Can an Intelligent Tutoring System Predict Math Proficiency as Well as a Standardized Test?

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

Abstract. It has been reported in previous work that students' online tutoring data collected from intelligent tutoring systems can be used to build models to predict actual state test scores. In this paper, we replicated a previous study to model students' math proficiency by taking into consideration students' response data during the tutoring session and their help-seeking behavior. To extend our previous work, we propose a new method of using students test scores from multiple years (referred to as cross-year data) for determining whether a student model is as good as the standardized test to which it is compared at estimating student math proficiency. We show that our model can do as well as a standardized test. We show that what we assess has prediction ability two years later. We stress that the contribution of the paper is the methodology of using student cross-year state test score to evaluate a student model against a standardized test.

1 Introduction

In many states there are concerns about poor student performance on new high-stakes standards-based tests that are required by No Child Left Behind (NCLB). For instance, the Massachusetts Comprehensive Assessment System (MCAS) administers rigorous standardized tests in English, math, history and science in grades 3–12. Students need to pass the math and English portions of the 10th grade versions in order to get a high school diploma without further remediation. 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 has 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 [12]. Some teachers make extensive use of practice tests and released test items to help 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[2]) 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.

[1] http://www.doe.mass.edu/mcas/2002/results/summary.pdf
[2] www.renlearn.com

The ASSISTment system (http://www.assistment.org) is an attempt to blend the positive features of both computer-based tutoring and benchmark testing. It helps students to work through tough problem by breaking the problem into 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 (e.g. [10]), for individual students and for groups, have been developed to help teachers and other stakeholders to better understand students' performance and progress. Furthermore, we have done research (e.g. [1], [2], [8], [9]) to train models that use the ASSISTment metrics to accurately predict state test scores.

In this paper, we replicated the study in [8] to model students' math proficiency for Spring 2005 by taking into consideration students' response data during the tutoring session and their help-seeking behavior. Among these studies on modeling student proficiency, different criteria have been used to evaluate the effectiveness of the student models. For instance, Beck & Sison [3] used relative closeness to real scores and mean absolute error (MAE); Anozie & Junker ([1]) evaluated their model using mean absolute deviation (MAD) which is essentially the same as MAE. Feng, Heffernan & Koedinger ([8]) reported R square and Bayesian Information Criterion (BIC) of the models we fitted. In this paper, we propose a new method of using students' cross-year test scores (from the year 2007) to evaluate the effectiveness of the student model on predicting student math proficiency. We will compare our prediction with the 2005 MCAS test on how well each correlated with the 2007 MCAS scores, assuming student math proficiency at 8^{th} grade and at 10^{th} grade are highly correlated with each other. The paper will be organized as follows. We will briefly describe the ASSISTment system in section 2 and then in section 3 present our approach, including the data we used for this study, the modeling process and the results. The analysis of the results will be discussed in the fourth section. Finally, we conclude in section 5 and bring up some future work.

2 ASSISTment - The Intelligent Tutoring System

Traditionally, the areas of testing (i.e. Psychometrics) and instruction (i.e., math educational research and instructional technology research) have been separate fields of research with their own goals. To address this issue, we built a web based e-learning system that integrates assistance and assessment. The system, named ASSISTment (e.g. [14], [15]), was built to offer instruction to students while providing a more detailed evaluation of their abilities to the teacher than is possible under current approaches. The key feature of ASSISTments is that they provide instructional support (i.e. assistance) in the process of assessing students. In Massachusetts, the state department of education has released ten years (1998-2007) worth of MCAS test items, over 300 items, which we have turned into ASSISTments by adding "tutoring." If students get an item (called the *original question*) correct they will be given a new item. If they get it wrong, they are provided with a small "tutoring" session where they are forced to answer a few (3 ~ 5) questions (called *scaffolding questions*) that break the problem down into steps to eventually get the problem correct. Students are allowed to ask for hints when they encounter difficulty answering a question. 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 three or four weeks. As

students interact with the system, the background logging system collects detailed data of student responses. 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 a "dynamic assessment" approach based on the data collected online. In particular, ASSISTments use the amount and nature of the assistance that students receive as a way to judge the extent of student knowledge limitations.

