Questions and Answers about Questions and Answers: Lessons from generating, scoring, and analyzing questions in a Reading Tutor for children

Jack Mostow
Project LISTEN (www.cs.cmu.edu/~listen)

AAAI Symposium on Question Generation keynote, Nov. 5, 2011, Arlington, VA

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grants R305A080628. The opinions expressed are those of the authors and do not necessarily represent the views of the Institute or the U.S. Department of Education.
Questions about Questions

**Target:** What does it take to answer the questions?

**Purpose:** Why ask the questions?

**Question type:** In what form will the questions be output?

**Answer type:** In what form will responses be input?

**Generation:** How construct questions, answers, distracters?

**Modality:** What channels will convey questions and answers?

**Assessment:** How score answers? How generate feedback?

**Evaluation:** How to tell how well questions serve purpose?
What can questions target?

Reading
- Decoding: In “Word Swap,” click on the “misread” word [Zhang ITS 08]

The lid sat in the loft.

was

Vocabulary
- Recall: Which means most like [word] [Aist 01]
- Recognize: Which word means [definition] [TICL 04]
- Remind: Definition cloze [Gates QG 11]
- Disambiguate: What does [word] mean here?

Knowledge
- Fact: Noiz arte du nahiduenak iritzia emateko [Aldabe QG 11]
- Skill: How many grams can a worker ant carry? [Williams QG 11]
What can questions target?

Reading
- Decoding: In “Word Swap,” click on the “misread” word [Zhang ITS 08]
- Comprehension:  Click on the missing word.  [Mostow TICL 04]

Vocabulary
- Recall: Which means most like <word>? [Aist 01]
- Recognize: Which word means <definition>? [TICL 04]
- Remind: Definition cloze [Gates QG 1]
- Disambiguate: What does <word> mean here? 

Knowledge
- Fact: Noiz arte du nahi duenak iritzia emateko aukera? [Aldabe QG 11]
- Skill: How many grams can a worker ant carry? [Williams QG 1]
- Concept: …

"I am helping to lay up food for the winter," said the Ant, "and ______ you to do the same."

- bother
- recommend
- chat
- grasshopper
What can questions target?

Reading

- Decoding: In “Word Swap,” click on the “misread” word [Zhang ITS 08]
- Comprehension: Click on the missing word. [Mostow TICL 04]
- Inter-sentential prediction: Which will come next? [Beck ITS 04]
- Monitor: Did that make sense?
- Self-question: Why was the country mouse surprised? [Chen AIED 09]
- Disengagement: hasty guessing [Beck AIED 05]

Vocabulary

- Recall: Which means most like <word>? [Aist 01]
- Recognize: Which word means <definition>? [TICL 04]
- Remind: Definition cloze [Gates QG 1]
- Disambiguate: What does <word> mean here?

Knowledge

- Fact: Noiz arte du nahi duenak iritzia emateko aukera? [Aldabe QG 11]
- Skill: How many grams can a worker ant carry? [Williams QG 1]
- Concept: …

Jack Mostow keynote 11/5/11
What can questions target?

Reading
- Decoding: In “Word Swap,” click on the “misread” word [Zhang ITS 08]
- Comprehension: Click on the missing word. [Mostow TICL 04]
- Inter-sentential prediction: Which will come next? [Beck ITS 04]
- Monitor: Did that make sense?
- Self-question: Why was the country mouse surprised? [Chen AIED 09]
- Disengagement: hasty guessing [Beck AIED 05]

Vocabulary
- Recall: Which means most like <word>? [Aist 01]
- Recognize: Which word means <definition>? [TICL 04]
- Remind: Definition cloze [Gates QG 11]
- Disambiguate: What does <word> mean here?

Knowledge
- Fact: Noiz arte du nahi duenak iritzia emateko aukera? [Aldabe QG 11]
- Skill: How many grams can a worker ant carry? [Williams QG 11]
- Concept: ...
How has Project LISTEN used questions?

1. Assess comprehension [Mostow et al., TICL 04]
2. Help comprehension [Beck et al., ITS 04]
3. Assess engagement [Beck, AIED 05]
4. Teach self-questioning [Mostow & Chen, AIED 09]
5. Model self-questioning [Chen et al., ITS 10]
6. Assess self-questioning [Chen et al., QG 11]
1. Assess comprehension [TICL 04]

Target: comprehend sentence
Purpose: assess comprehension while reading
Source: sentence in text
Question type: cloze
Answer type: multiple choice among 4 words from text
Generation: randomly pick sentence, word, and distracters
Modality: play recorded sentence and words; click on one
Assessment: original word? immediate correctness feedback
Evaluation: correlate against standard comprehension test
In a field one summer's day a Grasshopper was hopping about, chirping and singing to its heart's content.
"Why not come and chat with me," said the Grasshopper,
"instead of toiling and moiling in that way?"
"I am helping to lay up food for the winter," said the Ant, "and ______ you to do the same."

