A Case Study Empirical Comparison of Three Methods to Evaluate Tutorial Behaviors

Xiaonan Zhang[1], Jack Mostow[1], and Joseph E. Beck[2]

[1] Project LISTEN, School of Computer Science, Carnegie Mellon University
[2] Computer Science Department, Worcester Polytechnic Institute

Abstract. Researchers have used various methods to evaluate the fine-grained interactions of intelligent tutors with their students. We present a case study comparing three such methods on the same data set, logged by Project LISTEN's Reading Tutor from usage by 174 children in grades 2-4 (typically 7-10 years) over the course of the 2005-2006 school year. The Reading Tutor chooses randomly between two different types of reading practice. In assisted oral reading, the child reads aloud and the tutor helps. In "Word Swap," the tutor reads aloud and the child identifies misread words. One method we use here to evaluate reading practice is conventional analysis of randomized controlled trials (RCTs), where the outcome is performance on the same words when encountered again later. The second method is learning decomposition, which estimates the impact of each practice type as a parameter in an exponential learning curve. The third method is knowledge tracing, which estimates the impact of practice as a probability in a dynamic Bayes net. The comparison shows qualitative agreement among the three methods, which is evidence for their validity.

Keywords: educational data mining, randomized controlled trials, learning decomposition, knowledge tracing, evaluating tutor strategies.

1 Introduction

The behavior of an intelligent tutor affects its efficacy, so it is important to evaluate. One reason is to improve the tutor as part of data-driven iterative refinement. Another reason is to draw lessons for what behaviors to embrace or avoid in designing other tutors. The obvious way to evaluate alternative tutorial behaviors is to perform a controlled between-subjects comparison of different versions of the tutor, with each version employing a different behavior. However, such experiments may require many students and considerable time to achieve statistically reliable results. Is there a better way?

Fortunately, intelligent tutors can log detailed, longitudinal interactions, and experimentally vary the behaviors that affect those interactions. Analyzing the resulting data lets us evaluate tutorial behaviors. Such evaluation can test whether a behavior works, gauge how well it works, and compare alternatives.

Previous research has employed various methods to perform such analyses, but we are not aware of any studies whose express purpose was to compare alternative methods for

B. Woolf et al. (Eds.): ITS 2008, LNCS 5091, pp. 122–131, 2008.

evaluating tutor behavior. To help fill this gap, we present a case study that applies three analysis methods to the same data set, described in Section 2. Sections 3, 4, and 5 respectively describe each method as applied to the data. Finally, Section 6 summarizes results, conclusions, and contributions.

2 Case Study: Evaluate Two Modes of Practice in a Reading Tutor

We carried out our case study on data from Project LISTEN's Reading Tutor, which helps children learn how to read [1]. The Reading Tutor and the student take turns to pick a story, which is then displayed line by line on a computer screen. The Reading Tutor listens to the student read the story aloud, and uses automatic speech recognition (ASR) to track the student's position in the text, detect (some) mistakes, and measure the time to read each word. The Reading Tutor also provides various forms of assistance when the student gets stuck, or clicks for help. It logs its interactions and speech recognizer output into a database.

The analysis problem in our case study is to compare two modes of practice for children who are still learning the letter-sound mappings of English. The Reading Tutor uses an instructional activity adapted from published interventions [2-5] to teach these mappings in the context of isolated words. To exercise taught mappings in the context of connected reading, the Reading Tutor then presents practice text in one of two modes – choosing randomly between them each time, but using the same text either way.

One mode of practice is assisted oral reading. In this mode, the Reading Tutor displays each successive story sentence, e.g., *Sam sat on the mat*, and listens to the child read it aloud, giving help as necessary.

