Can automated questions scaffold children's reading comprehension?

Joseph E. Beck, Jack Mostow, and Juliet Bey[1]

Project LISTEN (www.cs.cmu.edu/~listen)
Carnegie Mellon University
RI-NSH 4213, 5000 Forbes Avenue, Pittsburgh, PA 15213-3890, USA
Telephone: 412-268-1330 voice / 412-268-6436 FAX
{Joseph.Beck, Jack.Mostow}@cs.cmu.edu

Abstract. Can automatically generated questions scaffold reading comprehension? We automated three kinds of multiple-choice questions in children's assisted reading:

1. *Wh-* questions: ask a generically worded *What/Where/When* question.
2. Sentence prediction: ask which of three sentences belongs next.
3. Cloze: ask which of four words best fills in a blank in the next sentence.

A within-subject experiment in the spring 2003 version of Project LISTEN's Reading Tutor randomly inserted all three kinds of questions during stories as it helped children read them. To compare their effects on story-specific comprehension, we analyzed 15,196 subsequent cloze test responses by 404 children in grades 1-4.

- *Wh-* questions significantly raised children's subsequent cloze performance.
- This effect was cumulative over the story rather than a recency effect.
- Sentence prediction questions probably helped (p = .07).
- Cloze questions did not improve performance on later questions.
- The rate of hasty responses rose over the year.
- Asking a question less than 10 seconds after the previous question increased the likelihood of the student giving a hasty response.

The results show that a computer can scaffold a child's comprehension of a text without understanding the text itself, provided it avoids irritating the student.

1 Introduction: Problem and Approach

In 2000, the National Reading Panel [10] sifted through the reading research literature to identify interventions whose efficacy is supported by scientifically rigorous evidence. We focus here on a type of intervention found to improve children's comprehension skills when performed by humans: asking questions. "Teachers ask students questions during or after reading passages of text. […] A question focuses the student on particular content and can facilitate reasoning (e.g., answering why or how)." [10]

[1] Now at University of Southern California Law School, Los Angeles, CA 90089.

Can such interventions be automated? Are the automated versions effective? How can we tell?

We investigate these questions in the context of Project LISTEN's Reading Tutor, which listens to children read aloud, and helps them learn to read [7]. During the 2002-2003 school year, children used the Reading Tutor daily on some 180 Windows™ computers in nine public schools.

The aspect of the 2002-2003 version relevant to this study was its ability to insert questions when children read. The Reading Tutor presented text incrementally, adding one sentence (or fragment) at a time. Before doing so, it could interrupt the story to present a multiple-choice question. It displayed a prompt and a menu of choices, and read them both aloud to the student using digitized human speech, highlighting each menu item in turn. The student chose a response by clicking on it. The Reading Tutor then continued, giving the student spoken feedback on whether the answer was correct, at least when it could tell. We tried to avoid free response typed input since, aside from difficulties in scoring responses, students using the Reading Tutor are too young to be skilled typists. In other experiments students average 30 seconds to type a single word. Requiring typed responses would be far too time-consuming.

This paper investigates three research issues:

- What kinds of automated questions assist children's reading comprehension?
- Are their benefits within a story cumulative or transient?
- At what point do questions frustrate students?

Section 2 describes the automated questions. Section 3 describes our methodology and data. Section 4 reports results for the three research issues. Section 5 concludes.

2 Interventions: Automated Question Insertion

First we had to generate comprehension questions. Good questions should help student comprehension. Skilled personnel might write good questions by hand. However, this approach would be labor-intensive and text-specific. The Reading Tutor has hundreds of stories, totaling tens of thousands of words. Writing good questions for every story, let alone every sentence, would take considerable time, and the questions would not be reusable for new stories.

Natural language understanding might be used to generate questions based on understanding the text. Although this approach might in principle provide good questions for any text, it would require non-trivial development effort to achieve high quality output, efficient performance, and robustness to arbitrary text.

Instead, we eschewed both the "brute force" and "high tech" approaches, and took a "low tech" approach. That is, we looked for ways to generate comprehension questions automatically, but without relying on technology to understand the text.

