Feng, M., Beck, J. and Heffernan, N. (2009). Using Learning Decomposition and Bootstrapping with Randomization to Compare the Impact of Different Educational Interventions on Learning. In Barnes & Desmarais (Eds.), Proceedings of the Second International Conference on Educational Data Mining. Cordoba, Spain, 2009.

Using Learning Decomposition and Bootstrapping with Randomization to Compare the Impact of Different Educational Interventions on Learning

Mingyu Feng, Joseph E. Beck and Neil T. Heffernan
{mfeng, josephbeck, nth}@wpi.edu
Computer Science Department, Worcester Polytechnic Institute

Abstract. A basic question of instructional interventions is how effective it is in promoting student learning. This paper presents a study to determine the relative efficacy of different instructional strategies by applying an educational data mining technique, learning decomposition. We use logistic regression to determine how much learning is caused by different methods of teaching the same skill, relative to each other. We compare our results with a previous study, which used classical analysis techniques and reported no main effect. Our results show that there is a marginal difference, suggesting giving students scaffolding questions is less effective at promoting student learning than providing them delayed feedback. Our study utilizes learning decomposition, an easier and quicker approach of evaluating the quality of ITS interventions than experimental studies. We also demonstrate the usage of computer-intensive approach, bootstrapping, for hypothesis testing in educational data mining area.

1 Introduction

The field of Intelligent Tutoring Systems (ITS) is often concerned with what type of educational intervention is more effective on promoting student learning. A handful of studies [e.g. 8, 11, 14, 15] have been conducted on comparing different variants of tutoring and feedback strategies, such as worked-out examples, tutored problem solving. One popular method of determining whether one type of instruction is more effective than the other is to run a randomized controlled study. Although the method is shown to be useful, 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 [2] introduced an approach called learning decomposition, an easy recipe to enable researchers to answer questions such as what type of practice is most effective for helping student to learn a skill. Instead of focusing on performance gain from pretest to posttest, learning decomposition leverages item-level data during a study and is concerned with how student performance changes while students are using the tutor. This approach is a modification of the learning curve analysis technique [12] that has been used in evaluating the efficacy of instructional contents. For instance, Koedinger and Mathan [9] compared learning outcomes associated with two types of feedback in the context of a spreadsheet tutor. Martin et al. [10] evaluated ITS using learning curves, and they also described the impact of changes in system's setup on the results of such analysis.

The ASSISTment system [13] is an online system that presents math problems to students who range from approximately 12 to 16 year olds in middle school or high school to solve. When a student has trouble solving a problem, the system usually

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 our database. In the ASSISTment system, when the authors create the instructional content, they may use different tutoring strategies. Razzaq et al. [14] reported a randomized controlled experiment that examined effects of the level of tutor-student interaction on helping students learn math skills. In this paper, we take a second look at the study and use a different approach to analyze the experiment: learning decomposition and bootstrapping with randomization test.

The goal of this paper includes 1) Comparing the relative impact of various educational interventions in the ASSISTment system by doing an item-level analysis. 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]. 3) There has been little prior use of bootstrapping with educational data [1]. We show how bootstrapping can be used with learning decomposition.

2 Methods

2.1 Experimental design

As mentioned in section 1, the experiment reported in [14] compared the efficacy of interventions with various levels of interactions. The experiment included three conditions: *scaffolding + hints*; *hints on demand*; *delayed feedback*. When a problem first appears on the screen, we refer to this as the "main question." If students answered the main question wrong, the "scaffolding + hints" (referred to as scaffold condition) condition forced them to do the scaffolding questions, which would ask them to complete each step required to solve a problem, and they must answer all scaffolding questions correctly to proceed. While in the "hints on demand" (referred to as hint condition) these students only received a message indicating their answer was wrong, and the hint messages, which would tell them the same information without expecting an answer to each step, would only appear when they press the Hint button on the screen. The third condition was a delayed feedback condition (referred to as delayed condition) where students got no immediate feedback from the tutor (even if they answered the question wrong) until they have finished all of the problems in the experiment, whereupon they received worked out solutions to all of the problems.

