Feng, M., Heffernan, N.T., & Koedinger, K.R. (2006b). **Predicting state test scores better with intelligent tutoring systems: developing metrics to measure assistance required**. In Ikeda, Ashley & Chan (Eds.). *Proceedings of the 8th International Conference on Intelligent Tutoring Systems*. Springer-Verlag: Berlin. pp. 31-40. 2006.

Predicting State Test Scores Better with Intelligent Tutoring Systems: Developing Metrics to Measure Assistance Required

Mingyu Feng[1], Neil T. Heffernan[1], and Kenneth R. Koedinger[2]

[1] Computer Science Department, Worcester Polytechnic Institute
Worcester, MA, USA
{mfeng, nth}@wpi.edu
[2] Human-Computer Interaction Institute, Carnegie Mellon University
Pittsburgh, PA, USA
{koedinger}@cmu.edu

Abstract. The ASSISTment system was used by over 600 students in 2004-05 school year as part of their math class. While in [7] we reported student learning within the ASSISTment system, in this paper we focus on the assessment aspect. Our approach is to use data that the system collected through a year to tracking student learning and thus estimate their performance on a high-stake state test (MCAS) at the end of the year. Because our system is an intelligent tutoring system, we are able to log how much assistance students needed to solve problems (how many hints students requested and how many attempts they had to make). In this paper, our goal is to determine if the models we built by taking the assistance information into account could predict students' test scores better. We present some positive evidence that shows our goal is achieved.

1 Introduction

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. To help resolve this dilemma, assistance and assessment are integrated in a web-based intelligent tutoring system ("ASSISTment") that offers instruction to students while providing detailed evaluation of their abilities to the teachers. In the 2004-2005 school year some 600+ students used the system about every two weeks to practice their skills on 8th grade Math items. These students were presented with randomly selected Massachusetts Comprehensive Assessment System (MCAS)[1] test items. If students got the *original item* correct they were given a new one, otherwise they were provided with a small "tutoring" session where they were forced to answer a few *scaffolding questions* that broke the problem down into steps. By doing this, the ASSISTment system is able to differentiate students who get the

[1] http://www.doe.mass.edu/mcas.

same original item wrong at first but need different levels of tutoring to get the problem correct eventually. For instance, suppose Tom, Dick and Harry all got the same original item wrong, but Tom needed to only ask for one hint to finish the whole item, Dick had to ask for 5 hints on one question and took a very large amount of time answering, while Harry needed no help on any of the scaffolding questions. Given these students asked for different amount of instructional assistance, we could expect Harry's MCAS score higher than Tom and Dick's. Essentially, our assistance metrics (measuring hint requests, timing information etc. as discussed in Section 3.1) are partial credit metrics and this paper asks if we can do a better job of predicting MCAS score using these assistance metrics. For those who are interested in knowing if students learn from the computer, please see [7] and [8], as this paper models primarily students' learning due to their classroom instruction.

Providing instructional assistance in the process of assessing students is the key feature of the ASSISTments. The hypothesis is that the ASSISTments can do a better job of assessing student knowledge than practice tests or other on-line testing approaches by using a "dynamic assessment" approach, thus providing a more precise prediction of student performance on the MCAS test. Feng, Heffernan and Koedinger [5] showed that by introducing the assistance students required as parameters, we were able to construct a better fitted regression model to predict students' performance on MCAS than simply using their performance on original items or on paper and pencil tests. Meanwhile, the longitudinal analysis approach[2] [9] has been applied to track student learning over time. In this paper, we propose a new method of MCAS score prediction by combining these two parts (i.e. regression model fitting plus longitudinal analysis). Specifically, our research question is:

Research question: Can we make a more precise prediction of students' performance on the MCAS by using assistance data longitudinally?

