How who should practice: Using learning decomposition to evaluate the efficacy of different types of practice for different types of students

Joseph E. Beck[1] and Jack Mostow[2]

[1]Computer Science Department, Worcester Polytechnic Institute
[2]Robotics Institute, Carnegie Mellon University
joseph.beck@EducationalDataMining.org, mostow@cs.cmu.edu

Abstract. A basic question of instruction is how much students will actually learn from it. This paper presents an approach called *learning decomposition*, which determines the relative efficacy of different types of learning opportunities. This approach is a generalization of learning curve analysis, and uses non-linear regression to determine how to weight different types of practice opportunities relative to each other. We analyze 346 students reading 6.9 million words and show that different types of practice differ reliably in how efficiently students acquire the skill of reading words quickly and accurately. Specifically, massed practice is generally not effective for helping students learn words, and rereading the same stories is not as effective as reading a variety of stories. However, we were able to analyze data for individual student's learning and use bottom-up processing to detect small subgroups of students who did benefit from rereading (11 students) and from massed practice (5 students). The existence of these has two implications: 1) one size fits all instruction is adequate for perhaps 95% of the student population using computer tutors, but as a community we can do better and 2) the ITS community is well poised to study what type of instruction is optimal for the individual.

Key words: learning decomposition, educational data mining, learning curves, bottom-up processing

1 Introduction

The goal of this paper is to investigate how different types of practice affect a student's progress in learning a skill. Specifically, we utilize an approach, learning decomposition [1], as a means of leveraging fine-grained interaction data collected by computer tutors and present a case study of applying the technique to the domain of reading. The goal is twofold: 1) Be able to make claims that are interesting to domain researchers, and 2) Develop a technique for analyzing tutor log data that applies to other domains and tutors. The first goal should not be underestimated; if we make discoveries about how students learn a domain that remain limited to those students using computer tutors that would be an unfortunate result. Only a small minority of students use computer tutors, so if we wish our research to have broad impact then finding a means of explaining our results to those outside of the ITS community is essential. To address these issues we present an approach that uses learning curves to measure the relative impact of various types of learning events.

The two most common types of learning curves are exponential and power curves. In this paper we discuss exponential curves as they have been shown to more accurately model individual observations [2] and are simpler analytically. However, the approach we present can be trivially adapted to work with power curves. The standard form of the exponential learning curve can be seen in Equation 1a. The free parameter *A* represents how well students perform on their first trial performing the skill; *e* is the numerical constant (2.718), the free parameter *b* represents how quickly students learn the skill, and *t* is the number of practice trials the learner has had at this skill. This model can be solved with any non-linear regression package (we use SPSS 11.0).

$$performance = A * e^{-b*t} \qquad performance = A * e^{-b*(\mathrm{B}*t_{1/}+t_2)}$$

Equation 1. (a) Exponential model of practice, (b) Learning decomposition model of practice

By simply using one parameter, *t*, to represent the number of prior trials, learning curves assume that all types of practice are equally valuable. But what if all types of practice are not equally valuable? For example, we could believe that the subject will learn better the first time he practices the skill that day, and rather than simply lumping all of the learning opportunities together as *t*, we can create two new variables t_1 and t_2. The variable t_1 represents the number of learning opportunities where it was the first time the learner practiced the skill that day; t_2 represents the number of practice opportunities where the learner has already practiced the skill that day. This method of factoring learning opportunities into various types of practice does not change the amount of prior practice to student has had; $t= t_1+t_2$ since learning opportunities are either the first one of the day or are not.

The basic idea of *learning decomposition* is to find how to weight two types of learning opportunities to construct a best fitting learning curve. Equation 1b shows a learning curve model designed to find how to weight the two types of practice. Similar to standard learning curves, we estimate the *A* and *b* parameters. However, we also estimate a new parameter, *B,* that represents the relative impact of the first learning opportunity of a day relative to learning opportunities occurring later in the same day. Note that t_2 does not receive a weight of its own, as it is assumed to be worth 1.0 learning opportunities. That t_2 has this implicit weight does not affect the conclusions we draw from the model as our goal is only to estimate relative efficacy the two types of practice.

