Feng, M., Heffernan, N.T., Beck, J. (In press). Using learning decomposition to analyze instructional effectiveness in the ASSISTment system. In Proceedings of the 14th International Conference on Artificial Intelligence in Education (AIED-2009).

Using Learning Decomposition to Analyze Instructional Effectiveness in the ASSISTment System

Mingyu Feng[1], Neil Heffernan and Joseph E. Beck
Dept. of Computer Science, Worcester Polytechnic Institute
100 Institute Road, Worcester, MA 01609

Abstract. A basic question of instruction is how effective it is in promoting student learning. This paper presents a study determining the relative efficacy of different instructional content by applying an educational data mining technique, learning decomposition. We use logistic regression to determine how much learning caused by different methods of presenting same skill, relative to each other. We analyze more than 60,000 performance data across 181 items from more than 2,000 students. Our results show that items are not all as effective on promoting student learning. We also did preliminary study on validating our results by comparing them with rankings from human experts. Our study demonstrates an easier and quicker approach of evaluating the quality of ITS contents than experimental studies.

Keywords. Evaluation, student modeling, learning decomposition, learning curves, educational data mining, item response theory

Introduction

The field of Intelligent Tutoring Systems 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 tutor. Many Intelligent Tutoring Systems have similar items that share the same prerequisite knowledge and target the same set of skills. In [5], we referred to such a group of items as Group of Learning Opportunities (GLOP). Naturally, a question arises: can we tell which item is the most effective at causing learning?

One popular method of determining whether one type of instruction is more effective than the other, or whether one tutor is more beneficial than another on helping student learn a skill, is to run a randomized controlled study. A major problem with the controlled study approach is that it can be expensive. A study could involve many users (in each condition), be of considerable duration, and require the administration of pre/post tests. To address this problem, Beck [1] introduced an approach called *learning decomposition*, an easy recipe to enable researchers to answer research question such as what type of practice is most effective for helping student to learn a skill. In this paper, we applied learning decomposition to leverage fine grained interaction data collected in the ASSISTment system [6].

[1] Corresponding Author.

The ASSISTment system is an online system that presents math problems to students of approximately 13 to 16 year old in middle school or high school to solve. When a student has trouble solving a problem, the system provides instructional assistance to lead the student through by breaking the problem into scaffolding steps, or displaying hint messages on the screen, upon student request. Time-stamped student answers are logged into the background database. In ASSISTment system, items in a GLOP often have different surface features; also, the authors may use different tutoring strategies to when creating the instructional content. Both of these factors may influence the amount students learn from an item.

The goal of this paper includes 1) Estimating and comparing the relative impact of various tutoring components in the ASSISTment system. 2) Presenting a case study of applying the learning decomposition technique to a domain, mathematics, other than reading where the technique has been shown to be valuable ([1,2,3]).

1. Methodology

Beck [1] introduced the idea of learning decomposition. It extends the classic exponential learning curve by taking into account the heterogeneity of different learning opportunities for a single skill. The standard form of exponential learning curve can be seen in Equation 1. In this model, parameter *A* represents students' performance on the first trial; *e* is the numerical constant (2.718); parameter *b* represents the learning rate of a skill, and *t* is the number of practice the learner has at the skill.

$$performance = A * e^{-b*t}$$

Equation 1. Standard exponential learning curve model

$$performance = A * e^{-b*(B*t_1+t_2)}$$

Equation 2. Learning decomposition model

The model as shown in Equation 1 does not differentiate different types of practice, but just counts up the total number of previous opportunities. In order to investigate the difference two types of practice (I and II), the learning opportunities are "decomposed" into two parts in the model in Equation 2 in which two new variables t_1 and t_2 are introduced in replace of *t*, and $t = t_1 + t_2$. t_1 represents the number of previous practice opportunities at one type I; and t_2 represents the number of previous opportunities of type II. The new parameter *B* characterizes the relative impact of type I trials compared to type II trials. The estimated value of B indicates how many trials that one practice of type I is worth relative to that of type II. For example, a B value of 2 would mean that practice of type I is twice as valuable as one practice of type II, while a B value of .5 indicates a practice of type I is half as effective as a practice of type II. The basic idea of learning decomposition is to find an estimate of weight B that renders the best fitting learning curve.

