Using Automated Within-Subject Invisible Experiments to Test the Effectiveness of Automated Vocabulary Assistance

Greg Aist and Jack Mostow[1]

[1] Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh PA 15213
aist@cs.cmu.edu, mostow@cs.cmu.edu

Abstract. Machine learning offers the potential to allow an intelligent tutoring system to learn effective tutoring strategies. A necessary prerequisite to learning an effective strategy is being able to automatically test a strategy's effectiveness. We conducted an automated, within-subject "invisible experiment" to test the effectiveness of a particular form of vocabulary instruction in a Reading Tutor that listens. Both conditions were in the context of assisted oral reading with the computer. The control condition was encountering a word in a story. The experimental condition was first reading a short automatically generated "factoid" about the word, such as "cheetah can be a kind of cat. Is it here?" and then reading the sentence from the story containing the target word. The initial analysis revealed no significant difference between the conditions. Further inspection revealed that sometimes students benefited from receiving help on "hard" or infrequent words. Designing, implementing, and analyzing this experiment shed light not only on the particular vocabulary help tested, but also on the machine-learning-inspired methodology we used to test the effectiveness of this tutorial action.

1 How can tutors learn?

Good human teachers learn what works best for which students in different contexts. In contrast, automated tutors generally learn little if anything from their interactions with students. This may be one reason why their effectiveness – though sometimes surpassing conventional classroom instruction – still lags behind individual human tutoring. Yet an automated tutor could potentially learn from individual interaction with many more students than a human could tutor in a lifetime.

Learning to tutor better means learning to make better tutorial choices. Automated tutors embody many choices, such as which task to pose next, or what help to give. Such choices may be built in to the tutor design, or computed at runtime. These choices are presumably crucial to educational effectiveness. For example, a study of one-on-one human tutoring [Juel, 1996] found that successful tutor-student dyads engaged in a significantly different distribution of activities than less successful dyads.

However, individual tutorial choices are difficult to evaluate. First, establishing student improvement may be challenging. Second, attributing a particular improvement to a specific tutorial decision is also difficult. Testing improvement immediately after a tutorial intervention may reduce interactions between interventions, but may also allow recency effects to dominate the experimental outcome. Assigning credit for outcomes to specific tutorial decisions can also be simplified if outcomes can be factored in terms of which tutorial actions are likely to affect them. In particular, in this paper we assume that vocabulary gains are the sum of independent gains on individual words, affected only by exposure to that word. This assumption is a simplifying approximation, because learning about one word such as *tusk* may help a student learn about another word such as *elephant*. Nonetheless, assuming independence of vocabulary gains lets us relate outcomes on individual words to the choices involving the student's encounters with that word. Finally, even if a strong model relates instructional interventions to educational outcomes, conventional between-student controlled experiments can compare only a few alternatives out of the astronomical number of possible combinations, given the expense of assigning a large enough sample of students to each condition to obtain statistically informative results. Moreover, such comparisons reveal only which design works better overall for which students. They do not characterize the specific contexts in which each compared behavior works best. We need more efficient methods for learning to make better tutorial choices.

Learning from experience involves trying out different choices and evaluating their effects. Thus a tutor that learns needs to systematically explore different tutorial choices, assess their effects on students' educational gains, and apportion credit for those effects among the series of choices that led to them. Analogously, a spoken dialog system that learns to improve the quality of its interactions needs to explore alternative responses, assess their effects on customer satisfaction or other outcome measures, and infer when to use each response [Walker et al. 1997, Singh et al. 1999].

We have therefore been exploring a novel methodology for evaluating tutorial methods, made possible by automated tutors. In this paradigm, which we call "invisible experiments," an automated tutorial agent randomly selects from a set of felicitous (context-appropriate) behaviors, and records the machine-observable effects of each such decision on subsequent dialog. Aggregating over many randomized trials then enables us to evaluate effects of different conversational behaviors on human-tutor dialog. Conventional experiments assess the overall effects of a particular tutor design on tutorial effectiveness. In contrast, invisible experiments offer a controlled assessment of the fine-grained effects of a given tutorial choice in various contexts, compared to what would happen otherwise. Thus they illuminate why and when a choice succeeds.