We had presented preliminary estimates of students' MCAS scores as a single column in one of our online teacher reports [4], the "Grade Book" report. The prediction was made based only upon student response on the original items. So it can not distinguish Tom, Dick and Harry in the example above. Besides, the predicted value was generated cumulatively: all past data was utilized equally while *time*, an important factor on student learning, was ignored. A positive answer to the research question would help us improve our reports.

Students' monthly performance on the original items was selected as the variable whose change we tracked longitudinally in our former work, while in this work we created two new variables **original_predicted_score** and **assistance_predicted_score** by applying regression models. The calculation of the two variables will be discussed in detail in Section 3. To answer our research question, we ran a longitudinal analysis to

[2] "Singer and Willet" style longitudinal data analysis is an approach for investigating change over time. It allows us to learn a slope (i.e., learning rate) and intercept (i.e. an estimate of incoming knowledge) for the group as a whole and for each individual student. This is achieved by fitting a multilevel model that simultaneously builds two sub-models, in which level-1 sub-model fits *within-person* change and describes how individuals change over time and level-2 sub-model tracks *between-person* change and describes how these changes vary across individuals This method extends well to allow us to ask questions like "Is student learning different in different schools, for different teachers or different classes?").

track the change of these two variables over time, obtained the prediction of students' MCAS scores in May 2005 and then compared the accuracy of the models as measured in *Median Absolute Difference* (MAD) – the average of absolute residual of the predicted score and students' real score in MCAS. In Section 3.5, we present the evidence that shows using the **assistance_predicted_score,** which took into account the amount of assistance students required, we did a better job estimating students' MCAS score.

2 Related Work

Other researchers have been interested in trying to get more assessment value by comparing traditional assessment (students getting an item marked wrong or even getting partial credit) with a measure that shows how much help they needed. Campione et al. [3] compared traditional testing paradigms against a dynamic testing paradigm. Grigorenko and Sternberg [6] 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 not to be making progress, would be given a prewritten hint. If the student was still not making progress, another prewritten hint was presented and the process was repeated. In this study they wanted to predict learning gains between pretest and posttest. They found that static testing was not as well correlated (R = 0.45) with student learning data as with their "dynamic testing" (R = 0.60) measure. Campione et al. suggested that this method could be effectively done by computer, but, as far as we know, their work was not continued. Beck et al. [2] examined using speech recognition measures (speed and correctness of reading) and student help requests to estimate reading proficiency and showed that a model can do a better job at estimating proficiency when taking into consideration student help requests. Luckily, the ASSISTment system provides an ideal test bed as it already provides a set of hints to students. In [5], we extended and tested Campione's hypothesis and replicated their finding of ASSISTment-style measures (including but not limited to student help requests) being effective and even better assessors.

3 Approach

Our new approach of MCAS score prediction combines assistance students required and the effect of time. It contains the following steps: a) Split the 494 students into training and testing sets; b) train regression models on the training set and obtain the variables entered in the models and their associated coefficients; c) apply regression models to the testing set and calculate the values of the variables **original_predicted_score** and **assistance_predicted_score** for each student for every month; d) longitudinally track student knowledge using original_predicted_score and assistance_predicted_score as an outcome variable; e) predict student MCAS score in May 2005 given the result of step d); f) compare the two outcome variables based on the MCAS score prediction result and answer the research question.

3.1 Data Source

For 2004 - 2005 school year, we collected data from 494 students who were using the ASSISTment system from September 16, 2004 through May 16, 2005 for an average of 249 minutes and finished an average of 135 items. Given the fact that the MCAS test was given on May 17, 2005, it would be inappropriate to use data after that day for the purpose of predicting MCAS scores. We also excluded data from the students' first day of using the ASSISTment system since they were learning how to use the system at that time. Though more than 600 students used our system, we were only able to collect integral data for 494 students as MCAS scores and/or the results of the paper practice test were missing for the rest. A student's raw MCAS score is out of 54 points, where each correct multiple choice or short answer question earns a point and a full correct answer to open response questions[3] earns 4 points. The paper practice test (we will refer to as *pretest)* was administered in September 2004. Students were asked to finish the test in two periods over two days (totally 80 minutes) and scores of this test were shown to be a significant predictor of MCAS scores in [5].

