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 is presumed that if you have a limited amount of time for assessing student knowledge, that you should give a test that does not waste time giving students feedback. Feng and Heffernan (2010) compared two such conditions and found the counter-intuitive results that the condition that let them get feedback was actually superior (not statistically reliable) than the “test” condition, in which student did about double the number of problems. The gain in assessment value for the tutor condition came in the form of the addition measure like number of hints a student needed, and how many attempts they had to make before they got the problem correct. Trivedi, Pardos, and Heffernan (2011) analyzed the same data set and found a more powerful way of using the information available in the dynamic testing condition (i.e., the tutor condition) by applying clustering techniques. They found a reliable improvement over the work in Feng and Heffernan (2010) that made the advantage of tutoring appear a decisively better choice if you want to assess student (and obviously a better choice if you care about student learning from feedback). The weakness in Feng & Heffernan (2010) was in its being a simulated condition. In this study, we explain that weakness, and run a new randomized control trial to see if the main effects would replicate. In this paper, we attempted to combine the two studies and eliminate the “weakness by running a randomized controlled experiment in a tutoring system with participants from two different grades, 7th and 8th. 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 found that 7th graders and 8th graders behaved differently while working within the system. 8th graders asking for significantly more hints than 7th graders when working on the same set of assignments, which appeared to be related to “gaming”.

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

1. INTRODUCTION

In Feng, Heffernan & Koedinger (2009), we reported the counter-intuitive results that data from an intelligent tutoring system can better predict state test scores if it considered the extra measures collected while providing the students with feedback other than just looking at student’s correct rate on the problems, as did in traditional test context. These included metrics such as number of hints that students needed to solve a problem correctly and the time it took them to solve it (called assistance metrics). This highlights the power of the extra “assistance” measures, which 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 than 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 the amount of 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 that was used by Feng & Heffernan (2010) and introduced a more sophisticated method by which we can ensemble together multiple models based upon clustering students. We showed that the assessment quality as determined by the assistance metrics is a better estimator of student knowledge than traditional tests. Although the findings from Feng & Heffernan (2010) and Trivedi et al. (2011) are encouraging, it is worth mentioning that the predictions were made based upon 40 minutes of historical log data from a tutoring system (“ASSISTments”) 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. The simulated situation may be different from a real computer-based testing condition in several ways. Among other things, students may behave differently when assistance is available and students may perform differently when there is a time constraint as moderated by student test anxiety (Hill, 1984; Onwuegbuzie & Seaman, 1995). Additionally, potentially students would have spent more than 40 minutes working in the system considering the tutoring portion following main problems. In order to address these concerns, in this paper, we report our findings from a randomized controlled study that was run early 2011 in a middle school in Worcester in central Massachusetts.

2. BACKGROUND AND LITERATURE REVIEW
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. However, student assessment is very important. In the US, the No Child Left Behind Act of 2001 has exerted accountability pressures on school administrators, teachers and students. The accountability pressure has led to increased focus on benchmark assessments and practice tests on top of the usual end-of-chapter testing. The hope is that such assessment will help determine what instruction or remediation is needed to raise student achievement and consequently raise their test scores on the high-stakes year-end exam. 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, since giving such assessment is not meant to help students learn, but is mainly focused on being able to tell teachers and principals about who needs help on what. As a matter of fact, although the aim of benchmark testing is to improve classroom instruction and ultimately improve student achievement, Henderson, Petrosino, Guckenburg, & Hamilton (2007a, 2007b) found no significant difference on middle school math scores between schools who used quarterly benchmark assessments and comparison schools.

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.
While challenging, accurately assessing student’s performance level is a critical task in an intelligent tutoring system since that the system needs this knowledge estimate in order to calibrate the amount of practice that students require for skill mastery and, thus, adapts the educational interaction effectively to the specific needs of the individual student.

