

Automated Comprehension Assessment in a Reading Tutor

Jack Mostow, Brian Tobin, Andrew Cuneo

Project LISTEN¹, RI-NSH 4213, 5000 Forbes Ave, Carnegie Mellon University, USA
mostow@cs.cmu.edu
<http://www.cs.cmu.edu/~listen>

Abstract

Can vocabulary and comprehension assessments be generated automatically for a given text? We describe the automated method used to generate, administer, and score multiple-choice vocabulary and comprehension questions in the 2001-2002 version of Project LISTEN's Reading Tutor. To validate the method against the Woodcock Reading Mastery Test, we analyzed 69,326 multiple-choice cloze items generated in the course of regular Reading Tutor use by 364 students in grades 1-9 at seven schools. Correlation between predicted and actual scores reached $R=.85$ for Word and Passage Comprehension.

Key words

Assessment of reading comprehension and vocabulary, multiple-choice cloze tests, curriculum-based assessment, validation, Woodcock Reading Mastery Test, Reading Tutor

1 Introduction

Project LISTEN's Reading Tutor listens to children read, and helps them learn to read (Mostow & Aist, 2001). To balance learner control with tutorial guidance, the Reading Tutor takes turns with the student to pick from its hundreds of stories, and uses students' assisted reading rate to adjust the story level it picks (Aist & Mostow, in press), ranging from kindergarten to grade 7 (disguised as K, A, B, ..., G to avoid embarrassing poor readers). Thus every student reads a different set of stories. Can we assess their comprehension automatically without writing (let alone validating!) comprehension questions for every story by hand? That is, how can we assess comprehension of given texts automatically to trace students' developing vocabulary and comprehension skills?

Existing assessments of children's vocabulary and comprehension such as the Woodcock Reading Mastery Test (Woodcock, 1998), Spache's *Diagnostic Reading Scales* (Spache, 1981), and Gray Oral Reading Tests (Wiederholt & Bryant, 1992) use comprehension questions developed by hand for specific text passages. In contrast, *curriculum-based measurement* (Deno, 1985) assesses students based on material they use in the course of normal instruction. One curriculum-based measurement approach to assessing comprehension is to prompt readers to retell what they read. Although such a prompt is easy to automate, scoring oral responses is not, because automated speech recognition is too inaccurate on such unpredictable speech.

Some researchers have generated *cloze* tests mechanically from a given text by replacing one or more words with blanks to fill in. For example, free software for one such "cloze test involves taking a document (or a document sample) of about 250 words and deleting every fifth word (these seem to be the canonical numbers,) leaving a blank in its place. The reader is then asked to fill in the missing words. In technical writing we use this as a test of readability. The idea is that there should be sufficient (local) redundancy in a document to allow a reader to score in the 50-60% range. Used in this way it measures the writer not the reader" (Drott). The cloze method was applied to the Pascal programming language to measure students' comprehension of computer programs

¹This work was supported by the National Science Foundation under Grant Number REC -9979894. 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 other members of Project LISTEN who contributed to this work, and the students and educators at the schools where Reading Tutors recorded data.

(Entin, 1984). However, most words in English text (apart from highly predictable function words) are too hard to guess from context (Beck, McKeown, & Kucan, 2002).

1.1 Approach

To overcome the unpredictability problem, we decided to make our cloze items be multiple-choice instead of fill-in-the-blank. This tactic replaced one problem with another. We no longer had to score arbitrary student responses, but in addition to choosing which words to turn into cloze items, we also had to generate appropriate distractors for each item.

The relationship of the distractors to the correct target word affects the difficulty of the question. For example, if they are the same part of speech, then the item requires semantic processing to ensure a correct answer. Matching by word class was one approach tried in (Coniam, 1997), using an automatic part of speech tagger. Though not perfect, automated tagging works well enough for this purpose if its occasional errors are tolerable.