Equation 2 factors the learning opportunities into two types, but the decomposition technique can generalize to *n* types of trials by replacing t with $B_1*t_1 + B_2*t_2 + \ldots + t_n$. Thus, parameter B_i represents the impact of a type *i* trial relative to the "baseline" type *n*.

Various metrics can be used as an outcome measurement of student performance. For instance, Beck ([3]) chose to model student's reading time since it is a continuous

variable. When it comes to a nominal variable, e.g. dichotomous (0/1) response data, a logistic model should be used. Now learned performance, (i.e. *performance* in Equation 2), is reflected by odds ratio of success to failure. Equation 3 represents a logistic regression model for learning decomposition.

$$performance = \frac{P(correct\_answer)}{P(wrong\_answer)} = A * e^{-b*(B*t_1+t_2)} = e^{a+b*(B*t_1+t_2)}$$

Equation 3. Logistic models for learning decomposition

Equation 3 can be transformed to an equivalent form as below:

$$\ln(\frac{P(correct\_answer)}{P(wrong\_answer)}) = a + b*(B*t_1 + t_2)$$

Different approaches have been established to track students' progress in learning. One technique by Koedinger and colleagues is called Learning Factors Analysis (LFA) [4]. LFA has been proposed as a generic solution to evaluate, compare, and refine many potential cognitive models of learning. Since student performance is often represented by a dichotomous variable, logistic regression models have been used as the statistical model for evaluation. Although both LFA and learning decomposition are concerned with better understanding student learning, and both use logistic models, they have different assumption. LFA assumes all trials cause the same amount of learning but the skills associated with each trial may vary, while learning decomposition assumes the domain representation is constant but different types of practice cause different amounts of learning.

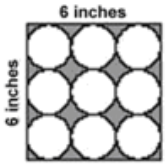
The reminder of the paper explores applying learning decomposition approach to answer questions about how students' acquisition of math skills is impacted by different instructional items, and various tutoring strategies.

2. Approach

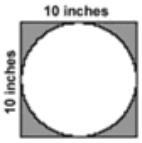
In [5], we reported our work on finding out whether students are reliably learning from the ASSISTment systems; and from which GLOPs. In this section, we take a closer look at each GLOP and investigate which item is most effective at causing learning in each GLOP. Rather than have explicit experimental and control groups, our approach in this paper is to examine how students' performance change based on which item in the GLOP they have just finished.

Our subject manner expert picked 181 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 GLOPs**.** Besides, the expert excluded groups of items where learning would be too obvious or too trivial to be impressive. Also, GLOPS had to be of at least size two. The selected 181 items fall into 39 GLOPs with the number of items in each GLOP varies from 2 to 11. The items are a fair sampling of the curriculum and cover knowledge from all of the five major content strands identified by the Massachusetts Mathematics Curriculum Framework: Number Sense and Operations; Patterns; Relations and Algebra; Geometry, Measurement, and Data Analysis; and Statistics and Probability. It sampled relatively heavily on the strand Patterns, Relations & Algebra. Items in the same GLOP were collected into the same section of ASSISTments, and seen in random order by students.

Each student potentially saw 39 different GLOPs that involve different 8th grade math skills (e.g. fraction-multiplication, inducing-functions). 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. We collected data for this analysis from Oct. 31, 2006 to Oct. 11th, 2007. Over 2000 8th grade students participated in the study. We exclude cases where the student only finished one item in the GLOP. We ended up with a data set of 54,600 rows, with each row representing a student's attempt at an item. 2,502 students entered into our final data set, mostly from Worcester, Massachusetts area. Each student on average worked on 22 items. 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.



What is the area of the shaded part of this figure? Assume $\pi = 3.14$.



What is the area of the shaded part of this figure? Assume $\pi = 3.14$.

What is the area of the shaded region in the figure above? (Use 3.14 for pi.)

