Content learning analysis using the moment-by-moment learning detector

Sujith M. Gowda[1], Zachary Pardos[2], Ryan S.J.d. Baker[1]

[1]Department of Social Science and Policy Studies, Worcester Polytechnic Institute, Worcester, MA USA
[2] Department of Computer Science, Worcester Polytechnic Institute, Worcester, MA USA
{ sujithmg, zpardos, rsbaker }@wpi.edu

Abstract. In recent years, it has become clear that educational data mining methods can play a positive role in refining the content of intelligent tutoring systems. In particular, efforts to determine which content is more and less effective at promoting learning can help improve tutoring systems by identifying ineffective content and cycling it out of the system. Analysis of the learning value of content can also help teachers and system designers create better content by taking notice of what has and has not worked in the past. Past work has looked solely at student response data in doing this type of analysis; we extend this work by instead utilizing the moment-by-moment learning model, P(J). This model uses parameters learned from Bayesian Knowledge Tracing as well as other features extracted from log data to compute the probability that a student learned a skill at a specific problem step. By averaging P(J) values for a particular item across students, and comparing items using statistical testing with post-hoc controls, we can investigate which items typically produce more and less learning. We use this analysis to evaluate items within twenty problem sets completed by students using the ASSISTments Platform, and show how item learning results can be obtained and interpreted from this analysis.

Keywords: Educational data mining, item sequencing, learning gains

1 Introduction

The last several years have begun to see a shift in the sources of intelligent tutor content. As recently as five years ago, most intelligent tutor content was authored in programming development kits, and took considerable work to create – according to one estimate, it takes over 200 hours of a Ph.D.-level researcher's time to create one hour of student-usable content [16]. However, the recent advent of tools for rapid problem authoring by non-programmers [cf. 1, 13] has begun to change this practice. In fact, some intelligent tutoring systems are being authored via crowd-sourcing methods, where a wide range of individuals can contribute problems and content. For example, in the ASSISTments Platform [10], many problems and associated tutoring for those problems are now authored by teachers.

adfa, p. 1, 2011.

The move toward a wider base of content developers presents both opportunities and challenges. A wider developer base enables new content to be created more quickly and more responsively than traditional approaches. However, assuring and maintaining quality is a greater challenge when content is being created by a wider range of individuals, many of whom do not have explicit training in creating intelligent tutoring systems. (Though this is an opportunity in itself, as some teachers may have innovative new ideas for problem content that are better than current approaches). Also, as community-authored content grows rapidly, it is not feasible for small research teams to continually vet new content.

Given rapidly expanding content of uncertain quality, one approach to assuring and maintaining quality is to use educational data mining to vet content. The data produced by students as they use a tutoring system can provide indicators of which problems are most effective. Work in this area can build off of prior approaches to determine which pedagogical strategies lead to better learning experiences for students. For example, Beck and colleagues [6] used learning decomposition methods to study the effectiveness of different learning strategies for different groups of students. Chi and VanLehn [7] used reinforcement learning to study this same issue.

The approach proposed in [6] was adopted by Feng et al. [11], who used learning decomposition to determine that problems had varying efficacy within the ASSISTments Platform. This approach used logistic regression to analyze the future performance associated with having received a specific problem. Similarly, Pardos and Heffernan [17] addressed this same issue with a model based on Bayesian Knowledge-Tracing. Pardos et al. showed that models based this framework could be modified to measure the learning probability of individual items within particular knowledge components (KCs). Pardos suggested that item learning effects can be measured so long as the order of the items within a KC is randomized per student. Given randomization of item order, the sets of items can be analyzed as a quasi-randomized controlled trial.

These approaches provide actionable information on which problems are most effective and least effective. However, they are somewhat limited in terms of their sensitivity. First, assessments of problem effectiveness are dependent on performance in immediately subsequent problems; if those problems are of varying difficulty, there may be substantial noise in estimations of learning effectiveness. In addition, correctness does not take into account all of the information about a student action; other aspects of student performance have also been shown to predict knowledge and learning [cf. 9].

To address this possible limitation and create a richer indicator of the differential learning associated with different problems, we adopt an alternate paradigm for measuring learning: the moment-by-moment learning model [4]. This model is designed to specifically assess the learning that occurs within a specific problem. Instead of assessing the current degree of latent knowledge, it assesses the degree of knowledge learned at a specific moment using a function of the aspects of the student's actions on that problem (such as speed of response and use of help features).