2.1 Generic wh- questions

Teachers can improve children's reading comprehension by training them to generate questions [10], especially generic *wh*- (e.g. *wh*at, *wh*ere, *wh*en) questions [11]. Accordingly, we developed a few generic questions that we could reuse in (virtually) any context: *Who? What? When? Where? Why? How? So?* Each of these questions is almost always applicable, and very often useful. The last question, short for *So what?*, was suggested by Al Corbett, as a short way to ask the larger significance of the current sentence. Not only should asking these questions stimulate comprehension, but also asking them enough might train students to ask them themselves.

First we had to make the questions usable. Our initial attempts failed, in informative ways. Our first thought was to insert one-word questions to elicit free-form spoken responses, which we would not attempt to recognize; their purpose was to stimulate, not to assess, comprehension. However, not every w*h*- question makes sense in every context. We feared that asking nonsensical questions would confuse children.

We tried to overcome this problem by asking the meta-question, *Click on a question you can answer, or click Back to reread the sentence*: *Who? What*? *When*? *Where? Why*? *How? So?* This approach was a step in the direction of training students to generate questions and would hopefully stimulate children's metacognition.

However, when we "kid-tested" this meta-question at a July 2002 reading lab, children found it too confusing, as evidenced by prolonged inaction or by asking the lab monitor for help. We attributed these difficulties to several problems, which we addressed as follows. To avoid cognitive overload caused by the number of questions, we abandoned the meta-question approach and had the Reading Tutor randomly choose which question to ask. The task was too hard for young children with poor comprehension, so we restricted questions to stories at a grade 3 level or harder; comprehension interventions seldom start before grade 3 [10]. The one-word questions were too short to map clearly to the context, so we rephrased the prompts to make them more explicit, at the suggestion of LISTENer June Sison. The questions were too open-ended to suggest answers, so we changed them to be multiple-choice instead of free-form. Usability testing at an August 2002 reading lab indicated that children understood the revised questions:

What part of the story are you reading now? the end; the beginning; the middle

What has happened so far? a problem has been solved; a mistake; a problem; a problem is being solved; a meeting; an introduction; facts were given; nothing yet; I don't know

Has this happened to you? It happens to me sometimes; It has happened to someone I know; It has never happened to me; This is a silly question!

What could you learn from this? How not to do something; Some new words; How to solve a problem; How to do something; New facts about a subject; A new way of saying something; I don't know

When does this take place? in the present; in the future; in the past; It could happen in the past; I can't tell

Where does this take place? in an apartment; in a house; in an ancient kingdom; anywhere; in outer space; indoors; in a forest; nowhere; on a farm; in the water; outdoors; I can't tell

We also added questions limited to particular genres, e.g., *Who* for fiction. We didn't think of any good generic multiple choice *Why* questions.

2.2 Sentence prediction questions

One way to stimulate or test comprehension of a text is to ask the reader to unscramble it. We operationalized this idea as a sentence prediction task in the form of the multiple-choice question *Which will come next?* The three response choices were the next three sentences of the story, in randomized order.

The sentence prediction task had an advantage over the generic *wh-* questions in that the Reading Tutor knew which answer was correct. This information enabled it to give immediate feedback by saying (in recorded human speech) either *Way to go!* or *Not quite.*

2.3 Cloze questions

A third kind of multiple-choice question was a "cloze" (fill-in-the-blank) prompt generated from a story sentence by deleting a word, e.g. *Resources such as fish are renewable, as long as too many are not taken or ______. coral; damage; market; destroyed.* The choices consisted of the missing word plus three distractor words chosen randomly from the same story, but so as to have the same general type as the correct word:

- "sight" words (the most frequent 225 words in a corpus of children's stories)
- "easy" words (the top 3,000 except for sight words)
- "hard" words (the next 22,000 words), and
- "defined" words (words explicitly annotated with explanations).

The Reading Tutor automatically generated, inserted, and scored such multiple choice cloze questions [8]. The 2002 study had used these automatically generated questions to assess comprehension. Students' performance on such questions predicted their performance on the vocabulary and comprehension subtests of the Woodcock Reading Mastery Test with correlations better than 0.8.

This study also found careless guessing, indicated by responding sooner than 3 seconds after the prompt. In an attempt to reduce guessing, we modified cloze questions to provide explicit feedback on correctness: *Alright, good job!* or *I don't think so*.

