because they did not know congruence versus not knowing perimeter, versus not being able to set up and solve the equation. As a matter of logging, the student is only marked as getting the item correct if they answered the questions correctly before asking for any hints or encountering scaffolding.



The area of each square in the figure is 16 square units.
What is the perimeter of the figure?

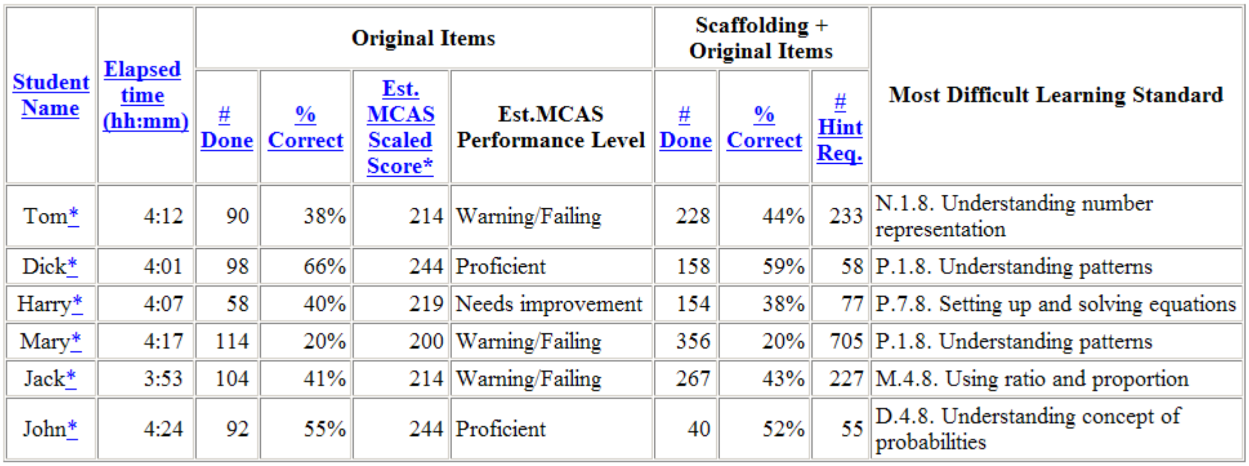
Figure 2. The original question of item 27 of 1999 MCAS test

Figure 2 shows the original question of another ASSISTment built for item 27 of the 1999 MCAS test. The ASSISTment provides two scaffolding questions. The first one asked "What is the length of one side of a square in the figure?" and the second says "Now you have enough information to find the perimeter of the figure. What do you think it is?" In the fine grained skill model developed by our colleagues at WPI, the original question was tagged with 2 skills: "Perimeter" and "Area"; the first scaffolding question is associated with "Perimeter" and the second one "Area".

In the first year the ASSISTment system was launched, 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. Since then the number of users has expanded every year and more than 3000 students from Massachusetts used the system during the school year of 2007-2008.

3 Reporting in the ASSISTment System

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. The reporting in the ASSISTment System (Feng & Heffernan, 2007) has been 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. 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	Elapsed time (hh:mm)	Original Items				Scaffolding + Original Items			Most Difficult Learning Standard
		# Done	% Correct	Est. MCAS Scaled Score*	Est.MCAS Performance Level	# Done	% Correct	# Hint Req.	
Tom*	4:12	90	38%	214	Warning/Failing	228	44%	233	N.1.8. Understanding number representation
Dick*	4:01	98	66%	244	Proficient	158	59%	58	P.1.8. Understanding patterns
Harry*	4:07	58	40%	219	Needs improvement	154	38%	77	P.7.8. Setting up and solving equations
Mary*	4:17	114	20%	200	Warning/Failing	356	20%	705	P.1.8. Understanding patterns
Jack*	3:53	104	41%	214	Warning/Failing	267	43%	227	M.4.8. Using ratio and proportion
John*	4:24	92	55%	244	Proficient	40	52%	55	D.4.8. Understanding concept of probabilities

Figure 3. The Grade book report for a real class (with fake student names)

The "Grade Book", shown in Figure 3, 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 problems he has done and his percent correct, our estimation of his MCAS scaled score and performance level[5]. Besides presenting information on the item level, it also summarizes the student's actions in "ASSISTment metrics": how many scaffolding questions have been done, the 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. In addition, the "ASSISTment metrics" tell more about students' actions than just 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 were "gaming the system" (Baker, Corbett, & Koedinger, 2004; Walonoski, & Heffernan, 2006). Note, however, that the values in the MCAS prediction columns in Figure 3 do not yet use ASSISTment metrics, but is simply a function of the student's percent correct on original items. We are working on updating this prediction (and the associated "performance level" column) based on the results to be shown in later sections of this paper.

In Figure 3, we see that these students have used the system for about 4 hours. "Dick" has finished 98 original items and only asked for 58 hints. Most of the items he solved correctly and thus our prediction of his MCAS performance was high. The last column indicates that he has made the greatest number of errors on questions that have been tagged with the learning standard "P.1.8 understanding patterns" that represents the first learning standard in the category of *Pattern, Relations, and Algebra* as required for 7-8th grade students by Massachusetts Department of Education (2000). Teachers can also see "Mary" has asked for too many hints (705 hints, far more than others). Noticing this, the teacher might discuss with Mary why she is asking for so many hints and encourage her to do more thinking on her own. 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. The "Grade Book" is so detailed that a student commented: "It's spooky, he's watching everything we do".

[5] Please refer to section 4.1 for a more detailed description of MCAS scaled score and performance level.

We feel that we have developed some state-of-the-art online reporting tools that will help teachers and researchers be better informed about what their students know. The prediction of MCAS performance provided in the reports can be further improved and that is a key goal of our analysis.

4 Assess Students Better by Looking into the Interactive Tutoring Session

In our past research, we reported a correlation between our prediction for the 68 students who have used our system in May 2004 and their real MCAS raw score (r = .7) (Razzaq et al., 2005). We are continually refining our prediction function based on new data. We developed a group of dynamic metrics, first reported in Feng, Heffernan & Koedinger (2006), to measure the amount of assistance students need to solve a problem, and showed that a student's test score can be better predicted using these metrics and the data of student interaction with the ASSISTment system than just relying on student responses to the original questions during the school year 2004-2005. In this work, we will reuse the dynamic metrics, and we will explore new models and replicate the same study on a new school year's data (2005-2006).

