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)?

Mingyu Feng[1] and Neil Heffernan[2]

[1] SRI International, Menlo Park, CA 94025
[2] Worcester Polytechnic Institute, Worcester, MA 01609
mingyu.feng@sri.com, nth@wpi.edu

Abstract. Dynamic assessment (DA) has been advocated as an interactive approach to conducting assessments to students in the learning systems. Sternberg and others proposed to give students tests to see how much assistance it takes a student to learn a topic; and to use as a measure of their learning gain. To researchers in the ITS community, it comes as no surprise that measuring how much assistance a student needs to complete a task successfully is probably a good indicator of this lack of knowledge. However, a cautionary note is that conducting DA takes more time than simply administering regular test items to students. In this paper, we report a study analyzing 40-minutes data of totally 1,392 students from two school years. The result suggests that for the purpose of assessing student performance, it is more efficient for students to take DA than just having practice items.

Keywords: Dynamic assessment, assessment in learning system.

1 Introduction

In the past twenty years, much attention from the Intelligent Tutoring System community has been paid to improve student learning while the quality of assessment has not been emphasized as much. In the US, state tests are causing many schools to give extra tests. It would be great if ITSs could be used to do the tests, so that no time from instruction is taken away. 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. Can ITSs, if given the same amount of time, be better assessors of students (while also providing the benefit of helping students learn during that time period)?

Assessing students accurately without interfering with learning is an appealing but also a challenging task. Dynamic assessment (DA, also called dynamic testing) [3] has been advocated as an interactive approach to conducting assessments to students. DA uses the amount and nature of the assistance that students receive to judge the extent of student knowledge limitations (e.g. [3], [4], [5]) or measures student learning potential (e.g. [2]). ITSs are perfect test beds for DA as they naturally lead students into a tutoring process to help students with the difficulties. We [1] have

V. Aleven, J. Kay, and J. Mostow (Eds.): ITS 2010, Part II, LNCS 6095, pp. 309–311, 2010.

collected extensive information to assess students dynamically in a computer)-based tutoring system (http://ASSISTments.org). In this system if a student has trouble solving a problem (the **main** item, the system provides instructional assistance by breaking the problem into a few **scaffolding** steps, or displaying **hint** messages on the screen upon request. Although DA has been shown to be effective predicting student performance, yet there is a cautionary note: since students are allowed to request assistance, it generally takes longer to finish a test using the DA approach than using a traditional testing approach.

2 Methods

Fundamentally, in order to find out whether DA was worth the time, we would want to run a study comparing the assessment value of the following two conditions: Static assessment condition (A): students were presented with one static test item and were requested to submit an answer; Dynamic assessment condition (B): students were presented with one static test item followed by a DA portion where they could request help. Since ASSISTments had collected data with the information needed, we chose to compare predictions made based on log data from 40 minutes of time across *simulated* conditions that were similar but not exactly the same as above: **Simulated static assessment condition (A'):** 40 minutes of student work selected from existing log data on only main items; **Dynamic assessment condition (B'):** 40 minutes of work selected from existing logged response data on both main items and the scaffolding steps and hints. Such a simulation study not only saved time, but also allowed us to compare the same student's work in different conditions, which naturally rules out the subject effect. We chose to use student's end of year state accountability test score as the measure of student achievement.

We considered two data sets, one from the 2004-2005 school year with 628 students, and the other from 2005-2006 school year of 764 students. The online metrics for dynamic testing that measures student accuracy, speed, attempts, and help-seeking behaviors are simply:

- Main_Percent_Correct – students' percent correct on main questions
- Main_Count - the number of main items students completed.
- Scaffold_Percent_Correct - students' percent correct on scaffolding questions.
- Avg_Hint_Request - the average number of hint requests per question.
- Avg_Attempt - the average number of attempts students made for each question.
- Avg_Question_Time - on average, how long it takes for a student to answer a question, whether original or scaffolding, measured in seconds.

The last five metrics are DA style metrics and were not measured in traditional tests. We ran stepwise linear regression to use the metrics described above to predict student state test scores (the dependent variable). For condition A', the independent variable of the simple linear regression model was *Main_Percent_Correct;* while for condition B', it changed to be the collection of the DA style metrics.

Looking at the results, we noticed that in both years, students finished more test items in the 40 minutes in static condition than in dynamic condition (22 vs 11 in the first year; 31 vs. 13 in the second year). We examined the parameters in the linear regression models esp. for the dynamic condition. The first three parameters entered the models were the same in both years (with the order changed a little bit).

Scaffold_Percent_Correct was the most significant predictor in the first year while in the second year, it was Main_Percent_Correct. Also, in the later year 2005-2006, Avg_Attempt was considered as a significant predictor while in the first year it was Avg_Hint instead. We also chose to use Bayesian Information Criterion (BIC) to compare the generalization quality of the model. In both years, the R squares of the model from the dynamic condition were higher, and BICs were significantly lower than those of the static condition, suggesting DA condition did a statistically significantly better job at predicting state test scores than the static condition did. We also conducted 5-fold cross validation on the 2004-2005 data and noticed variables in the trained regression models of the DA condition were consistent across the 5 folds validation. So we took the average of coefficients from the five trained regression models and applied the average model on the full data set. Mean absolute difference (MAD) was calculated as a measure of prediction accuracy. The average model from the simulated static condition and the DA condition produced MAD of 9.01 (out of 54) and 8.7 respectively. The paired t-test suggested that there was a marginally significant difference (p=0.10).

Based on the results, we conclude that dynamic assessment is more efficient than just giving practice test items. DA can produce more accurate assessment of student math performance, even limited by using the same amount of testing time. This is surprising as students in the dynamic assessment do few problems.

3 Conclusion

In this paper, we compared DA against a tough contrast case where students were doing assessment all the time in order to evaluate efficiency and accuracy of DA in a tutoring system. This paper eliminates the cautionary note about dynamic assessment that says DA will always need a longer time to do as well at assessing students, which further validates the usage of tutoring systems for assessment.

Acknowledgements

We would like to acknowledge funding from the US Department of Education, the National Science Foundation, the Office of Naval Research and the Spencer Foundation. All of the opinions expressed in this paper are those solely of the authors and not those of our funding organizations.

References

1. Feng, M., Heffernan, N.T., Koedinger, K.R.: 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), 243–266 (2009)
2. Fuchs, L.S., Compton, D.L., Fuchs, D., Hollenbeck, K.N., Craddock, C.F., Hamlett, C.L.: Dynamic assessment of algebraic learning in predicting third graders' development of mathematical problem solving. Journal of Educational Psychology 100(4), 829–850 (2008)
3. Grigorenko, E.L., Sternberg, R.J.: Dynamic testing. Psychological Bulletin 124, 75–111 (1998)
4. Sternburg, R.J., Grigorenko, E.L.: All testing is dynamic testing. Issues in Education 7, 137–170 (2001)
5. Sternburg, R.J., Grigorenko, E.L.: Dynamic testing: The nature and measurement of learning potential. Cambridge University Press, Cambridge (2002)