Using learning decomposition to analyze student fluency development

Joseph E. Beck
Machine Learning Department
Carnegie Mellon University
Pittsburgh, PA 15213
joseph.beck@gmail.com

Abstract. This paper introduces an approach called *learning decomposition* to analyze what types of practice are most effective for helping students learn a skill. The 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 are able to show that different types of practice differ reliably in how quickly students acquire the skill of reading words quickly and accurately. Specifically, massed practice is generally not effective for helping students learn words, but may be acceptable for less proficient readers. Rereading the same story is generally not as effective as reading a variety of stories, but might be beneficial for more proficient readers.

Keywords: learning curves, effect of practice, learning decomposition, reading

INTRODUCTION TO LEARNING CURVES AND LEARNING DECOMPOSITION

The goal of this paper is to investigate how different types of practice affect a student's progress in learning a skill. Specifically, we present a new approach, learning decomposition, as a means to leverage 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 should generalize 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 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 AIED/ITS communities 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.

For tracking student performance, learning curves (e.g. (Ebbinghaus, 1885; Newell & Simon, 1972)) are a well established technique that provide a means for linking practice at a skill to improvements in performance. Recently in the AIED/ITS community, there has been work on using learning curves to try to determine how instruction impacts which strategy students will employ in problem solving (Koedinger & Mathan, 2004), and on distinguishing the impact of various versions of a computer tutor (Martin, Koedinger, Mitrovic, & Mathan, 2005).

The two most common types of learning curves are exponential and power curves. In this paper we discuss exponential curves as they seem better suited to predicting individual observations (Heathcote, Brown, & Mewhort, 2000) 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 1. 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).

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

Equation 1. Exponential model of practice

The standard learning curve methodology assumes that all trials of learning a skill are equivalent. It is not difficult to imagine theories that predict that certain types of practice may be more beneficial to learning than others. For example, we could believe that the subject will learn better the first time he practices the skill that day, and later trials that same day will have less benefit for learning. Rather than simply lumping all of the trials

together as t, we can create two new variables t_1 and t_2. The variable t_1 represents the number of trials 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 encountered the skill that day. This method of factoring trials into various types of practice does not change the amount of prior practice to student has had; $t = t_1 + t_2$ since trials are either the first one of the day or are not.

By simply using one parameter, t, to represent the number of prior trials, learning curves assume that both types of practice are equally valuable and the number of both trials should simply be added together. What if both types of practice are not equally valuable? Suppose the first learning opportunity of the day is twice as valuable as ones that occur later in the same day. In that case, a better summary of past student opportunities would be $t = 2*t_1 + t_2$ as it would better fit the performance data and more fairly equate the past opportunities of a learner. Imagine one learner who practiced the skill 10 times on 10 different days, while a second learner practiced the skill twice a day for 5 separate days. The classic model would have both learners with t=10; this refined model would describe the first learner as having 20 units of practice and the second learner as 15. If the first practice opportunity of the day really is twice is good then this refined description of practice is better.

The basic idea of *learning decomposition* is to find how to weight two types of trials to construct a best fitting learning curve. Equation 2 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, β, that represents the relative impact of the first trial of a day relative to trials 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 trials. 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. I.e. if we assumed that t_2 had a weight of 1.0 and found that β was 2.0, then if we had instead assumed that t_1 had a weight of 1.0 we would have found that β was 0.5. Both results are equivalent ways of stating that practice of the first type is twice as valuable as practice of the second type.

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

Equation 2. Learning decomposition model of practice

The parameter β is very interpretable: it is how many trials learning opportunities that are characterized as t_1 are worth relative to those characterized as t_2. If β>1 then trials of type t_1 are better for learning than those of type t_2. If β<1 then the opposite is true, and if β=1 then neither trial type 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 trials by ones that occur on Monday, Wednesday, or Friday vs. those that occur on Tuesday and Thursday (assuming no trials on weekends). For this decomposition we would hopefully get $\beta \approx 1$, as we have no reason to believe the day of the week matters for learning. Thus, the technique of learning decomposition is very broad. All it requires is some way of splitting the data into two or more disjoint sets that encompass all of the data.