2 Project LISTEN's Reading Tutor

Project LISTEN's Reading Tutor listens to children read aloud, and helps them learn to read [Mostow & Aist CALICO 1999]. The Reading Tutor displays one sentence at a time to the student, listens to the student read all or part of the sentence aloud, and responds expressively using recorded human voices. The design of the Reading Tutor addresses the educational goal of learning to read, while balancing motivational factors such as confidence, challenge, curiosity, and control [Lepper et al. 1993]. Therefore, the Reading Tutor lets children choose stories from a variety of genres, including nonfiction, fictional narratives, and poems.

3 Teaching vocabulary during assisted oral reading

Part of helping children learn to read is helping them make the most of encounters with new vocabulary. As children transition from learning to read into reading to learn, they must be able to understand what they read. Having a good vocabulary is essential to reading for understanding. One of the best ways to teach vocabulary is to have the student read. However, encountering a word in a sentence may not be sufficient to ensure that a student learns the meaning of the word. We wanted to explore ways of augmenting text to help children learn words better than they would from the unaugmented text.

4 Experiment: Does automated vocabulary assistance help?

We conducted an experiment to test if augmenting text with information about words would help children learn the meanings of those words better than they would have from the text alone. In Fall 1999, 60 children in six grade 2-3 classrooms read stories using a version of the Reading Tutor modified to provide extra vocabulary help on some words. We augmented some words in stories the child was reading with synonyms (X means Y), antonyms (X is the opposite of Y) or hypernyms (X is a kind of Y). For a given child, some of the words were augmented and others were left unaugmented to serve as a control group. These synonyms, antonyms, or hypernyms were retrieved from WordNet [http://www.cogsci.princeton.edu/~wn/w3wn.html], a lexical database. The words selected were words with only one or two senses in WordNet. (We selected these words in order to handle any text without first having to sense-tag polysemous words.) The next time the child logged in (typically the next day), the computer presented multiple-choice vocabulary probes.

To summarize briefly the form of the experimental trials:

- Context – a student using the Reading Tutor encounters a new word in a story, such as “assistance” in:
  “but no one paid any heed to his cries, nor rendered any assistance.”
- Treatment – explain some new words but not others:
  “Maybe assistance is like aid here. Is it?”
- Test – next day, automatically generate a multiple-choice question:
  “Which do you think means the most like assistance:
  carrying out; help; saving; line?”

In this example, “assistance” was the target word, “aid” was the comparison word, and “help” was the expected answer in the multiple-choice question presented the next day. In some questions, the expected answer was the same as the comparison word presented the day before, and might therefore be easier due to lexical memory effects.

We now discuss several aspects of this experiment in more detail.

4.1 Assigning words to conditions for vocabulary assistance

For each student, half of the target words were randomly assigned to an “extra help” condition, and the rest of the target words to a control (no extra help) condition. Assignments of words to conditions was done just prior to displaying the sentence with the target word, while the student was reading the story with help from the Reading Tutor. When the student encountered a previously unseen target word, the new word was randomly assigned to either the experimental (context + factoid) or control (context only) condition for that student. By having an open-ended set of target words instead of a fixed list, we allowed for the addition of new material by teachers, students, or the Project LISTEN team without disrupting the study design. The assignments of words to conditions were intended to persist throughout the student’s history of Reading Tutor use to enable us to look for longer-term effects of multiple exposures to a word. Unfortunately, due to a flaw in the software the assignments were not saved to disk. We therefore analyzed only a student’s first day of experience with a word, and the subsequent vocabulary question.

4.2 Constructing and displaying vocabulary assistance

While the student read a story, the Reading Tutor displayed vocabulary help for the “extra help” target words as short “factoids” derived from WordNet. We kept factoids short to minimize disruption to story flow. Also, to keep factoids from overwhelming the story, we intended the Reading Tutor to display extra help on a given word stem at most once per student per day, regardless of how many times he or she encountered that word in text. Vocabulary help was displayed "just in time" before displaying a sentence containing the target word. Such help was omitted for story titles (represented in the Reading Tutor as the first sentences of stories) to avoid confusion.