Dynamic assessment (da, or sometimes called dynamic testing, Grigorenko & Sternberg, 1998) 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 which is normally not available in traditional practice test situations as a way to judge the extent of student knowledge limitations. Even before the computer supported systems become popular, much work has been done on developing “testing metrics” for dynamic testing (Grigorenko & Sternberg, 1998; Sternberg & Grigorenko, 2001, 2002) to supplement accuracy data (wrong/right scores) from a single setting. 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. Grigorenko and Sternberg (1998) reviewed relevant literature on this topic and expressed enthusiasm for the idea. 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 learn than traditional static tests. 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 the dynamic testing paradigm, learners are offered increasingly more explicit prewritten hints in response to incorrect responses. In this 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. They also suggested that this dynamic method could be effectively done by computer, but never pushed toward to conduct such studies using a computer system. The most unique information from DA information about the learner’s responsiveness to intervention (Fuches et al. 2007) in the tutoring system. There have been a few studies that pay attention to such unique information. For instance, recently Fuches and colleagues (Fuches et al., 2008) employed da in predicting third graders' development of mathematical problem solving.

3. METHODS

3.1 ASSISTments, the test bed

Traditionally, the areas of testing (i.e., psychometrics) and instruction (i.e., math educational research and instructional technology research) were separated fields of research with their own goals. The ASSISTments Platform (Razzaq et al., 2005) is an attempt to blend the positive features of both computer-based tutoring and benchmark testing. The online system presents math problems to students of approximately 13 to 16 years old in middle school or high school to solve. 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 shown in Fig.1. Although the system is web-based hence accessible in principle anywhere/anytime, students typically interact with the system during one class period in the schools’
computer labs every three or four weeks. As students interact with the system, time-stamped student answers and student actions are logged into the background database. The hypothesis is that ASSISTments can do a better job of assessing student knowledge limitations than practice tests or other online testing approaches by using the DA approach based on the data collected online.

3.2. Experimental Design

The experiment included two conditions. Condition A (called Test condition) and Condition B (Called Tutor condition). Condition A, the Test condition, mimicked the traditional computer-based test situation where students were presented with problems one after another without any feedback given. Condition B follows the normal ASSISTments approach as shown in Fig. 1. In this condition, students were also administered problems one after another but the system provides tutoring on problems that students have difficulty with by introducing scaffolding steps and/or providing hint messages upon student’s request. For condition B, the Tutor condition, students would have to reach a correct solution for current problem before moving on to the next one. Students’ response to every problem and associated scaffolding questions, if available, were collected by the ASSISTments system. Their help requests and other interactions with the system were also recorded with timestamps.

3.3. Setting and Participants:

The experiment was run early 2011 in regular math classrooms in a middle school in central Massachusetts. 392 students from 7th or 8th grade class (ages 12-13 years) participated in the experiment and were randomly assigned to either of the two conditions with 195 students in the Test condition and 197 in the Tutor condition. The experiment was run in one class period, about 45 minutes. These students have been using ASSISTments system regularly since September, 2010. Due to class schedules and climate conditions, students from different classes completed the study on different days across two weeks.

3.4 Materials
The materials used for the experiment were organized in two “big” problem sets, one for the Test Condition and one for the Tutor Condition. The problems 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 project. The problems in the two problem sets were the same except that the immediate feedback and tutoring function were turned off for the problem set assigned to students in the Test condition. Each set contained 212 ASSISTments (i.e. main problems). The problem set for the Tutor condition also contains a total of 701 scaffolding questions, associated with the 212 main problems. Problems in each set were fully randomized while being administered to students. All of the five strands in MCAS curriculum framework were covered. Fig. II shows the distribution of the coverage by problems. Comparing to MCAS test, the problem sets addressed heavier on Number Sense and Operations yet lighter in Geometry and Patterns, Relations and Algebra. Yet, since the problems were randomized with a set and students were randomly assigned to conditions, we don’t expect this difference on strand coverage will have an impact on our results.