The parameter *B* is very interpretable: it is how many trials that learning opportunities of type t_1 are worth relative to those of type t_2. If $B>1$ then learning opportunities of type t_1 are better for learning than those of type t_2. If $B<1$ then the opposite is true, and if $B=1$ then neither type of learning opportunity is preferable. Although the example presented is about first practice opportunity of the day vs. later ones, it is possible to split the data in any way that may be interesting. We could split learning opportunities by those that occur on Monday, Wednesday, or Friday vs. those that occur on Tuesday and Thursday. For this decomposition we would hopefully get $B\approx1$, as we have no reason to believe the day of the week matters for learning. Thus, the technique of learning decomposition is broadly applicable.

The remainder of this paper explores applying learning decomposition to answer some questions about how children acquire reading skills. However, the approach itself is applicable to a variety of learning tasks and possible ways to decompose learning.

2 A case study: applying learning decomposition to the domain of reading

The goal of this case study is to show how to apply learning decomposition to an actual data set and draw scientifically useful conclusions. We are trying to better understand how students learn to read by analyzing performance data about individual words recorded by the Reading Tutor [3] during the 2003-2004 school year. Rather than have explicit experimental and control groups, our approach is to examine how student progress in reading words quickly and accurately varies based on which type of practice he has had at the word. These data include 346 students from the Pittsburgh area attempting to read 6.9 million words. The student readings were scored by an automated speech recognizer (ASR). The ASR is far from perfect, and for that year detected approximately 25% of student misreadings and scored 4% of student correct readings as incorrect [4]. The ASR also records how long students took to read a word. Our general logging mechanism also records when students request help. Furthermore, all entries are time stamped so we know the relative temporal relations between events. Students used the tutor from September 2003 through May 2004 with a median usage of 5.9 hours.

We now show how we integrate student help, speed, and correctness into a single outcome measure of learning; explain what we believe constitutes a learning opportunity for a word; and finally show how we decomposed the learning opportunities into their component parts.

2.1 Creating an outcome to measure learning

There are a variety of approaches for representing student performance at reading fluency. We choose to model the student's reading time since it is a continuous variable and best able to track student progress; help requests and accuracy are binary and so cannot improve smoothly. Although it is possible to aggregate help requests and accuracy to create a continuous learning curve, we did not perform such aggregations as one goal of the research is to use individual observations (rather than aggregate descriptions) to construct our learning curves. It is a known potential pitfall that aggregate learning curves may not describe the learning trajectory of actual individual learners [5]. Therefore, fitting individual data points can produce a more authentic model of student learning

Although reading time is continuous, it is misleading to use it as an outcome and ignore accuracy and help requests. Our approach was to use the student's reading time as an outcome measure. However, when the student either asked for help or skipped the word, or the word was scored as incorrect by the ASR, then that word was assigned a reading time of 3.0 seconds. Also, words whose reading time was greater than 3.0 seconds were capped at 3.0 seconds. The penalty of 3.0 seconds is on the high end of reading times as only 0.1% of time exceeded this threshold, but not overly so as to be an unfair penalty.

2.2 What constitutes a learning opportunity?

Given that help can cause a short-term boost in student performance, a natural question is what other types of events can cause a similar effect? If our goal is to measure student *learning*, we should try to exclude such data from our learning curve construction. One example of such short-term scaffolding is that if a student reads a word and then shortly thereafter reads that same word again, we should be skeptical that the second reading really demonstrates the student's knowledge of the word (as opposed to just retrieving it from short term memory). Therefore, to model student reading development we only consider as an outcome variable his first encounter with a word on a particular day.

However, we do count subsequent encounters later in the day as opportunities to *learn* the word. Table 1 illustrates our approach. For the first encounter, the student requests help and then reads the word quickly. Since the student requested help, the outcome is set to 3.0 seconds. For the next learning opportunity, since it is the same day, that reading does not count as an outcome. Similarly, the next learning opportunity's performance is also ignored. However, note that column labeled "Overall" in the prior encounters field, which tracks the student's experience with this word, has been incremented to account for these two exposures.