In order to generate automated vocabulary help for arbitrary text, we could not rely on manual tagging of word senses. Some target words had more than one sense in WordNet, or might be used in a sense unknown to WordNet. Therefore, a factoid had to avoid a blanket assertion like "Assistance means aid," because it might explain the wrong sense of a polysemous word. (Does "aid" really explain "assistance" in "Directory Assistance"?). Instead, we hedged the phrasing: "Maybe assistance is like aid here… Is it?" We hoped that such hedging might have the fringe benefit of improving metacognition by encouraging the student to think about the meaning of the word in context.

The Reading Tutor constructed the text of a factoid from templates by filling in blanks for the target word and the comparison word, as illustrated here:

ugly may be the opposite of beautiful. Is it here?	(antonym)
cheetah can be a kind of cat. Is it here?	(hypernym)
Maybe assistance is like aid here… Is it?	(synonym)

Before displaying a sentence containing a target word assigned to the experimental condition, the Reading Tutor displayed the factoid in a yellow call-out box displayed on top of the original story, so as to indicate that it was not part of the story proper. After the student read the factoid (with the Reading Tutor's normal assistance), this box vanished, and the Reading Tutor displayed the sentence containing the target word just explained by the factoid. Of course, factoids were not presented for target words in the control condition.

4.3 Assessing the effectiveness of factoid assistance

The next time a student logged in (one day or more after the target word was seen) the Reading Tutor asked vocabulary questions for each of the target words encountered – both "extra help" and "control" words. For example, the Reading Tutor generated this question for the target word *snack*:

Which of these do YOU think means the most like snack?
meal
hash
spring roll
drink down

(The correct answer is 'meal'.)

4.4 Collection of data from factoid-assistance vocabulary experiment

The 1999 version of the Reading Tutor used in this experiment wrote information on student-Reading Tutor interactions directly to a database. Each classroom had a single Reading Tutor with the database for 10-12 students in that classroom. Project LISTEN staff copied these databases onto portable media and brought them back to the laboratory for analysis. Earlier versions of Reading Tutor had recorded detailed ASCII text logs from which information on student-Reading Tutor interactions had to be extracted and imported into a relational formation on student-Reading Tutor interactions had to be extracted and imported into a relational database (SQL Server™) for further analysis. We had found that post facto analysis of log files was time-consuming and error prone, and were determined to avoid such ad hoc analysis if at all possible.

4.5 Does factoid assistance result in better answers to vocabulary questions?

In order to assess the effectiveness of factoid assistance, we compared student performance on the experimental (context + factoid) condition to student performance on the control (context + factoid) condition.

We analyzed the results for three groups of words encountered during fall 1999:

1. all of the words,
2. the subset of words with only one sense in WordNet,
3. a set of words which would allow detection of a non-lexical effect (giving the help “X means Y” and then asking a multiple-choice question with expected answer Z).

Students’ individual performance on the vocabulary questions ranged between 23% and 80%, compared to a chance level performance of 25%. It might appear that a solid analysis would aggregate students’ performance on vocabulary questions that came after factoids, and compare the resulting ratio of correct answers to the corresponding ratio for questions that followed control exposures. Unfortunately, such an analysis is vulnerable to Simpson’s paradox [Romano and Siegel 1986]. Briefly stated, “It is not necessarily true that averaging the averages of different populations gives the average of the combined population” [Weisstein 2000]. Since different students have varying performance rates on vocabulary questions, aggregating all the data together confounds the identity of the student with the effect of the extra help on answering the question correctly.