In this paper, we apply the moment-by-moment learning model to a group of problem sets from the ASSISTments Platform. We then conduct statistical analysis to

determine the degree to which different problems have different moment-by-moment learning across students, and study the problems associated with the largest and smallest degree of moment-by-moment learning in two data sets.

2 Data

The data used in this analysis comes from the ASSISTments Tutoring Systems [10], with data drawn from the 2009-2010 school year. The students were from 7th and 8th grade Algebra classes with ages 12-14. The 8,519 students in the data set were drawn from 108 schools, primarily in Massachusetts. Students used the software for one class day approximately every two weeks throughout the school year, completing a range of problem sets involving different mathematical skills. The system provided instructional assistance to troubled students by breaking the original problem into scaffolding steps or displaying hint messages on-screen, upon student request. The ASSISTments tutoring system allows teachers to control the ordering of the problems within a problem set, choosing between a pre-chosen order, or random order. In this paper, we analyze a subset of the data drawn from students using random order problems within a problem set, selecting only problems that are associated with at least one cognitive skill.

There were a total of 78,558 student actions, made by 3,169 students on 1,170 problems, for whom the problem order was set to random and each problem was associated with at least one skill. There were some problems that were associated with more than one skill. For these problems, we treated them as representing evidence for each skill equally and with full credit assignment to each skill (i.e. a problem with three skills was treated the same as three problems, one tied to each of the three skills). Within the data set, there were a total of 945 skill-problem sets, out of which we selected 20 skill-problem sets that had the highest number of student actions, giving a final data set with 20,760 student actions produced by 2,210 students on 80 problems.

3 Detecting learning using moment-by-moment learning model

In this section we describe the moment-by-moment learning model developed by Baker and colleagues [4]. This model estimates the probability that a student learned a skill at a specific problem step, termed P(J). Recent results have argued in favor of this model's face validity; derivatives of this model can successfully predict students' final knowledge as assessed by Bayesian Knowledge Tracing [4], and can successfully predict students' preparation for future learning [5]. Bayesian Knowledge-Tracing (BKT) is a well-established approach for modeling student knowledge within an intelligent tutoring system [8]. BKT uses a four-parameter two-node dynamic Bayesian network to probabilistically assess the knowledge of a student for a specific skill. We use P(J) values in this analysis to assess the amount that students typically learn from each problem within a randomly ordered problem set.

3.1 Development of P(J) model

The P(J) model was developed using a two-step process, the same procedure used in [4]. First, training labels to detect moment-by-moment learning were generated for each problem step in a tutor data set. The labels were generated by applying Bayes' Rule to knowledge estimates from a traditional BKT model, in combination with the information about the correctness of the next two problem-solving actions of the student on items involving the same skill. Next, a set of predictor features was generated using past tutor data to form a training data set. The predictor feature set included 4 categories of features: 1) Action correctness, this category included features like is the action correct, incorrect or hint request, 2) Step interface type included feature that are based on type of interface widget involved, like is the problem multiple choice or just a single choice, 3) Response times, this categories included features that are derived from the amount of time taken to complete problem-solving steps, and 4) Problem solving history included features that characterize the student's problem-solving history in the tutor. These predictor features date back to the development of "gaming the system" detectors for Cognitive Tutors [3]. In addition to these features, skill difficulty related features were also included to increase the goodness of the model [12]. Linear regression was conducted within Rapidminer 4.6 [15] to develop models to predict P(J). This resulted in a set of numerical predictions of P(J), moment-by-moment learning, for each problem-solving step. The cross-validated correlation between the model and the original training labels was 0.449.

4 Overall Comparison of Problem Effectiveness

With the outputs of the P(J) detector, it is possible to assess the learning effectiveness of each problem in each skill-problem set. We do so by obtaining the set of values of P(J) for each problem, across students. We can then search for particularly poor problems and particularly effective problems. We analyze this in two ways. First, we conduct a one-way ANOVA to determine whether there are overall differences in the mean value of P(J) between problems in the same skill-problem set. Next, we attempt to determine if each skill-problem set has a single problem that is either better or worse than all other problems in the skill-problem set, an indicator that this problem is particularly effective or ineffective. It should be noted that the P(J) value is capturing the combined learning value of the problem and its tutoring (scaffolds and hints). Results are summarized in Table 1.