2.3 Learning components of fluency development

For reading, what types of practice are likely to be more (or less) effective for students' fluency development? There are many possible ways to think about what are ways of factoring apart learning opportunities at learning to read a word. We start with a known psychology principle: distributed practice is generally superior to massed practice for long term retention [6]. This general rule suggests a decomposition: we consider a learning opportunity as *distributed* practice if the student has not encountered the word in the preceding 16 hours. *Massed* practice would be times when the student encountered the word in the prior 16 hours (effectively during the same day). Table 1 shows how we decompose the prior encounters based on massed vs. distributed practice.

The other type of learning decomposition we performed was to examine whether reading the same story multiple times provides the same benefits as students reading different stories. This debate of wide- vs. re-reading has been ongoing in the reading community. We therefore decompose prior practice into learning opportunities where this student encounters this word while reading *new* material vs. *rereading* old stories. Since students can memorize a particular story, we only permit as an outcome variable the first time a student reads a particular story. However, analogous to how we handled massed practice learning opportunities, repeated readings of the same story count as learning opportunities for learning (in particular, the variable for *rereading* would be increased in each case).

To summarize, we only count the first opportunity each day as an outcome variable, and only if the student has not read this story in the past. However, we count all exposures to words as possible learning opportunities. To estimate the learning caused by different types of learning opportunities, we created four types:

1. RM represents **r**ereading-**m**assed learning opportunities. I.e. cases where the student has already read the story in the past and is seeing the word a second (or greater) time today.
2. RD represents **r**ereading-**d**istributed learning opportunities. I.e. cases where the student is rereading the story but has not seen the word earlier today.
3. NM represents **n**ew-**m**assed learning opportunities; cases where students are reading a story for the first time and have read the word previously today.
4. ND represents **n**ew-**d**istributed learning opportunities; students have not seen this story before and have not read the word previously today.

Table 1 . Decomposing prior learning opportunities as massed and distributed practice

Day	Helped?	Reading time (seconds)	Prior encounters			Outcome (seconds)
			Overall	Distributed	Massed	
1	Yes	0.5	0	0	0	3.0
1	Yes	1.5	1	1	0	-
1	No	1.3	2	1	1	-
2	No	3.8	3	1	2	3.0
3	No	1.7	4	2	2	1.7
3	No	1.2	5	2	3	-

Our model of reading development is shown in Equation 2. The term *A*, represents first trial performance, and *b* is the rate of learning. For brevity, the model presented omits some terms such one to control for word length (since reading time is correlated with word length) and another to control for amount of prior assistance. The remainder of the model is a learning decomposition model to simultaneously estimate the impact of massed- vs. distributed-practice and wide- vs. re-reading. Note that RM, RD, NM, and ND account for all possible trials, and are thus equal to *t*. The goal is to find best-fitting values of the *r* and *m* parameters to find the relative impact of rereading and massed practice, respectively, on student reading development.

$$readingTime = A * e^{-b*(r*m*RM + r*RD + m*NM + ND)}$$

Equation 2. Simplified model for examining effect of practice schedule and type of reading

Again, there are many possible ways to decompose learning opportunities. We chose two that were motivated by existing theories of learning and a current debate in the reading literature.

3 Results

To train the model, we had 959,455 learning opportunities (i.e. a student's first attempt at reading a word on a particular day) and a total of 6.9 million words read. For each of the 346 students in our data set who read at least 20 words in the Reading Tutor during the 2003-2004 school year, we fit the model shown in Equation 2 to each student's data (i.e. we had 346 estimates of each parameter—one for each student). Table 2 shows the median parameter estimates for the effects of rereading and massed practice. The column labeled "overall" contains the median for the entire population. The next three columns are estimates by the bottom third (reading pretest score below a beginning first grader), middle third, and upper third (reading pretest above that of a second grader at mid year) of the student population.