In the other mode, called Word Swap, the Reading Tutor reads aloud, and the child provides feedback. Word Swap is based on an activity used by a human expert to teach children to attend to the correspondence between print and sound. First the Reading Tutor explains the task:

> *Good, careful readers make sure that what they say matches what they see. Let's play a game called Word Swap. The Reading Tutor will read the story to you, but it might read some words wrong. Click on the words that do not match what you hear!*

In Word Swap, the Reading Tutor picks a word at random from each sentence, e.g., *sat*, and replaces it with some other random word from the story, e.g., *am*. It displays the modified sentence, e.g. *Sam am on the mat*, but plays the narration of the original sentence, so as to deliberately "misread" the replaced word. (The Reading Tutor uses recorded human speech, so it is easier to modify the displayed text of the sentence than its spoken narration.) The student's task is to click on the "misread" word. When the student clicks on the "misread" word *am*, the Reading Tutor replaces it with the correct word *sat* and says *Right! This says am, not sat.* If the student clicks on a correctly read word, the Reading Tutor says, *no, the Tutor read that word right!*

Which is more effective – assisted reading or Word Swap? To study this question, we define "effective" in terms of how well students do on the words in the story when

they read them again later. We measure performance in reading an individual word (in context) based on how long the student takes to read the word, whether the student clicks on the word for help, and whether the speech recognizer accepts the word as read correctly. We compute this information from the Reading Tutor's log data. Ideally we would also measure how well the child attends to spelling-sound correspondence when reading the word – the goal of Word Swap. However, we have not defined or automated such a measure, in part because the very signs that may indicate such attention (slow reading and frequent self-corrections) may merely indicate poor reading.

The 2005-2006 Reading Tutor logged 2669 encounters of letter-sound practice passages by 174 students in grades 2-4. The 1311 encounters in assisted reading mode comprised 76,326 words. The 1358 instances of Word Swap totaled 83,421 words. To avoid ceiling effects, we exclude the most common 200 English words from the dataset, leaving 31,216 word encounters under assisted reading conditions, and 37,028 under Word Swap conditions, respectively. We now discuss the three methods we used to evaluate the effects of these encounters.

3 RCT Analysis

Randomized controlled trials (RCTs) manipulate experimental variables to test their effects on outcomes. Randomizing assignment to treatment ensures that statistically reliable effects are truly causal. Intelligent tutors can randomize tutorial decisions such as what type of practice, assistance, or feedback to provide, and log large numbers of randomized trials, as illustrated by experiments in the Reading Tutor [1] as well as other tutors [6]. Each trial has a context in which it occurred, the decision made, and its outcome [7]. Aggregating over many trials by many students lets us analyze how the decision affects the outcome.

The context of the RCTs analyzed in this paper is the point at which the Reading Tutor has just taught some letter sounds and the student encounters a word in a practice text. The decision is which mode of practice to give – namely, assisted reading or Word Swap. The Reading Tutor randomizes this decision within-subject and within-text. That is, each time the Reading Tutor finishes a letter-sound lesson, it makes this decision anew. Randomizing within-subject – that is, giving each student both types of practice – controls for individual differences among students. Likewise, randomizing within-text – that is, using the same set of texts for both modes of practice – controls for differences among texts. However, the Reading Tutor chooses the mode of practice for an entire text at a time, rather than for each individual word. We can treat the practiced words as separate trials, but they are not independent.

How to define outcome? To analyze which mode of practice results in better word learning, we define the outcome of each trial as the student's performance on a later encounter of the same word. (Practice on a word affects performance on that word much more than on other words [8].) If this encounter occurs in a story the student has read before, the student's performance may reflect remembering the story rather than reading the word. If the encounter occurs too soon, the student's performance may just reflect how recently the student or tutor has read the word. On the other hand, as time elapses, the trial's effect diminishes relative to other influences, such as classroom instruction.

Therefore we define its outcome as performance on the student's first encounter of the word 1-3 days after the randomized trial, provided it occurs in a new context.

As Section 1 explained, we measure performance on a word based on how long the student takes to read it, whether the student clicks on it for help, and whether the speech recognizer accepts it as read correctly. Table 1 defines the measures we use for RCT analysis. We represent undefined outcomes as null values.