For each group of words, we constructed a logistic regression model[1] (using SPSS software) with factors FACTOID (help or no help), ID (student id), ANSWER (right or wrong), and FACTOID*ANSWER (an interaction term to look for the effect of FACTOID on ANSWER.) In all three cases, FACTOID*ANSWER was not significantly greater than 0, indicating no significant difference between FACTOID conditions:

1. All words: N=3419, FACTOID*ANSWER = 0.07 +/- 0.07
2. Single-sense words: N=769, FACTOID*ANSWER = 0.17 +/- 0.14
3. Non-lexical effect: N=1637, FACTOID*ANSWER = 0.08 +/-1 0.11

(The numbers here are the parameters of the regression equation.)

What could have led to the lack of difference between conditions?

- Perhaps the help did not help after all.
- Perhaps the automatically generated questions were not clear enough to answer.
- Perhaps the students already knew (some of) the words.

[1] Essentially, the logistic regression model seeks to explain a set of observed data by estimating parameters of an equation. The simplest form of a linear regression equation relates two variables x and y with the equation y = ax + b. The equation for our model included terms for FACTOID, ID, ANSWER, and FACTOID*ANSWER, and sought to explain the ratio of correct answers in terms of several variables: whether a word had received help or not (FACTOID), who the student was (ID) to account for student variability, the overall difficulty of the questions (ANSWER), and the effect of help on whether a student got the question right (FACTOID*ANSWER).

5 Learning which words to augment with extra help

While there was no overall effect, was it possible to identify a subset of words for which the extra help was effective? We suspected that students may already have known some of the target words. Therefore, we had two different raters – the experimenter (GSA) and a certified teacher (MBS) – code the words into two categories: words hard enough to give help on, and words not worth giving help on. This coding was done with the raters unaware of the specific outcomes on the vocabulary questions.

For the set of words identified by GSA as hard, the FACTOID*ANSWER results were:

1. All words: N=1486, FACTOID*ANSWER=0.17 +/- 0.11 (not significant)
2. Single-sense words: N=367, FACTOID*ANSWER=0.36 +/- 0.22 (significant at 90%)
3. Non-lexical effect: N=734, FACTOID*ANSWER=0.21 +/- 0.18 (not significant)

For the set of words identified by MBS as hard, the FACTOID*ANSWER results were:

a. All words: N=908, FACTOID*ANSWER=0.23 +/- 0.15 (not significant)
b. Single-sense words: N=203, FACTOID*ANSWER=0.14 +/- 0.31 (not significant)
c. Non-lexical effect: N=425, FACTOID*ANSWER=-0.03 +/- 0.24 (not significant)

Intrigued by these results, we identified a set of words that were rare and thus more likely to be unknown to the students before the experiment began. We chose as the “rare” criterion any word that occurred 15 times or less in the Brown corpus (Kucera and Francis 1967), using the MRC psycholinguistic database for retrieval at http://www.psy.uwa.edu.au/MRCDataBase/uwa_mrc.htm.

For these rare words:

1. All words: N=1753, FACTOID*ANSWER=0.19 +/- 0.10 (significant at 90%)
2. Single-sense words: N=319, FACTOID*ANSWER=0.30 +/- 0.23 (not significant)
3. Non-lexical effect: N=894, FACTOID*ANSWER=-0.04 +/- 0.15 (not significant)

These results should of course be considered suggestive, due to the relatively low (90%) level of confidence, and the variability between different subsets of words examined. However, an overall picture is emerging for when automatically generated factoids may help kids learn vocabulary: Give help on a word if it is rare enough to likely be new to the student – and either give help only if the word is monosemous words, or on the context-appropriate sense of the word if the word is polysemous.

6 Lessons learned

The design, implementation, and analysis of this experiment revealed some important considerations for automating educational experiments. First, experiments designed to be invisible to the student are also invisible to a casual observer. Thus, during implementation subtle bugs may creep in, especially if observing the bug would require observing more than one session. Second, analysis of such experiments should not be left until after the experiment is completed, but should take place in parallel with the experiment design. Third, automated experiments can and should be usefully augmented with other techniques such as firsthand observation and manual exploration of the collected data.