We found that 12 sets out of 20 skill-problem sets had statistically significant differences in learning between problems. Within these 12 skill-problem sets, we studied whether there was a best and worst problem, using post-hoc methods. The Levene test [14] was used to determine if the P(J) values for each problem in a skill-problem set had equal variance or not, to avoid violating the assumptions of the post-hoc analysis methods. Tukey's test was used when equal variance was assumed, and Tamhane's T2 test was used when equal variance was not assumed. Given the post-hoc differences between problems, a problem was labeled a best problem if it had positive mean difference with all the other problems and was significantly different from all the

other problems in the skill-problem set. Similarly, a problem was labeled a worst problem if it had negative mean difference with all the other problems and was significantly different from all the other problems in the skill-problem set. According to this test, 7 of the 12 problem sets had a single problem that was substantially better or worse than all other problems.

Table 1. – ANOVA results of 20 skill-problem sets. ** = statistical significance of $p<0.05$

Skill-Problem Set	Total Actions	Best Problem	Worst Problem	F-test
ConversionOfFractionDecimalsPercents	867	---	---	F(1, 865) = 0.22
CountingMethods	752	No	Yes	F(2, 749) = 12.64**
Estimation	510	---	---	F(2, 507) = 0.52
FindingFractionsandRatio	849	Yes	Yes	F(1, 847) = 8.28**
HistogramasTableOrGraph	481	---	---	F(2, 478) = 1.721
LineOfBestFit	713	---	---	F(3, 709) = 1.60
Median	612	Yes	No	F(2, 609) = 9.76**
MultiplicationandDivisionIntegers	850	No	No	F(7, 842) = 28.16**
NumberLine	864	---	---	F(1, 862) = 2.89
PercentOf	1703	No	No	F(7, 1695) = 84.85**
PickingEquationandExpressionFromChoices	535	---	---	F(3, 531) = 0.54
PointPlotting-1	868	Yes	Yes	F(1, 866) = 8.59**
PointPlotting-2	520	---	---	F(1, 518) = 0.99
Proportion-1	1220	Yes	Yes	F(1, 1218) = 8.59**
Proportion-2	2716	Yes	No	F(4, 2711) = 41.47**
Proportion-3	1056	No	No	F(4, 1051) = 33.06**
PythagoreanTheorem	2174	No	No	F(12, 2161) = 6.65**
Range	810	No	Yes	F(2, 807) = 16.44**
Transformation	878	No	No	F(7, 870) = 5.30**
UnitConversionWithinaSystem	595	---	---	F(1, 593) = 0.80

4.1 Case study of individual problems and their tutoring

Of the 20 skill-problem sets, there were seven problems that were significantly better or worse than all other problems in the same skill-problem set. These seven are shown in Table 2. Since the learning value of an item is a latent measurement, we have no ground truth to compare it to in order to verify that the best or worst items detected

were in fact the correct ones. Instead, we present the problems chosen as best and worst as dictated by the P(J) item learning detector and see if the results have face validity, that is if the detector looks like it measured what we intended for it to measure. Due to space limitations, we focus on two skill-problem sets, comparing a problem that is significantly different from all other problems with another problem from the set. We select the two skill-problem sets among the four possible options that have the largest difference in P(J) between the problems with the highest and lowest P(J). To facilitate discussion of differences, we compare the significantly different problems to the problem at the other end of the range. Within the PDF version of this document, the reader can inspect the problems, by clicking on any of the IDs in Table 2. The hyperlinks lead to a public preview of the items on the ASSISTments system.

Table 2. skill-problem sets with significant learning items

Problem set	Best item ID	Worst item ID	Mean difference between P(J) values
Proportion-2	**15792**	**24642**	0.0183
Range	**27521**	**25796**	0.0127
Counting Methods	**24754**	**24752**	0.0106
Median	**1059**	**2239**	0.0090
Proportion-1	**15792**	**15844**	0.0049
PointPlotting-1	**12353**	**12354**	0.0048
FindingFractionsandRatios	**12375**	**12376**	0.0038

4.2 Case study of Proportion-2's best and worst problems

The skill-problem set Proportion-2 had the largest difference in P(J) between the best and worst problem among the four skill-problem sets with a significant best or worst problem, 0.018. In this skill-problem set, one problem had statistically significantly higher P(J) than all the other problems in the skill-problem set. The problems with the highest P(J) and lowest are shown in Figure 1.

