Comparison of Traditional Assessment with Dynamic Testing in a Tutoring System

MINGYU FENG
SRI International, USA
AND
NEIL T. HEFFERNAN
ZACHARY A. PARDOS
CRISTINA HEFFERNAN
Worcester Polytechnic Institute, USA

It would be great if ITS can also be used to do the benchmark assessment, so that no time from instruction is "stolen" to do extra assessments. It is presumed that given a limited amount of time for assessing, people should give a test but not wasting time giving students feedback. However, Feng and Heffernan (2010) compared two simulated conditions and found that the condition that lets students get feedback during a test was actually superior (not statistically reliable) at predicting student performance than the "test" condition, in which students did about double the number of problems. In this study, we address the weakness in Feng & Heffernan (2010) (i.e. simulated conditions) and run a new randomized control trial in a tutoring system with participants from two different grades, 7th and 8th to see if the main effect would replicate. Our results suggest that unlike our previous results 1) there is no reliable main effect across all students; 2) the dynamic testing condition works better with 7th graders than with 8th graders. We also find that 7th graders and 8th graders behaved differently while working within the system, which appeared to be related to "gaming".

Key Words and Phrases: assessment, tutoring system, dynamic metrics

1. INTRODUCTION

In the past twenty years, much attention from the Intelligent Tutoring System (ITS) community has been paid to improve the quality of student learning while the topic of improving the quality of assessment has not been emphasized as much. The accountability pressure from No Child Left Behind Act of 2001 in the U.S. has led to increased focus on benchmark assessments and practice tests on top of the usual end-of-chapter testing. Such practice assessments can give a crude estimate, but they are also accompanied with the loss of precious, limited instruction time that typically occurs during assessment. It would be great if intelligent tutoring systems could be used to do the benchmark assessment, so that no time from instruction is "stolen" to do extra assessments. However, since students learn from tutoring systems (e.g. Koedinger et al. 1997), many psychometricians would argue that let students learn while being tested will make the assessment harder since you are trying to measure a moving target. Thus, assessing students automatically, continuously and accurately without interfering with student learning is an appealing but also a challenging task.

In Feng, Heffernan & Koedinger (2009), we reported the counter-intuitive results that metrics from an intelligent tutoring system can better predict student's state test scores

Authors' addresses: M. Feng, SRI International, Menlo Park, California, USA. E-mail: mingyu.feng@sri.com; N.T. Heffernan, Department of Computer Science, Worcester Polytechnic Institute, Worcester, Massachusetts, USA. E-mail : nth@wpi.edu; Z.A. Pardos, Department of Computer Science, Worcester Polytechnic Institute, Worcester, Massachusetts, USA. E-mail: zparodos@wpi.edu; C. Heffernan, Department of Computer Science, Worcester Polytechnic Institute, Worcester, Massachusetts, USA. E-mail: ch@wpi.edu;

than traditional test context does. The metrics used include the number of hints that students needed to solve a problem correctly, the time it took them to solve it, the number of attempts that students made before correctly answering a question (called assistance metrics). This finding suggests not only is it possible to get reliable information during "teaching on the test", but also data from the teaching process actually improves reliability. However, there is a caveat that it takes more time for students to complete a test when they are allowed to request for assistance, which seems unfair for the contrast case. Feng & Heffernan (2010) addressed the caveat by controlling for time. We found that students did half the number of problems in a dynamic test setting (where help was administered by the tutor) as opposed to the static condition (where students received no help) and reported better predictions on the state test by the dynamic condition, but not a statistically reliable difference. Trivedi, Pardos, and Heffernan (2011) reanalyzed the same data set by introducing a more sophisticated method to ensemble together multiple models based upon clustering students. Although the findings from Feng & Heffernan (2010) and Trivedi et al. (2011) are encouraging, the predictions were made based upon 40 minutes of historical log data and the traditional test condition was simulated by only including students' first attempt on the main problems and discarding all information while they were being tutored, which may be different from a real computer-based testing condition in several ways because of factors such as time constraints, test anxiety, etc. (Onwuegbuzie & Seaman, 1995). In order to address these concerns, in this paper, we run a randomized controlled study in a middle school in central Massachusetts.