Table 1. Outcome measures used in RCT analysis

Measure	Definition
Accepted	The speech recognizer (ASR) recognized the word as read correctly
Asked help	The student clicked on the word for help in reading it
Credited	True if the ASR accepted the word without the student receiving help; false if the ASR rejected the word or the student requested help; undefined (and excluded from RCT analysis) if the ASR accepted the word after tutor-initiated help that masked whether the student knew the word
Latency [9]	The delay from the end of reading the previous word until starting to say the current word
Reading time	Latency plus the time to say the word, with this sum capped at 3 seconds to deal with outliers
Adjusted time [10]	Reading time for credited word; 3 seconds for uncredited word; undefined if credit is undefined

Sources of variance in word reading performance include student, word, story, and practice mode. Since words differ more than students (C. Perfetti, personal communication), we compare practice modes paired by story and word. That is, for a story word encountered in both assisted reading and Word Swap (generally by different students), we compare performance on each word after one mode of practice versus after the other, averaged across students. We compute the difference in a performance measure M as M(Word Swap) – M(assisted reading). We use a t-test, paired by story and word, to test whether performance differs significantly by practice mode, so the degrees of freedom (253) are one fewer than the number of such words.

As Table 2 shows, this difference is significantly greater than 0 for latency, reading time, and adjusted time. The positive difference means that students read words significantly slower after Word Swap than after assisted reading practice. Whether this is good news or bad news for Word Swap depends on *why* they read slower: are they paying better attention? or did they just learn the words less well? We can't tell.

Because this comparison does not control for student identity, one possible confounding factor is the difference between students who get one type of practice and students who get the other. However, since treatment assignment is randomized anew for each passage for each student, and for each word is based on different randomized subsets of students across stories, we assume we can ignore differences between these subsets. To test this assumption, we verified that reading proficiency (measured by a paper pretest) and reading level (estimated by the Reading Tutor) did not differ significantly between treatment conditions.

Besides pairing by word, we also tried pairing by student and averaging performance after each mode across the words the student practiced in that mode, but none of the differences were statistically reliable. This approach is more conservative because it controls for individual differences among students, and because each student's performance is independent of other students' performance. However, it is less powerful statistically because there were fewer students than words, and because it does not control for differences between the words practiced in different modes.

Table 2. Differences in word reading performance after assisted reading versus after Word Swap, paired by story and word and averaged over the students who practiced that story word in that condition

Outcome measure	**Outcome differences (Word Swap – assisted reading)**				**Paired t-test**
	Mean	Std. Dev.	95% Confidence Interval of the Difference		Sig. (2-tailed)
			Lower	Upper	
% accepted	0.000	0.175	-0.021	0.022	0.965
% asked help	0.015	0.171	-0.006	0.036	0.168
% credited	-0.007	0.203	-0.032	0.018	0.592
Latency	0.039	0.244	0.009	0.069	**0.011**
Reading time	0.060	0.347	0.017	0.103	**0.006**
Adjusted time	0.074	0.537	0.008	0.150	**0.028**

4 Learning Decomposition

Learning decomposition generalizes classic exponential learning curve analysis to estimate the relative benefit of different types of practice [10], and has now been used in several such analyses. In brief, it models each student's item performance data (in this case word reading times) as an exponential function of previous practice on the item. The model disaggregates practice into the number of encounters of each practice type (e.g., Word Swap or assisted reading), each weighted by a free parameter coefficient. Fitting the model to the data (e.g. in SPSS) yields parameter estimates that represent the relative value of each type of practice for that student. Averaging and bootstrapping the parameter estimates across students gives confidence intervals on the means and tells which differences between practice types are reliable.

We follow earlier work [10] in three respects. First, we measure performance using the adjusted time measure defined in Table 1 of Section 3, and exclude encounters where its value is undefined. Second, to exclude recency and story memorization effects as mentioned in Section 3, we measure performance only on a student's first encounter of a word each day, and only in a story that he or she has not read before. Third, we adopt the same general model form, including an additive term to represent the effect of word length. However, we use different practice types, namely assisted reading and Word Swap. Equation 1 shows the resulting model:

$$adjusted\ reading\ time = L * word\_length + A * e^{-b*(\#Reading + \beta * \#Swap)} \quad (1)$$

Our four model parameters mean roughly this:

- *L*: the increase in predicted word reading time for each additional letter in the word
- *A*: the predicted time to read a word with no prior practice in either condition
- *b*: learning rate
- β: the impact of a Word Swap encounter compared to an assisted reading encounter, whose impact we define to be 1

The input variables *#Reading* and *#Swap* count the number of prior encounters of the same word in assisted reading and Word Swap, respectively. These practice variables include *all* encounters of the word, not just the first encounter on each day or in each story.