Figure 1 shows the best problem (on the left) in the Proportion-2 skill problem set. This problem has a visual component (the figure of the triangle) and is multiple-choice. The choices contain possible fraction equalities and the student is asked to select the one that can be used to solve for X. The first hint shows the student that there is a small triangle within the larger one. The following three hints proceed to evaluate the three wrong choices and tell the student which part of the answer is wrong and why. The last hint shows the correct answer, explains why it is correct, and shows four other proportion equalities that would have also been correct. The total hint count in this problem is five. Due to space limitations, the figure only shows the first three hints. This highly effective problem has more than double as many hints as the comparison problem, and uses visuals and significantly more text to teach the concept of proportion. From this comparison, it is not immediately clear which of these differences is beneficial, but multiple hypotheses are now available for improving other problems in this skill-problem set. For problems with much lesser magni-

tude of P(J) difference, additional attributes of the problems and their help would likely need to be defined in order to tease out an explanation for the more subtle difference in learning.

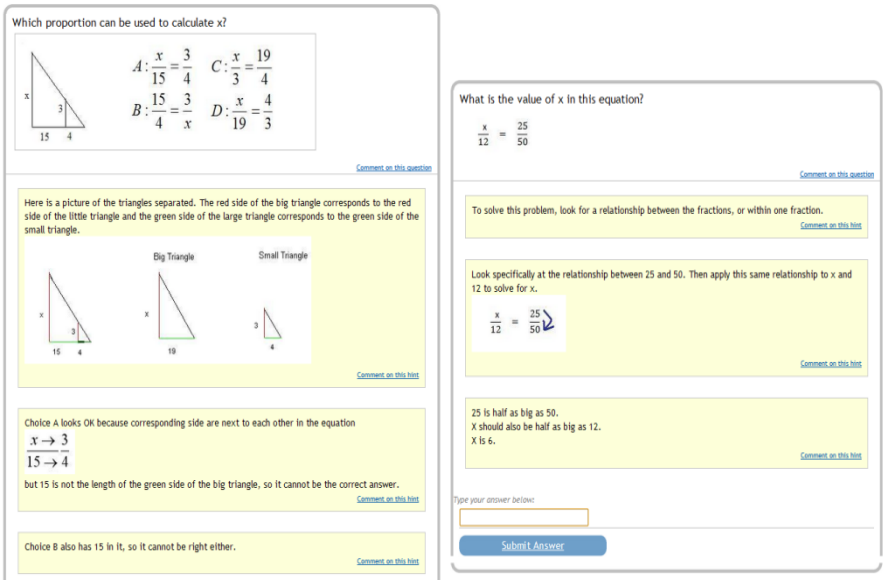

Fig. 1. Proportion 1: the (sig) best problem and its tutoring (left) and the worst problem (right)

The problem to the right in Figure 1 shows the worst P(J) problem, which asks the student to solve for X where X is part of a fraction equal to another fraction. The tutoring for the problem gives the student a first hint which suggests the student observe the relationship between the numerator and denominator of the fraction on the right side of the equation and apply this relationship to the fraction on the left side to determine X. The second hint explicitly tells the student the relationship between numerator and denominator which is that the numerator is half of the denominator of the fraction on the right side of the equation. The third hint is a bottom-out hint, and gives the student the answer.

4.3 Case study of Range, best and worst problems

The skill-problem set Range had the second-largest difference in P(J) between the best and worst problem, 0.013. In this skill-problem set, one problem had statistically significantly lower P(J) than all the other problems in the skill-problem set.

Figure 2 shows the worst problem in this skill-problem set (on the right), which asks for the range of the points scored in the table. This problem contains three scaffolds that in turn prompt the student for the maximum and minimum scores observed, and then re-asks the original question. Each of the scaffolds contains two hints. The first hint suggests the student look at the table for the answer and the second hint

shows another picture of the table with the relevant row highlighted. The total number of hint in this problem was six.

The problem to the left in Figure 2 shows the best problem, which also shows a two column table but asks which of four multiple choice statistics has the highest value. This problem has six scaffolds. The first prompts the student to count the number of animals listed. The next four scaffolds teach the student how to compute the mean, median, mode, and range using the table in the problem. The last scaffold re-asks the original question. There are 20 hints in this problem.