3 Approach

3.1 Data Source

The 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. The data set contains online interaction data of 417 8th grade students 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. Since the purpose of this study is comparing the prediction power of ASSISTment with MCAS and there are 39 questions in the MCAS tests, we excluded the data of the 25 students who have done fewer than 39 questions in ASSISTments. The 392 students in the final data set on average have worked in the system for around 7 days (one session per day) with 40 minutes of practice per session. They finished 147 items on average (standard deviation = 60 items) and at maximum 319 items.

In this paper, student models will be built based on the online data to predict student math proficiency at the end of the school year 2004-2005 as measured by the 2005 MCAS test scores. And we will evaluate our prediction against the 2007 MCAS test to answer the research question: Can we model student proficiency as well as the standardized test?

3.2 Modeling

Our effort on predicting student math proficiency starts from **a "lean" model**, an Item Response Theory (IRT)-style ([7], [16]) model within the ASSISTment dataset only. As a start place, we used one-parameter IRT model (the Rasch model[3]), the straightforward model with just student proficiency and item difficulty parameters (and only on original questions). 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, we hope to acquire a more reliable estimate of the parameters. We fitted the model in BILOG-MG 3.0[4] and added an extra column in the data set that we described in Section 3.1 to record the student proficiency estimates for the 392 students.

---

[3] In the Rasch model, the probability of a specified response is modeled as a logistic function of the difference between the person and item parameter. In educational tests, item parameters pertain to the difficulty of items while person parameters pertain to the ability or attainment level of people who are assessed.

[4] http://www.scienceplus.nl/scienceplus/main/softwareshop/bilogmg.jsp

The lean model accounts for how students performed on the original questions but totally ignores how students interacted with the system. Much work has been done in the past 10 years or so on developing "online testing metrics" for dynamic testing (or dynamic assessment) [11] to supplement accuracy data (wrong/right scores) for characterizing student proficiency. 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. Bryant, Brown and Campione ([4]) compared traditional testing paradigms against a dynamic testing paradigm. Grigorenko and Sternberg ([11]) 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. In Bryant, Brown and Campione's study they wanted to predict learning gains between pretest and posttest. They found that student's predicted learning gains were not correlated ($R = 0.45$) with static ability score as well as their "dynamic testing" ($R = 0.60$) score. It was suggested that this method could be effectively done by computer. Feng, Heffernan & Koedinger ([8]) continued the dynamic testing approach and developed a group of metrics that measures students' accuracy, speed, attempts and help-seeking behavior. We reused these metrics in this study. Namely, the metrics are:

- ORIGINAL_COUNT - the number of original items students have done. This measures students' attendance and on-task-ness. This measure 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 original items, students' performance on scaffolding questions is also a reasonable reflection of their knowledge. For instance, students who failed on original items simply because of their lack of ability of forming problem-solving strategies will probably answer all 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 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.
- HINT_COUNT - the total number of hints students got.
- AVG_HINT_COUNT - the number of hint messages students got averaged over all questions

- BOTTOM-OUT_HINT_COUNT - the total number of bottom-out[5] hint messages students got.
- 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 students made for each question.
- AVG_ITEM_TIME - on average, how long does it take for students to finish a problem (including all scaffolding questions if students answered the original questions incorrectly).
- AVG_QUESTION_TIME - on average, how long does it take for a student to answer a question, 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 all these measures would be negatively correlated with MCAS scores.

Now that we have the online metrics, we ran a backwards regression analysis ($F > .10$) to predict 2005 MCAS test scores. In addition to the metrics themselves, quadratic terms and interaction between the metrics are all introduced into the model initially to address the fact that there may exist non-linear relationships between the test score and the online metrics. We will refer to this model as **the online model.**

We did not include student percent correct on original questions here as it is a static metric that mimics paper practice tests by scoring students as either correct or incorrect on each item. The student proficiency parameter estimated by the lean model will work for the purpose of scoring students plus the advantage that it takes into consideration the fact that item difficulty varies. Therefore, in the next step, we combine the proficiency parameter with the online metrics together and fit another model in SPSS by doing backwards regression ($F > .10$). We name this model **the mixed model**.

Thus, we have constructed three models: the lean model, the online model, and the mixed model and each model addresses different aspects of student performance. In the next section, we will evaluate the models.