4. ANALYSIS AND RESULTS

4.1 Measures and preliminary analysis

We picked MCAS test score from May, 2010 as the outcome measure for prediction. 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. Data for the rest 72 students were eliminated. Table I compared students in the two experiment conditions on several measures to ensure the two conditions were balanced. Specifically, we examined in each condition and in the two grades, the number of students in each condition, student average MCAS score, standard deviation on MCAS scores, the total number of problems students have finished in the 45 minutes of experiments and the total time they actually spent on solving problems. We didn’t notice any significant difference between the two conditions except that students in the Test Condition finished more problems, which was

---

1 For those who are confused, yes, we are “predicting” history data. One reason is that 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.
expected since they didn’t have to deal with scaffolding questions and reach the correct answer.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Grade</th>
<th># of students</th>
<th>avg. MCAS</th>
<th>std. MCAS</th>
<th># of problems done</th>
<th>avg. total minutes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test</td>
<td>7th grade</td>
<td>79</td>
<td>38</td>
<td>12</td>
<td>18</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>8th grade</td>
<td>81</td>
<td>37</td>
<td>11</td>
<td>16</td>
<td>31</td>
</tr>
<tr>
<td>Tutor</td>
<td>7th grade</td>
<td>76</td>
<td>39</td>
<td>11</td>
<td>9</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>8th grade</td>
<td>84</td>
<td>35</td>
<td>12</td>
<td>12</td>
<td>31</td>
</tr>
</tbody>
</table>

4.2 Metrics

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. As always, we admit that these metrics might not be the best metrics to measure how students have interacted with the system as they are not fine-grained, individualized, nor time sensitive. Yet, we also believe that adding these metrics into prediction models grasped the essence that student-system interaction information matters for more accurate prediction of student proficiency, although they have been largely ignored in traditional benchmark tests and even modern computerized tests, unfortunately. Results from our previous studies have showed the value of these metrics.

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.

For clarity, our data file for those in the Test condition held one row per student, his/her state test score, the number of questions the student completed, and finally the student’s percent correct across those questions. The data file for the Tutor condition held all of those metrics from the Test condition plus the “online” metrics mentioned above (two that seems most salient are the average number of hints the student required per problem, the average number of attempts made per problem). The complete data file will be able available in archive.

4.3. Modeling

We applied various models including regression, random forest, and clustering to compute for the Test and Tutor conditions predictions of student state test scores based upon metrics collected in ASSISTments.

Feng & Heffernan (2010) fit a stepwise linear regression model using the dynamic assessment features as independent variables to make a prediction on the MCAS scores.

---

2 The complete data will be made available in archive form just as we have done for Feng & Heffernan (2010). See here to get the data http://teacherwiki.assistment.org/wiki/Feng2009#Follow_up_paper
Yet, we found only a marginal statistical significance on the accuracy of prediction when comparing the Tutor condition against the static condition. Although a stepwise regression is helpful to find a parsimonious model, it works best when there is a linear relationship among the independent variables and dependent variables, which might not be true for our data (i.e. online metrics and the state test scores, respectively). Additionally, our subject population has students of various proficiency levels and they may interact with the system differently, for instance, in terms of help-seeking behaviors. Thus, it is thus worth attempting to fit models that can capture non-linear relationships and to fit different models for different groups of students.

Random forests (Breiman, 2001) algorithm was also applied to this data set. It is a method for training and combining predictions made by decision trees. The algorithm trains a specified number of trees, commonly chosen to be 50 or higher. Each tree selects a portion of the features at random and resamples the original dataset with replacement. Each decision tree is then trained on its own data and can be used to make a prediction of unseen data. The predictions of all the trees are combined by uniform averaging and the result is the final prediction of the Random Forests algorithm.

Trivedi, Pardos & Heffernan (2011) introduced a new method for optimizing prediction accuracy using clustering. It was shown to be powerful when applied to the simulated data set as used by Feng & Heffernan (2010). This method uses the training data to define N clusters where N is varied from 1 to some cut-off. Various cut-offs were explored more in Trivedi et al. but for this paper a cut-off of 3 was chosen since it was the maximum number of clusters that were found before an empty cluster appeared. When the N clusters are formed, a model is trained for the data of each cluster. The model can be fit using any classifier but for this paper we used linear regression and random forests. When unseen data is predicted, the data is first assigned to one of the N clusters and then the model for that cluster is used to predict the data. As N is increased, additional predictions for the unseen datum are created. These additional predictions are averaged together to provide the final prediction.