4.1 Description of the Data

The first data we consider comes from the 2004 – 2005 school year, the first full year in which the ASSISTment system was used in classes in two middle schools in Massachusetts. At that time, the ASSISTment system contained a total of 493 main questions and 1216 scaffolds; 912 8th grade students' logs were maintained in the system over the time period from September to May. Of these, approximately 400 main questions and their corresponding scaffolds were in regular use by approximately 700 students. Although the system is web-based and hence accessible in principle anywhere/anytime, students typically interact with the system during one class period in the schools' computer labs every few weeks. Among these 700 students, we were able to obtain complete data for 417 of them. The data set contains online interaction data from the ASSISTment system, the results of 8th grade MCAS tests taken in May, 2005 and the results of 10th grade MCAS tests taken by the same group of students two years later, in May, 2007. We excluded the data of the 25 students who did less than 39[6] questions in ASSISTments. The 391 students in our final 04-05 data set have practiced mathematics problems in the ASSISTment system for a mean length of 267 minutes (standard deviation = 79) across about 9 sessions, finishing on average 147 items (standard deviation = 60).

The second data set we will use is from the 2005-2006 school year. About 3,000 students used the system during the year and among these, we collected a full data set for 616 students from Worcester Public Schools, including the online data from ASSISTments and their 8th grade MCAS test raw scores (MCAS test was taken in May, 2006). The students on average worked in the ASSISTment system for 196 minutes (standard deviation = 76), and finished an average of 88 items (at least 39 items, standard deviation = 42).

MCAS results for students are reported as scaled scores and performance levels that are calculated based on the raw scores. The raw score reflects the total number of points a student got from a MCAS test, ranging from 0 to 54. The range of scaled scores for MCAS tests is arbitrarily defined to be between a minimum of 200 and maximum of 280, which is further

[6] The number 39 was picked because there are 39 questions in the real 8th grade MCAS test each year.

divided into four performance levels, namely Warning/Failing (when the scaled score falls in the range of 200 ~ 218), Needs Improvement (220 ~ 238), Proficient (240 ~ 258), and Advanced (260 ~ 280). The purpose of scaled scores is to provide information about the position of a student's results within a performance level. Student raw scores, or the total number of points, on the MCAS tests are converted into scaled scores using a scaling procedure. Generally speaking, the MCAS scaling procedure involves two-step transformations: "*(1) non-linear monotonic transformations of the raw score points into theta metric*[7] *points, and (2) linear transformations of theta points into MCAS scaled score points.*" (MCAS technical report, 2001, p41) These transformations do not change the rank ordering of students, nor their performance level classifications. Also, three threshold raw scores determined by a standard setting procedure were used as the cut points for the four performance levels. The whole procedure tries to minimize the fluctuations in the thresholds from year to year, consistent with the slight shifts in the difficulty of the tests and the gaps in theta estimates between rounded raw scores.

In the reporting system, we represented student performance in terms of scaled score and performance level (as seen in Figure 3) because teachers and students are more used to interpreting these metrics than understanding the raw score. But since the raw score provides a finer grain differentiation between students, when we analyze our data, all of our procedures are judged on their ability to predict the MCAS raw scores.

In the following sections, we will first apply our approach on the 2004-05 data and then on the 2005-06 data to investigate whether the approach generalizes across years.

4.2 Developing Metrics to Measure Assistance Required

Much work has been done in the past 10 years or so on developing "online testing metrics" for dynamic testing (Grigorenko & Sternberg, 1998; Sternberg & Grigorenko, 2001, 2002) to supplement accuracy data (wrong/right scores) from a single sitting. Researchers have been interested in trying to get more assessment value by comparing traditional assessment (static testing; students getting an item marked wrong or even getting partial credit) with a measure that shows how much help they needed. Bryant, Brown and Campione (1983) compared traditional testing paradigms against a dynamic testing paradigm. Grigorenko and Sternberg (1998) reviewed relevant literature on the topic and expressed enthusiasm for the idea. 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. Sternberg & Grigorenko (2001, 2002) argued that dynamic tests not only serve to enhance students' learning of cognitive skills, but also provide more accurate measures of ability to learning than traditional static tests. In Bryant, Brown and Campione's study they wanted to predict learning gains between pretest and posttest. They found that student learning gains were not as well correlated (R = 0.45) with static ability score as with their "dynamic testing" (R = 0.60) score. It was suggested that this method could be effectively done by computer. Beck, Jia & Mostow (2004) were able to increase the within-grade correlation between their model and student performance on a fluency test significantly by adding help-seeking behavior in the computer Reading Tutor (Mostow & Aist, 2001). The ASSISTment system provides an ideal test bed to investigate dynamic assessment as it automatically provides students with feedback, scaffolding questions

[7] The "theta metric" refers to the student ability (θ) metric in Item Response Theory models.

and hints. So it is a natural way to extend and test prior work on dynamic assessment. On top of that, we want to determine whether other measures, such as student response speed, can increase predictive power.

We developed the following group of metrics that measure students' accuracy, speed, attempts and help-seeking behavior:

- Original_Percent_Correct – students' percent correct on original questions, which we often referred to as the "static metric".
- Original_Count - the number of original items students have done. This measures students' attendance and how on-task they are. This measure also reflects students' knowledge since better students have a higher potential to finish more items in the same period of time.
- Scaffold_Percent_Correct - students' percent correct on scaffolding questions. In addition to original items, students' performance on scaffolding questions is also a reasonable reflection of their knowledge. For instance, two students who get the same original item wrong may, in fact, have different knowledge levels and this may be reflected in the one may get more of the scaffolding questions right than the other.
- Question_Count - the number of questions (both original items and scaffolding questions) students have finished. Similar to Original_Count, this variable is also a measure of attendance and knowledge but given the fact that scaffolding questions show up only if students fail the original question, it is not obvious how this measure will correlate with students' MCAS scores.
- Hint_Request_Count - how many times students have asked for hints.
- Avg_Hint_Request - the average number of hint requests per question.
- Bottom-Out_Hint_Count - the total number of bottom-out[8] hint messages students got.
- Avg_Bottom_Hint - the average number of bottom-out hint messages students got per question.
- AvgFirstHintRequest_original – on average, how many times students requested hints before they made an attempt on an original question.
- AvgFirstHintRequest_scaffold – Similar to AvgFirstHintRequest_original, this variable measures on average, how many times students requested hints before they made an attempt on a scaffolding question.