We constructed 15 "online measures" that we think indicate the amount of assistance a student needs to get an item correct. These online measures are:

- original_percent_correct – students' percent correct on original items only, which we often referred to as "static metric". Apparently, this measure correlates positively with knowledge.
- original_count – the number of original items students have done. This measures students' attendance and on-task-ness. The metric also reflects students' knowledge since better students have a higher potential to finish more items in the same period of time.
- percent_correct – students' percent correct over **all** questions (both original items and scaffolding questions). In addition to the original items, students' performance on the scaffolding questions is also a reasonable reflection of their knowledge. For instance, students who failed on the original items simply because of their lack of ability of forming problem-solving strategies will probably answer all the scaffolding questions correctly.
- question_count – the number of questions (both original items and scaffolding questions) students have finished. Similar to original_count, this is also a measure of attendance and knowledge but given the fact that scaffolding questions show up only if students failed the original items, it is not straightforward 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.
- hint_count – the total number of hints students received.
- avg_hint_count – the number of hint messages students received averaged over all questions.
- bottom-out_hint_count – the number of bottom-out[4] hint messages students got.

[3] Open response questions are not supported by the ASSISTment system currently.

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

- avg_bottom_hint – the average number of bottom-out hint messages students got.
- attempt_count – the total number of attempts students made.
- avg_attempt – the average number of attempts made for each question.
- avg_question_time – on average, the length of time it takes for a student to answer a question, measured in seconds.
- avg_item_time – on average, the length of time it takes for students to finish a problem (including all scaffolding questions if students answered the original items incorrectly).

The ten measures above are generally all *ASSISTment style*, dynamic assessment metrics indicating the amount of assistance students need to finish problems and the amount of time they spend to finish items. Therefore, we hypothesize all these measures would be negatively correlated with MCAS scores.

- total_minutes – the total number of minutes students worked on items in the ASSISTment system. Just like original_count, this metric is an indicator of attendance. Our hypothesis is that this measure will positively correlate with MCAS score with regard to the result we reported in [7] that students learned in the ASSISTment system.

3.2 Constructing Training and Testing Data Set

Among the 494 students, we selected approximately 50% as training individuals to train up regression models, leaving 244 students in the testing set. For the training individuals, we created a file of 250 rows with one row per student. Each row includes variables representing their associated real MCAS score, the student's pretest scores, and 15 "online measures" which we think indicate the amount of assistance a student needs to get an item correct.

In contrast to the training set, data for the 244 testing individuals are organized in the "person-period" style [9] to facilitate longitudinal analysis. To run a longitudinal data analysis, the first thing to decide is a sensible metric for time. Because a student only worked on the ASSISTments for one period (about 20 to 40 minutes, varies among schools) every time they came to the lab, rather than treating visiting days as the metric for time, we collapsed all data in one month and used *month* as the level of granularity to measure time to achieve more stable learning-over-time data. This variable for time is called "CenteredMonth" since it is centered around September 16, 2004 and it runs from 0 to 7. Rows in which CenteredMonth equals 0 contain data from Sep. 16 to October 16, and rows where CenteredMonth equals 1 contain data from October 17 to November 16 and so on. The "person-period" structured dataset contains on average 5 data waves for each student and values of all the online measures for each CenteredMonth were calculated.

To analyze data longitudinally, another important thing to determine is an outcome whose values change systematically over time. As mentioned in Section 1, traditionally students' percent correct on the original items was treated as an outcome. To mimic the real MCAS score, we multiplied the percent correct by 54 (the full MCAS score), which makes the outcome range change to 0~54. We will refer to this variable as **plain_predicted_score** to emphasize the fact that it is computed directly from

students' monthly performance on original items without any correction. In addition, two new variables, referred to as **original_predicted_score** and **assistance_predicted_score**, will be calculated by applying the regression models that were trained using the training data set. The calculation of the variables will be discussed in detail in the following sections. All three predicted scores will be used as the outcome variable individually in our longitudinal data analysis and results will be compared.