Another basis for matching is word frequency, also tried by (Coniam, 1997). That is, choose distractors in the same frequency range as the target word. We adopted this approach, using a table of word frequencies derived by former Project LISTENER Greg Aist from a corpus of children's stories. We used four frequency ranges:

- “Sight words”: the most frequent 225 words in our table, approximately the same as the “Dolch list,” which cover over half the word tokens in English text, and are therefore emphasized in early reading.
- “Easy words”: the most frequent 3000 words (a heuristic cutoff) in our table, excluding the top 225.
- “Hard words”: all 25,000 words in our frequency table except for the top 3000.
- “Defined words”: story words explicitly annotated as warranting explanation to the student. This category is not defined in terms of frequency, so it overlaps with the other ranges. The Reading Tutor explained only some of the defined words, as part of a separate experiment we have not yet analyzed.

The Reading Tutor gives students reading assistance on difficult words and sentences. To avoid frustrating them by depriving them of such assistance, we decided to have the Reading Tutor read the cloze questions aloud by playing back the already-recorded human narrations of the sentences, minus the deleted words.

We chose distractors from other words in the story, rather than from a general lexicon, for several reasons:

- *Voice matching*: Unlike written cloze tests (Coniam, 1997), we needed to consider which voices spoke the redacted sentence, the target, and the distractors. If the sentence voice matched the target but not the distractors, children could answer based just on the voices.
- *Word recency*: Students might be biased toward picking words they have encountered recently, in particular earlier in the story – unlike most words from a general lexicon. Choosing distractors from the story makes them as likely to have appeared in the story as the test word.
- *Social acceptability*: Words chosen from an unrestricted lexicon may be offensive. Choosing words from the story means that they have already been judged acceptable by whichever adult added the story.

In using cloze items to assess vocabulary and comprehension, another decision was when to present them – before, during, and/or after the story. Presenting a cloze item prior to a story would take it out of context. Presenting a cloze item after a story might help test what students retained from the story, but would take extra time. We decided to insert occasional cloze questions in a story just before displaying a sentence to read.

2 Examples of cloze questions generated

Figure 1 illustrates a cloze question in the “hard word” category. The Reading Tutor says, “click on the missing word,” and reads aloud the cloze prompt at the top of the screen. Then it reads each choice aloud, highlighting its background in yellow as it does so. If the student has not clicked yet, the Reading Tutor reads the list again.

What does this item test? Three of the choices happen to be verbs, including the correct answer “recommend.” Information about part of speech can rule out “grasshopper,” improving the odds of guessing correctly to 1 in 3, but additional knowledge, such as semantics, is required to distinguish among the remaining choices.

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

bother

recommend

chat

grasshopper

Figure 1: Example of a "hard word" cloze question

Goodbye  Jane Student has read 8 minutes today. Jane (level E) has read "The Ant And The Grasshopper" (level C) 0 times, and 102 words and 1 stories as of April 4, 2002 at 04:58:35 PM. **Back**  **Go** 

Project LISTEN Reading Tutor Version: Jan 8 2002 16:24:56 F Instructions: The Reading Tutor expects the student to read the whole sentence
Copyright 1995-1999
Carnegie Mellon University
U.S. Patent No. 5,520,838

Help
Say: "I am helping to
Read together: "I ar
Play back last..."