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.

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. There are a variety of decisions to be made about how to approach our data; some of those decisions are principled, some are informed guesses, and others are little more than a hunch.

We now describe the data set we use for our case study; 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.

We are trying to better understand how students learn how to read by analyzing performance data about individual words recorded by the Reading Tutor (Mostow & Aist, 2001) 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 650 students, mostly from the Pittsburgh area, attempting to read 6.9 million words. Students used the tutor from September 2003 through May 2004 with a median usage of 5.9 hours, and improved their fluency from an average of 35 words per minute to an average of 62 words per minute read correctly as measured by a human scored paper test.

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 (Banerjee, Beck, & Mostow, 2003). 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.

Creating an outcome to measure learning

Given that we are investigating how students learn to read words by using learning decomposition, we must still decide what to consider an outcome variable. That is, what automatically collected marker should we use to track student reading development?

There are a variety of approaches for representing student performance at reading fluency. Students improve at reading in a variety of ways: they read words more quickly, read with fewer mistakes, ask for less help, and have appropriate prosodic inflection. Since we are only able to measure the first three of these items, we focus on those to measure progress. 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. In fact, the conditions under which the learning curve of the population is of the same form as curves for individuals is quite restricted (Brown & Heathcote, 2003). Therefore, fitting individual data points can produce a more authentic model of student learning. It is important to keep in mind that fitting individual data points tends to result in a poorer model fit than fitting a curve to student aggregate performance (since the aggregation smoothes out noise in the data).

Although reading time is continuous, it is misleading to use it as an outcome and ignore accuracy and help requests. Imagine one student who always asks for help when he encounters a challenging word and then mimics the tutor. This student would have a low reading time and would appear to be doing well. Then imagine a second student who pauses and tries to work the word out for himself. This student would not appear to be doing as well even though both students could have equal proficiency at reading. Therefore, some means of combining help and reading time is needed. A similar argument can be made for accuracy; a student who simply skips over challenging words will have a rather different learning curve from a student who puzzles out challenging words. In effect, student strategy choice results in a differential censoring of trials as the student who skips words is missing observations.

Our approach was to use the student's reading time as an outcome measure. However, when the student either asked for help, or the word was scored as incorrect by the ASR, or the student skipped the word, 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. Overall, it seems a reasonable way to distill a variety of student strategy choices into a single number for reading time. There are a variety of possible ways to combine help, correctness, and timing data. Our approach seems reasonable.

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. We do not use later encounters on the same day to track the student's reading development since performance on those encounters is contaminated by short term scaffolding effects of the prior encounter(s). This concern about short-term performance improvements is similar to the rationale we

have used in the past for ignoring immediate performance to estimate the impact of help (Heiner, Beck, & Mostow, 2004).

However, we do count subsequent encounters later in the day as opportunities to *learn* the word. As an example to explain this differential treatment of later day encounters, consider a student who encounters the word "divided" in ten consecutive sentences (there are stories in the Reading Tutor where this actually occurs). By the fifth time the student has to read the word he will have an extremely fast reading time. However, we do not believe the student has necessarily learned the word that well. Much of the improved performance comes from having the word immediately available in short term memory. So, the observed performance is not necessarily an indicator of how much the student has learned, and we therefore do not use those observations as an outcome for our learning curve. However, it is possible that the student continues to learn the word on the successive readings on the same day. Therefore, we count the exposure as an encounter to 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 trial, since it is the same day, that reading does not count as an outcome. Similarly, the next trial's performance is also ignored. However, note that the prior encounters field, which tracks the student's experience with this word, has been incremented to account for these two exposures.

Table 1. History of student encounters with word "elephant"

Day	Asked for help?	Reading time (seconds)	Prior encounters	Outcome (seconds)
1	yes	0.5	0	3.0
1	yes	1.5	1	-
1	no	1.3	2	-
2	no	3.8	3	3.0
3	no	1.7	4	1.7
3	no	1.2	5	-

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 trials at learning to read a word. We start with a known general psychology principle: distributed practice is generally superior to massed practice for long term retention (Ebbinghaus, 1885). This general rule suggests a decomposition: we consider a trial 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. We used 16 rather than 24 hours as the cutoff since the tutor was used during the school day and we wanted to avoid situations where the student encounters a word at 10:00 a.m. one day and then at 9:30 a.m. the next day. The essential construct is "did the trial occur on a later day?" Table 2 shows has we decompose the prior encounters based on massed vs. distributed practice.