Brandão & Oakhill [4] asked children *Do you know why?* to probe – and stimulate – their comprehension. We adapted this question to follow up cloze questions on "defined" words. After a correct answer, the Reading Tutor added, *That's right! Do you know why?* If not, or after an incorrect answer, it asked, *Which phrase fits better in the sentence?* The two choices were short definitions of the correct word and a distractor.

3 Methodology: Experimental Design, Data, and Analysis

The next problem was how to tell if asking automated questions improved students' comprehension. We couldn't simply test whether their comprehension improved over time, because we expected it to improve as a consequence of their regular classroom instruction. A conventional between-subjects experiment would have compared two versions of the Reading Tutor, one with the questions and one without, in terms of their impact on students' gains in comprehension skills. However, such experiments are costly in time, personnel, and participants.

Project LISTEN had previously addressed this difficulty by embedding within-subject experiments in the Reading Tutor to evaluate various tutorial interventions. For example, one experiment evaluated the effect of vocabulary assistance by randomly explaining some words but not others, and administering multiple choice questions the next day to see if students did better on the explained words [1]. These experiments assumed that instruction on one word was unlikely to affect performance on another word – i.e., that vocabulary knowledge can be approximated as a collection of separately learned atomic pieces of knowledge that do not transfer to each other.

In contrast, instruction on a general comprehension skill violates this non-transfer assumption. We therefore decided to look for *scaffolding* effects instead of *learning* effects. Students who are ready to benefit from comprehension strategies but have not yet internalized them should comprehend better with the intervention than without it. We therefore look for a difference between assisted and unassisted performance.

As in [8], we segmented student-tutor interaction sequences into episodes with measurable local outcomes. We hypothesized that if the intervention were effective, students would perform better on cloze questions for awhile thereafter – for how long, we didn't know; perhaps the next few sentences.

3.1 Within-subject randomized-dosage experimental manipulation

The question-asking experiment operated as follows. Before each sentence, the Reading Tutor randomly decided whether to insert a question, and if so, of what kind. Thus the number and kinds of questions varied randomly from one story reading to another. The Reading Tutor inserted questions only in new stories, not in stories students were rereading, where they might therefore remember answers based on prior exposure.

The three kinds of questions differed slightly in when they could occur. Such differences between experimental conditions can introduce bias if not properly controlled. To avoid confusing poor readers, the Reading Tutor inserted *wh-* questions, sentence prediction questions, and "defined word" cloze questions only in stories at and above level C (roughly grade 3). However, it asked other cloze questions in stories at all levels. Also, some *wh-* questions were genre-specific. For example, the Reading Tutor inserted *Who* questions in fiction, which it could assume had one or more characters, but not in non-fiction and poetry, which can violate that assumption.

To avoid sample bias we needed to compare data generated under the same conditions. For example, it would be unfair to compare fiction-specific *wh-*questions to null interventions in other genres. We therefore excluded data from stories below level C and genre-specific *wh-* questions, leaving "3W": *what, when,* and *where.*

3.2 Data set

The data set for this paper came from eight public schools that used the Reading Tutor throughout the 2002-2003 school year, located in four Pittsburgh-area school districts, urban and suburban, low-income and affluent, African-American and Caucasian. Reading Tutors at each school used a shared database on a server at that school. Each night these servers sent the day’s transactions back via the Internet to our lab to update a single aggregated database. We mapped research questions onto MySQL queries as described in [9]. We used SPSS and Excel to analyze and visualize query results. A bug in the Reading Tutor’s mechanism for assigning students to appropriate story levels affected data for fall 2002 [3], so we restricted our data set to the 2003 data.

Of 404 users the Reading Tutor logged as having read stories at level C or higher in 2003, 252 students had moderate usage – that is, at least one hour and at least 10 sentences. There were 56 first-graders, 96 second-graders, 50 third-graders, 17 fourth-graders, and 33 students for whom we did not know their grade.