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



Figure 2: Sentence for student to read

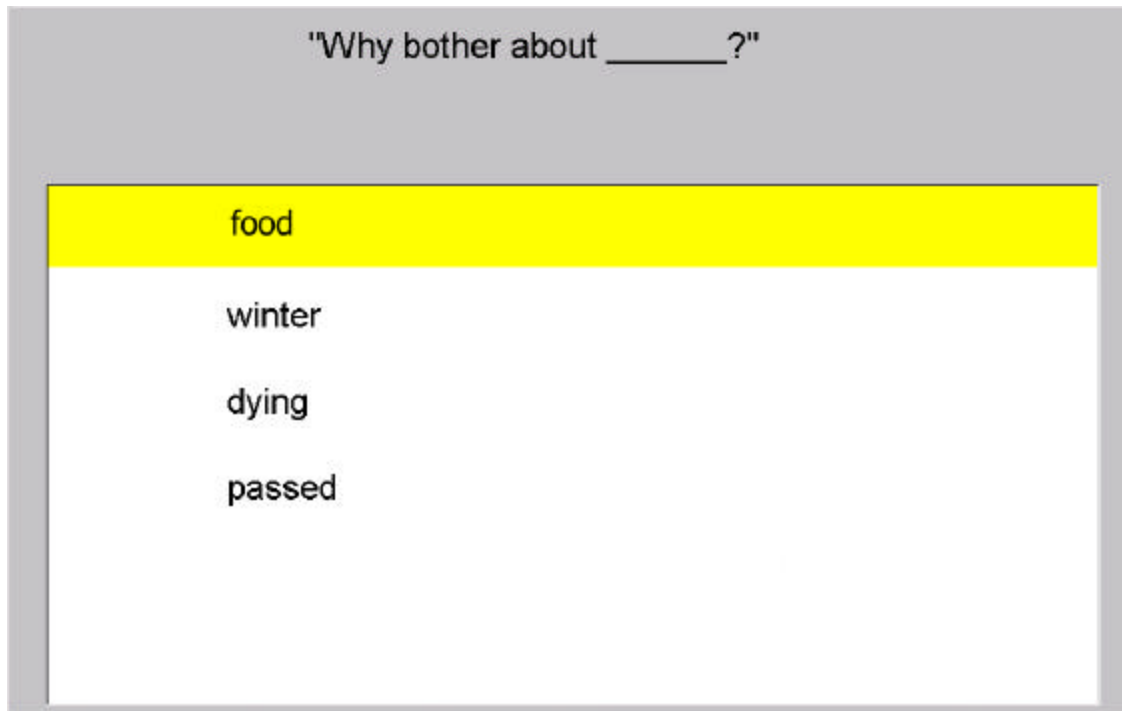


Figure 3: Example of an "easy word" cloze question that tests intersentential processing

Table 1: More examples of cloze items, by word type and story level (K=kindergarten, C=gr 3, G=gr 7)

Word Type	Story Level	Cloze Prompt	Choices	Correct Answer
Sight Words	K	Fruit is _____ to eat.	good, be, for, do	good
	C	_____ people kill them to get their skins to make coats and other things.	more, some, other, world	some
	G	By 1911, Carnegie had given away a huge amount of money, _____ 90 percent of his fortune.	who, than, about, think	about
Easy Words	K	Do you _____ to eat them?	nine, want, wish, big	want
	C	When cheetahs _____ they seem not to touch the ground.	close, word, run, ago	run
	G	It was _____ work, and they did not live in the same place for long.	united, large, hard, became	hard
Hard Words	K	Do _____ have a nose?	beak, cake, bake, hens	hens
	C	In _____ the cheetahs got a share of their master's food.	baby, reward, fur, tricks	reward
	G	Throughout his life, _____ Carnegie loved to read.	2,000, donation, international, Andrew	Andrew
Defined Words	C	And the very next day, the _____ had turned into a lovely flower.	grain, lily, walnut, prepare	grain
	G	Roadside diners and drive-ins _____ to auto tourists.	mobile, necessity, luxury, catered	catered

When the student clicks on a choice, the Reading Tutor does not give explicit feedback, but goes on to display the actual text sentence for the student to read, implicitly indicating the correct answer, as Figure 2 shows. Thanks to having just heard the Reading Tutor read the cloze item, the student presumably reads the sentence faster than she would have on her own. Consequently, the net time cost of inserting the cloze question should

actually be somewhat less than the time it takes to administer, and may even be negative for a student who would have read the sentence very slowly otherwise. The time cost of assessment is an issue to the extent that time spent on assessment detracts from time spent on educationally valuable practice and instruction. It is conceivable that the cloze activity itself builds comprehension skills, though (Johns, 1977) found no such effect.

Though widely used, cloze tests are sometimes criticized for not assessing the ability to integrate information across a text passage. However, comparison with a measure expressly designed to assess such across-sentence information integration suggested that “cloze scores may reflect intersentential comprehension sufficiently to warrant their continued use in assessment,” based on a study of 281 fifth graders (McKenna & Layton, 1990).

