Feng, M., Heffernan, N., Beck, J, & Koedinger, K. (2008) Can we predict which groups of questions students will learn from? In Baker & Beck (Eds.). Proceedings of the 1st International Conference on Education Data Mining. pp.218-225. Montreal, 2008.

Can we Predict which Groups of Questions Students will Learn from?

Mingyu Feng[1], Neil Heffernan[1], Joseph E. Beck[1], Kenneth Koedinger[2]

[1]{mfeng, nth}@wpi.edu, joseph.beck@educationaldatamining.org
Computer Science Department, Worcester Polytechnic Institute
[2]koedinger@cmu.edu
Human Computer Interaction Institute, Carnegie Mellon University

Abstract. In a previous study ([4]), we used the ASSISTment system to track student knowledge longitudinally over the course of a schools year, based upon each student using our system about a dozen times during the course of the year. This result confounded learning from the computer system with students learning from their sitting their normal class. In this work, we look to see if students were reliably learning from their time spent with the *computer in a single day*. Our result suggests that students performed better later in the same computer session on similar skills, which indicates students are learning from using ASSISTments. However, learning is rather uneven across groups of skills. We test a variety of hypotheses to explain this phenomenon and found that the automated approaches we tried were unable to account for the variation. However, human expert judgments were predictive as to which groups of skills were learnable.

1 Introduction

The field of educational data mining is often concerned with how to model student learning over time. More often than not, these models are concerned with how student performance changes while students are using the computer. In this project we look to see if learning from the computer system was happening over time, trying to separate out learning from the classroom. We then wanted to investigate to see if we could predict on which knowledge components students were systematically learning.

The ASSISTment project was funded to see if it was possible to teach students effectively while assessing student performance accurately at the same time. We have reported the results of our analysis of the assessment value of our system in [4] . The results indicated that student performance is on average reliably increasing during the course of the year. But since most students only use the system every other week, it is unclear how much credit should go to ASSISTments vs. classroom instruction.

In order to track students' progress in learning, different approaches have been established. Corbett & Anderson [3] employed a diagnostic approach called *knowledge tracing* that models students as an overlay of the ideal production rules. They proposed a two-state (learned & unlearned) learning model that allows student knowledge state to transit from one to the other probabilistically. One technique by Koedinger and colleagues is called Learning Factors Analysis (LFA) [2] takes advantage of the Power Law of Learning (see [7]) to fit student performance to a power function reflecting decreasing error rates over time. LFA has been proposed as a generic solution to evaluate

and compare many potential cognitive models of learning. Since student performance was often represented by a dichotomous variable, logistic regression models have been used as the statistical model for evaluation (e.g. [2] , [6]). In terms of related work on investigating the reasons of learning, Vanlehn et al. [8] explored the problem of what causes learning by contrasting cases where tutoring does or does not result in learning. In this study, we will investigate whether students learn within ASSISTments. We conducted a focused analysis of a subset of items and tracked how student performance on these items changes during the same ASSISTment session. We will explore the possible reasons of why on some sets of problems students learned or failed to learn.

2 Methodology

2.1 Experimental Design

Our hypothesis was that students were learning groups of items that share the same background knowledge requirement. Our subject manner expert picked 182 items out of the 300 8th grade (approximately 13 to 14 years old) math items in ASSISTment. Items that have the same deep features or knowledge requirements, like approximating square roots, but have different surface features, like cover stories, were organized into a **Group of Learning OPportunity (GLOP).** Besides, the expert excluded groups of items where learning would be too obvious or too trivial to be impressive. Singleton items were not selected either. The selected 182 items fall into 40 GLOPs with the number of items in each GLOP varies from 2 to 11. The items cover knowledge from all of the five major content strands identified by the Massachusetts Mathematics Curriculum Framework, relatively heavy on the strand Patterns, Relations & Algebra. Items in the same group were collected into the same section of ASSISTments, and seen in random order by students. Each student potentially saw 40 different GLOPs that involve different 8th grade math skills (e.g. fraction-multiplication, inducing-functions, symbolization articulation) in random order. Figure 1 shows three items in one GLOP that are about the concept "Area." All these problems asked students to compute the area of the shaded part in the figures. It is worth pointing out that all the GLOPs were constructed by focusing on the content of the items before the analysis done in this paper.