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

Although one may argue for other indicators, e.g. students' help requests and response times, we simply choose to use the correctness of student's first attempt to an item as an outcome measure of their performance. Table 1 shows a sequence of time-ordered trials of student A on items in GLOP 1, together with the correctness of each response. The student finished all 5 items in the GLOP. He managed to solve the first, the forth and the fifth item but failed on the second and the third one. So, for teaching the math skill involved in GLOP 1, which item is likely to be more (or less) effective for student proficiency development? In order to answer this question, we adopt the idea of learning decomposition. Each item in a GLOP is considered as a different type of practice; then students' practice opportunities are factored into practice at each individual item. Since each student has at most one chance at an item, the number of previous opportunities at each item is either 0, indicating the student has not worked on that item, or 1, indicating the student has finished that item before. And Table 2 shows the corresponding data after the trials are decomposed into component parts. Rather than counting the number of previous encounters, we instead count the number of prior encounters of each item in the GLOP.

Table 1. Raw response data of student A

Student ID	Item ID	Timestamp	Previous trials (t)	Correct?
A	1045	11/7/2007 12:30:31AM	0	1
A	1649	11/7/2007 12:31:15 AM	1	0
A	1263	11/7/2007 12:43:40 AM	2	0
A	1022	11/7/2007 12:46:09 AM	3	1
A	1660	11/7/2007 12:48:20 AM	4	1

Table 2. Decomposed response data of student A

Student ID	Item ID	Trials (t)	Correct?	Prior encounters				
				Item 1022	Item 1045	Item 1263	Item 1649	Item 1660
A	1045	0	1	0	0	0	0	0
A	1649	1	0	0	1	0	0	0
A	1263	2	0	0	1	0	1	0
A	1022	3	1	0	1	1	1	0
A	1660	4	1	1	1	1	1	0

Given the data in Table 2, in order to determine the influence of each item on student learning we use a logistic regression model. We use a logistic model since our data are dichotomous. By fitting a logistic regression model, we seek to model the odds of giving a correct answer as

$$\frac{P(correct\_answer)}{P(wrong\_answer)} = \exp(A + \sum_{i \in D} B_i * t_i)$$

Equation 4. Logistic regression model for examining effect of practice on different items

Here A is the intercept of the regression model. The remainder part, $\sum_{i \in D} B_i * t_i$, represents a learning decomposition model that simultaneously estimate the impact of all items in a GLOP. *D* is a space of all items in a GLOP. The space is different for different GLOPs. Coefficient B_i represents the amount of learning caused by item *i* (or learning rate of item *i*). Generally, a positive estimate of B_i suggests students tend to perform better on later opportunities after they encountered item *i*; or students have learned from the instructional assistance provided on item *i* by the ASSISTment system that they worked on earlier by answering the scaffolding questions or by reading hint messages. t_i represents the number of prior encounters of item *i*. Note that, the t_i's account for all possible trials, and are thus equal to *t*. When the model parameters are estimated from data, the *B* parameters indicate the relative impact of different items on student math skill development.

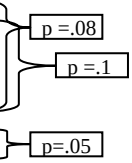
3. Results

We fit the model shown in Equation 4 to data of each GLOP separately in the statistics software package R (see www.r-project.org). To account for variance among students and items, student IDs and item IDs are also introduced as factors. By taking this step we account the fact that student responses are not independent of each other, and properly compute statistical reliability and standard errors. After the model is fitted, it outputs estimated coefficients for every item in each GLOP. Table 3 reports the estimated value of the *B* parameters of items in two GLOPs, GLOP 1 and GLOP 4, and also the standard error, order descending by the value of B in each GLOP. We can see that among all the items in GLOP 1, Item 1022 has the largest positive impact on student skill development: .464 in scale of *logit* (although the logit (log-odds) scale is not the most common one, it has the property that the item with the largest B coefficient will result in the largest learning gain, and an increase of .464 in logit scale is approximately equivalent to an increase of .116 in the probability of giving a correct

answer). Unfortunately, the model has determined that working on Item 1649 does not help student learning, indicated by a negative value of B although the value is not reliably lower than zero.