3.3 Building Regression Models Based on Training Data

For a long time, we have observed that the ASSISTment system was consistently under-predicting student performance due to the following reasons. Firstly, when building the ASSISTments, authors changed the type of many questions from multiple choice to text input questions, which makes the ASSISTments on average harder than the actual MCAS items. Secondly, the ASSISTment system always allows students to ask for hints, which to some degree prevents students from trying their best to get the solution. Since hint requests were treated as false responses, this feature could impact students' evaluation. Thirdly, students did not take the ASSISTments as seriously as a real, high-stakes test such as the MCAS and finally they may behave differently when working on a computer because they like or dislike computers [1]. Therefore we want to take advantage of regression models to adjust the predicted scores.

First of all, we checked the correlations between MCAS scores and all independent variables (pretest and the 15 online measures) in the training dataset. All these factors except attempt_count turned out to be significantly correlated with MCAS scores ($p < 0.05$). The highest correlation ($r = 0.742$) occurs between MCAS score and pretest scores. Among all the online measures, original_percent_correct correlates best with MCAS score ($r= 0.709$). And the sign of the correlations verified our hypothesis about the relationships between the online measures and the MCAS score.

Table 1. Regression Models

Model	Parameter	Un-std. Coeff.	Std. Coeff.
Original_ Regression_ Model	(Constant)	4.753	
	pretest	.764	.496
	original_percent_correct	27.869	.367
Assistance_ Regression_ Model	(Constant)	26.035	
	pretest	0.64	.415
	percent_correct	24.205	.307
	avg_attempt	-10.56	-.202
	avg_hint_request	-2.283	-.125

We ran stepwise linear regressions to predict students' real MCAS scores using pretest scores plus original_percent_correct, and pretest scores plus all of the online measures respectively. The models, named Original_Regression_Model and Assistance_Regression_Model, are summarized in Table 1.

The interpretation of Table 1 is straightforward. Because of the lack of space we will only present the interpretation for Assistance_Regression_Model.

- Every one point increase in the pretest adds 0.64 points to the prediction of MCAS score. This is also the most significant parameter in both of the models according to standardized coefficients.
- It was percent_correct, not original_percent_correct that entered the model, which indicates that students' response to scaffolding questions should not be ignored when evaluating their knowledge. One percent increase on the percent correct earns student 0.24 points in the predicted MCAS score.
- The coefficient of the parameter avg_attempt is negative and thus consistent with our hypothesis about this measure. On average, if a student needs one more attempt to reach a correct answer for an item, he/she will lose 10.56 points in his/her predicted MCAS score.
- Similar to avg_attempt, avg_hint_request is also negatively correlated with MCAS score. The difference is that students' predicted score will be penalized for only 2.28 points for every hint request averaged over all questions.

3.4 Tracking Two Outcomes Longitudinally

Given Table 1, we constructed the following formulas to compute values for the two new variables that represent student knowledge in a certain month:

original_predicted_score = 4.753 + pretest * 0.764 + original_percent_correct * 27.869

assistance_predicted_score = 26.035 + pretest * 0.64 + percent_correct * 24.205 - avg_attempt * 10.56 -avg_hint_request * 2.283

It is worth pointing out that using the above formula, **assistance_predicted_score** takes into account student performance on scaffolding questions together with the amount of assistance, in particular, the number of attempts and hints, students need on average to get an item correct.