The data set includes a total of 23,372 questions, consisting of 6,720 3W questions, 1,865 sentence prediction questions, and 15,187 cloze questions. Table 1 shows the mean and maximum number of questions of each kind seen by each student. The minimum is not shown because it was zero for each kind. While reading new stories at levels C-G (approximately grades 3-7), students were asked a 3W, prediction, or cloze question about once every 4 minutes or 10 sentences, on average.

Table 1. Questions asked (per-student mean and maximum)

	Number		**Per minute**		**Per sentence**	
	Mean	Max	Mean	Max	Mean	Max
3W	24.9	164	0.0865	0.60	4%	19%
Prediction	7	40	0.0235	0.12	1%	6%
Cloze	35.7	278	0.1232	0.60	5%	24%
All	67.7	472	0.2332	1.20	10%	40%

3.3 Cloze performance as outcome variable

To measure students’ fluctuating comprehension of stories as they read, we used available data – their responses to the inserted questions. We did not know which answers to the *wh-* questions were correct, some questions can have multiple correct answers, and some questions could not even be scored by a human rater (e.g. *Has this happened to you?*).

The sentence prediction and cloze questions were both machine-scorable. In fact the Reading Tutor gave students immediate feedback on responses to them. But did they really measure comprehension?

To make sure, we validated students’ performance on each kind of question against their Passage Comprehension pretest scores. Performance on sentence prediction questions averaged only 41% correct. To test their validity as a measure of comprehension, we correlated this percentage against students’ posttest Passage Comprehension, excluding students with fewer than 10 non-hasty sentence prediction responses. The correlation was only 0.03, indicating that they were not a valid test of comprehension. In contrast, Mostow *et al.* [8] had already shown that performance on automated cloze questions in the 2001-2002 version of the Reading Tutor predicted Passage Comprehension at R=.5 for raw % correct, and at R=0.85 in a model that included the effects of item difficulty of story level and word type. We didn’t regenerate such a model for the 2003 data, but we confirmed that it showed a similar correlation of raw cloze performance to test scores.

Note that the same cloze question operated both as an intervention that might scaffold comprehension, and as a local outcome measure of the preceding interventions. We use the terms "cloze intervention" and "test question" to distinguish these roles.

Figure 1 shows the number of recent interventions before 15,196 cloze test items. We operationalize "recent" as "within the past two minutes," based on our initial analysis, which suggested a two-minute window for effects on cloze performance.

3.4 Logistic regression model

To test the effects of 3W, prediction, and cloze interventions on students' subsequent comprehension, we constructed a logistic regression model [6] in SPSS to predict the correctness of their responses to test questions.

To control for differences between students, we included student identity as a factor in the model. Omitting student identity would ignore statistical dependencies among the same student's performance on different items. Including student identity as a factor accounts for statistical dependencies among responses by the same student, subject to the assumption that responses are independent given the ability of the student and the difficulty of the item. This "local independence" assumption is justified by the fact that each test question was asked only once, and was unlikely to affect the student's answers to other test questions. We neglect possible dependency among test responses caused by a common underlying cause such as story difficulty.

To control for differences in difficulty of test questions, the model included the type of cloze question, according to the type of word deleted -- "sight," "easy," "hard," or "defined". An earlier study [8] had previously found that word type significantly affected cloze performance.

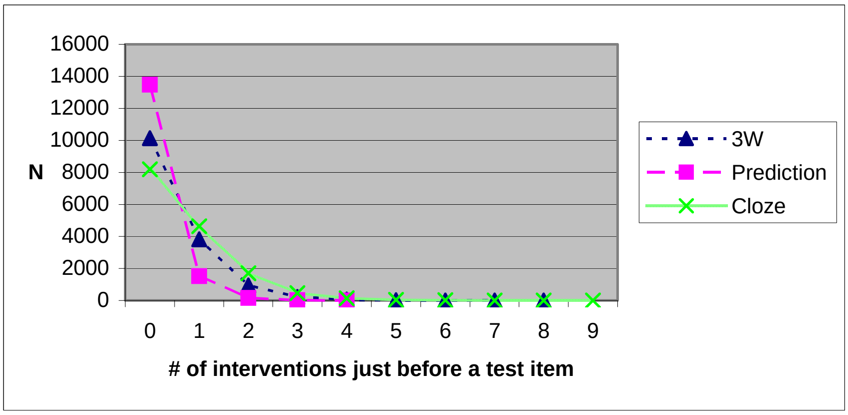

Figure 1. Histogram of # recent interventions

To represent cumulative effects of different types of questions, our model included as separate covariates the number of 3W, prediction, and cloze interventions, since the start of the current story. Our initial analysis had suggested that cloze performance was higher for two minutes after a 3W question. To model such recency effects, we added similar covariates for the number of 3W and cloze interventions in the two minutes preceding the current test question.