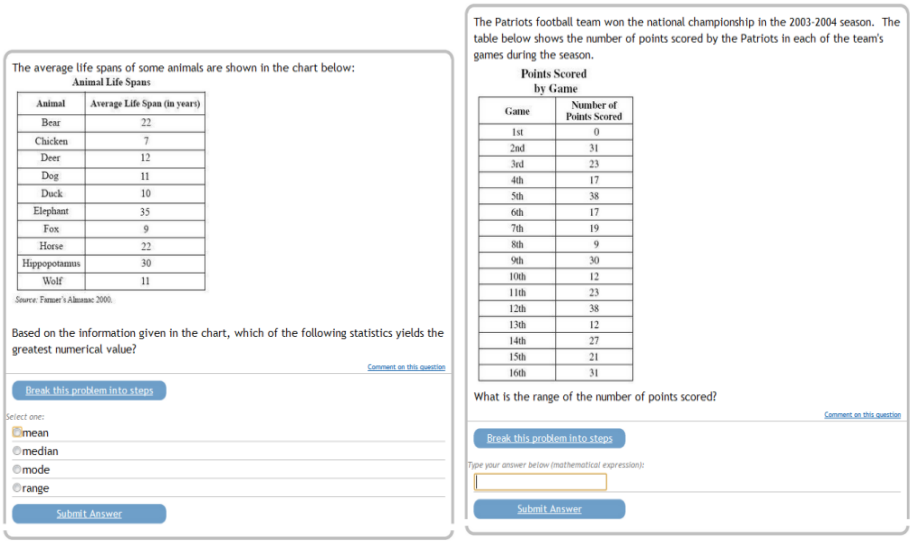

Fig. 2. Range: the best problem (left) and the (sig) worst (right)

Comparing the two problem's tutoring of the skill of range, there does not appear to be anything strikingly deficient about the significantly worst problem's tutoring. However, the most significant difference in the content between the worst and best problem is that the best problem contains three additional skills (mean, median, and mode) while the worst problem only contains range. A look at the Q-matrix for both problems revealed that the best problem was indeed tagged with four skills while the worst problem was tagged with only a single skill. Since P(J) is computed based on the relative learning value of the problems in a set, it appears that P(J) has detected a skill difference between problems. The tutoring of the problem that teaches and requires only the skill of range has little chance of providing the requisite knowledge to solve a problem that requires mean, median, mode, and range; however the four-skill problem has the tutoring to provide the requisite knowledge for the single-skill problem which would explain significant P(J) difference between problems.

5 Discussion

We have shown how the moment of learning detector can be applied to evaluate the relative learning value of problems in a set and how statistical tests can be run to determine if there are problems which are significantly better or worse. We conducted a case study of problem pairs in two skill-problem sets which showed the most significant differences in P(J) in order to investigate if differences could be plainly observed by viewing the problems and their tutoring approaches.

Several avenues exist for further research in the area of learning value analysis. Firstly, the method could be applied at the skill-problem set level to detect which problem sets pertaining to a common skill are providing the most learning value. This analysis would require a dataset where the order of problem sets, at least within a skill, were randomized per student. A second area for further study is a more stringent validity test. Face validity tests are subjective and fall far short of confirming that the claimed underlying construct is being accurately measured. A gold standard validity test would be a randomized controlled trial where individual problems were tested for learning gain with a pre/post-test design. The existence of a significantly higher or lower learning gain problem could be identified and compared to the findings of the P(J) learning value detector and other learning item analysis techniques.

Acknowledgements: We would also like to thank Lisa Rossi for valuable comments and suggestions. This research was supported by the National Science Foundation via the Pittsburgh Science of Learning Center, grant award #SBE-0836012, and by a "Graduates in K-12 Education" (GK-12) Fellowship, award number DGE0742503. We would like to thank the additional funders of the ASSISTments Platform found here: http://www.webcitation.org/5ym157Yfr