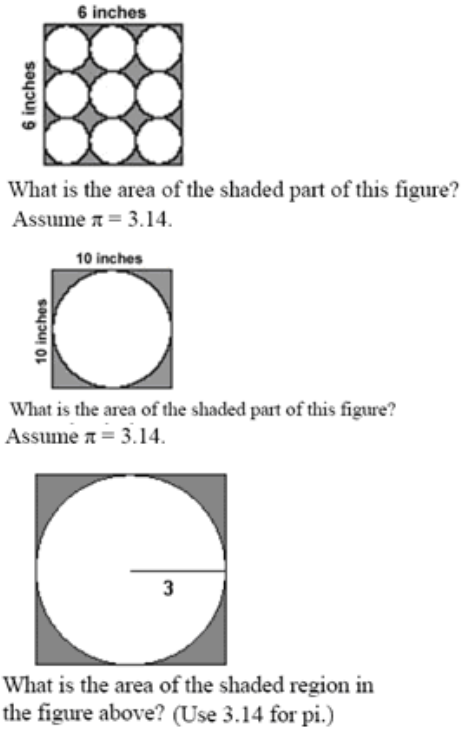

Figure 1. A sample GLOP that addresses the skill "Area"

We assessed learning by comparing student performance the first time they were given one item from a GLOP with their performance when they were given more items (also more opportunities) from the same GLOP in the same day. If students tend to perform

better on later opportunities of items in a GLOP, it indicates that they may have learned from the instructional assistance provided on items by the ASSISTment system that they worked on earlier by answering the scaffolding questions or by reading hint messages. There is controversy over whether same-day learning opportunities should be used as evidence of learning. For example, Beck [1] thought repeated trials were not indicative of learning. He chose not to use later encounters on the same day in the Reading Tutor since performance on those encounters is not a reflection of student knowledge but just retrieved from short term memory. Our domain (mostly 8th grade multi-step math problems) is more complex than reading and the items in a GLOP usually have different surface features. Solving these problems is not simple retrieval of an answer from a previous question. And even if it wasn't more complex, our later day trials are horribly confounded by classroom instruction due to low density of usage (every other week). In this paper, we chose to analyze the response data on the same day to eliminate the confound of learning happening because of classroom instruction **between** two ASSISTment sessions.

2.2 Sample of the data we used for analysis

We collected data for this analysis from Oct. 31, 2006 to Oct. 11th, 2007. 2000+ 8th grade students participated in the study. We defined participation in a GLOP as answering two or more questions in it, and excluded students who participated in less than five GLOPs to make sure each student has at least 10 data points. We ended up with a data set of 42,086 rows, with each row representing a student's attempt at an item. 777 students entered into our final data set, and each student on average worked on 54 items across 14 GLOPs. Table 1 shows a small sample of our data. In particular, Table 1 shows two students' performance on two GLOPs, 16, and 3 (partly). We use the column "correct?" to indicate whether the student answered the question correctly or not. The value will be 1 where he succeeded; otherwise, it is set to be zero. The first student worked on three problems (i.e. had 3 opportunities to learn) from GLOP 16 on April 2nd, 2007 starting from 9:39AM. He failed the first two but managed to solve the last one.

Table 1. Sample data showing two students' performance on two GLOPs

Student ID	GLOP ID	Question ID	Date	Time	Correct?	Opportunity
30296	16	1069	4/2/2007	9:39:20	0	1
30296	16	231	4/2/2007	9:42:19	0	2
30296	16	1512	4/2/2007	9:53:11	1	3
30300	3	2267	4/25/2007	7:48:15	0	1
30300	3	2244	4/25/2007	7:58:47	1	2

3 Research Question 1: Do students learn from ASSISTments?

We first attempted to determine whether the system effectively teaches. To answer the first research question if students are learning from ASSISTments, we ran a logistic regression to study the relationship between student performance (i.e. their responses to items) and the number of opportunities the student has on a GLOP. In our method, the dependent variable is student response to a question and we account for the difference of

student math proficiency by including the student as one of the predictor variables. Similarly, we include the question as another predictor with regard to the fact that questions in one GLOP may vary in difficulties. The regression formula is

Equation 1. Logistic regression model

$$\ln(\frac{p_{ij}}{1-p_{ij}}) = \alpha_i * Student_i + \beta_j * Question_j + \gamma * Opportunity\#$$

Where p_{ij} is the probability that the student *i* will answer question *j* correctly

Opportunity# indicates how many opportunities the student *i* has on a particular GLOP.