2. LITERATURE REVIEW

Dynamic assessment (DA, Grigorenko & Sternberg, 1998; Sternberg & Grigorenko, 2001, 2002) has been advocated as an interactive approach to conducting assessments to students in the learning systems as it can differentiate student proficiency at the finer grained level. Different from traditional assessment, DA uses the amount and nature of the assistance that students receive as a way to judge the extent of student knowledge limitations. Campione and colleagues (Bryant, Brown & Campione, 1983; Campione & Brown, 1985) took a graduated prompting procedure to compare traditional testing paradigms against a dynamic testing paradigm in which learners are offered increasingly more explicit prewritten hints in response to incorrect responses. 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. Recently, Fuches and colleagues (Fuchs et al., 2008) employed DA in predicting third graders' development of mathematical problem solving.

3. METHODS

3.1 ASSISTments, the test bed

The ASSISTments platform (Razzaq et al., 2005) is an attempt to blend the positive features of both computer-based tutoring and benchmark testing. In ASSISTments, if a student gets an item (the **main** item) right, they will get a new item. If a student has trouble solving a problem, the system provides instructional assistance to lead the student through by breaking the problem into a few **scaffolding** steps (typically 3~5 per problem), or displaying **hint** messages on the screen (usually 2~4 per question), upon student request. As students interact with the system, time-stamped student answers and student actions are logged into the background database.

3.2. Experimental Design

The experiment included two conditions, Test condition and Tutor condition. The Test condition mimicked the traditional computer-based test situation while the Tutor Condition follows the ASSISTments approach as described above. 392 students from 7th or 8th grade classes were randomly assigned to conditions. The experiment was run in one class period, about 45 minutes. Due to class schedules and climate conditions, students from different classes completed the study on different days across two weeks. The materials used for the experiment were selected from released Massachusetts Comprehensive Assessment System (MCAS) test items and the scaffolding questions and hint messages built in by subject matter experts from ASSISTments. The problem set contained 212 randomly organized problems.

4. ANALYSIS AND RESULTS

4.1 Measures and preliminary analysis

MCAS test scores from May, 2010 were used as the outcome measure for prediction[1]. After linking student data collected from ASSISTments with their state test scores, we ended up with a total of 320 students, 160 in each condition. We examined the two experiment conditions on several measures (e.g. student average/standard deviation on MCAS score, the total number of problems finished in the experiments, and the total time they actually spent on solving problems) to ensure the two conditions were balanced. No significant difference was noticed except that students in the Test Condition finished more problems as expected. We reused the online metrics for dynamic testing that measures student accuracy, speed, attempts, and help-seeking behaviors (Feng, Heffernan & Koedinger, 2009; Feng & Heffernan, 2010), including measures on students' percent correct on main problems, which we often referred to as the "static metric", the number of main problems students completed, students' percent correct on scaffolding questions, the average number of hint requests per question, the average number of attempts students made for each question, how long it takes for a student to answer a question, whether main or scaffolding, measured in seconds, how often a student reached the "bottom-out" hints that revealed the correct answer, etc. Among these metrics, the Test condition used only measures on students' percent correct on main problems and the number of main problems students completed during the experiment, while the Tutor condition used all of the metrics. Additionally, during our modeling process, the metrics were normalized so that they were all on the same scale. [2]

4.2. Modeling and results

We followed Feng & Heffernan (2010) to fit a stepwise linear regression model using the dynamic assessment features as independent variables to make a prediction on the MCAS scores. In order to capture non-linear relationships and to fit different models for different groups of students, we also introduced random forests algorithm and clustering technique, following Trivedi et al. (2011). Using random forests (Breiman, 2001) algorithm *N* decision trees will be trained with each tree selecting a portion of the features at random and resamples the original dataset with replacement. Each decision tree is then used to make a prediction of unseen data. The predictions of all the trees are combined by uniform averaging to give the final prediction of the Random Forests algorithm. Using clustering technique, the training data is first used to define *N* clusters and then a specific

[1] For those who are confused, yes, we are "predicting" history data. We don't want to wait till September 2011 to do the analysis. Another reason is that our analysis of MCAS data of students from Worcester, Massachusetts in the past 6 years shows there is a high correlation (0.8~0.9) between students' state test scores across years.