Figure 3 shows an example of a cloze question that exercises such integration. Although the words themselves are in the “easy” range, the question itself is challenging because only one choice (“passed”) can be ruled out based on its part of speech or other information local to the sentence. The desired choice (“winter”) depends on the preceding context, namely the sentence in Figure 2. “Food” is reasonable, but not what the author wrote.

Table 1 gives additional examples of cloze items for the four word types. Only levels C and higher (grade 3 and above) have “defined” words. We chose one random example of each type from levels K, C, and G to suggest their relative difficulty. Notice that question and word length both tend to increase with story level, and that distractors chosen from the same story are often semantically related to the target word.

3 Implementation: Automated Generation of Test Items

We now describe how the Reading Tutor generates and presents cloze questions, starting with how it decides which category of cloze question to insert when. The Reading Tutor has an event-driven control architecture. An `ev_before_new_sentence` event fires before each sentence of a story. Possible responses to that event are the four categories of cloze questions, plus “Do Nothing” (which just goes on to display the new sentence).

The Reading Tutor chooses probabilistically from these responses, according to the following weights:

```
defined words=10000
Easy Words=100
Hard Words=100
Sight Words=100
Do Nothing=400
```

The effect of the high weights for the “defined words” category is to prefer it at almost every opportunity – that is, when at least one word in the sentence and least three other words in the story are marked as defined.

If the chosen response fails – for example, if no word in the sentence is marked as defined – then the Reading Tutor picks another response probabilistically until it finds one it can fire, possibly Do Nothing.

Figure 4 shows the actual code to generate cloze items for “easy” words. We represent responses in a language we developed to express activities in a concise form that we can understand and the Reading Tutor can execute. We include the code for precision – it’s the representation we ourselves consult when in doubt – and to convey the generality of the specification, and the conciseness afforded by some key constructs of the activity language.

One such construct provides “smart” filtered random selection without replacement. Thus the statement

```
Set_variable test_word a_sentence_word _WHERE_ ...
```

randomly selects from the sentence a word whose frequency puts it in the top 3000 words but not the top 225. The next three such statements randomly select different story words in the same frequency range.

The generators for the other three categories are similar except that the “defined” category uses a different filter:

```
Set_variable word_definition a_definition_of test_word
```

This filter succeeds only if `test_word` is annotated in the story as having a definition. So must the distractors.

The meaning of a word may differ from one story to another, but seldom within the same story. Therefore, the Reading Tutor associates word definitions with stories, rather than with specific instances of a word in a story.

The definitions themselves are not used in the cloze questions, but are used in vocabulary preview activities inserted before the story to introduce new words using various methods (such as giving a definition or a synonym), and in post-test activities inserted after the story, which we plan to use to evaluate the effectiveness of those methods.

The Cloze function returns a copy of the sentence, substituting a blank for each instance of the test word. To avoid unacceptable test items, this function enforces some commonsense constraints by failing if any is violated:

- Sentences must be at least four words long.
- Sentences must be complete, that is, start with a capitalized word and end with a period, question mark, or exclamation point.
- To prevent truncation when the cloze item is displayed as a prompt, it must not exceed 100 characters.

The `If_new_story` test fails if the student has finished this story before, in which case the generator fails.

The next portion of Figure 4 specifies how the cloze item is presented. `USE_PICKER` says to use the Reading Tutor's generic "talking menu" mechanism for multiple-choice questions. This mechanism displays and speaks a prompt and a list of possible choices to click on. For cloze items, the Reading Tutor first says "Click on the missing word." Next it reads the sentence aloud, minus the test word, by playing the appropriate portions of the recorded sentence narration. Then it reads aloud the displayed menu of choices, presented in randomized order, consisting of the test word and the three distractors.