Using Equation 1, we build a model for each individual student. After excluding models for which the fitting procedure fails due to sparse data, we take medians of the remaining 140 models as the overall parameter estimates. We use medians instead of means in order to deemphasize outliers in the noisy individual estimates. We also derive the 95% confidence interval for each parameter using non-parametric bootstrapping [11]. Table 3 shows the result.

Table 3. Overall parameter estimates (± 95% confidence interval)

Parameter	*L*	*A*	*b*	β
Estimate	0.0615 ± 0.022	0.7035 ± 0.0775	-0.0515 ± 0.015	0.125 ± 0.1147

The confidence interval for β shows that it is significantly less than 1, which implies that Word Swap has significantly less impact than assisted reading in reducing (adjusted) word reading time. However, β is also reliably (though just barely) greater than 0, implying that Word Swap also reduces word reading time.

5 Knowledge Tracing

Knowledge tracing [12] infers a student's knowledge of a skill from observations of the student's performance on that skill. Knowledge tracing incrementally updates the probability K_n that the student knows a given skill at time step n, according to a dynamic Bayes net model with the following parameters:

- *knew*: Probability K_0 that the student already knew the skill prior to instruction
- *learn*: Probability of acquiring the skill from a single practice
- *forget*: Probability of losing a known skill
- *guess*: Probability of answering correctly without knowing the skill
- *slip*: Probability of answering incorrectly despite knowing the skill.

To investigate how different modes of practice influence student knowledge, we introduce another node *Practice Mode (PM)* into the basic knowledge tracing model, as Figure 1 shows.

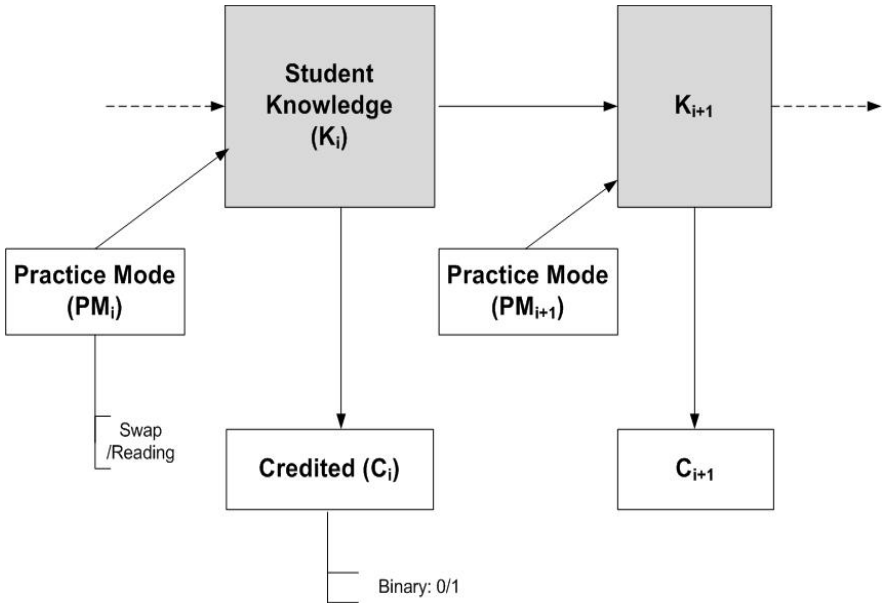

Fig. 1. Knowledge tracing model extended with a binary-valued "Practice Mode" node

This model assumes that the probability of a student's learning a skill depends on practice mode. To measure student performance in assisted word reading, we use *credited* (see section 3 for definition). For Word Swap steps, however, since we do not have observations of a student's reading the word, *credited* is unobservable. The extended model has more parameters: *is_reading* is the probability that the practice mode is assisted reading; K_0*_swap* and K_0*_reading* are the probability that the student already knew the word prior to any practice, conditioned on whether the first practice was Word Swap or assisted reading; *learn_swap* and *learn_reading* are the respective probabilities of acquiring the skill from a Word Swap or assisted reading practice opportunity; and finally, *forget_swap* and *forget_reading* are the respective probabilities of losing a skill after a Word Swap or assisted reading practice opportunity. The parameters *guess* and *slip* remain the same as in the basic knowledge tracing model.