In this experiment students were presented an assignment with two pretest problems organized in one pretest section, four experiment problems in one experiment section, and four post-test problems in the posttest section that addressed the topic of interpreting linear equations, an 8^{th}-grade (approximately 13-year old) math skill. Two of the pretest problems were repeated in the post-test. Problems in the same section were shown in random order. Students were randomly assigned to the three conditions with equal probability. There were 366 eighth grade students from the Worcester Public Schools in Worcester, Massachusetts who participated in the experiment: 131 students were in honors level classes and 235 were in regular math classes. For the analysis in this paper, we exclude students who got both pretest problems correct (assuming they have mastered the skill), and those who did not finish all problems in the experiment. This leaves 300

students in our data set, with 101 in the delayed condition, 106 in the hint condition and 93 in the scaffold condition. We check to make sure students in all three conditions do not differ on their incoming knowledge. The mean and 95% confidence interval of average pretest and posttest scores for the three groups are listed in Table 1, and Table 2.

Table 1. Statistics of students' performance on pretest

Condition	Mean	Std. Err	95% confidence interval
Delayed	0.342	0.023	[0.297, 0.387]
Hint	0.354	0.022	[0.311, 0.397]
Scaffold	0.323	0.025	[0.274, 0.372]

Table 2. Statistics of students' performance on posttest

Condition	Mean	Std. Err	95% confidence interval
Delayed	0.381	0.025	[0.332, 0.430]
Hint	0.368	0.024	[0.321, 0.415]
Scaffold	0.341	0.025	[0.292, 0.390]

2.2 Approach

2.2.1 Introducing learning decomposition

Beck [2] introduced the idea of learning decomposition that 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 opportunities 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 $t = t_1 + t_2$[1]. 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

[1] Interestingly, t1 + t2 does not have to equal t, as shown in [16] and as we will show in this paper.

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*.

2.2.2 Decomposing learning opportunities

Now that we have described the model of learning decomposition, we want to "decompose" students' learning opportunities in our data set in order to fit such a model. Various metrics can be used as an outcome measurement of student performance. For instance, Beck [4] chose to model student's reading time since it is a continuous variable. 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 a problem as an outcome measure of their performance. A "1" in the data indicates the student got a problem correctly on the first attempt, and thus proceeded to the next problem without getting any instructional assistance, while a "0" means he failed on the first try and received certain type of tutoring from the system, depending on which condition the student has been assigned into.

When it comes to a nominal variable, in our case, 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.

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

Equation 3. Logistic models for learning decomposition

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

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

Where α, γ are the new representation of students' initial knowledge and their learning rates of a skill on the logistic scale. Now that we have determined our outcome variable and functional form of the model, all that remains is to decompose learning opportunities into components. We split student trials into four groups largely on the basis of experimental condition as below. Therefore, the number *n* is equal to 4 in this analysis.

- *hint_wrong_trial* (t_h) indicates the number of prior wrong trials that a student in the hint condition had encountered
- *scaffold_wrong_trial* (t_s) counts the number of prior wrong trials that a student in the scaffold condition had made before.
- *delayed_wrong_trial* (t_d) is similar to the other two variables but for students in the delayed condition. However, it is specially calculated such that the prior encounters will not increase until the student was presented the explanations for all the problems in order to address the fact that the learning actually happened at the moment when the explanations were shown. Note that by doing so we assume that simple exposure to the content does not cause learning. It is also worth pointing out that although the approach of learning decomposition itself does not require the administration of pretests and posttests, in this particular analysis, we do need the results of posttest to be able to detect the impact of explanations (in the delayed feedback condition) on student learning.

- *Others* (t_o). Because what we really care about is the relative effectiveness of the different tutoring interventions during the experiment, we did not differentiate students' practice trials on pretest, posttest and trials where they gave a correct answer to the experiment problem. Instead, all these trials are combined together into the group *others*. Actually, since the number trials on pretest and posttest are the same for all students, it is the correct trial on experiment problems that matter in this group.