α_i, β_j and γ are the coefficients for the corresponding predictors $Student_i$, $Question_j$ and *Opportunity#.*

The model is very similar to LFA models except that skills are not included as factors since we are investigating generalized learning over all GLOPS. We ran a multinomial logistic regression treating student and question as factors and opportunity as a covariate. The regression coefficient estimated by the model, corresponding to the number of learning opportunities (γ), is .03 ($p < .001$). This result suggests that in general, students performed reliably better as they have more chances of practicing on the same GLOP. This result suggests that in general, students performed reliably better as they have more chances of practicing on the same GLOP. The coefficient in the logistic regression model indicates that students will improve by 0.03, on a logit scale, for each practice opportunity. This learning corresponds to approximately a 0.8% improvement in performance for each problem practiced, a rather small effect. In Massachusetts, MCAS test scores are categorized into four performance levels (namely warning, need improvement, proficient, and advanced). According to the results of 2006 MCAS test, students need to earn 13 more points (24% of the full score) to jump from need improvement to proficient which is required by the federal movement based on NCLB standards to graduate from high school. Theoretically, if students can gain 0.7% for each learning opportunity, they will fulfill the 24% improvement by solving 31 problems in ASSISTments. It should be noted that there may be a selection effect in this experiment in that better students are more likely to do more problems in a day and therefore more likely to contribute to this analysis. Also there is a limitation with the model that all GLOPs are assumed to produce the same amount of learning, which may not be true as we will show later.

A positive answer to the first question allows us to claim that students are learning from working in ASSISTments and the learning results are generalized across the 40 GLOPs. Then, we stepped further to explore if all of the GLOPs are equally effective at promoting learning. The answer is “no”, which is not surprising anyway since the items in different GLOPs vary on several aspects (e.g. focusing on various skills; built by authors with differing teaching experience using various teaching strategies, etc.). In summary, out of the 40 GLOPs, the amount of learning per opportunity is statistically reliably higher than zero on 11 of them. 2 GLOPs caused marginally reliable learning and 16 caused unreliable learning. And there is non-reliable “un-learning” for the remaining 11 GLOPs, suggesting that not much learning occurred when students worked on these GLOPs.

4 Research Question 2: Why students learned or failed to learn?

Now that we have shown that learning varies among GLOPs, we will explore the reasons for this variation. We are not only interested to know which category each GLOP falls in and but also curious why. Particularly, we want to investigate why students did not show learning on certain GLOPs. Our four hypotheses are:

1. H1: Learning transfer from harder items to easier items, or students tend to learn more by doing harder items than by doing easier items. Presumably, if a student learned to solve a hard item, he then should be able to do better on an easier item that requires similar skills. However, the converse is not necessarily true.
2. H2: Knowledge transfer occurs within GLOPs of items that use similar skills. We can never know exactly how a student internally represents a problem and what the exact skills a student applied to solve a problem. But if a GLOP is well-focused in what it covers, presumably students should show more learning within it.
3. H3: The "learnability" of the skills required by GLOPs varies. Our statistics show that each ASSISTment provides about 2 minutes of instruction. It can be hard to teach some skills effectively, for instance, *symbolization articulation*, in such a short period. Such skills require deep understanding and more practice to be able to apply and transfer, whereas some other skills such as *area* are more teachable since students only need to be reminded to apply the area formula.
4. H4: The efficacy of instructions has an impact on learning results. We can easily imagine that some GLOPs have better teaching efficacy than others. The quality of the scaffolding questions and hint messages can differ from one item to another as authors used a tutoring strategy that are more, or less, effective than others.