Given this data set, we fit mixed-effect models ([9], also referred to as *multilevel linear models* in sociological research) on the testing data set and continuously track **original_predicted_score** and **assistance_predicted_score** respectively. The modeling was conducted in SPSS. In [5], *school* was discovered to be a significant predictor of both students' initial knowledge status and learning rates. Hence here we introduced school as a predictor again. To facilitate our discussion, we will refer to the two models as *Original_Mixed_Model* when original_predicted_score was picked as the outcome variable and *Assistance_Mixed_Model* when assistance_predicted_score was used as the outcome variable respectively. Each model gave two parameters for any individual student, intercept (representing initial knowledge status in the first month) and slope (denoting learning rate across the 8 months).

3.5 Which is the Best Model That will Predict MCAS Scores?

Recall that our research question asked whether a more precise prediction can be achieved by taking into account the assistance information. To investigate this ques-

tion, we computed the MAD result from the above models. Naturally, the predicted scores for the last month (i.e. CenteredMonth = 7) were adopted as the predicted MCAS score.

With predicted MCAS scores available, we can calculate MAD for both models. For the Original_Mixed_Model, we got a MAD of 6.20, with a standard deviation equal to 4.72 while for the Assistance_Mixed_Model the MAD is 5.533 with standard deviation being 4.40. Consequently, we claim that the Assistance_Mixed_Model, by utilizing the dynamic online metrics, helps to improve the correctness of the prediction on MCAS score. The paired t-test comparing absolute residuals of each student indicates the improvement is statistically significant ($p = 0.011$).

3.6 More Results

Sharp readers may have noticed that in Section 4.2, no quadratic terms or interactions between factors were included 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 and assistance_predicted_score computed. Though the R^2 of the non-linear model is higher than that of the Assistance_Regression_Model, it led to significantly larger MAD. The non-linear model probably over-fitted the training data and was thus disregarded. In both regression models presented in Table 1, pretest was a significant parameter. We wondered how much the tutoring and assistance information can help without pretest because pretest scores are not always available every school year. We replicated the whole process without using pretest. A comparison of evaluation measures to corresponding values in the above sections shows that pretest is an important predictor and without it, the precision of prediction degrades; meanwhile, the model involving tutoring and assistance information still exhibits its superiority and the difference in MAD is almost statistically significant ($p = 0.055$).

3.7 Can We Do Better, or Are We Done?

In Section 4.4, we presented that we achieved a MAD of 5.533 when predicting MCAS score using the Assistance_Mixed_Model, which is about 10.2% of the full score. To see how good the prediction is, we compare this prediction to the prediction reached by 3 other approaches as measured by MAD scores.

Among other things, pretest scores alone could be used for prediction purposes. So we did a simple regression to predict student's real MCAS scores using associated pretest scores and ended up with a MAD of 6.57 that was statistically significantly higher ($p<0.05$) than the 5.533 scores from Section 4.4.

For a second comparison we looked at the predictions in the "Grade Book" reports to teachers on our current web site (Shown in Figure 1). The prediction was primitive and was simply a linear function of percent correct on original items. For students in the testing data set, this approach gave a MAD equal to 7.47.

In yet a third comparison, we can compare it to using the *plain_predicted_score* as an outcome variable in the longitudinal analysis which brought on a MAD of 9.13. Obviously, all three of these comparisons show higher MAD values, thus indicated that they are not as good at predicting MCAS scores.

Note that the comparison between pretest-prediction-method and the ASSISTment approach confounds total time during the assessment (80 vs. 249 minutes) in the sense that it took only about 80 minutes to do the paper and pencil pretest. However, we argue that this is a fair comparison, because our schools (6 schools have adopted the system this year) say they are willing to use the ASSISTments often because they think that students are learning during their use of the ASSISTment web site.