Main effects. We found that rereading had a coefficient of 0.49 for the entire student population. In other words, rereading a story only results in 49% as much learning as reading a story for the first time. So if a student reread a story twice that would result in as much learning as reading a new story for the first time ($2 * 0.49 = 0.98 \approx 1.0$). Therefore, our results suggest that students learn to read words better when they read a wide selection of stories rather than read the same story multiple times, and this trend holds for all of the levels of student proficiency that we examined. For massed practice, the picture was even more bleak. Overall, students learned very little from multiple opportunities to practice a word on the same day, with high proficiency students deriving almost no benefit at all from the exposure. Seeing the word again is almost a complete waste of time for these students.

Table 2. Median parameter estimates for learning decomposition model

	Overall (N=346)	Low proficiency (N=118)	Medium proficiency (N=106)	High proficiency (N=122)
Reread (r)	0.49	0.71	0.42	0.33
Massed (m)	0.19	0.36	0.28	0.02

We report median rather than mean scores due to difficulties with accounting for outliers. For example, student rereading parameters range from -1754 to 14211. Clearly those extreme value are outliers and would bias the mean. However, it is difficult to determine exactly what constitutes an outlier. For example, 3.0 is an unlikely level of benefit from rereading, should we disallow that? How about 2.0? Rather than inventing an arbitrary cutoff, we instead use the median and treat improbably high values as a vote that the true value of the parameter is higher than 1.0.

For the rereading parameter, 95 students had an *r* parameter that was reliably less than 1.0, while only 7 had a parameter estimate that was reliably greater than 1.0. Using a sign test gives $p \approx 10^{-17}$, thus the majority of students have less effective learning as a result of rereading. For massed practice, 177 students had an *m* parameter that was reliably less than 1.0, with only 6 students having an *m* parameter reliably greater than 1.0 ($p \approx 10^{-35}$). Thus, the majority of students benefit less from massed practice. This result for massed practice is not novel, as there has been ample research in psychology investigating spaced practice effects. However, it serves as a sanity check on our results: if this aspect our investigation disconfirmed 120 years of psychology research we should be hesitant about accepting our other results.

Which students benefit from rereading and massed practice? One benefit of estimating a per-student parameter for the effect of rereading and massed practice is that it enables fine-grained detection of student subgroups who benefit. Although Table 2 shows that low proficiency students do not benefit from rereading or from massed practice, there is a definite trend with weaker readers receiving relatively more benefit than more proficient readers. Perhaps there are subgroups of low proficiency students who are benefiting, but we cannot detect them since they are averaged in with a larger group?

Our approach is to treat the problem as one of classification via logistic regression. For our dependent variable, if the student's *r* parameter exceeded 1.1, we treated that student as benefiting from rereading; if it was below 0.9 the student is considered to benefit from wide reading. Values between 0.9 and 1.1 were not considered conclusive evidence either way, and those students were not used for the classification process. By compressing the student's *r* parameter into a single bit of information, we greatly reduce the inaccuracy of poorly estimated per-student parameters, and instead focus on the simpler task of finding commonalities between students whose *r* parameter indicates they would benefit from rereading. The act of creating a classifier results in a smoothing of the noisy data, and any reliable predictors are indicative of a subgroup that benefits. (We performed the identical procedure for the *m* parameter to determine if students would benefit from massed practice.) For the independent variables, we used the student's gender, grade, learning support status (yes, no, or not known), and words read correctly per minute in a paper-based fluency pretest. We chose these variables as they are easily available to classroom teachers or other classroom policy makers.