In this paper, we will test the first two hypotheses and leave the last two as future work. We plan to invite more content experts to help us identify the learnability of the related skills and to evaluate the quality of the ASSISTments by looking closely at the scaffolding questions and hint messages.

4.1 Do students learn more from harder items or easier items?

Noticing learning varies among GLOPs, the first thing we did is to explore the relationship between the amount of learning and the easiness of a GLOP (measured by the average difficulty of items in the GLOP). We calculated the rank-order correlation and got a coefficient of .333 ($p = .036$, $N=40$), which indicates that students learned more on harder GLOPs than on easier ones. We asked ourselves: why is this? A quick answer is that there is more room to grow for harder items. Or, maybe students just learn more from harder items than easier items.

Beck [1] introduced an approach called *learning decomposition* to analyze what type of practice was most effective for helping students learn a skill. The approach is a generalization of learning curve analysis, and uses regression to determine how to weight different types of practice opportunities relative to each other. In this paper, we apply learning decomposition to our data set to investigate how students acquire math skills: will their practice on harder items produce more learning? To test our first hypothesis, we added two columns to our data set. One column, entitled "easier_before_current",

represents how many items the student has seen in the same GLOP are easier than the current item. The other column, entitled "harder_before_current", indicates how many items were seen that are harder than the current one. We measure the easiness of the items using the item parameter given by a one-parameter Item Response Theory model (i.e. Rasch model[1]). The Rasch model was trained over data collected in the system from Sept., 2004 to Jan., 2008, including responses to 2,700 items from more than 14,000 students. We include the two columns as covariates in the regression model.

Equation 2. Learning Decomposition Model

$$\ln(\frac{p_{ij}}{1-p_{ij}}) = \alpha_i * Student_i + \beta_j * Question_j + \gamma_e * easier\_before\_current + \gamma_h * harder\_before\_current$$

Where γ_e and γ_h represents the coefficients for the two new covariates respectively.

The model was fitted in SPSS 14.0. We noticed that the coefficients of the two covariates of *easier_before_current* and *harder_before_current* are very close to each other (γ_e = .032, and γ_h = .033). The coefficient for *easier_before_current* is fractionally but not reliably lower (p=.966), which suggested that students learn as much from easier items as from harder items, and thus our first hypothesis is rejected.

4.2 Does more learning occur in GLOPs that are more focused?

H2 is different than H1 in that it believes that transfer occurs within GLOPs that have similar difficulty questions (and therefore address similar skills based on our assumption). To test H2, we want to investigate the relationship between the amount of learning that happened in each GLOP and the cohesiveness of the GLOP in term of item difficulty and the skills that are needed to answer the items.

We used two approaches to quantify the cohesiveness of the GLOPs. The first metric is an automated measure that comes from a computer modeling process based on the assumption that if two skills, A and B, are better modeled by a single skill, then practice on either A or B symmetrically transfers to the other. During the modeling process, for each GLOP, we compared the BIC of two models. The first model treated each question as having a separate difficulty. The second model treated all questions as having the same difficulty, and thus had *number_of_questions_in_GLOP-1* fewer parameters. Presumably, if the cohesiveness of a GLOP is high, we should expect the second model to fit better on our data as measured by Bayesian Information Criteria (BIC) (or any model fitting criteria that penalizes for model complexity). We followed the same procedure and calculated the difference of BIC values between two models for each of the GLOPs.

The second metric is based on our subject matter expert's ranking of the cohesiveness of the GLOPs. As requested by us, our subject matter expert set the rating from 1 to 5. A fit

[1] 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.

of 1 or 2 means that the items are very different. A 3 means there are some flaws in the selection, a 4 means there are just a few inconsistencies and a 5 means they fit very well. According to the ranking of the expert, 18 GLOPs got a fit of 5, 10 were given a fit of 4, 7 GLOPs got 3 and the remaining 5 GLOPs scored 2.