For those readers who are familiar with ASSISTments vocabulary, it is also worth pointing out that although in the experiment there are three versions of the experiment problems with different associated interventions, one for each condition, we created one unified problem ID for all the three versions, since the main questions are the same.

Table 3. Decomposed response data of student A

Student ID	Section	Problem ID	Correct?	Previous trials (t)	Decomposed previous trials			
					Hint_ wrong_ trial (t_h)	Scaffold_ wrong_ trial (t_s)	Delayed_ wrong_ trial (t_d)	Others (t_o)
A	Pretest	Pre-1	1	0	0	0	0	0
A	Pretest	Pre-2	0	1	0	0	0	1
A	Exp	Exp2	0	2	0	0	0	2
A	Exp	Exp4	1	3	0	1	0	2
A	Exp	Exp1	0	4	0	1	0	3
A	Exp	Exp3	1	5	0	2	0	3
A	Posttest	Pre-1	1	6	0	2	0	4
A	Posttest	Post-2	0	7	0	2	0	5
A	Posttest	Pre-2	1	8	0	2	0	6
A	Posttest	Post-1	1	9	0	2	0	7

Table 3 shows a sequence of time-ordered trials of a student who was assigned in the *scaffolding* condition. The student finishes all three sections, fails on one of the pretest problems, but learns to solve the problem during the experiment as suggested by a correct answer to the same problem in the posttest. The right part of Table 3 shows the corresponding data after the trials are decomposed into component parts. Since the student is in the scaffolding condition, all values in the hint_wrong_trial and delayed_wrong_trial are zero. He solves the first encountered experiment problem wrong (row 3), which cause an increase on the value of scaffold_wrong_trial from zero to 1 (row 4). Again, he gets the third experiment problem wrong (row 5), and then the value of scaffold_wrong_trial increases from 1 to 2 in row 6. The value of trial for others just increases by one whenever a pretest problem, a posttest problem or a correct trial was encountered. For instance, the student answers the second encountered experiment problem correct (row 4), and thus the value of others increased by 1 (row 5). Limited by space, we only demonstrate the decomposition process for a student in the scaffold condition; the process for the hint condition would be identical. For the delayed feedback condition, since the student would not see the feedback until after all of the experimental trials, it is necessary to model that differently. In the delayed condition, the number of

delayed-wrong trials would stay as zero until it jumps to be 2 in row 7, since the student would have seen the two explanations after finishing the experimental questions. This problem requires a rather novel use of learning decomposition, and some care in accounting for when the learning opportunities actually occur.

2.3 Results

We fit the model shown in Equation 3 to the decomposed data in the statistics software package R (see www.r-project.org). To account for variance among students and items, student IDs and unified problem 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. Also, by fitting our model in this manner we do not suffer the scaling problems mentioned by [10] since all three conditions have the same intercept (i.e. *A* parameter). After the model is fitted, it outputs estimated coefficients for every condition, as shown in Table 4. The result suggests that the delayed feedback, estimated coefficient being 0.720, is more effective at helping student learn the skill than the other two conditions, esp. the scaffolding condition for which the coefficient estimate is 0.633. In prior work with this experiment [14], the authors reported that they did not find any main effect. It is possible to use the estimated coefficients (B) and standard errors in Table 4 to perform a statistical z-test, as we did in [7]. However, there is a bit of serendipity: the first author was conducting some exploratory analyses using resampling to see how stable the parameter estimates really were. It appeared that there was little overlap between the estimates for the scaffold and delayed conditions. Therefore, we decide to test this approach formally using bootstrapping [5] and randomization tests [6].