In this paper, 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/multiple 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 produce the results of the analysis for every model. Each dataset was randomly split into 5 bins of similar size. There were five phases to the cross-validation. In each phase, one fold served as the test set and the other four folds served as the training set. At the end of the validation a list of cross-validated predictions were produced as a result of each fold serving as the test set. The same 5 bin split of data was used to analyze each prediction method.

4.4 Results

Below in Table II, we report for both conditions the Pearson correlation coefficient between fitted values and the actual state test scores. RMSEs from the five fitted models are represented in Fig. 3. Surprisingly, 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 what Trivedi et al. (2011) have found. Additionally, when comparing the
results for two conditions, we noticed that there was a trend in favor of the Test condition over the Tutor condition, which was also contradictory 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.

Table II. Correlations between fitted values and actual MCAS scores

<table>
<thead>
<tr>
<th>Correlation</th>
<th>Test condition</th>
<th>Tutor condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear Regression</td>
<td>0.5671</td>
<td>0.4960</td>
</tr>
<tr>
<td>Stepwise Regression</td>
<td>0.5890</td>
<td>0.5403</td>
</tr>
<tr>
<td>Random Forests</td>
<td>0.5660</td>
<td>0.5232</td>
</tr>
<tr>
<td>Clustering with Random Forests</td>
<td>0.5677</td>
<td>0.5674</td>
</tr>
<tr>
<td>Clustering with Linear Regression</td>
<td>0.5927</td>
<td>0.4524</td>
</tr>
</tbody>
</table>

Fig. 3. RMSE from five fitted models

4.5. Post experiment analysis

The fact that the results from this experiment varied from our previous findings 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 (i.e. 9 vs. 18). Yet, for the 8th graders, the gap was not as big (i.e. 12 vs. 16). Our subject matter expert confirmed that the content was appropriate for both grades. Then, what’s so different between the 7th graders and 8th graders? In order to answer this question, we explored various teacher reports in ASSISTments that were generated based upon students’ responses to all the assignments that have been made to the students before the experiment. Among other things, we found that 8th graders overall requested for significantly more hints than 7th graders did (effect size = 0.4, p < 0.001), 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. This finding encouraged us to treat “gamers” differently from regular users during the modeling process. Yet, we didn’t know of other ways to effectively and efficiently filter out the gamers, so we just chose to model 7th and 8th graders separately in the follow-up analysis.

In the post-experiment analysis, we repeated the modeling process as described in section 4.3 for 7th graders and 8th graders respectively. We found 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).

The results from Feng & Heffernan (2010) were widely exciting, as it appeared we could have our cake (better learning cause by students getting feedback) and eat it too (better assessment though the use of our interesting prediction methodology). To be clear, it was almost too good to be true. Maybe it was. We are not sure. We failed to replicate the effect in a real randomized trial. This might have to do with the fact that our first study was a fluke. But we doubt that. There are several key differences between this study and the previous one. First of all, in the paper we have a null result, not a negative finding. This point seems very important. Normally null results are not even reported as reason from them is difficult. If we had replicated the main effect reliably, we would have had quite a lot to say. But 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. In order to deem a null result publishable, we need to look to see if the experiment was well powered, or at least as well done as the prior work.

The three most salient differences between this study and our prior work (i.e., Feng & Heffernan, 2010; Trivdi, Pardos & Heffernan, 2011) seems that, firstly, in our prior work, we could run paired t-test as we had a simulated 40 minute conditions for the Test condition (see Feng & Heffernan, 2010 for details). Yet this experiment was not a between-subject design. Between-subject design 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 our 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, even though this experiment was conducted in one class period, there was no guarantee that students all committed to working on ASSISTments problems for 40 minutes. As a matter of fact, we have reported that 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 introduced Trivedi et al. often require larger amounts of data to build a competent model.

The post experiment analysis was promising although not significant. This result might have to do with the simple partition of students by grades. For future analysis, we would look for more sophisticated algorithms to separate out gamers.

With all factors considered, we suggest 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