The remainder of Figure 4 logs the test word, the distractors, the cloze sentence, and the student's answer. The

```
As_comment Cloze-Pre-Sentence Easy Word Intervention
Set_variable test_word a_sentence_word _WHERE_ (WordRank(test_word) <
3001) _WHERE_ (WordRank(test_word) > 225)
Set_variable d_1 Distractor(test_word) _WHERE_ (WordRank(d_1) < 3001)
_WHERE_ (WordRank(d_1) > 225)
Set_variable d_2 Distractor(test_word) _WHERE_ (WordRank(d_2) < 3001)
_WHERE_ (WordRank(d_2) > 225)
Set_variable d_3 Distractor(test_word) _WHERE_ (WordRank(d_3) < 3001)
_WHERE_ (WordRank(d_3) > 225)
Set_variable test_cloze Cloze(test_word)
If_new_story
USE_PICKER
As_spoken_only Click on the missing word
As_spoken_prompt test_cloze
As_randomized_choices
As_correct_answer test_word
d_1
d_2
d_3
Store_result_as student_answer
ev_finish_step {
As_experiment_log_file Cloze-Pre-Sentence Intervention Easy Word
as_experiment_log_entry test word <test_word> distractors <d_1, d_2, d_3>
sentence <test_cloze> answer <student_answer>
```

generic logging utility automatically includes additional context information such as timestamp and student ID.

4 Evaluation

How accurately can we assess students' vocabulary and comprehension using our automatically generated items? To answer this question, we analyzed items from 99 Reading Tutors used daily in eight schools in 2001-2002.

4.1 Data set

A perl script on each Reading Tutor ftps logged data back to Carnegie Mellon every night, where it is parsed into a form we can import into the SPSS statistical analysis package. The following analysis is restricted to the 364 students we individually pretested on the Woodcock Reading Mastery Test (WRMT) before they used the Reading Tutor. These 364 students are from 65 different classes in grades 1-9 at seven schools in the Pittsburgh area. We excluded data from students who used four Reading Tutors at an eighth school in North Carolina because they did not take the WRMT.

Our data comes from 69,326 cloze items presented to the 364 students over the weeks of October 3, 2001, through March 12, 2002. The amount of data per student varies widely, depending on how much they used the Reading Tutor, how fast they read, and how much of their data was successfully sent back. The students escaped 729 (1.1%) of the items by clicking *Goodbye*, 361 (0.5%) by clicking *Back*, and 265 cases (0.4%) by waiting long enough for the Reading Tutor to time out, but they answered the remaining 67,971 (98.0%).

4.2 Item repetition

The relevance of item response theory to this data is limited because few students saw the same item, even if they read the same story. The 67,971 items answered include 16,942 distinct cloze prompts as defined just by sentence and test word. The number of distinct items is even larger if we distinguish different sets of distractors for the same prompt, which may make it much harder or easier. 41.1% of the prompts occurred only once, 80.2% occurred 4 or fewer times, and 94.4% occurred 10 or fewer times. The only prompts presented more than 41 times (up to 206 times) were for defined words, because each story had at most a few, and they were tested whenever possible.

To exclude stories the student has read before, the Reading Tutor inserts cloze items only if the student has not previously finished the story. However, students did not finish every story they started, and sometimes read a story they had started reading before, for example if they were in the middle of the story when their time was up. This situation was especially frequent at higher levels, where stories were longer. In fact some level G stories were too long to finish in one session, and students kept having to start over from the beginning of the story at the next session. This problem was sufficiently frustrating to outweigh the risk of introducing bugs by deploying new Reading Tutor functionality in mid-year. In December we modified the deployed Reading Tutor to let students resume where they left off the last time, if they so chose.

Due to rereading stories they had not finished, or to clicking *Back* to return to a previous sentence, students sometimes encountered the same prompt more than once. 2,313 prompts were seen twice, 406 were seen once, and three prompts were seen 9 times. However, these events were relatively rare, especially once we added the resume feature. In 61,475 (90.4%) of the cases, the student encountered the prompt only once.