Table 4. Coefficients of logistic learning decomposition model

Coefficients	Estimate (B)	Std. Error	z value	Pr(>|z|)
Others	-0.235	0.034	-6.816	9.33e-12 ***
Hint_wrong_trial	0.706	0.091	7.760	8.52e-15 ***
Scaffold_wrong_trial	0.633	0.103	6.175	6.62e-10 ***
Delayed_wrong_trial	0.720	0.054	13.224	< 2e-16 ***

Bootstrapping is a modern, computer-intensive, general purpose approach to statistical inference, falling within a broader class of resampling methods [5]. It involves the construction of a number of resamples of the observed dataset by random sampling with replacement from the original data set; and each resample is independent (conditioned on the original sample) and identically distributed. Although bootstrapping was developed as techniques for parameter estimation, it can be used for hypothesis testing as well. In general, first we make a null hypothesis. Then we draw repeated samples from the original data set under the condition that the null hypothesis is true, and then we reject the null hypothesis if the statistic computed from the observed dataset is unlikely under the null hypothesis, or otherwise retain the null. In this particular analysis, the hypothesis we would like to test is "The delayed feedback strategy promotes learning more or less effectively than the scaffold (or hint) strategy." Correspondingly, the null hypothesis

would be "There is no difference on learning promotion between the delayed and scaffolding strategies."

Specifically, we follow the following steps to test our hypothesis.

Step 1: Decide on a metric to measure the relative effectiveness between delayed feedback and scaffolding strategies. For this example, we choose the difference between the estimated coefficients of Delayed_wrong_trial and Scaffold_wrong_trial.

Step 2: Calculate the metric on the original data. The results in Table 3 provides B(Delayed_wrong_trial) – B(Scaffold_wrong_trial), equal to .087.

Step 3: Bootstrap the original data with randomization to construct samples where the null hypothesis is true

Repeat *N* times {

Repeat M times (*M* = the number of students in our original data set) {

Sample data of one student (with replacement) from the original data;

Randomly allocate the student into one of the three conditions: delayed, scaffold, or hint by changing the "Condition" label of each data point

Re-compute the number of prior trials for the student according to the newly assigned condition;

}

Train logistic learning decomposition model on the re-sampled data, and record B(Delayed_wrong_trial) – B(Scaffold_wrong_trial) ;

}

In our case, we pick the repeated times *N* to be 500, and *M* is 300 as there are 300 students in our data set.

Step 4: Check how likely our original result is under the null hypothesis, and reject or retain the null hypothesis. After the bootstrapping process, we obtain a list of difference between Delayed_wrong_trial and Scaffold_wrong_trial, totally 501 cases including our original result. Then we rank the list descending, and found that the original result was at the 95 percentile, the 25^{th} in the ranking order, which suggests that the probability of the original result has a probability of less than 5%. Although it is tempting to think we have $p < 0.05$, this methodology is actually conducting a one-tailed test. Thus, the two-tailed value is $p = 0.1$. Therefore, we have a marginally reliable result that *delayed feedback* is better than *scaffold + hint*, and giving students delayed feedback seems causing more learning than requesting them to finish a series of scaffolding questions.

To complete the story, we repeat the same process compare the other two pairs: delayed vs. hint conditions, and scaffold vs. hint conditions, but find that they are comparable to each other at helping students learning the math skill in ASSISTments.

3 Conclusion

This paper explored the research question of measuring the instructional effectiveness of different tutoring interventions, using the learning decomposition technique. We found that presenting students with delayed feedback works better than breaking problems into scaffolding questions. We also used bootstrapping with randomization to test the statistical reliability of the finding.

Typically, there are two reasons for the usage of learning decomposition (or any educational data mining technique). The first is repurposing a previous experiment's data to answer a new question. The second is using EDM techniques to "zoom in" and detecting subtle effects that previous approaches failed to report. Previous works on learning decomposition [3, 4, 16] have been focusing on the first reason, while in this paper we focus on the second reason through an item level analysis and bootstrapping.