[2] The complete data will be available at http://teacherwiki.assistment.org/wiki/Feng2009#Follow_up_paper

classifier will be trained for the data of each cluster. During the predicting stage, an unseen data was first assigned to one of the three clusters and then the classifier for that cluster is used to make a prediction. We combined approaches mentioned above and fit five different models for each of the two data sets, one for the Test condition and one for the Tutor condition. Namely, the models we applied were a) linear regression, b) stepwise regression, c) random forests with 50 trees, d) clustering method with random forests as classifier, e) clustering method with linear regression as classifier. As mentioned before, for each model, the MCAS state test score was used as dependent variable and the computed online metrics as predictors. 5-fold cross-validation was run to evaluate the results of the analysis for every model.

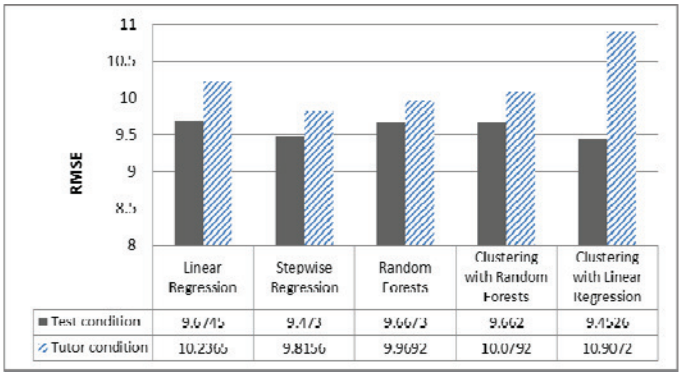

Fig. 1. RMSE from five fitted models

We report for both conditions the RMSEs from the five fitted models in Fig. 3. We observed that clustering method with linear regression worked best for the Test condition and has produced the highest correlation and lowest RMSE. However, it appeared to be the least effective model for the Tutor condition with lowest correlation coefficient and the highest RMSE, which was contradictive to Trivedi et al. (2011). Additionally, we noticed that there was a trend in favor of the Test condition over the Tutor condition, which was also contradictive to what we have found before. We then conducted a series of t-tests between residuals generated from the five models for both conditions, but failed to find any significant difference.

The results from this experiment varied from our previous findings, which made us ponder what made the change. One thing we observed was that during the experiment, the 7th graders in the Tutor condition finished only a half of the problems as completed by those in the Test condition (9 vs. 18). Yet, for the 8th graders, the gap was not as big (12 vs. 16). Our subject matter expert confirmed that the content was appropriate for both grades. Yet, we found that 8th graders overall requested for significantly more hints than 7th graders did (effect size = 0.4, $p < 0.001$) after examining students' response data to all assignments before the experiment, which appeared to be related to students "gaming" the system (Baker et al., 2004). We speculated that gaming behaviors, such as requesting for hints for every question, always reaching out to the bottom-out hint, would be a big detriment to the dynamic assessment approach as the approach depends heavily on the interaction between students and the system, esp. their help-seeking behavior and response speed. We chose to model 7th and 8th graders separately and repeated the modeling process as described in section 4.3 for 7^{th} graders and 8^{th} graders respectively. We found out that in 7th grade the trend was in favor of the Tutor condition while in 8th grade was in favor of the Test condition, which was aligned with our speculation, but again neither difference was significant.