Table 2 . Decomposing prior trials as massed and distributed practice

Day	Asked for help?	Reading time (seconds)	Prior encounters		Outcome (seconds)
			Distributed	Massed	
1	Yes	0.5	0	0	3.0
1	Yes	1.5	1	0	-
1	No	1.3	1	1	-
2	No	3.8	1	2	3.0
3	No	1.7	2	2	1.7
3	No	1.2	2	3	-

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 decompose the *t* parameter into trials 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 trials, repeated readings of the same story count as trials for learning (in particular, the variable for *rereading* would be increased in each case).

To simultaneously model rereading and massed practice, we created four possible trials 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 time (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.

Our model of reading development is shown in Equation 3. The first term, *l*word_length*, represents that longer words can take longer to read. The second term, *w*wordID*, accounts for more proficient readers reading text more quickly by modifying the term, *A*, representing first trial performance. This model also includes a term, *h*, to represent the amount of past help the student has received on this word. The *w* parameter controls for the student's overall reading proficiency; the *h* parameter controls for his knowledge of this particular word. 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. The goal is to find best-fitting values of the *r* and *m* parameters to find the relative impact of different types of practice on student reading development.

$$readingtime = l * word\_length + (w * wordID + A) * e^{-b*(r*m*RM + r*RD + m*NM + ND + h*\#helps)}$$

Equation 3. Model for examining effect of practice schedule and type of reading

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

RESULTS

We fit our model using SPSS's non-linear regression method. Computationally this approach is quite efficient, taking about fifteen minutes to estimate coefficients with 770,858 outcome data points. Table 3 shows the best fitting parameter estimates with standard errors for each of the terms in Equation 3. The column labeled "overall" are the estimates for the entire population. The next three columns are estimates by the bottom third, middle third, and upper third of the student population. Low proficiency students had a test score of less than 2.3 on the Woodcock Reading Mastery Word Identification subtest (2.3 corresponds to a 2nd grader in the 3rd month). High proficiency students had scores greater than 3.1. Medium proficiency students were those in between.

Table 3. Parameter estimates (± standard error) for learning decomposition model both overall and disaggregated by student proficiency

	Overall	Low proficiency	Medium proficiency	High proficiency
L	0.113 ± .001	0.111 ±.001	0.138 ± .001	0.112 ± .001
A	1.612 ± .005	1.701 ±.011	2.454 ± .03	0.825 ± .001
W	-0.231 ±. 001	-0.185 ±.005	-0.638 ± .011	-0.073 ± .001
B	0.053 ±. 001	0.029 ±.001	0.076 ± .003	0.054 ± .003
Help (h)	-1.93 ± .05	-2.29 ±.11	-1.98 ± .10	-5.62 ± .42
Reread (r)	0.53 ± .03	0.75 ± .06	0.97 ± .08	1.16 ± .17
Massed (m)	0.56 ± .05	0.98 ± .08	0.35 ± .04	0.24 ± .06

Since *L* has a positive coefficient (0.113), it indicates that longer words take longer to read. Since *W* has a negative coefficient (-0.231), it indicates that more proficient readers read words more quickly. The *A* parameters suggests that students take about 1.61 seconds to read a word (using our modified outcome with a penalty of 3.0 seconds for mistakes and help requests) when encountering it for the first time. The last three lines in Table 3 are of the most interest as they deal with the impact of help and the learning decomposition.

The impact of help on learning was a bit perplexing: help was worth a negative number of trials! One interpretation is that help is actually hurting students somehow. In complex domains where there are multiple representation and learning one way to solve a problem can interfere with another, instruction hurting student performance is possible. However, given the nature of the Reading Tutor's help this interpretation is unlikely. A more likely reason is that help indicates that students are having difficulty with this word, especially since

78% of help is student requested. In cases where a student requests help, he performs roughly 2 trials worse than on similar words on which he doesn't need help. I.e. if the student had 10 previous encounters with the word and 2 help requests, we expect he would perform comparably to a student who had 6 previous encounters with the word. So the help result is not about the impact of help, but what it says about the student's knowledge of this word.