Of the 235 students for whom we had complete demographic and testing data, the rereading classifier found a subgroup of 11 students for whom it thought rereading would benefit. The student's learning support status (p=0.00003) and fluency (p=0.04) were both reliable predictors in the model. Only 24 students were noted as having learning support, of those the model felt 4 would benefit from rereading (as opposed to predicting benefit for 7 of 122 for students who were known to not be receiving learning support). The students who would benefit from rereading also had a sharply lower fluency: 44 words per minute as opposed to 56 words per minute for those who would not benefit. There were similar trends for the students who would benefit from massed practice. The classifier found a subgroup of just 5 students who would benefit from massed practice, with the only reliable differences being the student's grade, with a mean of 2.15 from those who do not benefit vs. 4.6 for those who do (p=0.013) and fluency, with a mean reading rate of 55 wpm for those who do not benefit vs. 47 for those who do (p=0.04). Although not statistically reliable, it is interesting to note that all 5 of these students were categorized as receiving learning support. As a general trend, those who benefited from massed practice and from repeated reading were older, less proficient readers who were tagged as requiring learning support (but who were still able to operate the Reading Tutor software effectively). Gender was not a reliable predictor in either model.

4 Limitations and future work

This paper explored two learning decompositions, but there is a large space of possible ways of splitting the data. Automating the construction and evaluation of possible decompositions is a fruitful avenue of research. One crucial problem that must be overcome is finding some method for seeding this search space with expert knowledge. Expert knowledge both reduces the size of the space and biases the results so that it better fits with what is known. The output of educational data mining can certainly improve computer tutors, but if that is all it does that would be unfortunate. As a field, we have several novel methodological hammers that are unavailable to domain researchers who aren't using these approaches. We must find ways to transfer what we learn to the broader research community. By hand selecting two major hypotheses of learning and reading, we manually biased the search to have output that will (hopefully) have high impact. Can we have automated search that produces results that are equally shareable?

Given that we have a set of high-level decompositions to perform, we still need to operationalize them. For our analysis, massed practice was equated with seeing a word a 2nd time on the same day. However, there are many ways to view "massed practice." Perhaps it means more than 5 practice attempts within 3 minutes? Maybe the first 2 attempts on the same day aren't massed but subsequent ones are? There is a wide space of possible ways to instantiate the theory. How do we know which is best? Searching across instantiations of distributed practice is itself a large search space. Can we afford to search both the space of good decompositions and the ways to instantiate the decompositions?

One hybrid approach is to accept that the high-level decompositions should come from humans who will (hopefully) use existing theory (e.g. mass vs. distributed practice) to generate decompositions, and spend computational resources on exploring that search space to find a good way to operationalize the decompositions. Such an approach would seem to draw on the strengths of humans and computers. Science is a social process and we need results that fit with existing theory (perhaps to disconfirm it) for domain researchers to take it seriously [7]. Spending time searching for new theory may not be productive if no one understands the results. However, a specific instantiation of an expert generated decomposition should be understandable. For example, if instead our model of reading used the "5 practice attempts within 3 minutes" definition of massed practice, the results wouldn't be any harder or easier for others to understand. Such a hybrid approach seems a promising route forward.

We would like to make statements of the form "Rereading is less helpful for developing reading proficiency than wide reading." Unfortunately, our data were not gathered from randomized trials, but rather are observational in nature. However, by estimating practice effects per-student, we are able to make stronger claims than might be expected. For example if Chris has a rereading parameter of 0.8, that means when rereading he learns 80% as much *as Chris would normally learn*. By having the student act as his own control, we remove other constant factors that could act as confounds for our result. For example, if less proficient students are the ones doing more rereading, our result would not be biased by the fact. In general, by building per-student models we control for all *trait* information about the learner such as language aptitude, memory capacity, interest in learning, etc. However, we do not control for transitory *state* information. For example, if students only reread when they are tired after a poor night's sleep (and are presumably less able to learn), that is a possible confound for our results.

Finally, our results were based on a set of assumptions about which words (first attempt at reading a word for the day) and which stories (first time the student read the story) were most indicative of learning. These assumptions represent a best-effort on the part of the authors. However, a sensitivity analysis of specific the results are to our assumptions would be helpful. We performed one such analysis and found that, qualitatively, the results were similar whether we looked at the first time a word was read in a day or the first time in a week.