- bother
- recommend
- chat
- grasshopper

- Reading Tutor reads question and choices aloud
- Target and distracters are words from same story
"I am helping to lay up food for the winter," said the Ant, "and recommend you to do the same."
What do cloze questions test?

"Why bother about ______?"

food

winter

dying

passed
Oct.01 – Mar.02 evaluation data

Reading Tutor asked 69,326 automated cloze questions

- 364 students in grades 1-9 at 7 schools
- 20-758 questions per student (median 136)
- 98% of questions answered – else Goodbye, Back, or timeout
- 24-88% correct (median 61%); 25% = chance

How much guessing? Hasty responses [J. Valeri]:

- 3,078 (4.5%) faster than 3 seconds, only 29% correct
- 3.9% per-student mean, but below 1% for most students
- Guessing rose from 1% in October to 11% by March
Reliability: Guttman split-half test

Split N responses into two halves
- Match by word and story level

Reliability increases with N
- .83 for N≥10 (338 students)
- .95 for N≥80 (199 students)
What affects cloze difficulty?

Similarity of distractors to answer
- Part of speech [Hensler & Beck, ITS 06]
- Semantic class
- Consistency with local context
- Consistency with inter-sentential context

Vocabulary level of answer and distracters
- “Sight words” = 225 most frequent words of English [Dolch list]
- “Easy words” = 226...3,000
- “Hard words” = 3,001...25,000
- “Defined words” = marked as warranting explanation

Text level of story
- Grade K, 1, 2, 3, 4, 5, 6, 7

Cloze performance at 4 word levels x 8 text levels predicts Woodcock Reading Mastery Test comprehension (R = .84)
2a. Help comprehension [ITS 04]

**Target:** comprehend text by questioning

**Purpose:** scaffold comprehension while reading

**Source:** none

**Question type:** *Wh-*

**Answer type:** multiple choice

**Generation:** scripted generic questions and choices

**Modality:** play recorded prompt and words; click on one

**Assessment:** none

**Evaluation:** test efficacy on ensuing cloze performance
Generic *Wh*-questions: initial

Generic prompt (meta-question):
- *Click on a question you can answer, or click Back to reread the sentence*

Generic 1-word questions as choices:

Evaluation: failed 2002 user test
- Meta-question confusing
- 1-word questions too vague to map to text
Generic *Wh*-questions: revised

What has happened so far?

- facts were given
- a problem is being solved
- a problem has been solved
- a problem
- an introduction
- a mistake
- a meeting
- nothing yet
- I don't know
- nothing yet
- I don't know
- nothing yet
- I don't know
- nothing yet
2b. Help comprehension [ITS 04]

Target: comprehend text by predicting
Purpose: scaffold comprehension while reading
Source: none

Question type: Which will come next?
Answer type: multiple choice
Generation: answer = next sentence; distracters = 2 following
Modality: play recorded prompt and sentences; click on one
Assessment: original sentence? immediate correctness feedback

Evaluation: 41% right; test efficacy on ensuing cloze performance
Sentence prediction

Joe Valeri has read 3.0 minutes today. Joe (level C) has read "Sentence Prediction Intervention" (level 0) 0 times, and 2478 words and 10 stories as

Which will come next?

"One doesn't get the offer of a beautiful princess and half a kingdom every day."
He might have a hundred horsemen, too, to back him up.
"That's the very thing for a man like me," thought the little tailor.
2003 evaluation data

Reading Tutor asked 23,372 randomly inserted questions
- 252 students in grades 1-4+? who used for an hour or more
- 6,720 generic *what, when, where*
- 1,865 *Which will come next?* sentence prediction
- 15,187 cloze
- On **average**, one question every 10 sentences
# Evaluation results