One problem with fitting the data to a knowledge tracing model, however, is that the observed student performance can correspond to an infinite family of possible model parameter estimates [13]. Following [14], we address this problem by specifiying a plausible initial value for each parameter, and encoding domain knowledge as Dirichlet priors on the parameters to bias the model fitting procedure. We specify an order-2 Dirichlet distribution as two positive numbers α1 and α2, which correspond roughly to the number of positive and negative examples seen. For example, we use α1=9 and α2 = 6 for K_0*_swap*. These values mean roughly that the Dirichlet prior for K_0*_swap* is generated from 9 cases of the student already knowing a skill, and 6 cases of the student not knowing it, when the first practice is Word Swap. The expected value of K_0*_swap* is 9/(9+6) = 0.6.

We set the initial parameters, as well as α1 and α2, by examining histograms from previous knowledge tracing experiments, getting similar values to those in [14]. The

first three columns in Table 4 show the name and initial value of each parameter, as well as the α1 and α2 values of the Dirichlet priors. Notice that they avoid any bias toward either Word Swap or assisted reading. We refrain from specifying Dirichlet priors for the *learn* and *forget* parameters, so as not to prejudice the search through the model space.

Table 4. Initial values, Dirichlet priors, and aggregated estimates of the parameters in the knowledge tracing model

Parameter	Initial Value	Dirichlet (α1, α2)	Mean	Std. Dev.
is_reading	0.5	N/A	0.683	0.237
K_0*_swap*	0.66	(9,6)	0.599	0.019
K_0*_reading*	0.66	(9,6)	0.655	0.061
Guess	0.64	(17,9)	0.670	0.041
Slip	0.07	(1,15)	0.028	0.020
learn_swap	0.14	N/A	0.258	0.187
learn_reading	0.14	N/A	0.566	0.360
forget_swap	0.0014	N/A	0.014	0.086
forget_reading	0.0014	N/A	0.011	0.087

To investigate which practice mode helps more to learn a word, we treat the ability to read each word as a distinct skill. Then we build a model for each word using observations of many students' encounters of that word, using Bayes Net Toolkit for Student Models (BNT-SM) [15]. After excluding the cases where model construction fails due to sparse data (e.g. the word was encountered very few times, or in only one treatment condition), we get 259 word-specific models, across which we average the parameter estimates. The last two columns of Table 4 show the mean and standard deviation for each parameter.

A t-test, paired by word, shows no significant difference between *forget_swap* and *forget_reading*. In contrast, *learn_reading* is significantly larger than *learn_swap* ($p<0.01$). That is, students are likelier to acquire a word from assisted reading practice than from Word Swap practice.

6 Conclusion

This paper explored three methods to evaluate tutorial behaviors: RCT analysis, learning decomposition, and knowledge tracing. It reports a case study in the context of Project LISTEN's Reading Tutor, to test whether assisted reading and Word Swap practice differ in how well they help students learn words.

One result of this endeavor is to confirm that knowledge tracing can usefully be adapted to evaluate the impact of different tutor behaviors. Previous work [16, 17] used this approach to evaluate the same mode of practice with versus without tutor help. Here we evaluate two different modes of practice, each with a different task for the student, and consequently different types of performance to observe.

In comparing evaluation methods, we have two basic questions. First, did their results agree? Yes, all three methods indicate that assisted reading beat Word Swap on one or more of our measures. Though the three methods differ in input, output, and model form, the qualitative consistency of their results provides some empirical evidence for the validity of the results and the methods.

Second, were some methods more sensitive than others? If methods A and B agree, and A is more sensitive than B, we expect A to achieve statistical significance on more comparisons than B does. We see no such pattern. The methods agree qualitatively, but not on which measures show statistically significant differences between the two modes of practice. Clarifying the empirical behavior and relative utility of these methods will require comparing them on additional data sets from diverse domains.

The results imply that assisted reading is more effective than Word Swap at helping students learn to read words quickly, accurately, and independently. They do not necessarily imply that Word Swap is inferior for its intended purpose of teaching children to attend to the correspondence between print and speech. Indeed, conceivably children read words more slowly after Word Swap than after assisted reading because it actually succeeded. One challenging direction for future work is to develop an automated measure of such attention.