This paper has described a particular invisible experiment that we performed to evaluate vocabulary assistance in Project LISTEN’s Reading Tutor. The experiment was deliberately simple in at least two ways: We compared vocabulary assistance versus none, to see whether it helped at all, and on which words. Also, the experiment wasn't fully invisible; to assess student learning, we inserted multiple-choice vocabulary questions in the Reading Tutor. Subtler invisible experiments to evaluate other aspects of the Reading Tutor will be reported elsewhere [Mostow & Aist, forthcoming chapter in Forbes & Feltovich; Aist & Mostow, manuscript under revision]. The Reading Tutor automatically chose and executed different tutorial behaviors and recorded their effects, which we then analyzed off-line. This experiment is therefore just one step towards a tutor that "closes the loop" by designing, performing, and analyzing its own experiments, and adjusting its behavior based on their results.

Acknowledgements

This paper is based upon work supported in part by the National Science Foundation under Grant Nos. MDR-CDA-9616546, REC-9720348, and REC-9979894, and by the first author's National Science Foundation Graduate Fellowship and Harvey Fellowship. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the National Science Foundation or the official policies, either expressed or implied, of the sponsors or of the United States Government.

We thank current and previous members of Project LISTEN who made substantive contributions to this work. In particular, Cathy Huang and Hua Lan wrote Perl and SQL code to extract information on student-Reading Tutor interactions from log files and import the information into a relational database (SQL Server™) for further analysis. Finally, we thank Principal Joe Fiori and the teachers and students at Centennial Elementary School in McKeesport, PA, who participated in this experiment.

References

1. Aist, G. & Mostow, J., manuscript under revision.
2. Kucera, H and Francis, W. N. (1967). *Computational Analysis of Present-Day American English*, Brown University Press, Providence, Rhode Island, 1967.
3. Juel, C. (1996) What makes literacy tutoring effective? *Reading Research Quarterly* 31(3), pp. 268-289.
4. Lepper, M. R., Woolverton, M., Mumme, D. L., and Gurtner, J. (1993) Motivational techniques of expert human tutors: Lessons for the design of computer-based tutors. In S. P. Lajoie and S. J. Derry (Eds.), Computers as Cognitive Tools, 75-105. Hillsdale, NJ: Erlbaum.
5. Mostow, J. & Aist, G. S. (PUI 1997). When speech input is not an afterthought: A reading tutor that listens. Proceedings of the Workshop on Perceptual User Interfaces. Banff, Canada, Reprinted in Proceedings of the Conference on Automated Learning and Discovery (CONALD98), June 11-13, 1998, Carnegie Mellon University, Pittsburgh, PA.
6. Mostow, J. & Aist, G. S. (CALICO 1999). Giving help and praise in a reading tutor with imperfect listening – because automated speech recognition means never being able to say you're certain. CALICO Journal 16(3), 407-424. Special issue (M. Holland, Ed.), Tutors that Listen: Speech recognition for Language Learning, 1999.
7. Mostow, J., & Aist, G. (2000). Evaluating tutors that listen: An overview of Project LISTEN. In K. Forbes & P. Feltovich (Eds.), forthcoming book on AI and education. AAAI Press.
8. Romano, J. P., and Siegel, A. F. (1986) *Counterexamples in Probability and Statistics*. Wadsworth & Brooks/Cole Statistics/Probability Series.
9. Singh, S., Kearns, M. S., Litman, D. J., and Walker, M. A. (1999) Reinforcement Learning for Spoken Dialogue Systems. In *Proceedings of NIPS*99*. http://www.research.att.com/~diane/nips99.ps
10. Walker, M. A., Litman, D. J., Kamm, C. A., and Abella, A. (1998) Evaluating Spoken Dialogue Agents with PARADISE: Two Case Studies. In Computer Speech and Language, Vol. 12, No. 3, 1998. http://www.research.att.com/~diane/csl98.ps

11. Weisstein, E. W. 2000. Simpson’s Paradox. Topic in *Eric Weisstein’s World of Mathematics*. Wolfram Research. http://mathworld.wolfram.com/SimpsonsParadox.html

See also Project LISTEN’s web page at http://www.cs.cmu.edu/~listen