However, we treated recent sentence prediction questions differently, because they revealed the next three sentences, thereby giving away the answer to test questions on those sentences. To exclude such contamination, we screened out from the data set any test response closely preceded by a sentence prediction question. Consequently, our model had no covariate for the number of recent sentence prediction questions, because it was always zero.

The model included three covariates to represent possible temporal effects at different scales. To model improvement over the course of the year, we included the month when the question was asked. To model changes in comprehension over the course of the story, we included the time elapsed since the story started. To model effects of interruption, we included the time since the most recent Reading Tutor question.

4 Results

Table 2 shows which predictor variables in the logistic regression model affected cloze test performance. As expected, student identity and test question type were highly significant. The beta value for a covariate shows how an increase of 1 in the value of the covariate affects the log odds of the outcome. Thus the increasingly negative beta values for successive test question types reflect their increasing difficulty. These beta values are not normalized and hence should not be compared to measure effect size. The *p* values give the significance of each predictor variable after controlling for the other predictors.

Table 2. Logistic regression model

	Beta	*p*
Student identity	.	0.000
Type of (cloze) test question	0	0.000
Sight	0	
Easy	-0.03	
Hard	-0.19	
Defined	-1.04	
# 3W questions	0.05	0.023
# sentence prediction questions	*0.08*	*0.072*
# cloze questions	-0.005	0.765
# recent 3W questions	*-0.07*	*0.074*
# recent cloze questions	0.02	0.548
Time of year (month)	-0.01	0.551
Time since start of story (minutes)	-0.013	0.137
Time since last intervention (sec)	-0.001	0.036

4.1 What kinds of questions assisted children's reading comprehension?

According to the logistic regression model, 3W questions had a positive effect (beta = .05, p = .023) and sentence prediction had a possible effect (beta = .08, p = *.072*). Cloze interventions had no effect (beta = -.005, p = .765), lending credence to our local independence assumption. These results cannot be credited simply to the time spent so far reading the story, which had a *negative* though insignificant effect (beta = -.013, p = .137) on cloze performance. We conclude that *3W questions boosted comprehension enough to outweigh the cost of disrupting reading*.

Generic questions force readers to carry more of the load than do text-specific questions. Is this extra burden on the student's working memory worthwhile [5] or a hindrance [2]? Generic 3W questions, which let students figure out how a question relates to the current context, had a positive effect. Cloze interventions, which are sentence-specific and more explicitly related to the text, did not.

What about feedback? One might expect questions to help more when students are told if their answers are correct. One reason is cognitive: the feedback itself may improve comprehension by flagging misconceptions. Another reason is motivational: students might consider a question more seriously if they receive feedback.

Despite the lack of such feedback, 3W questions bolstered comprehension of later sentences. Despite providing such feedback, cloze interventions did not help. Evidently the advantages of 3W questions sufficed to overcome their lack of feedback.

4.2 Were the benefits within a story cumulative or transient?

We had previously [3] considered only the effect of an intervention on the very next test item. Our logistic regression model now revealed the effect of recent 3W questions was actually negative, and only marginally significant. Recent cloze interventions had no effect. In summary, the *benefits of 3W questions were cumulative.*

Figure 2 shows how cloze performance varied with the number of preceding questions of each type. To reduce noise, cases with fewer than 30 observations are omitted. The y values are raw % correct, not adjusted for any of the logistic regression variables, so they must be interpreted with caution, but suggest that 3W beats cloze after 4 questions.

4.3 At what point did questions frustrate students?

The temporal portion of the logistic regression model shows that cloze performance fell over the year, over the story, and (significantly) right after an intervention. Why?