AvgFirstHintRequest_original and AvgFirstHintRequest_scaffold are only available for 2005-06 data. We hypothesize the two metrics are both a reflection of student knowledge and their confidence. Students who either do not understand the question or are uncertain of their response may request hints before making an attempt. Therefore, our hypothesis is that both metrics will be negatively correlated with MCAS scores without other variables in the model. If Original_Percent_Correct is already in the model, then AvgFirstHintRequest_original might be positively correlated because it may be better to know you don't know (and those ask for a hint) than not know and just guess.

- Attempt_Count - the total number of attempts students made across all original and scaffolding questions.
- Avg_Attempt - the average number of attempts students made for each question.

[8] Since the ASSISTment system does not allow students to skip problems, to prevent students from being stuck, most questions in the system were built such that the last hint message almost always reveals the correct answer. This message is referred to as "Bottom-out" hint.

- Avg_Item_Time - on average, how long it takes for students to finish a problem (including all scaffolding questions if students answered the original questions incorrectly).
- Avg_Question_Time - on average, how long it takes for a student to answer a question, whether original or scaffolding, measured in seconds.
- Total_Minutes - how many total minutes students have been working on items in the ASSISTment system. Just like Original_Count, this metric is an indicator of the attendance.

The ten measures from Hint_Request_Count to Avg_Question_Time in the above list are generally all ASSISTment style metrics (or the assistance metrics), which indicate the amount of assistance students need to finish problems and the amount of time they spend to finish items. Therefore, the hypothesis is that these measures will generally be negatively correlated with MCAS scores (though, for instance, Attempt_Count or Total_Minutes may be positively correlated because they are partially determined by time on task).

Among these measures, “Original_Percent_Correct” is a static metric that mimics paper practice tests by scoring students as either correct or incorrect on each item, while the rest of the measures are dynamic assessment metrics that measure the amount of assistance students need before they get an item correct, how fast they answered the questions, etc.

We have been presenting some of these online measures in our reports to teachers (See Figure 3). Particularly, student *Mary* used the system 4 hours and 17 minutes, finished 114 items with 20% correct. She went through 356 scaffolding questions with 20% correct and asked for 705 hints, which is excessive compared to her classmates.

Given the data set, our goal was to see if we can reliably predict students' MCAS raw scores and to evaluate how well on-line use of the ASSISTment System, can help in the prediction. To achieve the goal, we performed a stepwise linear regression using the online measures as independent variables to predict students' MCAS scores.

4.3 Modeling

First, we present the Pearson correlations between MCAS raw scores and all the independent variables in Table 1 to give some idea of how these factors are directly related to MCAS score. The first column shows the correlation between the factors and the 2005 MCAS test scores in the 2004-05 data set, and the second column shows the correlation between the factors and the 2006 MCAS test scores. All these factors except Attempt_Count and Avg_Question_Time turned out to be significantly correlated with MCAS score at the 0.01 level (2-tailed). In general, students who maintained higher percent correct, finished more items, requested for less help and solved problems faster in ASSISTment, tended to have a high MCAS score at the end of the year, which is consistent with our hypothesis.

Table 1. Correlations

Correlations		Factors	04-05 data/ 2005 MCAS (391 students)	05-06 data/ 2006 MCAS (616 students)
Online Computer Metrics	Static metric	Original_Percent_Correct	.786	.689
	Attendance	Original_Count	.477	.394
		Total_Minutes	.241	.202
	Assistance Style Metrics (dynamic metrics)	Scaffold_Percent_Correct	.721	.647
		Question_Count	.163	.132
		Hint_Request_Count	-.465	-.265
		Avg_Hint_Request	-.689	-.536
		Bottom_Out_Hint_Count	-.423	-.258
		Avg_Bottom_Hint	-.584	-.477
		Attempt_Count	**.036**	**.026**
		Avg_Attempt	-.398	-.593
		Avg_Question_Time	**-.062**	**-.026**
		Avg_Item_Time	-.391	-.299
		AvgFirstHintRequest_original	NA	-.347
		AvgFirstHintRequest_scaffold	NA	-.351

Our effort to predict student MCAS scores starts with **a "lean" model**, an Item Response Theory (IRT)-style model using ASSISTment data only. As a starting point, we used a one-parameter IRT model (the Rasch model, also called 1-PL IRT model) (van der Linden & Hamilton, 1997), the straightforward model with just student proficiency and item difficulty parameters (and only on original question data). In the simple Rasch model, the probability of a correct response is modeled as a logistic function of the difference between the student and item parameter. To train the lean model, we used a data set of all data collected in the ASSISTment system from Sept., 2004 to Jan., 2008, including responses to 2,797 items (only on original questions) from 14,274 students. By including more student response data from the four years, we hope to acquire a more reliable estimate of the item parameters, and thus a more stable estimate of the student proficiency.

Some readers may argue for multi-parameter models, such as 2-PL or 3-PL IRT models, that are expected to perform better on parameter estimation for such data. However, the number of data points per item in our data set varies tremendously (mean = 276, and standard deviation = 463). And there are 424 items for which we have less than 10 data points. Therefore, we stick with the simplest 1-PL Rasch model.

We fitted the Rasch model in BILOG-MG 3.0 (Zimowski, Muraki, Mislevy & Bock, 2005) and obtained an estimate of every student's proficiency trait score (usually called the student parameter), ranging from -4 to 4. The trait score was then transformed to the probability of a correct response on an item of average difficulty (i.e. item difficulty parameter equal to .5). We named this probability "IRT_Proficiency_Estimate" and added an extra column in the data set that we described in Section 4.1 to record the student proficiency estimates. Finally, we did a linear regression to predict MCAS scores using student proficiency estimate to put them on the same scale.

The lean model is based on how students performed on the original questions but ignores how students interacted with the system. Using the online assistance metrics, we ran a stepwise regression[9] analysis to predict MCAS test scores. We will refer to this model as **the assistance model.** Original_Percent_Correct was not included in the assistance model in order to distinguish it from the lean model and to show how the interaction information alone can do in estimating student performance. It is worth highlighting that the assistance model, which does not use the assessment information from the original questions is fundamentally different from traditional assessment model, represented by the lean model which only uses the original question information. By contrasting the assistance model against the lean model, we can isolate the assessment value of the tutor-student interaction session.