After obtaining the two metrics, we continued to analyze the relationship between the cohesiveness of the GLOPs and the amount of learning that happened in each of them. First, we calculated the rank-order correlation between the automated metric and the amount of learning (given by Equation II) but did not find a significant relation ($r = .13$, $p = .94$). We then discretized coherence into 3 coherence bins: high, medium and low and performed a one-way ANOVA to explore whether there were any differences in the amount of learning, but found no main effect ($F = .676$, $p = .515$). After that, we did the same analysis using the expert ranking of the cohesiveness. The rank-order correlation between fit and amount of learning is equal to .322 ($p = .045$). Yet the ANOVA shows no main effect of fit ($F = 1.573$, $p = .213$). Further more, instead of using five groups, we merged all GLOPs with fit less than 5 into one category named “non-perfect-fit” as a contrast to the ones with “perfect-fit” and ran an independent sample t-test to compare the mean between the two categories. The result suggested that there is statistically reliably more learning happening in GLOPs of perfect fit ($t = 2.311$, $p = .030$).

To complete the third side of the triangle of learning/automated coherence metric/expert ranking, we also computed the correlation between our two metrics of fit/cohesiveness and found out that they do not correlate with each other ($r =-.198$, $p=.22$), which means that an automated measure and an expert's judgment differ. In conclusion, H2 was supported by the expert’s judgment but not by the result of data mining.

5 Future work and Conclusions

In terms of caveats and recognized limitations, we want to first acknowledge that we don’t have control group to compare the learning result against to. Also, if student performance systematically varies over time apart from learning, our model is not able to account for it. For instance, if students experienced a ramp up effect of doing better over time, this could explain away our results. Similarly, if students get fatigued over a class period we would be underestimate the learning effect.

As a future work, we want to look to see how the items would be grouped by some automated method such as Q-matrix algorithm or LFA.

In conclusion, we presented evidence that suggests there is learning within ASSISTments. More interestingly, we found that the learning differed across the groups of items. We tested a variety of hypotheses to explain this phenomenon and found that the automated approaches we tried were unable to account for the variation. However, human expert judgments were predictive as to which groups of skills were learnable.

The contribution of the paper lies in two aspects. First, we looked at when learning occurs in an intelligent tutoring system and examined a variety of hypotheses on why learning happens. While these hypotheses seemed intuitive, they were not supported by our analysis. Second, student modeling research typically accounts for the amount of learning due to a practice opportunity, but generally does not try to take into account of

learning outside the tutor (such as classroom instruction, homework, etc.). Using this type of analysis that focuses on within-session learning, we isolated the effect to those caused by our tutor.

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.

References

[1] Beck, J.E. (2006). Using learning decomposition to analyze student fluency development. *Proceedings of the Workshop on Educational Data Mining at the 8th International Conference on Intelligent Tutoring Systems*. Jhongli, Taiwan. Pages 21-28.

[2] Cen, H., Koedinger, K., & Junker, B. (2006). Learning factor analysis – A general method for cognitive model evaluation and improvement. In *Proceedings of the 8th International Conference on Intelligent Tutoring Systems*. Springer-Verlag: Berlin. pp. 164-175.

[3] Corbett, A. T. & Anderson, J. R. (1995). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, 4, 253-278.

[4] Feng, M., Heffernan, N.T, Koedinger, K.R. (2006). Addressing the testing challenge with a web-based e-assessment system that tutors as it assesses. In *Proceedings of the 15th International World Wide Web Conference*. pp. 307-316. ACM Press: New York, NY. 2006.

[5] Koedinger, K. R., Anderson, J. R., Hadley, W. H. & Mark, M. A. (1995) Intelligent tutoring goes to school in the big city. In *Proceedings of the 7th Conference on Artificial Intelligence in Education*, pp. 421-428. Charlottesville, VA: Association for the Advancement of Computing in Education.

[6] Leszczenski, J. M. & Beck J. E. (2007). What's in a Word? Extending Learning Factors Analysis to Model Reading Transfer. In *Proceedings of the Educational Data Mining workshop held at the 14th International Conference on Artificial Intelligence in Education.*

[7] Newell, A. & Simon, H.A. (1972). *Human Problem Solving.* Englewood Cliffs, NJ: Prentice-Hall.

[8] VanLehn, K., Siler, S., Murray, C., & Baggett, W. (1998). What makes a tutorial event effective? In M. A. Gernsbacher & S. Derry (Eds.), *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society* (pp. 1084-1089). Hillsdale, NJ: Erlbaum