We found that rereading had a coefficient of 0.53 for the entire student population. In other words, rereading a story twice would have roughly the same amount of learning as reading a new story for the first time ($2 * 0.53 = 1.06 \approx 1.0$). 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. However, the picture changes when we examine the estimates obtained for each level of student proficiency. There is a possible trend of more proficient readers benefiting from rereading. Students in our high proficiency category showed 16% more benefit from rereading compared to wide reading. However , this difference is not statistically reliable.

We found that, for the entire population, massed practice had a coefficient of 0.56, similar to that for rereading. So, in general, massed practice is not helpful for learning how to read and students should read stories that do not repeat words many times. However, this term showed an opposite trend as rereading: less proficient readers were not harmed (nor were they helped) by massed practice while more proficient readers were held back by encountering words *en masse*.

LIMITATIONS AND FUTURE WORK

The relatively high standard error (SE) for this model was rather surprising. We could reject the null hypothesis that the parameters representing rereading and massed practice were equal to 1.0 for the overall student populations. However, for high proficiency students and rereading we cannot determine whether a difference of 16% for high proficiency students in wide- vs. re-reading is real or a statistical fluke. Given that we used roughly a quarter-million data points (one-third of the data set) to estimate the parameter for this population, such a null result is a bit discouraging. How much data do we need to get reliable estimates?

One possible reason for the high SE is the relatively weak model fit. The R^2 for the model in Equation 3 is 11.0%. This fit is within the range of fitting individual trials of 5.1% to 65.8% reported by (Heathcote et al., 2000), and is not a sign the model is of the incorrect form. It is important to keep in mind that fitting individual points will have a lower R^2 than fitting aggregate statistics. For example, a simple exponential model that models the average reading time but ignores differences in word length, reading proficiency, and type of exposure, achieves an R^2 of 76%. It is unfortunate that while fitting individual observations is more mathematically sound and enables more sophisticated approaches, it results in weaker model fits. Determining some means of modeling individual observations but improving model fit is a high priority, as it should reduce the SE in the parameter estimates and enable more refined learning decompositions.

Automating the search of the decomposition space is another fruitful avenue of research. One crucial problem that must be overcome is finding some method for seeding the search space with expert knowledge. Expert knowledge both reduces the size of the search 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 operatioanlize 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) and spend the computational resources on exploring that search space to find a good way to operationalize the theory. Such an approach would seem to

draw on the strengths of both resources. Science is a social process and we need results that fit with existing theory (perhaps to disconfirm it). Spending time searching for new theory may not be productive if no one understands the results. However, a specific instantiation of the theory 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.

Another limitation of the work is that we can say that massed practice appears to be no better than, and in some cases is worse, than distributed practice. However, we cannot state that no students would benefit from massed practice. That less proficient readers had a coefficient of 0.98 is suggestive: perhaps the very least proficient readers would benefit from massed practice? Our current analysis provides no way of knowing. One conceptually straightforward extension is to not specify *a priori* the probable types of students (e.g. upper, middle, and lower thirds) but to instead estimate the decomposition results per student and note the students for whom massed practice was beneficial vs. those for whom it was not. The task would then be to find ways of classifying students to find general properties that predict when massed practice will be effective. One stumbling block is that disaggregating by students decimates the sample size and results in very noisy parameter estimates. Two approaches to address this problem are to either overcome the noisy parameter estimation process, or to instead live with the noise and treat the parameters as noisy estimates and still try to do the classification.