Table 3. Coefficients of logistic regression model for items in GLOP 1 and GLOP 4

GLOP ID	Item ID	B (Coefficient) (higher is better)	std. err
1	1022	0.464	0.257
1	1660	0.414	0.247
1	1045	0.127	0.254
1	1263	0.011	0.241
1	1649	-0.176	0.261
4	2264	0.707	0.225
4	2236	0.079	0.236
4	9086	-0.014	0.232
4	2239	-0.236	0.237
4	2274	-0.274	0.240



It looks like the items vary in their instructional effectiveness in helping student learn the skill(s) associated with a GLOP. But the standard errors are relatively large, too. Therefore, given any pair of item *i* and item *j*, we perform z-test to determine whether coefficients for item *i* and item *j* in the logistic regression model are statistically significantly different (p<.05). The z-score is calculated using $z = (B_i - B_j) / \sqrt{stderr_i^2 + stderr_j^2}$, assuming a normal distribution. As shown next to Table 3, we fail to reliably tell the difference among the top 4 items in GLOP 1 (# of items = 5; # of students = 531; # of data points = 2,256), but only find marginal difference (.05 < p <.1) between Item 1022 and item 1649, and between Item 1660 and Item 1649. However, we succeed to detect difference between the top 2 items in GLOP 4, Item 2264 and Item 2236 (p =.05) (# students = 652; #data points = 2,573).

Ideally, we would like to come up with a partial order of items in a GLOP that reflects which item caused the most learning, which one comes next, and which item is least effective. Figure 2(a) illustrates the partial order relationship among items in GLOP 1. In the diagram, an arrow connecting two items suggests there is a reliable difference between the instruction impact of the two items; and the arrow points to the less effective item. Additionally, the higher an item locates in the diagram, the larger the estimated value of B is. We follow the same process to acquire orders of items in all 39 GLOPs in our data set. The partial order diagram for items in GLOP 4 is shown in Figure 2(b).

One question is what types of items lead to more learning, easier or harder ones? Presumably, we hypothesized if a student learned to solve a hard item, he/she then should be able to do better on an easier item that requires similar skills. However, the converse is not necessarily true. In [4], we tested this hypothesis and our results suggested students learn as much from easier items as from harder items. Thus, those results suggested rejecting the hypothesis. In this paper we replicate the investigation: we directly estimate the amount of learning caused by each item (B parameter), and also since item ID was used as a factor in the logistic regression model, we get one parameter each item that reflects the easiness of the item. Interestingly, this time we are able to find a significant correlation between the two values (the amount of learning vs. easiness) of the 181 items (Spearman's rho = .192, p = .010), which suggests that, although the effect is not large, in general students did learn more by doing easier items. This result makes sense from the perspective of cognitive development. A hypothesis

proposed by a cognitive scientist, Kenneth Koedinger (personal communication), was that easy questions were easy because they only required the usage of a single skill, or fewer closely related skills. Hard questions were hard because they involved multiple (or extra) skills. The extra required skills were not intrinsic to the GLOP, and thus practice on them should not be helpful on other items in the GLOP. But easy items forced students to focus on the crucial part of the GLOP. Thus, practicing these items helped students to perform better later in the same GLOP.

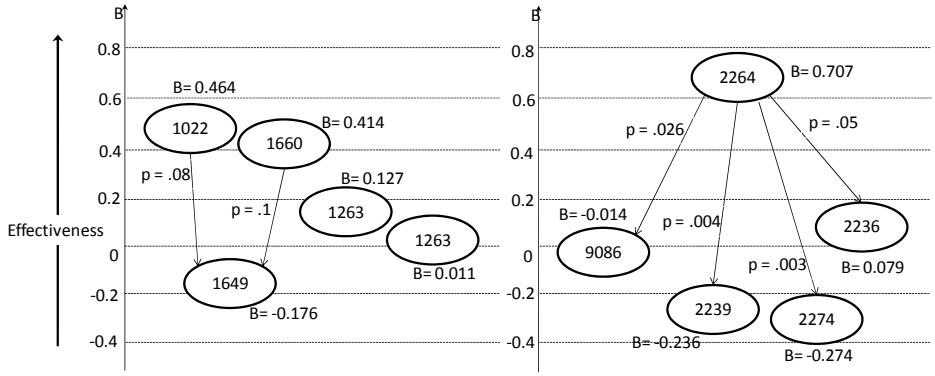


Figure 2. (a) Partial order relationship of items in GLOP 1; (b). Partial order relationship of items in GLOP 4.