We explored a third model that combines the student proficiency parameters, estimated by the lean model, with the online assistance metrics. New parameter values for the assistance metrics were again fit in a stepwise regression. We name this model **the mixed model**. We did not include student percent correct on original questions here because these values were used to find the best fitting student proficiency parameters. Using student proficiency has the advantage that it takes into consideration the difficulty of each item a student attempts.

As mentioned before, Beck, Jia & Mostow (2004) were able to increase the within-grade correlation between their model and student performance on a fluency test significantly by adding help-seeking behavior in the computer Reading Tutor. In order to see how much the information on the amount of assistance that students needed to solve a problem can help in the ASSISTment system, we constructed another model, which we will refer to as **the help model**. In Beck et al. (2004), two features were constructed to describe student help requests: the percentage of words on which the student clicked for help and the percentage of sentences on which the student requested help. While in this work, we have developed more features, including Hint_Request_Count, Avg_Hint_Request, Bottom_Out_Hint_Count, Avg_Bottom_Hint. Recall that all these hint related metrics are significantly correlated with MCAS test scores. Thus, to build the help model, we ran a stepwise linear regression to predict student MCAS scores using the student proficiency parameter estimated by the lean model and all of the hint related metrics.

Thus, we constructed four models: the lean model, the assistance model, the mixed model and the help model. Each model addresses different aspects of student performance. The modeling process has been replicated using data from both 2004-2005 and 2005-2006 school years. In the next section, we will evaluate the models.

4.4 Result Analysis and Model Evaluation

4.4.1 Did the dynamic metrics help us building a more predictive model?

In Table 2, we summarize the four models that have been built using 2004-2005 data, for which we selected different groups of independent variables (IV) for regression. For each model, we reported how many variables entered into the model, how well the model fit data and how the

[9] We set probability of F to enter <= .05; probability of F to remove >= .10 for all stepwise regression analysis we did.

predicted scores correlated with the real MCAS scores. SPSS automatically displays R Square and Adjusted R Square in the output. Because our models have different numbers of parameters and are not always nested[10], we needed a method to compare the generalization quality of our models. We chose to use the Bayesian Information Criterion (BIC) for this purpose and applied the formula for linear regression models introduced by Raftery (1995, p135), which is different from what is typical used for calculating BIC but most convenient for linear regression models:

$$BIC = n\log(1-R^2) + p\log(n)$$

where

n: the sample size (for the 2004-2005 data case, n = 391; for the 2005-2006 data, n=616)

log: natural logarithm

p: the number of independent variables included in each model (not including intercept)

The MCAS test score predictions of the best fitting models were compared with actual 2005 MCAS test scores to calculate mean absolute deviation (MAD), which has been used as the measure to evaluate the student models in prior works (e.g. Anozie & Junker, 2006; Ayers & Junker, 2006; Feng, Heffernan & Koedinger, 2006; Feng, Beck, Heffernan & Koedinger, 2008). In this paper we follow those works and compute MAD for the models. We also report the correlation between the adjusted predicted scores[11] by each model and student real MCAS test scores. The absolute deviations (AD) of each model for each of the 391 students were compared with each other, two models at a time, using paired t-tests (the *p*-values for comparisons of successively better models are shown in the first column of Table 2).

As we can see in Table 2, the help model correlates reliably better with MCAS score than the lean model, suggesting that in ASSISTments, student help-seeking behavior is helpful in improving MCAS score prediction. The lean model does not correlate with MCAS score as well as the assistance model, which indicates that MCAS score can be better predicted by using features reflecting student assistance requirement, effort, attendance, etc, even if we ignore student responses to original questions. Additionally, we can improve our prediction of MCAS score further on top of the assistance model by combining the student proficiency parameter together with the online metrics that describe how students interacted with the system, such as the number of attempts students need, and how long a student need to answer a question (See Table 3). The improvement is statistically reliable in terms of MAD[12], BIC[13] and the correlations[14].

[10] Two models are nested if they both contain the same terms and one has at least one additional term.

[11] The adjusted predicted score is calculated by doing "leave-one-out" cross validation in SPSS.

[12] See the chapter of "Evaluating hypotheses" in Mitchell (1997).

[13] Raftery (1995) discussed a Bayesian model selection procedure, in which the author proposed the heuristic of a BIC difference of 10 is about the same as getting a p-value of p = 0.05.

[14] We did significance testing of the difference between two dependent correlations from a single sample using the applet online at http://www.quantitativeskills.com/sisa/statistics/correl.htm.

Table 2. Model Summary (based on 2004-2005 data)

Model	MAD	R Square	BIC	#variables	Correlation with 2005 8th grade MCAS
The lean model	6.40	.537	-295	1	.733
p=.031					
The help model	6.13	.585	-326	3	.765
p=.004					
The assistance model	5.46	.674	-402	6	.821
p=.001					
The mixed model	5.04	.707	-450	5	.841

As the best fitted model, the mixed model took into account students' performance on both original items (indicated by the IRT_Proficiency_Estimate) and scaffolding questions (indicated by the feature Scaffold_Percent_Correct), together with the other online metrics, and the model correlates with MCAS scores fairly well (r = .841). With five variables (one less than in the assistance model), it gained a 0.033 increment in the R^2 value, 48 points lower on BIC value, and a significant better correlation with MCAS scores, which means the mixed model is significantly better than other models. Variables entered the mixed model in the following order: IRT_Proficiency_Estimate, Scaffold_Percent_Correct, Avg_Question_Time, Avg_Attempt and Avg_Hint_Request. Among these variables, IRT_Proficiency_Estimate and Scaffold_Percent_Correct have positive coefficients, and Avg_Question_Time, Avg_Attempt and Avg_Hint_Request have negative correlation coefficients. Notice that student percent correct on original questions (IRT_Proficiency_Estimate) was considered the most significant predictor since it entered the final model earlier than other factors. However, the information from the tutoring session is helpful too, which indicates that performance on sub-steps, the speed, and the help-seeking behavior are additional, reliable indicators of students' level of knowledge acquisition. The mixed model is presented in Table 3 and the interpretation of the model is straightforward:

- Every point increase in IRT_Proficiency_Estimate (ranging from 0 to 1) adds 26.80 points to the prediction of MCAS score.
- Every one percent increase in Scaffold_Percent_Correct adds 20.42 points to the prediction of MCAS score.
- Average students' predicted score will reduce 0.170 points for every extra second students spent to finish a question.
- On average, if a student needs one more attempt to reach a correct answer for an item, s/he will lose 10.50 points in his/her predicted MCAS score
- On average, if a student requests one more hint, s/he will lose 3.22 points in his/her predicted MCAS score.

