Can automated questioning help children's reading comprehension?

Joseph E. BECK, Jack MOSTOW, Andrew CUNEO, and Juliet BEY
Project LISTEN (www.cs.cmu.edu/~listen)
Robotics Institute
Carnegie Mellon University
Pittsburgh, PA 15213-3890, USA

Abstract. We present an automated method to ask children questions during assisted reading, and experimentally evaluate its effects on their comprehension. In 2002, after a randomly inserted generic multiple-choice *What/Where/When* question, children were likelier to correctly answer an automatically generated comprehension question on a later sentence. The positive effects of such questions vanished during the second half of the study in 2003. We hypothesize why.

1. Introduction: Problem and Approach

Teachers can improve children's reading comprehension by training them to generate questions [1], especially generic *wh-* (e.g. *Wh*at, *wh*ere, *wh*en) questions [2]. We describe and evaluate automated scaffolding for this skill in Project LISTEN's Reading Tutor [3].

The aspect of the Reading Tutor most relevant to this study is its ability to insert questions when children read. The Reading Tutor displays a story incrementally, adding one sentence (or fragment) at a time. Before doing so, it can interrupt the story to present a multiple choice question. It displays a prompt and a menu of choices, and reads them both aloud to the student, highlighting each menu item in turn. The student chooses an item by clicking on it. The Reading Tutor then proceeds, optionally giving the student spoken feedback on whether the answer was correct, at least in cases where it can tell.

We cast the generic *wh-* questions in multiple-choice form, e.g. *When does this take place? in the present; in the future; in the past; It could happen in the past; I can't tell.* User tests showed that children understood them better than shorter, less explicit questions.

To test the scaffolding effects of such questions on children's reading comprehension, we measured their performance on story-specific questions asked shortly thereafter. For this purpose, we used multiple-choice "cloze" (fill-in-the-blank) questions generated automatically from a story sentence by deleting a word. The choices consist of the missing word plus three distractor words. E.g. *Why bother about _____?* Choices: *food; winter; dying; passed.*

The distractor words are chosen randomly from the same story, but constrained to have the same general type as the correct word: "sight" words (the most frequent 225 words in a corpus of children's stories analyzed by former LISTENer Greg Aist), "easy" words (the top 3000 except for sight words, "hard" words (the next 22,000 words), and "defined" words (words explicitly annotated with explanations). A previous study [4] showed that these four types of questions are successively harder, and that children's performance on them predicted their performance on standard measures of *general* comprehension ability with correlations surpassing 0.8. We hypothesized that if a *wh*-question assisted comprehension of the *specific* text at hand, it would make the reader likelier to answer the next cloze question correctly.

Thus our experimental design was as follows. The randomized experimental manipulation was to occasionally insert a *wh-* question or a cloze question of any type (sight, easy, hard, defined). Each randomly inserted cloze question defined the outcome of one trial. The independent variable was the intervention immediately preceding the cloze question. The intervention could be a *wh-* question; null, if the cloze question was the first question inserted in the story; or another cloze question. Thus a cloze question could be the intervention for a trial as well as the outcome of the previous trial. Hereafter we will use "cloze intervention" and "test question" to distinguish these two roles.

The purpose of the *wh-* questions was not to *assess* comprehension, but to *assist* it. If test question performance was higher after *wh-* questions, we could infer that they helped students comprehend. We wouldn't know if they were improving students' comprehension over time, but we'd have evidence of near transfer in the sense of improved performance on nearby sentences – that is, past the point in the text where the *wh-* question was inserted.

As of the 2002-2003 school year, 216 Reading Tutors were used daily in nine public schools by 427 children in grades K-4 (typically ages 5-10) and sent each day's transactions back at night via Internet to our lab to update a single aggregated database, enabling us to formulate research questions as MySQL queries, analyze the results in SPSS, and visualize them in Excel.

On December 18, 2002 we fixed a bug that was causing the Reading Tutor's student model to promote some students to stories too hard for them. Therefore, we split the test questions into those occurring before December 18 and those occurring after January 5, 2003 (to give time after the patch for the student model to correct itself). *Wh-* questions can only occur in stories that are grade level 3 or higher. Of the 427 students using the Reading Tutor, 288 of them saw a cloze question in a story rated level 3 or higher. Before the December 18 patch, for stories that could contain cloze and *wh-* questions, the Reading Tutor selected stories at an average grade level of 5.2 (second month of the fifth grade, which is appropriate for students roughly 10 years old). After January 5, this level decreased on average to 3.9.