<table>
<thead>
<tr>
<th>Variable</th>
<th>Helps/Hurts</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td># 3W questions</td>
<td>☻</td>
<td>0.023</td>
</tr>
<tr>
<td># sentence predictions</td>
<td>☻</td>
<td>0.072</td>
</tr>
<tr>
<td># cloze questions</td>
<td>=</td>
<td></td>
</tr>
<tr>
<td># recent 3W questions</td>
<td>☹</td>
<td>0.074</td>
</tr>
<tr>
<td># recent cloze questions</td>
<td>=</td>
<td></td>
</tr>
<tr>
<td>Time since prior question (sec)</td>
<td>☻</td>
<td>0.036</td>
</tr>
<tr>
<td>Time since start of story (sec)</td>
<td>☹</td>
<td>0.14</td>
</tr>
</tbody>
</table>
Effect of recent questions

- Response time < 3 sec indicates disengagement
- Use to track (dis-)engagement [Beck AIED 05]
Effect of question type

![Graph showing the effect of question type on cloze performance. The x-axis represents the number of prior questions, and the y-axis represents cloze performance percentage. The graph includes lines for 3W, Sentence Pred, and Cloze, showing an increase in performance with more prior questions.](image-url)
# QG costs (lowest → highest)

<table>
<thead>
<tr>
<th>Generation</th>
<th>None</th>
<th>Trivial</th>
<th>Simple</th>
<th>Complex</th>
<th>Manual</th>
</tr>
</thead>
<tbody>
<tr>
<td>Question</td>
<td>Generic</td>
<td>Cloze</td>
<td>NLP</td>
<td>Scripted</td>
<td></td>
</tr>
<tr>
<td>Answer</td>
<td>Given</td>
<td>= text</td>
<td>NLP</td>
<td>Scripted</td>
<td></td>
</tr>
<tr>
<td>Distracter</td>
<td>Generic</td>
<td>From text</td>
<td>NLP</td>
<td>Scripted</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Modality</th>
<th>Menu</th>
<th>Text</th>
<th>TTS, visual</th>
<th>Prerecorded</th>
</tr>
</thead>
<tbody>
<tr>
<td>Output</td>
<td>None</td>
<td>Click</td>
<td>Keyboard</td>
<td>ASR, gesture</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Assessment</th>
<th>None</th>
<th>M/C</th>
<th>= answer</th>
<th>NLP</th>
<th>Manual</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scoring</td>
<td>None</td>
<td>Generic</td>
<td>answer</td>
<td>NLP</td>
<td>Scripted</td>
</tr>
<tr>
<td>Feedback</td>
<td>None</td>
<td>Generic</td>
<td>answer</td>
<td>NLP</td>
<td>Scripted</td>
</tr>
</tbody>
</table>
# QG costs (lowest → highest)

<table>
<thead>
<tr>
<th>Generation</th>
<th>None</th>
<th>Trivial</th>
<th>Simple</th>
<th>Complex</th>
<th>Manual</th>
</tr>
</thead>
<tbody>
<tr>
<td>Question</td>
<td>Generic</td>
<td>Cloze</td>
<td>NLP</td>
<td>Scripted</td>
<td></td>
</tr>
<tr>
<td>Answer</td>
<td>None</td>
<td>Given</td>
<td>= text</td>
<td>NLP</td>
<td>Scripted</td>
</tr>
<tr>
<td>Distracter</td>
<td>None</td>
<td>Generic</td>
<td>From text</td>
<td>NLP</td>
<td>Scripted</td>
</tr>
</tbody>
</table>

## Modality

<table>
<thead>
<tr>
<th>Modality</th>
<th>Output</th>
<th>Input</th>
<th>Assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Menu</td>
<td>Click</td>
<td>None</td>
</tr>
<tr>
<td></td>
<td>Text</td>
<td>Keyboard</td>
<td>= answer</td>
</tr>
<tr>
<td></td>
<td>TTS, visual</td>
<td>ASR, gesture</td>
<td>NLP</td>
</tr>
<tr>
<td></td>
<td>Prerecorded</td>
<td>Transcribed</td>
<td>Manual</td>
</tr>
</tbody>
</table>

## Assessment

<table>
<thead>
<tr>
<th>Assessment</th>
<th>Scoring</th>
<th>Feedback</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td></td>
<td>M/C</td>
<td>Generic</td>
</tr>
<tr>
<td></td>
<td>= answer</td>
<td>answer</td>
</tr>
<tr>
<td></td>
<td>NLP</td>
<td>NLP</td>
</tr>
<tr>
<td></td>
<td>Manual</td>
<td>Scripted</td>
</tr>
</tbody>
</table>