Table 3. Variables and coefficients of the mixed model (based on 2004-2005 data)

Order	Variables Entered	Coefficient	Standardized Coefficient	t	Sig.
0	(Constant)	32.414		6.136	0.000
1	IRT_Proficiency_Estimate	26.800	0.443	10.364	0.000
2	Scaffold_Percent_Correct	20.427	0.283	4.436	0.000
3	Avg_Question_Time	-0.170	-0.212	-6.941	0.000
4	Avg_Attempt	-10.500	-0.178	-5.485	0.000
5	Avg_Hint_Request	-3.217	-0.149	-2.175	0.030

The variables and corresponding coefficients of the final help model and the final assistance model are shown in Table 4 and Table 5. As we can see in Table 4, in addition to the IRT_Proficiency_Estimate, the average number of hint requests per question (Avg_Hint_Request) and the total number of hint requests (Hint_Request_Count) are also significant predictors of student performance on the MCAS test. Comparing Table 5 with Table 3, we notice that the assistance model is not nested within the mixed model. Without the IRT_proficiency_estimate, the step-wise regression added two other variables to the assistance model: original_count (total number of problems finished) and attempt_count (total number of attempts).

Table 4. Variables and coefficients of the help model (based on 2004-2005 data)

Order	Variables Entered	Coefficient	Standardized Coefficient	t	Sig.
0	(Constant)	17.021		6.816	0.000
1	IRT_Proficiency_Estimate	31.389	0.519	10.430	0.000
2	Avg_Hint_Request	-8.967	-0.416	-7.019	0.000
3	Hint_Request_Count	0.008	0.160	3.284	0.001

Table 5. Variables and coefficients of the assistance model (based on 2004-2005 data)

Order	Variables Entered	Coefficient	Standardized Coefficient	t	Sig.
0	(Constant)	39.698		7.124	0.000
1	Scaffold_Percent_Correct	16.383	0.227	3.348	0.001
2	Original_Count	0.114	0.572	7.775	0.000
3	Attempt_Count	-0.029	-0.473	-6.712	0.000
4	Avg_Question_Time	-0.131	-0.162	-3.941	0.000
5	Avg_Hint_Request	-6.933	-0.322	-4.774	0.000
6	Avg_Attempt	-5.349	-0.091	-2.236	0.026

We replicated the same modeling process on the data from the school year of 2005-2006. As mentioned before, the data is for 616 8th grade students from Worcester Public Schools. All of the four models were fitted to this data set and we summarize the results in

Table 6.

Table 6. Model Summary (based on 2005-2006 data)

Model	MAD	R Square	BIC	#variables	Correlation with 2006 8th grade MCAS
The lean model	5.77	0.615	-581	1	.784
p<.001 The help model	5.46	0.669	-656	4	.818
p=.693 The assistance model	5.39	0.666	-630	7	.816
p<.001 The mixed model	4.89	0.728	-763	6	.853

The results are similar to those obtained using the 2004-2005 data set where the mixed model is still the best fitting and most generalizable model. The adjusted predicted score correlated with the 2006 MCAS test scores at .853, which is reliably higher than the correlation between the other models and the MCAS scores. The ranking of the other three models is the same as the ranking we got from using the 2004-2005 data except that the difference between the help model and the assistance model is no longer reliable (p = .693). The assistance model and the mixed model are still reliably better than the lean model (p < .001).

Some readers may have noticed that no quadratic terms or interactions between factors were included in our models when building regression models. As a matter of fact, we suspected that there might be a non-linear relationship between the online measures and MCAS scores and therefore such a regression model was also trained. We obtained a much more complicated model; the models with quadratic and interaction terms fit (by BIC values) the best on training data of both years, yet they are not statistically reliablely better than the mixed models where no quadratic or interaction terms are included. With 25 variables entered (mostly interaction terms), the model becomes hard to interpret and potentially overfits[15] our data. Both for clarity and because our goal is generalization (e.g., better cross validation and BIC values) not just fit, the mixed model is our preferred model for MCAS score prediction.

4.4.2 Model validation

Now that we have trained models on data from two different years, we want to compare the models and see how different they are from each other. The coefficients of the mixed model trained based on 2005-2006 data is shown in Table 7. Comparing the model here with the model trained using 2004-2005 data as shown in Table 3, we noticed that the model is entered with 6 variables, yet the mixed model trained using 2004-2005 data was more parsimonious (5 variables). And it is worth pointing out that the assistance metrics that measure student help-seeking behaviors did not enter the model as they did for the 04-05 model (shown in Table 3). In both models, the student proficiency parameter estimated by the IRT model and student percent correct on scaffolding questions are the top two predictors.

[15] We applied the model trained based on 2004-2005 data on the data from 2005-2006 and the predicted value does not correlate with 2006 MCAS score very well. The correlation is at the same level as the lean model.

Table 7. Variables and coefficients of the mixed model (based on 2005-2006 data)

Order	Variables Entered	Coefficient	Standardized Coefficient	t	Sig.
0	(Constant)	3.284		2.224	0.027
1	IRT_Proficiency_Estimate	32.944	0.530	13.657	0.000
2	Scaffold_Percent_Correct	21.327	0.309	8.544	0.000
3	Question_Count	0.072	0.652	8.891	0.000
4	Avg_Question_Time	-0.102	-0.173	-5.498	0.000
5	Avg_Item_Time	0.045	0.154	5.432	0.000
6	Total_Attempt	-0.044	-0.550	-7.570	0.000

Since the models are not the same across years, the models might be overfitting the training data. Presenting BIC that penalizes a model by the number of variables entering the model and using adjusted predicted score that is calculated using "leave-one-out" cross validation have given us some protection against overfitting. On top of that, we now will investigate the issue further and explore how well the models generalize by validating the model using data in the same year and across years. We will evaluate the model fitting using the correlation and the MAD on the testing data.