2. Analysis and Future Work

We compare student performance on test questions preceded by no intervention, those preceded by a *wh-* question, and those preceded by a cloze intervention. Since story level affects ability to answer test questions, we disaggregate the data by whether the question occurred before or after the patch. We also disaggregate by test question type. Table 1 shows how performance varies by time of year, type of intervention, and type of cloze question. Cell means are well-estimated, with N ranging from 99 to 528 (mean of 251).

Table 1. Average proportion correct on cloze items disaggregated by time and type of cloze question

	Before December 18			After January 5		
Test question type	Null	*wh-*	Cloze	Null	*wh-*	Cloze
Sight	0.61	0.62	0.68	0.66	0.65	0.64
Easy	0.56	0.59	0.57	0.63	0.66	0.66
Hard	0.50	0.56	0.52	0.61	0.54	0.65
Defined	0.36	0.38	0.38	0.39	0.43	0.46

One general pattern in Table 1 is a general monotonic decrease in performance as cloze questions get more difficult. Cloze questions involving sight words are easiest, those involving words labeled as specifically needing to be defined for the student are hardest. Performance after January 5 is higher than for questions before December 18.

Whether *wh-* questions had an impact is a more complex issue. Before December 18, *wh-* questions had a distinct advantage over test questions without a preceding intervention. After January 5, *wh-* questions showed no advantage over test questions without a preceding intervention. We are not sure what caused this perceived change in effectiveness, and are still gathering and analyzing data. We have 3 hypotheses:

1. **The null hypothesis.** Perhaps the advantage *wh-* questions had over questions with no intervention was simply a statistical fluke? We have roughly twice as much data from after January 5 as from before December 18, so the lack of effect in the newer, larger dataset suggests there is no benefit to *wh-* questions. However, the change in the Reading Tutor's behavior with modeling students and selecting stories makes this conclusion only possible, not certain.
2. ***Wh-* questions provide context-dependent scaffolding.** Scaffolding [5] is effective when students are working on material that is too complex for them to solve alone. In the first half of the year, stories were more challenging and students could have had more difficulty understanding the material. In this context, *wh-* questions provided needed support. After January 5, when stories were chosen at an appropriate level of difficulty, the support provided by the *wh-* questions was not needed. To explore this idea, we have data from a variety of story-choice strategies: 1) in 2002 stories were too difficult, 2) in 2003 the Reading Tutor presumably selected appropriate stories, and 3) due to the Reading Tutor's turn-taking mechanism half of the stories are selected by the student. Analyzing how student performance after *wh-* prompts varies based on story level, **relative to student reading level** as assessed by paper tests, provides a method to determine whether the scaffolding hypothesis is true. If so, the effect of *wh-* prompts should be strongest for stories that are challenging for the student. We do not yet have paper-test data in analyzable form.
3. **Students have learned how to incorporate the *wh-* strategies.** It is possible that *wh-* prompts initially help students through scaffolding, but students learn from the prompts and begin using those comprehension strategies while reading. Once students have internalized using the *wh-* prompts, there is no further benefit to displaying them. This hypothesis accounts for the general improvement in performance on test items over time. It also accounts for the lack of a measured effect of *wh-* prompts in 2003: students have internalized the strategies so there is no measurable effect of further presentation. We can determine if learning effects are responsible by constructing curves examining test item accuracy vs. number of *wh-* prompts.

Acknowledgements: This work was supported by NSF under IERI Grant REC-9979894. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[1] 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.

[2] Roshenshine, B., Meister, C. and Chapman, S., Teaching students to generate questions: A review of the intervention studies. *Review of Educational Research*, 1996. **66**(2): p. 181-221.

[3] Mostow, J. and Aist, G., Evaluating tutors that listen: An overview of Project LISTEN, in *Smart Machines in Education*, P. Feltovich, Editor. 2001. p. 169-234.

[4] Mostow, J., Tobin, B. and Cuneo, A., Automated Comprehension Assessment in a Reading Tutor. Proceedings of the *Proceedings of the ITS 2002 Workshop on Creating Valid Diagnostic Assessments*. p. 52-63. 2002

[5] Doolittle, P.E., Vygotsky's Zone of Proximal Development as a Theoretical Foundation for Cooperative Learning. *Journal on Excellence in College Teaching*, 1997. **8**(1): p. 83-103.