In Section 4, we found we had reduced the MAD to 5.533, but can we do better? Should we be dissatisfied unless we can get a MAD of zero? We want to investigate what a reasonable comparison should be. Ideally, we wanted to see how good one MCAS test was at predicting another MCAS test. We could not hope to do better than that. We did not have access to data for a group of kids that took two different versions of the MCAS test to measure this, but we could estimate this by taking students' scores on MCAS, randomly splitting the test in half, and then using their score on the first half to predict the second half. We excluded open response questions from the MCAS 2005 test and kept the remaining 34 multiple-choice and short answer questions with regard to the fact that open response questions are not supported in the ASSISTment system. Then the 34 items were randomly split into two halves and student performance on one half was used to predict their performance on the other half. This process was repeated 5 times. On average, we got MAD of 1.89, which is about 11% of the full score (17 points with one point for each item). Thus we drew the conclusion that using the new approach, our prediction of MCAS score is as good as the real MCAS test itself, with the caveat that only 34 items were utilized in the process here, while our prediction models were built based on students' work on 135 ASSISTment items over eight months.

4 Conclusion and Future work

In this paper, we continued our work in [5] and proposed a new method of MCAS score prediction by integrating timing information and the amount of assistance a student needs. To evaluate the method we compared this new method to some traditional methods. Evidence was presented that the new method did a better job of predicting student knowledge than traditional methods which only looked at students' performance on original items because items can be broken down into steps and students' responses to those steps are taken into consideration in the prediction. As our future work, we will evaluate the method further using this year's data and improve the teacher reporting system utilizing the new method.

Acknowledgements

This research was made possible by the US Dept of Education, Institute of Education Science, "Effective Mathematics Education Research" program grant #R305K03140, the Office of Naval Research grant #N00014-03-1-0221, NSF CAREER award to Neil Heffernan, and the Spencer Foundation. All of the opinions in this article are those of the authors, and not those of any of the funders.

We thank Professor Brian Junker from Carnegie Mellon University for helpful comment on this work. This work would not have been possible without the assistance of the 2004-2005 WPI/CMU ASSISTment Team including Brian Junker at CMU, Andrea Knight, Abraao Lourenco, Michael Macasek, Goss Nuzzo-Jones, Kai Rasmussen, Leena Razzaq, Steven Ritter at Carnegie Learning, Carolyn Rose at CMU, Terrence Turner, Ruta Upalekar, and Jason Walonoski.

References

1. 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 and Education*, 57-64.
2. Beck, J. E., Jia, P., Mostow, J. (2004). Automatically assessing oral reading fluency in a computer tutor that listens. In *Technology, Instruction, Cognition and Learning*, 2, 61-81.
3. Campione, J.C., Brown, A.L., & Bryant, N.R. (1985). Individual differences in learning and memory. In R.J. Sternberg (Ed.). In *Human abilities: An information-processing approach*, 103-126. New York: W.H. Freeman.
4. Feng, Mingyu, Heffernan, N.T. (2005). Informing Teachers Live about Student Learning: Reporting in the ASSISTment System. *The 12th Annual Conference on Artificial Intelligence in Education Workshop on Usage Analysis in Learning Systems*, 2005.
5. Feng, M., Heffernan, N.T, Koedinger, K.R. (in press). Addressing the Testing Challenge with a Web-Based E-Assessment System that Tutors as it Assesses. *The 15th International World Wide Web Conference*, Scotland. http://web.cs.wpi.edu/~mfeng/pub/www06.pdf
6. Grigorenko, E.L & Sternberg, R.J. (1998). *Dynamic Testing.* In *Psychological Bulletin*, 124, 75-111.
7. Razzaq, L, Feng, M., Nuzzo-Jones, G., Heffernan, N.T. et al. (2005). The ASSISTment Project: Blending Assessment and Assisting. In *Proceedings of the 12th Annual Conference on Artificial Intelligence in Education.*
8. Razzaq, L., & Heffernan, N. T. (in press). Scaffolding vs. Hints in the ASSISTment System. *The 8th International Conference on Intelligent Tutoring Systems*.http://nth.wpi.edu/pubs_and_grants/ITS2006/Submissons/Leena/razzaq.doc
9. Singer, J. D. & Willett, J. B. (2003). *Applied Longitudinal Data Analysis: Modelling Change and Occurrence*. Oxford University Press, New York.