Acknowledgments. The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305B070458 to Carnegie Mellon University, by the National Science Foundation under ITR/IERI Grant No. REC-0326153, and by the Heinz Endowments. The opinions expressed are those of the authors and do not necessarily represent the views of the Institute, the U.S. Department of Education, the National Science Foundation, or the Heinz Endowments. We also thank the educators, students, and LISTENers who helped generate and analyze our data.

References

1. Mostow, J., Aist, G.: Evaluating tutors that listen: An overview of Project LISTEN. In: Forbus, K., Feltovich, P. (eds.) Smart Machines in Education, pp. 169–234. MIT/AAAI Press, Menlo Park (2001)
2. Beck, I.: Reading Today and Tomorrow (Teachers' Editions for Grades 1 and 2). Holt and Co., Austin (1989)
3. Beck, I., Hamilton, R.: Beginning reading module. American Federation of Teachers, Washington (2000) (Original work published 1996)
4. Cunningham, P.M., Cunningham, J.W.: Making Words: Enhancing the invented spelling-decoding connection. The Reading Teacher 46(2), 106–114 (1992)
5. McCandliss, B., Beck, I.L., Sandak, R., Perfetti, C.: Focusing attention on decoding for children with poor reading skills: Design and preliminary tests of the word building intervention. Scientific Studies of Reading 7(1), 75–104 (2003)
6. Razzaq, L., Heffernan, N.T., Lindeman, R.W.: What level of tutor feedback is best? In: Proceedings of the 13th Conference on Artificial Intelligence in Education, IOS Press (2007)

7. Mostow, J., Beck, J.: Some useful tactics to modify, map, and mine data from intelligent tutors. Natural Language Engineering (Special Issue on Educational Applications) 12(2), 195–208 (2006)
8. Zhang, X., Mostow, J., Beck, J.E.: All in the (word) family: Using learning decomposition to estimate transfer between skills in a Reading Tutor that listens. In: AIED 2007 Educational Data Mining Workshop, Marina del Rey (2007)
9. Mostow, J., Aist, G.: The sounds of silence: Towards automated evaluation of student learning in a Reading Tutor that listens. In: Proceedings of the Fourteenth National Conference on Artificial Intelligence (AAAI 1997), pp. 355–361. American Association for Artificial Intelligence, Providence (1997)
10. Beck, J.: Using learning decomposition to analyze student fluency development. In: Ikeda, M., Ashley, K.D., Chan, T.-W. (eds.) ITS 2006. LNCS, vol. 4053. Springer, Heidelberg (2006)
11. Cohen, P.R.: Empirical Methods for Artificial Intelligence, p. 405. MIT Press, Cambridge (1995)
12. Corbett, A., Anderson, J.: Knowledge tracing: Modeling the acquisition of procedural knowledge. User modeling and user-adapted interaction 4, 253–278 (1995)
13. Beck, J.E., Chang, K.: Identifiability: A Fundamental Problem of Student Modeling. In: Conati, C., McCoy, K., Paliouras, G. (eds.) UM 2007. LNCS (LNAI), vol. 4511. Springer, Heidelberg (2007)
14. Beck, J.E.: Difficulties in inferring student knowledge from observations (and why you should care). In: Proceedings of the AIED 2007 Workshop on Educational Data Mining, Marina del Rey, pp. 21–30 (2007)
15. Chang, K.-m., Beck, J., Mostow, J., Corbett, A.: A Bayes Net Toolkit for Student Modeling in Intelligent Tutoring Systems. In: Proceedings of the 8th International Conference on Intelligent Tutoring Systems, Jhongli, Taiwan (2006)
16. Jonsson, A., Johns, J., Mehranian, H., Arroyo, I., Woolf, B., Barto, A., Fisher, D., Mahadevan, S.: Evaluating the Feasibility of Learning Student Models from Data. In: Educational Data Mining: Papers from the AAAI Workshop, pp. 1–6. AAAI Press, Pittsburgh (2005)
17. Chang, K.-m., Beck, J.E., Mostow, J., Corbett, A.: Does Help Help? A Bayes Net Approach to Modeling Tutor Interventions. In: AAAI 2006 Workshop on Educational Data Mining, Boston (2006)