The first thing we did is cross validating our model (the mixed model) using the same year's data. For this purpose, we allow SPSS to randomly select 50% of the 2004-2005 data as the training cases and trained the mixed model on the 203[16] selected training cases. We were glad to see that the mixed model correlated with the 2005 MCAS score with r equal to .842 (p < .01). The resulting model was very similar to the one shown in Table 3, where the same variables entered the model and the sign of all the coefficients were in the same direction as before, - only the values of the estimated coefficients changed slightly. We then fit the model on the 188 testing cases of the same year, and found out that the model fit well on the testing data (r = .837, MAD = 5.25). In the next step, we trained the model on the second half of the data and tested on the first half, and got similar result. Thus, we verified that our mixed model worked very well on the data of the same school year (i.e. the same group of students, and the same MCAS tests, etc.)

Now that we have validated our model inside a year, we further test its validity using data across years. We use the 2005-2006 data as the testing set for the model trained over the 2004-2005 data and vice versa and we will do the validation on the mixed model as it is the preferred model[17]. It turned out that both models work fairly well on the testing set. As presented in Table 8, the predicted scores tested on the 2005-2006 data correlated with the 2006 MCAS with the correlation coefficient equal to .827 and the correlation is .824 for the model tested on 2004-2005 data. And both correlations are significant at the .01 level. Therefore, even though the models constructed based on data from different years are not quite the same, they are both quite

[16] Sharp readers may notice 203 is not exactly 50% of 391, the total number of students in the data set. The fact is, SPSS finished the 50% sampling and returned 203 rows, and we decided not to change it.

[17] Since the range of MCAS raw score is 0 to 54, if the model predicted a score larger than 54 on the testing data, we only assigned a score of 54; and if the model predicted a negative score, we assigned a score of zero.

predictive of student end-of-year exam scores. The strong correlations did a good job of protecting us from overfitting the data.

However, we noticed that the MADs on testing data sets are relatively high. Given that the correlations are solid, we suspected that the models were overall under-predicting or over-predicting on the testing data sets. Therefore, we generated the scatter plots for both testing sets (shown in Figure 4). As we can see from the scatter plots (the fit line is shown), the model trained over 2005-2006 data under-predicted the 2005 MCAS scores for the students who used the ASSISTment system during the school year 2004-2005; while the model trained over the 2004-2005 data over-predicted the 2006 MCAS score for our users during 2005-2006. Why? Presumably, if our models were really overfitting, we probably would not get a solid correlation; and since the IRT_Proficiency_Estimate in the mixed model can account for student difference, we speculate that if the difficulty of the MCAS tests shifted in the two years that would directly have an impact on student raw scores[18] though their performance level wouldn't be affected because of the equalization procedure as described in section 4.1.

Table 8. Results of testing the mixed models on a different year's data

Year of training data	Year of Testing data	MAD on Testing data	Correlation with MCAS scores on testing data	MAD after correction
2004-2005	2005-2006	5.82	.827	5.16
2005-2006	2004-2005	6.55	.824	5.80

To investigate whether the difficulty of the MCAS tests changed from the year 2005 to 2006, we computed the average percent correct of all the students from Worcester for both tests. Overall, the students from Worcester got 42% correct on the 2005 MCAS test and 46% correct on the 2006 MCAS test which is statistically reliably higher than that of the 2005 test suggested by the two sample t-test of the two groups of students ($p < .05$). This indicates that the 2006 MCAS test was reliably easier than the 2005 MCAS test for students from Worcester. The result was confirmed by the fact that the threshold raw scores shifted 3 points higher in the year 2006 – students were required to score higher in 2006 to achieve the same performance level that their score would have received in 2005. We consider this finding as an explanation of why the 2006 MCAS scores were under-predicted by the 2004-2005 mixed model while the 2005-2006 model over-predicted the 2005 MCAS score. Given this, we tried a very simple correction on the predicted score by subtracting 3 points from the scores predicted by the 2005-2006 model and adding the same to the scores predicted by the 2004-2005 model. This simple correction lowered the MADs by .7 points in both cases. Namely, the MAD drops from 5.89 to 5.16 when we test our model on 2005-2006 data, and MAD drops from 6.49 to 5.80 when we use 2004-2005 data as the testing set.

[18] Recall that we have been using MCAS raw scores as the dependent variables in our models which are affected by the variant in the difficulty of the tests.

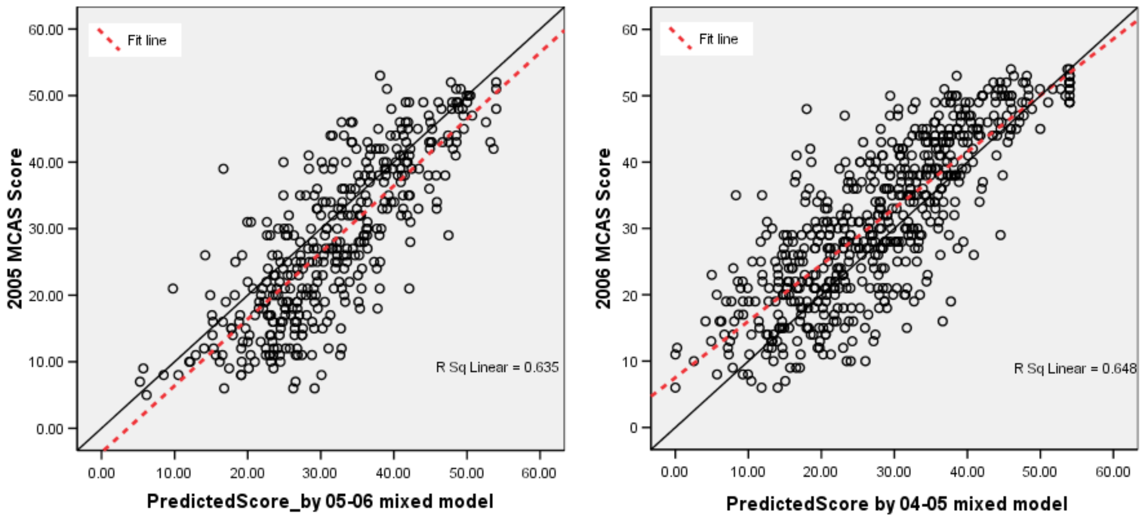


Figure 4. Scatter plot of predicted scores on testing data vs. MCAS raw scores

In summary, our models work fairly well on both the same year's data and data from different years. There are a number of possible reasons why the regression modeling process picked up different groups of predictors for data logged in different years including random variation and differences in the logging system, the user interface and the problem sets from one year to the next. Nevertheless, both models generalized and fit well on the testing set with some offsets caused by the changing of the MCAS tests themselves. With regard to the validation result, we claim that our dynamic testing approach can reliably improve the prediction of student end-of-year test scores by taking into consideration the help-seeking behavior, the speed and the number of attempts needed to answer questions.