Figure 3 shows how the blowoff rate changed after any Reading Tutor intervention. The x-axes show the time in seconds since the previous question. As the axis labels reflect, we binned into 2-second intervals. The blowoff rate spiked at nearly 90% for cloze questions asked too soon. Within 20 seconds, the blowoff rate decayed back to an asymptotic level of about 12%.

We analyzed how often students avoided answering questions. The "blowoff rate" measured the percentage of hasty responses. Prior analysis of cloze questions [8] had shown that students who responded in less than 3 seconds performed at or near chance level and were probably not seriously considering the question. The blowoff rates in 2003 were 23% for 3W questions, 12% for sentence prediction, and 11% for cloze (computed not just on the subset used in the logistic regression). The higher blowoff rate for 3W questions might be due to their lack of immediate feedback. As Table 3 shows, the overall percentage of hasty cloze responses rose over time.

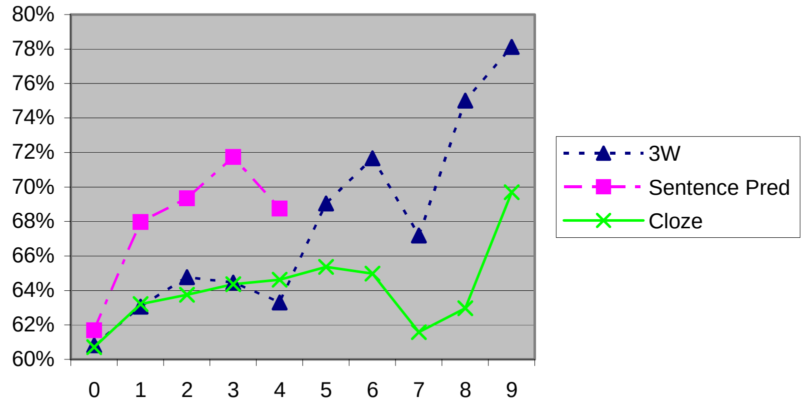

Figure 2. Cloze performance versus number of preceding questions of each type

Table 3:. Changing rate of hasty cloze responses in spring 2003

Month	Blowoff rate	# test items
Jan	9.9%	3255
Feb	11.4%	3224
Mar	16.0%	3889
Apr	15.7%	3120
May	18.7%	1612

In summary, *frustration with inserted questions, as measured by how often students responded too hastily to give them careful thought, rose over the course of the year and spiked when one question followed another by less than 10 seconds.*

5. Conclusion: Contributions and Lessons

This paper contributes interventions, evaluation, and methodology.

We reported three automatic ways to ask multiple-choice comprehension questions. Developing these methods involved adapting, user-testing, and generalizing methods used by human teachers. Generic *wh-* questions adapt a method found effective by the National Reading Panel. Sentence prediction questions resemble manually created unscrambling tasks. We augmented a previously reported method [8] for cloze question generation, adding feedback and *Do you know why?* follow-up probes.

We evaluated the effect of these questions on student comprehension as measured by subsequent cloze test questions. The 3W questions we evaluated had a significant positive effect, which was cumulative rather than a recency effect. The sentence prediction questions had a probable effect, and the cloze questions had no effect. Future work should study how effects vary by student level, text difficulty, and question type.

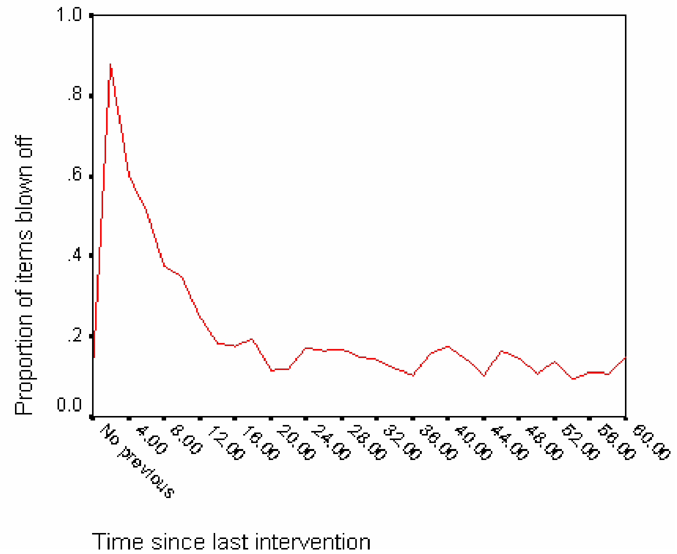

Figure 3. Blowoff rate versus time (in seconds) since previous question

We analyzed student frustration as shown by hasty responses. Such avoidance behavior was likelier when less than 10 seconds elapsed between questions.