4. Future work and conclusions

This paper explored the research question of measuring the instructional effectiveness of different problems, and associated tutoring, using the learning decomposition technique. One area of interest following this work is how to validate these numbers. One approach is to use human raters. While doing the learning decomposition analysis, we invite two human experts to review the items in 25 GLOPs that include less than 4 items. The human experts are asked to come up with a ranking of the items based on which items (including the scaffolding questions and the hint messages) they think will produce the most learning. It took the experts about 3 hours to finish ranking all 68 items. We examine the pair-wise correlation between rankings of the two human experts, and the ranking rendered by the learning decomposition approach. It turns out that the rankings of the experts correlate with each other significantly (Spearman's rho = .238, p = .049). Yet, we can only find reliable correlation between one rater's ranking and the learning decomposition ranking (Spearman's rho = .323, p = .007). Notice that this correlation is even stronger than that of the two human raters, which provides some evidence for the validity of our results. However, overall, the inter-rater reliability is relatively low, so we will need to try harder to obtain stronger evidence. Another approach is to make use of synthetic data by using a computer simulation study. The benefit of using a simulation is that we would know the ground truth about the effectiveness of each item. Furthermore, running a simulation study would allow us to better understand the power of the learning decomposition approach, such as how big the differences between the learning impacts of two items need to be for this approach to be able to detect it? How many students, data points the learning decomposition approach requires to tell a given difference between two items? Another open issue is related to the generalization of the approach. There are other factors that we have not yet explored such as the variant item effectiveness for students of different knowledge

levels. Applying the methodology in other domains, esp. ill-defined domains, possibly involves analyzing more other factors. The results may be affected by the organization of GLOPs as well. Currently, our GLOPs are manually constructed. It would be interesting to see how the items would be grouped by some automated method such as Q-matrix algorithm (Barnes, 2005) or LFA, and then how our results will be impacted by a new grouping method.

At this moment, we are able to tell which item is the best for student learning. But there is a caveat that since for most of the items we analyze, both the main question and the instructional content differs. Therefore, we do not know for sure whether an item caused more learning because the tutoring feedback is of high quality or just because the main question of the item is easier.

In this paper we showed how learning decomposition can be applied in the domain of mathematics to use observational data to estimate the effectiveness of different tutoring content. It provides an evidence that the learning decomposition is not domain specific and generally applicable to a variety of ITS that focus on different domains. Our analyses show in the ASSISTment system, some contents have more impact on student math skill development while some contents are not so effective. We suspect this result is not specific to ASSISTments, and other tutors have items that vary greatly in educational effectiveness. Our study demonstrates another, low cost, approach of evaluating ITS contents other than experimental study. Potentially, our approach and the results can be used to examine the quality of instructional contents in a learning system, and thus improve the overall learning impact.

Acknowledgements

This research was made possible by the U.S. Department of Education, Institute of Education Science (IES) grants #R305K03140 and #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] Barnes, T. (2005). Q-matrix Method: Mining Student Response Data for Knowledge. In Beck. J (Eds). *Educational Data Mining: Papers from the 2005 AAAI Workshop*.

[2] Beck, J.E. (2006). Using learning decomposition to analyze student fluency development. *ITS2006 Educational Data Mining Workshop* 2006. Jhongli, Taiwan.

[3] Beck, J.E. (2007). Does learner control affect learning? In *Proceedings of the 13th International Conference on Artificial Intelligence in Education.* pp. 135-142.

[4] Beck, J.E. (2008). How who should practice: Using learning decomposition to evaluate the efficacy of different types of practice for different types of students. *Proceedings of the 9th Intelligent Tutoring System Conference*. pp. 353-362.

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

[6] 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. Montréal 2008.

[7] 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*, pp. 555-562. Amsterdam: ISO Press.