5 CONCLUSION & IMPLICATIONS

In this paper, we addressed the assessment challenge in the ASSISTment system, which is a web-based tutoring system that serves as an e-learning and e-assessment environment. We focused on the assessment ability of the system and evaluated it by mining our log data and comparing with standardized test results. Some evidence was presented that the online assessment system did a better job of predicting student knowledge by being able to take into consideration how much tutoring assistance was needed, how fast a student solves a problem and how many attempts were needed to finish a problem.

Traditional assessment usually focuses on students' responses to test items and whether they are answered correctly or incorrectly. It ignores all other student behaviors during the test (e.g., response time). However, in this work, we take advantage of a computer-based tutoring system to collect extensive information while students interact with the system. Our results showed that the assistance model that includes no assessment result on the main problems leads to significantly better predictions than the lean model that is based on the assessment results alone. While the mixed model leads to the best predictions, the relative success of the assistance model over the lean model highlights the power of the assistance measures. Not only is it possible to get reliable test information while "teaching on the test", data from the teaching process actually improves reliability.

A critic may argue that it is not fair to have the lean model as a contrast case as students were not spending all their time on assessment. Whether or not the ASSISTment system would yield better predictions than such a tougher contrast case, where students only spend on-line time on assessment and not on instruction, is an open question worthy of further research. However, we would remind that critic that such a contrast would leave out the instructional benefit of the ASSISTment system and, moreover, might not be as well received by teachers and students.

The more general implication from this research suggests that continuous assessment systems are possible to build and that they can be quite accurate at helping schools get information on their students. This result is important because it provides evidence that reliable assessment and instructional assistance can be effectively blended. These results with the ASSISTment system open up a the possibility of a completely different approach to assessment that is contentious in nature in suggesting students may not need to spend any time on formal paper and pencil tests. Many states are moving towards adopting "*value added*" assessments, so that they can track the *value added* by teachers and schools. *Value added* is possible because you have year to year state assessments so you can see the average learning gain for students per year, and attribute those gains to teachers and schools. Such systems could benefit from data that is collected every few weeks, instead of once a year, thereby allowing schools to more quickly figure out *what works* at increasing student learning. Because the ASSISTment system teaches while it assesses, it makes collecting test data almost transparent to the students and teachers. One might be concerned that using the ASSISTment system may take longer than taking a paper practice test. However, unlike paper tests, the ASSISTment system is contributing to instruction (Razzaq et al., 2005; Razzaq, Heffernan, & Lindeman, 2007; Feng, Heffernan, Beck & Koedinger, 2008). While every minute spent on a paper test takes away a minute of instruction, every minute on the ASSISTment system contributes to instruction.

We end with a tantalizing question: Are we likely to see states move from a test that happens once a year, to an assessment tracking system that offers continuous assessment (Computer Research Association, 2005) every few weeks? While more research is warranted, our results suggest that perhaps the answer should be yes.

ACKNOWLEDGEMENTS

We would like to acknowledge Dr. Joseph Beck for his insightful comments on this research.

This research was made possible by the U.S. Department of Education, Institute of Education Science (IES) grants, "Effective Mathematics Education Research" program grant #R305K03140 and "Making Longitudinal Web-based Assessments Give Cognitively Diagnostic Reports to Teachers, Parents, & Students while Employing Mastery learning" program grant #R305A070440, the Office of Naval Research grant # N00014-03-1-0221, NSF CAREER award to Neil Heffernan, and the Spencer Foundation. All the opinions, findings, and conclusions expressed in this article are those of the authors, and do not reflect the views of any of the funders.

REFERENCES

[1] Anozie N., & Junker B. W. (2006). Predicting end-of-year accountability assessment scores from monthly student records in an online tutoring system. In Beck, J., Aimeur, E., & Barnes, T. (Eds). *Educational Data Mining: Papers from the AAAI Workshop.* Menlo Park, CA: AAAI Press, pp. 1-6. Technical Report WS-06-05.

[2] Ayers E., & Junker B. W. (2006). Do skills combine additively to predict task difficulty in eighth grade mathematics? In Beck, J., Aimeur, E., & Barnes, T. (Eds). *Educational Data Mining: Papers from the AAAI Workshop.* Menlo Park, CA: AAAI Press, pp. 14-20. Technical Report WS-06-05.

[3] Baker, R. S., Corbett, A. T., Koedinger, K. R. (2004). Detecting Student Misuse of Intelligent Tutoring Systems. In James, C.L., Vicari, R.M., & Paraguacu, F. (Eds.). *Intelligent Tutoring Systems: 7th International Conference ITS 2004, Maceió, Alagoas, Brazil Proceedings*. Berlin, Germany: Springer-Verlag Berlin Heidelberg. pp. 531-540.

[4] Baker, R. S., Roll, I., Corbett, A. T., Koedinger, K. R. (2005). Do Performance Goals Lead Students to Game the System? In *Proceedings of the 12th International Conference on Artificial Intelligence in Education*. Amsterdam, Netherlands. pp. 57-64.

[5] Beck, J. E., Jia, P., & Mostow, J. (2004). Automatically assessing oral reading fluency in a computer tutor that listens. *Technology, Instruction, Cognition and Learning*, 2, 61-81.

[6] Beck, J. E., & Sison, J. (2006). Using knowledge tracing in a noisy environment to measure student reading proficiencies. *International Journal of Artificial Intelligence in Education*, 16, 129-143.

[7] Black, P. & Wiliam, D. (1998a). Assessment and classroom learning. *Assessment in Education: Principles, Policy and Practice*, 5,7-74.

[8] Black, P., & Wiliam, D. (1998b). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, 80(2): 139-149.

[9] Boston, C. (2002). The concept of formative assessment. *Practical Assessment, Research & Evaluation*, 8(9).

[10] 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*, New York: W. H. Freeman. pp. 103-126.

[11] Corbett, A.T. & Bhatnagar, A. (1997). Student modeling in the ACT Programming Tutor: Adjusting a procedural learning model with declarative knowledge. *User Modeling: Proceedings of the Sixth International Conference on User Modeling UM97 Chia Laguna, Sardinia, Italy*. New York: Springer-Verlag Wein. pp.243-254.