Our evaluation methodology incorporated an interesting approach to the challenge of evaluating the effects of alternative tutorial interventions. The within-subject design avoided the sample size reduction incurred by conventional between-subjects designs. The randomized dosage explored the effects of different amounts of each intervention. The logistic regression model controlled for variations in students, item difficulty, and time.

Our analyses illustrate some advantages of networked tutors and storing student-tutor interactions in a database. The ability to easily combine data from many students and analyze information as recent as the previous day is very powerful. Capturing interactions in a suitable database representation makes them easier to integrate with other data and to analyze [9].

One theme of this research is to focus the AI where it can help the most, starting with the lowest-hanging fruit. Rather than trying to generate sophisticated questions or understand children's spoken answers, we instead focused on when to ask simpler, generic questions. There are many ways to apply language technologies to reading comprehension. However, what ultimately matters is the student's reading comprehension, not the computer's. The Reading Tutor cannot evaluate student answers to some types of questions it asks, but can nevertheless assist students' comprehension.

Using the analysis methods presented here may one day enable it to measure in real-time the effects of those questions.

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 or the official policies, either expressed or implied, of the sponsors or of the United States Government. We thank the students and educators at the schools where Reading Tutors recorded data, and other members of Project LISTEN who contributed to this work.

References

1. Aist, G., *Towards automatic glossarization: Automatically constructing and administering vocabulary assistance factoids and multiple-choice assessment.* International Journal of Artificial Intelligence in Education, 2001. **12**: p. 212-231.
2. Anderson, J.R., *Rules of the mind*. 1993, Hillsdale, NJ: Lawrence Erlbaum Associates.
3. Beck, J.E., J. Mostow, A. Cuneo, and J. Bey. *Can automated questioning help children's reading comprehension?* in *Proceedings of the Tenth International Conference on Artificial Intelligence in Education (AIED2003)*. 2003.p. 380-382 Sydney, Australia.
4. Brandão, A.C.P. and J. Oakhill. *"How do we know the answer?" Children's use of text data and general knowledge in story comprehension.* in *Society for the Scientific Study of Reading 2002 Conference*. 2002.p. The Palmer House Hilton, Chicago.
5. Kashihara, A., A. Sugano, K. Matsumura, and T. Hirashima. *A Cognitive Load Application Approach to Tutoring*. in *Proceedings of the Fourth International Conference on User Modeling*. 1994.p. 163-168.
6. Menard, S., *Applied Logistic Regression Analysis.* Quantitative Applications in the Social Sciences, 1995. **106**.
7. Mostow, J. and G. Aist, *Evaluating tutors that listen: An overview of Project LISTEN*, in *Smart Machines in Education*, K. Forbus and P. Feltovich, Editors. 2001, MIT/AAAI Press: Menlo Park, CA. p. 169-234.
8. Mostow, J., J. Beck, J. Bey, A. Cuneo, J. Sison, B. Tobin, and J. Valeri, *Using automated questions to assess reading comprehension, vocabulary, and effects of tutorial interventions.* Technology, Instruction, Cognition and Learning, to appear. **2**.
9. Mostow, J., J. Beck, R. Chalasani, A. Cuneo, and P. Jia. *Viewing and Analyzing Multimodal Human-computer Tutorial Dialogue: A Database Approach*. in *Proceedings of the Fourth IEEE International Conference on Multimodal Interfaces (ICMI 2002)*. 2002.p. 129-134 Pittsburgh, PA: IEEE.
10. NRP, *Report of the National Reading Panel. Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction*. 2000, National Institute of Child Health & Human Development: Washington, DC.
11. Roshenshine, B., C. Meister, and S. Chapman, *Teaching students to generate questions: A review of the intervention studies.* Review of Educational Research, 1996. **66**(2): p. 181-221.