5 Contributions and conclusions

This paper makes several contributions both methodologically and scientifically. From a methodology standpoint, learning decomposition extends learning curve analysis to enable estimation of the impact of various types of learning opportunities. The learning decomposition approach is broadly applicable to a wide variety of learning phenomena and is not specific to reading. Furthermore, it is fast computationally and can be applied via a variety of off the shelf software packages. Finally, the output is easy to interpret and share with other researchers.

From a scientific standpoint, this work may resolve a debate in the reading community (is wide- or re-reading better and for whom). If the goal of our work is to have impact beyond our own tutors, finding modeling approaches that are easily understandable by other communities must be a priority. Our results on what type of reading practice helps the most have not yet been fully disseminated to the reading community so it is premature to comment on whether this approach will result in conclusions understandable to domain researchers. However, an earlier version of this work was presented at the 2005 and 2006 Scientific Studies of Reading Conferences and was well received.

The closest related research is learning factors analysis (LFA) (e.g. [8]) Both LFA and learning decomposition are concerned with better understanding student learning. LFA focuses on modifying the domain representation by adding, removing, or combining skills to create better fitting learning curves where the impact of various types of practice is assumed to be constant. Learning decomposition focuses on determining the impact of various types of practice, and assumes the domain representation is constant. A unified framework that simultaneously allows both the skills and impact of practice to vary would be desirable.

In conclusion, we have shown how learning decomposition can be applied to use observational data to estimate the effectiveness of different types of learning opportunities. Our analyses show that in the domain of reading, different types of practice are more effective than others. Specifically, reading new stories and spacing exposure to words is good for long-term learning. Although our case-study was in the domain of reading, there is nothing domain specific about the learning decomposition approach, and it is broadly applicable to a variety of ITS. Furthermore, the massed practice result has implications both for sequencing instruction and for student modeling in an ITS. If a student model is not discounting learning opportunities that are temporally near each other, it is probably overestimating student knowledge. Finally, our bottom-up approach of using classification to detect important student subgroups, rather than relying on *a priori* beliefs about what disaggregation are important, was able to detect a subpopulation of students who benefits from an otherwise less effective treatment. If fully realized, this capability to truly adapt an ITS's instruction to meet the needs of learners would be a large step forward an ITS.

Acknowledgements

This work was supported by the National Science Foundation, ITR/IERI Grant No. REC-0326153. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the National Science Foundation

References

1. Beck, J.E. *Does learner control affect learning?* in *Proceedings of the 13th International Conference on Artificial Intelligence in Education*. 2007. p. 135-142 Los Angeles, CA.
2. Heathcote, A., S. Brown, and D.J.K. Mewhort, *The Power Law Repealed:The Case for an Exponential Law of Practice.* Psychonomics Bulletin Review, 2000: p. 185-207.
3. Mostow, J. and G. Aist, *Evaluating tutors that listen: An overview of Project LISTEN*, in *Smart Machines in Education*, P. Feltovich, Editor. 2001, MIT/AAAI Press: Menlo Park, CA. p. 169-234.
4. Banerjee, S., J. Beck, and J. Mostow. *Evaluating the Effect of Predicting Oral Reading Miscues*. in *Proc. 8th European Conference on Speech Communication and Technology (Eurospeech 2003)*. 2003. p. 3165-3168 Geneva, Switzerland.
5. Brown, S. and A. Heathcote, *Averaging learning curves across and within participants.* Behaviour Research Methods, Instruments & Computers, 2003. **31**: p. 11-21.
6. Ebbinghaus, H., *Memory: A Contribution to Experimental Psychology*. 1885, New York: Teachers College, Columbia University.
7. Pazzani, M.J., J.S. Mani, and W.R. Shankle, *Acceptance by medical experts of rules generated by machine learning.* Methods of Information in Medicine, 2001. **40**(5): p. 380-385.
8. Cen, H., K. Koedinger, and B. Junker. *Learning Factors Analysis - A General Method for Cognitive Model Evaluation and Improvement.* in *Intelligent Tutoring Systems*. 2006. p. 164-175 Jhongli, Taiwan: Springer.