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. Although our model controls for several student properties, including overall proficiency, proficiency on this word (via help), and prior exposure to this word in the Reading Tutor, we should be careful before making causal inferences. An example of such a confound is the result with help. Students are not randomly assigned to a condition where some receive help and some do not; the self-selection of help creates a large bias in our data set leading to difficulties interpreting just what the negative help coefficient really meant. The question is whether such self-selection of rereading and massed practice causes similar biases. There are many possible causes for rereading: student preference, low reading proficiency, etc. However, to disconfirm our model result the cause of rereading would also have to cause low gains in learning how to read. Given that our model controls for test scores, it is difficult to come up with a plausible candidate. For massed practice the results are unlikely to be impacted by a latent variable since students and the tutor take turns picking stories (Mostow & Aist, 2001), so it is moderately difficult for students to cause massed practice to occur especially since students select stories based on the titles. It would be desirable to find a method to estimate the plausibility or bound the effect of such hidden selection biases.

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 trials. 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 standing 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 grokkable and directly transferable to 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. An earlier version of this work was presented at the 2005 Scientific Studies of Reading Conference and was well received. The model presented in this paper will be discussed at the SSSR 2006 conference. Feedback received there will guide development of the model for publication in a reading or psychology journal.

This paper also introduces an asymmetry in the process of constructing learning curves. Typically the same trials are used both to construct the dependent measure (in our case, reading time) and the independents (the prior number of trials). This work instead breaks that linkage by only permitting certain trials, those thought to be clean indicators of learning, to be used in the model fitting process. However, less pure trials (e.g. massed practice) are entered into the model to estimate their impact on the outcome. E.g. in Table 1 there are six

attempts to read the word "elephant," but only three of those trials would be used as dependents when constructing the learning curve (however, all prior trials would be entered as independents).

The closest related research is learning factors analysis (LFA) (e.g. (Cen, Koedinger, & Junker, 2005)). 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 the skills and impact of practice to vary would be desirable.

The fairly high standard errors for the parameter estimates suggest that the approach of learning decomposition is only feasible with big data. Thus, it is an excellent "secret weapon" for use by educational data miners for answers questions about learning and instruction.

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

Banerjee, S., Beck, J., & Mostow, J. (2003, September 1-4). Evaluating the Effect of Predicting Oral Reading Miscues. *Proc. 8th European Conference on Speech Communication and Technology (Eurospeech 2003)*,(pp. 3165-3168),Geneva, Switzerland.

Brown, S., & Heathcote, A. (2003). Averaging learning curves across and within participants. *Behaviour Research Methods, Instruments & Computers, 31*, 11-21.

Cen, H., Koedinger, K., & Junker, B. (2005). Automating Cognitive Model Improvement by A*Search and Logistic Regression. *Educational data mining workshop at National Conference on Artificial Intelligence*,(pp.

Ebbinghaus, H. (1885). *Memory: A Contribution to Experimental Psychology* (H. A. R. C. E. Bussenius, Trans.). New York: Teachers College, Columbia University.

Heathcote, A., Brown, S., & Mewhort, D. J. K. (2000). The Power Law Repealed:The Case for an Exponential Law of Practice. *Psychonomics Bulletin Review*, 185-207.

Heiner, C., Beck, J. E., & Mostow, J. (2004, June 17-19). Improving the Help Selection Policy in a Reading Tutor that Listens. *Proceedings of the InSTIL/ICALL Symposium on NLP and Speech Technologies in Advanced Language Learning Systems*,(pp. 195-198),Venice, Italy.

Koedinger, K. R., & Mathan, S. (2004). Distinguishing Qualitatively Different Kinds of Learning Using Log Files and Learning Curves. *ITS2004 Workshop on Analyzing Student-Tutor Interaction Logs to Improve Educational Outcomes*,(pp.

Martin, B., Koedinger, K., Mitrovic, T., & Mathan, S. (2005). On Using Learning Curves to Evaluate ITS. *Twelfth International Conference on Artificial Intelligence in Education*,(pp. 419-426),Amsterdam.

Mostow, J., & Aist, G. (2001). Evaluating tutors that listen: An overview of Project LISTEN. In K. Forbus & P. Feltovich (Eds.), *Smart Machines in Education* (pp. 169-234). Menlo Park, CA: MIT/AAAI Press.

Newell, A., & Simon, H. (1972). *Human Problem Solving*. Englewood Cliffs, N.J.: Prentice-Hall.