[5] 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.

3.3 Evaluating results using data from multiple years

Many works (e.g. [1, 2, 8, 9]) have been done to build student models to estimate end-of-year MCAS test scores and predicted scores were compared with actual 2005 MCAS test scores to calculate MAD which was used as the measure to evaluate the student models. In Table 1, we summarize the three regression models and present MAD of each model.

Table 1: Model summary

Model	MAD	R Square	BIC*	#variables entered final model	Correlation with 8th grade MCAS
The lean model	/				.731
The online model	4.76	.786	-437	28	.865
The mixed model	4.61	.794	-470	25	.871

*BIC: Bayesian Information Criteria, using the formula for linear regression models introduced in [13].

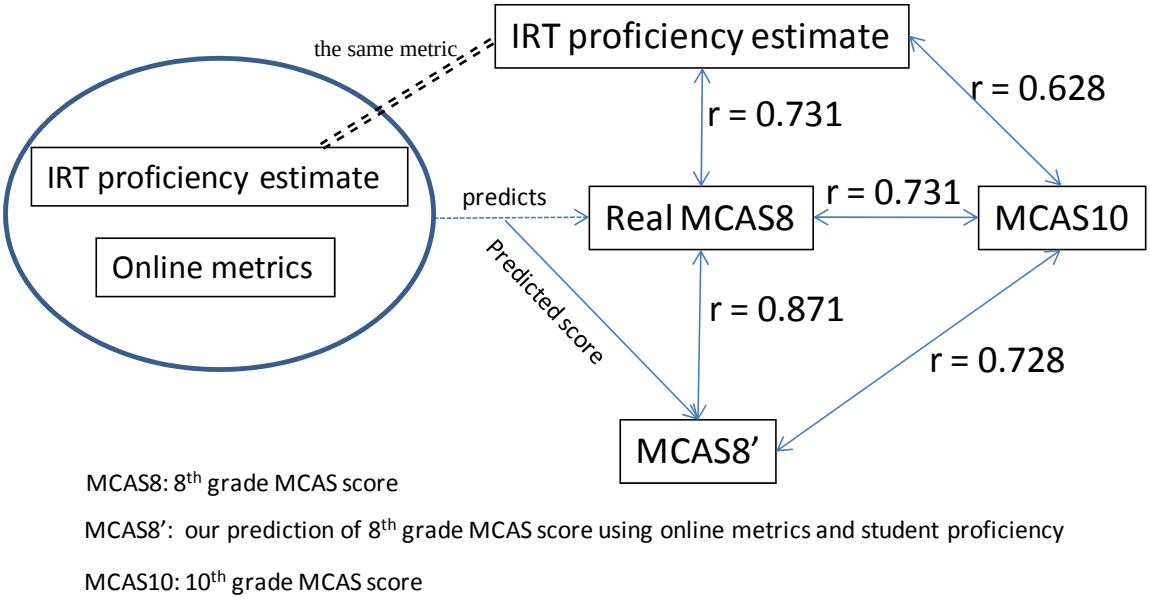
First of all, all of the correlation coefficients in the rightmost column in Table 1 are reliably different from zero, which indicate the data collected within the ASSISTment system can accurately predict student math proficiency at the end of the year. The lean model does not correlate with MCAS score as well as the online model, which indicated that student math proficiency can be better estimated by adding in features reflecting student assistance, effort, attendance, etc. The result is consistent with [8]. Additionally we can improve our prediction of MCAS score further by combining the student proficiency parameter together with the online metrics in the mixed model. The improvement is statistically reliably in terms of BIC[6] while the difference between the correlations is not significant.

A MAD equal to 4.61 is about 8.5% of the total score of the 8th grade MCAS math test, which is not bad. But should we be satisfied or we should push even harder until we can get a MAD of zero? An important question is what a good comparison should be. One reasonable thing to do is to compare to the MCAS, the standardized test itself. Ideally, we wanted to see how good one MCAS test was at predicting another MCAS test. We probably should not hope to do better than that. (But considering there is measurement error in the MCAS test, maybe we can!). We ran a simulation study by randomly splitting one test into two halves and used student scores on one half to predict the second half [9]. They reported MAD of 1.89 that is 11% of the maximum score on half of the test. Since their prediction error in [9] is very close to 11%, they claimed that their approach did as well as MCAS test on predicting math proficiency. In this paper, we propose a different approach to use student cross-year data for determining whether a student model is as good as the standardized test at estimating student proficiency. Assume student math proficiency in the 8th grade and in the 10th grade are highly correlated. Since the

[6] Raftery [[13] 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

measurement error is relatively independent due to the two years time interval between the tests, therefore, whichever (our student model or the MCAS test) better predicts 10th grade MCAS score is better assessing student math skill at 8th grade.