[12] Corbett, A. T., Anderson, J. R., & O'Brien, A. T. (1995) Student modeling in the ACT programming tutor. In Nichols, P., Chipman, S., & Brennan, R. (eds.), *Cognitively Diagnostic Assessment*. Hillsdale, NJ: Erlbaum.

[13] Computer Research Association. (2005). Cyberinfrastructure for Education and Learning for the Future: a Vision and Research Agenda. Final report of Cyberlearning Workshop Series workshops held Fall 2004 - Spring 2005 by the Computing Research Association and the International Society of the Learning Sciences. Retrieved from http://www.cra.org/reports/cyberinfrastructure.pdf on November 10th, 2006

[14] Embretson, S. E. (1992). Structured Rasch models for measuring individual-difference in learning and change. *International Journal of Psychology*. 27(3-4):372-372.

[15] Feng, M., Heffernan, N.T., & Koedinger, K.R. (2006). Addressing the Testing Challenge with a Web-Based E-Assessment System that Tutors as it Assesses. In Leslie A. Carr, David C. De Roure, Arun Iyengar, Carole A. Goble, Michael Dahlin (Ed.), *Proceedings of the Fifteenth International World Wide Web Conference*. New York, NY: ACM Press. pp. 307-316. Edinburgh, UK, 2006.

[16] Feng, M., & Heffernan, N. T., (2007). Towards Live Informing and Automatic Analyzing of Student Learning: Reporting in the Assistment System. *Journal of Interactive Learning Research.* 18 (2), pp. 207-230. Chesapeake, VA: AACE.

[17] Feng, M, Heffernan, N., Beck, J., & Koedinger, K. (2008). Can we predict which groups of questions students will learn from? In Baker & Beck (Eds.). *Proceedings of the First International Conference on Educational Data Mining*. pp. 218-225. Montreal, Canada, 2008.

[18] Feng, M, Beck, J., Heffernan, N., & Koedinger, K. (2008). Can an Intelligent Tutoring System Predict Math Proficiency as Well as a Standardized Test? In Baker & Beck (Eds.). *Proceedings of the First International Conference on Educational Data Mining*. pp. 107-116. Montreal, Canada, 2008.

[19] Fischer, G. & Seliger, E. (1997). Multidimensional Linear Logistic Models for Change. Chapter 19 in van der Linden, W. J. and Hambleton, R. K. (eds.) (1997). *Handbook of modern item response theory*. New York: Springer-Verlag.

[20] Grigorenko, E. L., & Sternberg, R. J. (1998). Dynamic Testing. *Psychological Bulletin*, 124, 75-111.

[21] Jannarone, R. J. (1986) Conjunctive item response theory kernels. *Psychometrika*, 55 (3): 357-373.

[22] Koedinger, K. R., Aleven, V., Heffernan. N. T., McLaren, B. & Hockenberry, M. (2004). Opening the Door to Non-Programmers: Authoring Intelligent Tutor Behavior by Demonstration. In *Proceedings of the 7th International Conference on Intelligent Tutoring Systems*, pages 162-173, Maceio, Brazil.

[23] Massachusetts Department of Education. (2000). Massachusetts Mathematics Curriculum Framework. Retrieved from http://www.doe.mass.edu/frameworks/math/2000/final.pdf, Nov. 6th, 2005.

[24] MCAS technical report (2001). Retrieved from http://www.cs.wpi.edu/mfeng/pub/mcas_techrpt01.pdf, August 5th, 2005.

[25] Mitchell, T. (1997). *Machine Learning.* Columbus, OH: McGraw-Hill.

[26] Mostow, J., & Aist, G. (2001). Evalutating tutors that listen: An overview of Project LISTEN. In P. Feltovich (Ed.), *Smart Machines in Education*. Menlo Park, CA: MIT/AAAI Press. pp. 169-234.

[27] Olson, L. (2004). State Test Programs Mushroom as NCLB Mandate Kicks In. In *Education Week,* Nov. 20th, pp. 10-14.

[28] Olson, L. (2005). Special report: testing takes off. *Education Week*, November 30, 2005, pp. 10–14.

[29] Raftery, A. E. (1995). Bayesian model selection in social research. In *Sociological Methodology*, 25, 111-163.

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

[31] Razzaq, L., Heffernan, N.T., & Lindeman, R.W. (2007). What level of tutor interaction is best?. In Luckin & Koedinger (Eds.). *Proceedings of the 13th Conference on Artificial Intelligence in Education*. Amsterdam, Netherlands: IOS Press. pp. 222-229. Los Angeles, CA, 2007.

[32] 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 *Proceedings of the 12th Annual Conference on Artificial Intelligence in Education*. Amsterdam, Netherlands: ISO Press. pp. 555-562. Amsterdam, 2005.

[33] Roediger, H.L. III, and Karpicke, J.D. (2006). The power of testing memory. *Perspectives on Psychological Science*. 1(3), pp. 181- 210.

[34] Sternburg, R.J., & Grigorenko, E.L. (2001). All testing is dynamic testing. *Issues in Education*, 7, 137-170.

[35] Sternburg, R.J., & Grigorenko, E.L. (2002). *Dynamic testing: The nature and measurement of learning potential*. Cambridge, England: Cambridge University Press.

[36] Tan, E. S., Imbos, T. & Does R. J. M. (1994) A distribution-free approach to comparing growth of knowledge. *Journal of Education Measurement*, 31 (1):51-65.Tatsuoka, K.K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. Journal of Educational Measurement, 20, 345-354.

[37] van der Linden, & W. J., & Hambleton, R. K. (eds.) (1997). *Handbook of modern item response theory*. New York, NY: Springer Verlag.

[38] Walonoski, J., & Heffernan, N.T. (2006). Detection and Analysis of Off-Task Gaming Behavior in Intelligent Tutoring Systems. In Ikeda, Ashley & Chan (Eds.). In *Proceedings of the 8th International Conference on Intelligent Tutoring Systems*. Berlin: Springer-Verlag. pp. 382-391. Jhongli, Taiwan, 2006.

[39] Zimowski, M., Muraki, E., Mislevy, R., & Bock, D. (2005). BILOG-MG 3 – Multiple-Group IRT Analysis and Test maintenance for Binary Items. Scientific Software International, Inc., Lincolnwood, IL. URL http://www.ssicentral.com/.