5. CONCLUSION

The results from Feng, Heffernan & Koedinger (2009) that started this line of work off were so exciting that they were cited in the National Educational Technology Plan (U.S. Department of Education, 2010) and were followed in Feng & Heffernan (2010) in a simulated study. However, we got a null result in a real randomized trial. Although reasoning from a null result is difficulty, we think it is important to share with the community this null result as it does bring a caution to the excitement of the prior work. The three most salient differences between this study and our prior work are, firstly, in our prior work, we could run paired t-test using simulated data (see Feng & Heffernan, 2010 for details). Yet this experiment was not a between-subject design that can be more powerful if the variation between students is very large compared to the variation caused by the conditions. Secondly, in this experiment we had many fewer students (320 vs. 1392 in the prior study), which again reduced the statistical power of analysis. Thirdly, in the prior study, since existing log data were reused to simulate conditions, we made sure that the data sets contained exactly 40 minutes of work of every student. However, in this study, there was no guarantee that students all committed to working on ASSISTments problems for 40 minutes. As a matter of fact, on average, students have done only 30 minutes of problem solving work, which provided fewer amounts of data for our analysis. However, more complex models, such as the clustering method often require larger amounts of data to build a competent model. With all factors considered, we conclude that the experiment should be repeated with more students but also have the students swap conditions so that we can make a within-subject comparison.

REFERENCES

Baker, R. S., Corbett, A.T., Koedinger, K. R., & Wagner, A. Z. (2004). Off-task behavior in the Cognitive Tutor Classroom: When Students “game the system”. *Proceedings of the ACM CHI 2004: Computer - Human Interaction*. (pp. 383-390). New York: ACM.

Breiman, L. (2001) Random forests. Machine Learning, 45(1):5-32, 2001.

Campione. J. C. & Brown. A. L. (1985). Dynamic Assessment: One Approach and some Initial Data. Technical Report. No. 361. Cambridge, MA. Illinois University, Urbana, Center for the Study of Reading. ED 269735.

Campione. J. C., Brown. A. L., & Bryant. N. R. (1985). Individual Differences in Learning and Memory. In R.J. Steinberg (Ed.). *Human Abilities: An Information Processing Approach,* 103-126,. New York, W. H. Freeman.

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. 19(3). 2009.

Feng. M., Heffernan. N. T., (2010). Can We Get Better Assessment From A Tutoring System Compared to Traditional Paper Testing? Can We Have Our Cake (better assessment) and Eat it too (student learning during the test)? In *Proceedings of the 3rd International Conference on Educational Data Mining*, 41-50.

Grigerenko, E. L. & Steinberg, R. J. (1998). Dynamic Testing. *Psychological Bulletin, 124, 75-111*

Fuchs. L. S., Compton. D. L. Fuchs. D., Hollenbeck. K. N., Craddock. C. F., & Hamlett. C. L. (2008). Dynamic Assessment of Algebraic Learning in Predicting Third Graders’ of Mathematical Problem Solving. *Journal of Educational Psychology*, 100(4), 829-850.

U.S. Department of Education (2011). National Educational Technology Plan 2010. http://www.ed.gov/technology/netp-2010

Onwuegbuzie, A. J., & Seaman, M. A. (1995). The effect of time constraints and statistics test anxiety on test performance in a statistics course. Journal of Experimental Education, 63, 115-124.

Razzaq. L., Feng M., Nuzzo-Jones. G., Heffernan. N. T., Koedinger. K. R., Junker. B., Ritter. S., Knight A., Aniszczyk. C., Choksey. S., Livak. T., Mercado. E., Turner. T. E., Upalekar R., Walonoski. J.A., Macasek. M. A., & Rasmussen. K. P. (2005). The Assistment Project: Blending Assessment and Assisting. In C. K. Looi, G. McCalla, B. Bredeweg, & J. Breuker (Eds). *Proceedings of the 12th International Conference on Artificial Intelligence in Education, Amesterdam.* ISO Press, pp 555-562.

Trivedi, S., Pardos, Z.A., Heffernan, N.T. (2011). Clustering Students to Generate an Ensemble to Improve Standard Test Score Prediction. Accepted by the 15th International Conference on Artificial Intelligence in Education. Christchurch, New Zealand.