Let define MCAS8' be the leave-one-out[7] predicted score for 8th grade MCAS that comes from our best student model, the mixed model; MCAS8 be the actual 8th grade MCAS score and MCAS10 be the actual 10th grade MCAS score. Then we asked the question: Can MCAS8' predict MCAS10 better than MCAS8 does? To answer the question, we calculated the correlation between the three metrics: MCAS8', MCAS8 and MCAS10, as presented in Figure 1.



MCAS8: 8th grade MCAS score

MCAS8': our prediction of 8th grade MCAS score using online metrics and student proficiency

MCAS10: 10th grade MCAS score

Figure 1: Correlation between IRT student proficiency estimate, MCAS8', MCAS8 and MCAS10

First of all, we want to point out that all correlations in Figure 1 are statistically reliable ($p < 0.001$). The student proficiency estimated by the lean model correlates with MCAS10 with r equal to .628. It does not do as well as MCAS8 and MCAS8' as we have expected. Even though, we think it is worth finding out and having this lean model, which is based on less data, as a contrast case. It is the most direct test of the question of whether ASSISTment use could essentially replace the 8th grade test. Both MCAS8 and MCAS8' are reliable predictors of MCAS10. MCAS8 correlates with MCAS10 with r equal to 0.731 while the correlation between MCAS8' and MCAS10 is fractionally lower (r = 0.728). A significance test[8] shows they are not statistically reliably different, which suggests that our student model can do as well as MCAS test on predicting the MCAS score two years later. Since both MCAS tests are measuring the student's math proficiency, it can be considered as the evidence that the student model is doing a good job estimating student math proficiency. At the very least, what our system is modeling is relatively stable across a two-year interval.

---

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

[8] The test is done online at http://www.quantitativeskills.com/sisa/statistics/correl.htm

4 Discussion

There has been a big interest on modeling student knowledge. Corbett & Bhatnagar [6] describes an early and successful effort to increase the predictive validity of student modeling in the ACT 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 [3] used knowledge tracing to construct a student model that can predict student performance at both coarse- (overall proficiency) and fine-grained (for a particular word in the reading tutor) sizes. Anozie & Junker [1] 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 developed in [8], and built several linear prediction models, predicting end-of-year raw MCAS scores for each month. In [8] we developed the metrics as listed in section 3.2 to measure the amount of assistance a student needs to solve a problem, how fast a student can solve a problem, etc. and showed these metrics helped us better assess students. The result in this paper reinforced our previous result as evaluated by a different approach.

In section 3, we describe the method of using student test data from multiple years to compare a student model to a standardized test. Two other approaches have been described in the literature. In [3], Beck & Sison found 3 tests that measures extremely similar constructs to the standardized test that they were interested in. They took the arithmetic mean of those tests as a proxy measure for the true score on the original measure. The pro of this method is that it can be done quickly while the con is that construct validity could be an issue. In [9], we ran a simulation study by "splitting" a standardized test into two parts and the prediction power of the standardized test (actually a half of the standardized test) is determined by how well student performance on one half of the test predicts their performance on the other half. Similarly to the "proxy" measure method in [3], the pro of the "splitting" method is the quickness but it also has some cons. Firstly, if there is measurement error for a particular day (e.g. a student is somewhat ill or just tired), then splitting the test in half will produce a correlated measurement error in both halves, artificially increasing the test's reliability relative to the measure we bring up in this paper (which is not based on data from the same day as the MCAS). Secondly, to do the splitting, it required assess to item level data which is not always available. In this paper, we propose a third method, which is a longitudinal approach. By going across years, we avoid this confound with measurement error, and get a fairer baseline. Though, we do admit that it takes longer time and harder effort to collect data across years (in our case, 3 years).