One open question is why bootstrapping plus randomization gives different results than the parametric method of using estimated coefficients and standard errors to derive an analytic p-value. We did a z-test using estimated coefficients and standard errors given in Table 4 and obtained $p = .4$. Typically computationally intensive techniques are *less* powerful than parametric ones, unless one or more of the parametric tests' assumptions have been violated. We are not sure where the problem lies, but suggest caution in interpreting standard error terms from logistic regression models using learning decomposition.

The contribution of the paper lies in three aspects. First, we found that there is a main effect in a randomized controlled study that delayed feedback tutoring strategy is more effective than giving students scaffolding questions in ASSISTments. While previous analysis using ANOVA failed to detect such an effect, we were able to do so by conducting an item level analysis using EDM techniques. Second, we showed how learning decomposition can be applied in the domain of mathematics to use observational data to estimate the effectiveness of different tutoring strategies. It provides evidence that the learning decomposition is not domain specific. This simple, low cost approach is generally applicable to a variety of ITS that focus on different domains for identifying variances in educational effectiveness of interventions. Also, our use of learning decomposition is novel in that we are careful to consider when various aspects of an intervention occur, and do not give credit for a learning opportunity that has not yet happened (the delayed-wrong condition). Third, the process described in this paper serves as a demonstration of how bootstrapping approach and randomization tests can be employed in the educational data mining field.

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] Beal, C. and Cohen, P. (2005). Comparing apples and oranges: Computational methods for evaluating student and group learning histories in intelligent tutoring systems. 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.

[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] Davison, A. C.; Hinkley, D. (1997). *Bootstrap Methods and their Applications.* Cambridge: Cambridge Series in Statistical and Probabilistic Mathematics.

[6] Edgington E.S. (1995). Randomization tests, 3rd ed. New York: Marcel-Dekker, 1995.

[7] Feng, M., Heffernan, N.T., Beck, J. (In press). 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 (AIED-2009). Brighton, UK, 2009.

[8] Kim, R, Weitz, R., Heffernan, N. & Krach, N. (accepted). Tutored Problem Solving vs. "Pure": Worked Examples. Accepted by Cognitive Science Society Annual 2009 Conference.

[9] Koedinger, K.R. and Mathan, S. (2004). Distinguishing qualitatively different kinds of learning using log files and learning curves. in ITS 2004 Log Analysis Workshop. 2004. Maceio, Brazil. p. 39-46.

[10] Martin, B. Koedinger, K., Mitrovic, A. and Mathan, S. (2005). On Using Learning Curves to Evaluate ITS. Proceedings of the 12th international conference on Artificial Intelligence in Education, AIED2005 Amsterdam, pp. 419-426

[11] Mathan, S. & Koedinger, K. R. (2003). Recasting the Feedback Debate: Benefits of Tutoring Error Detection and Correction Skills. In Hoppe, Verdejo & Kay (Eds.), Artificial Intelligence in Education:Shaping the Future of Learning through Intelligent Technologies. Proceedings of AI-ED 2003 (pp. 39-46). Amsterdam, IOS Press.

[12] Newell, A. & Rosenbloom, P.S. (1981). Mechanisms of skill acquisition and the law of practice. In J.R. Anderson (Ed.). Cognitive skills and their acquisition. Lawrence Erlbaum Associates: Hillsdale, NJ. p.1-56.

[13] 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.

[14] Razzaq, L., Heffernan, N. T., Lindeman, R. W. (2007). What Level of Tutor Interaction is Best? In Luckin & Koedinger (Eds.) Proceedings of the 13th Conference on Artificial Intelligence in Education. IOS Press. pp 222-229.

[15] Sweller, J. & Cooper, G. A. (1985). The Use of Worked Examples As a Substitute For Problem Solving In Learning Algebra. Cognition and Instruction, 2, 59–89.

[16] Zhang, X., Mostow, J. & Beck, J.E. (2007). All in the (word) family: Using learning decomposition to estimate transfer between skills in a Reading Tutor that listens. Educational Data Mining Workshop at the 13th International Conference on Artificial Intelligence in Education.