References

1. Aleven, V., McLaren, B.M., Sewall, J., Koedinger, K.R. (2006) The Cognitive Tutor Authoring Tools (CTAT): Versatile and Increasingly Rapid Creation of Tutors. Proceedings of the 8th International Conference on Intelligent Tutoring Systems, 61-70.
2. Baker, R.S.J.d., Corbett, A.T., Aleven, V. (2008) More Accurate Student Modeling Through Contextual Estimation of Slip and Guess Probabilities in Bayesian Knowledge Tracing. Proc. of the 9th International Conference on Intelligent Tutoring Systems, 406-415
3. Baker, R.S.J.d., Corbett, A.T., Roll, I., Koedinger, K.R. (2008) Developing a Generalizable Detector of When Students Game the System. User Modeling and User-Adapted Interaction, 18, 3, 287-314.
4. Baker, R.S.J.d., Goldstein, A.B., Heffernan, N.T. (in press) Detecting Learning Moment-by-Moment. To appear in International Journal of Artificial Intelligence in Education.
5. Baker, R.S.J.d., Gowda, S.M., Corbett, A.T. (2011) Automatically Detecting a Student's Preparation for Future Learning: Help Use is Key. Proceedings of the 4th International Conference on Educational Data Mining, 179-188.
6. Beck, J.E., Mostow, J. (2008) How who should practice: Using learning decomposition to evaluate the efficacy of different types of practice for different types of students. In:

Woolf, B.P., Aïmeur, E., Nkambou, R., Lajoie, S. (eds.) ITS 2008. LNCS, vol. 5091, pp. 353–362. Springer, Heidelberg
7. Chi, M., VanLehn, K., Litman, D. (2010). Do Micro-Level Tutorial Decisions Matter: Applying Reinforcement Learning to Induce Pedagogical Tutorial Tactics. The 10th International Conference on Intelligent Tutoring Systems, 224-234
8. Corbett, A.T., Anderson, J.R. (1995) Knowledge Tracing: Modeling the Acquisition of Procedural Knowledge. User Modeling and User-Adapted Interaction, 4, 253-278
9. Feng, M., Heffernan, N.T, Koedinger, K.R., (2006). Predicting State Test Scores Better with Intelligent Tutoring Systems: Developing Metrics to Measure Assistance Required, The 8th International Conference on Intelligent Tutoring System, 2006, Taiwan.
10. Feng, M., Heffernan, N.T., & Koedinger, K.R. (2009). Addressing the assessment challenge in an Intelligent Tutoring System that tutors as it assesses. The Journal of User Modeling and User-Adapted Interaction. Vol 19: p243-266
11. Feng, M., Heffernan, N.T., Beck, J. (2009). Using learning decomposition to analyze instructional effectiveness in the ASSISTment system. In Graesser, A., Dimitrova , V., Mizoguchi, R. (Eds.). Proceedings of the 14th International Conference on Artificial Intelligence in Education. Brighton, UK.
12. Gowda, S.M., Rowe, J.P., Baker, R.S.J.d., Chi, M., Koedinger, K.R. (2011) Improving Models of Slipping, Guessing, and Moment-by-Moment Learning with Estimates of Skill Difficulty. Proc. of the 4th International Conference on Educational Data Mining, 199-208.
13. L. Razzaq, J. Patvarczki, S. F. Almeida, M. Vartak, M. Feng, N. T. Heffernan and K. R. Koedinger. (2009) "The ASSISTment builder: Supporting the Life-cycle of ITS Content Creation," IEEE Transactions on Learning Technologies Special Issue on Real-World Applications of Intelligent Tutoring Systems, vol. 2, no. 2, pp. 157-166.
14. Levene, Howard (1960). "Robust tests for equality of variances". In Ingram Olkin, Harold Hotelling, et alia. Stanford University Press. pp. 278–292.
15. Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M., Euler, T. (2006). YALE: Rapid Prototyping for Complex Data Mining Tasks. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2006), 935-94
16. Murray, T. (1999). Authoring intelligent tutoring systems: An analysis of the state of the art. International Journal of Artificial Intelligence in Education, 10, 98-129.
17. Pardos, Z. & Heffernan, N. (2009) Detecting the Learning Value of Items in a Randomized Problem Set. In Dimitrova, Mizoguchi, du Boulay & Graesser (Eds.) Proceedings of the 2009 Artificial Intelligence in Education Conference. IOS Press. pp. 499-506.
18. Pardos, Z.A., Dailey, M. & Heffernan, N. (In Press) Learning what works in ITS from non-traditional randomized controlled trial data. In press International Journal of Artificial Intelligence in Education.