5 Future work and Conclusions

We will continue working on improving the online assistance metrics. For instance, since the number of hints available is different across problems and the amount of information released in each level of hint differs too, instead of simply summing-up or computing the mean value, we want to construct some weighting function to better measure the amount of assistance students requested to solve a problem. Another piece of work follows up is to predict fine grained knowledge across years. Since our model is clearly capturing

something that is predictive of student future performance, we are considering focusing on determining what predicts specific deficits in an area. The research question we want to answer will be: can an 8th grade student model be used to predict the student will have a problem with a specific 10th grade skill? Teachers will be glad to know the answer so that they can adjust their instruction to better help student knowledge learning.

In this paper, we replicated the study in [8], showing the online ASSISTment metrics are doing a good job at predicting student math proficiency. On top of that, we propose a new method for evaluating the predictive accuracy of a student model relative to the standardized test, using student standardized test scores across years (2005 through 2007). We found some evidence that we can model student math proficiency as well as the standardized test as measured by the new evaluation criterion. Additionally, we want to stress that this is a rather long-term prediction. The collection of the online data started in September, 2004; the 8^{th} grade MCAS score that we are predicting came in at the end of year 2005; while the 10^{th} grade MCAS score that we used to evaluate our prediction were available at the end of year 2007. We consider the new method as a main contribution of this paper as there are few results showing a student model is as good as a standardized test. We have shown that our model hits this level and have presented an alternative way of performing the comparison.

Acknowledgement

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.

Reference

[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] 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.

[4] Brown, A. L., Bryant, N.R., & Campione, J. C. (1983). Preschool children's learning and transfer of matrices problems: Potential for improvement. Paper presented at the Society for Research in Child Development meetings, Detroit.
[5] Corbett, A. T., Koedinger, K. R.,& Hadley,W. H. (2001). Cognitive Tutors: From the research classroom to all classrooms. In Goodman, P. S. (Ed.) Technology Enhanced Learning: Opportunities for Change. Mahwah, NJ: Lawrence Erlbaum Associates.
[6] Corbett, A.T. and Bhatnagar, A. (1997). Student modeling in the ACT Programming Tutor: Adjusting a procedural learning model with declarative knowledge. Proceedings of the Sixth International Conference on User Modeling. New York: Springer-Verlag Wein.
[7] Embretson, S. E. & Reise, S. P. (2000). *Item Response Theory for Psychologists*. Lawrence Erlbaum Associates, New Jersey.
[8] Feng, M., Heffernan, N.T., & Koedinger, K.R. (2006a). Addressing the Testing Challenge with a Web-Based E-Assessment System that Tutors as it Assesses. Proceedings of the Fifteenth International World Wide Web Conference. pp. 307-316. New York, NY: ACM Press. 2006.
[9] 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.
[10] Feng, M. & Heffernan, N. (2007). Towards Live Informing and Automatic Analyzing of Student Learning: Reporting in ASSISTment System. Journal of Interactive Learning Research. 18 (2), pp. 207-230. Chesapeake, VA: AACE.
[11] Grigorenko, E. L. & Sternberg, R. J. (1998). Dynamic Testing. In Psychological Bulletin, 124, pages 75-111.
[12] Olson, L. (2005). Special report: testing takes off. Education Week, November 30, 2005, pp. 10–14.
[13] Raftery, A. E. (1995). Bayesian model selection in social research. In Sociological Methodology, 25, pages 111-163.
[14] 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, 555-562. Amsterdam: ISO Press
[15] Razzaq, Feng, Heffernan, Koedinger, Nuzzo-Jones, Junker, Macasek, Rasmussen, Turner & Walonoski (2007). Blending Assessment and Instructional Assistance. In Nadia Nedjah, Luiza deMacedo Mourelle, Mario Neto Borges and Nival Nunesde Almeida (Eds). Intelligent Educational Machines within the Intelligent Systems Engineering Book Series . pp.23-49. (see http://www.isebis.eng.uerj.br/). Springer Berlin / Heidelberg.
[16] Van der Linden, W. J. & Hambleton, R. K. (eds.) (1997). Handbook of Model Item Response Theory. New York: Springer Verlag.