4.3 Descriptive statistics on performance by word type, story level, and grade

Table 2: Perstudent number of cloze items and percent correct, by grade and overall

Grade:		1	2	3	4	5	6	7	9	All
# students		35	78	47	72	35	29	17	2	315
# items	Mean	201	143	121	219	344	471	104	194	214
	Median	200	85	80	134	364	511	78	194	136
	Range	24-446	21-525	20-671	23-758	28-736	54-733	22-317	192-195	20-758
% correct	Mean	49%	58%	57%	63%	61%	67%	73%	55%	60%
	Median	49%	59%	57%	63%	64%	69%	75%	55%	61%
	Range	25-71%	24-88%	30-80%	40-88%	29-84%	32-86%	45-88%	51-59%	24-88%

17,566 “sight” word items, 17,092 on “easy” words, 12,010 on “hard” words, and 21,303 on “defined” words comprised the 67,971 responses analyzed. Better readers read higher level stories. Only stories at levels C (grade 3) and above had “defined” words. Both these confounds made performance vary non-monotonically with story level, from 60% of the 3,163 level K items, up to 69% of the 6,031 level C items, then down to 55% of the 7,061 level E items, and finally back up to 59% of the 22,569 level G items. These percentages are per category and level, not per student. To avoid the resulting skew toward more prolific readers, Table 2 shows per-student averages for the 315 students with at least 20 responses. Performance rose with grade, ranging from 24% (chance) to 88%, with a floor for the lowest 1st and 2nd graders but no ceiling, and fell with word difficulty, averaging 68% on “sight” words, 67% on “easy” words, 61% on “hard” words, and 42% on “defined” words.

We had hoped that performance on the cloze questions would reflect student progress, and were therefore surprised to see that the percentage of correct items actually *declined* gradually over time. Why? Were the questions getting harder? Further analysis showed that story level and question length (in characters) rose over time, and were negatively correlated with performance when we controlled for student. But another, more disturbing possibility was that the students guessed more often as time went on and they tired of cloze items.

4.4 Guessing

How much guessing was there? We don’t know how to tell in general, but (as LISTENer Joe Valeri suggested) we can distinguish the important case of “too-fast responses,” which we define as responding too quickly to do better than chance. Plotting the percentage correct against response time (to the nearest second) showed that of the 67,971 responses, 3,078 (4.5%) were faster than 3 seconds, only 29% of which were correct – almost but not quite at chance. The percentage correct varied by type, from 34% for sight words down to 27% for defined words, suggesting that most but not all of them were guesses.

As one might expect, this analysis indicates that some students guessed more than others did. The per-student rate of “too-fast responses” averaged 3.9% overall, but 1% or less for over half of the students. We tried to characterize which students guessed more. This behavior did not correlate with any pretests we examined, but increased over time from an initial rate of 1% for the week of October 3 to a peak of 11% for the week of February 28, confirming our fears somewhat. Perhaps praise for correct responses would reduce guessing.

4.5 Reliability

The reliability of a measure characterizes the consistency of its results. How well does a student’s performance on one half of an N-item test match performance on the other half? To answer this question, we split the test items for each of the 364 students into two halves randomly, matching by word category (sight, easy, hard, defined) and story level (K, A, B, C, D, E, F, G) to the extent possible (mismatching leftover unpaired items).

We used SPSS to compute the Guttman Split-half test of reliability. The resulting coefficient depends on the minimum permissible value of N, ranging from .83 for N=10 (338 students) to .95 for N=80 (199 students). Thus student performance on a sufficient number of cloze items is indeed highly reliable.

4.6 External Validity

Does performance on these cloze items really measure comprehension and vocabulary? We correlated it with established instruments – the Woodcock Reading Mastery Test (WRMT) and the Gray Oral Reading Test (GORT). This analysis is restricted to the 315 students with 20 or more cloze responses, of whom 222 took the GORT. The raw proportion correct correlated significantly ($p=.000$) with WRMT Passage Comprehension ($R=.51$), WRMT Word Comprehension ($R=.53$), and GORT ($R=.40$), but much less than these tests do with each other: these two WRMT subtests each correlate at $R=.83$ with GORT, and at $R=.91$ with each other.

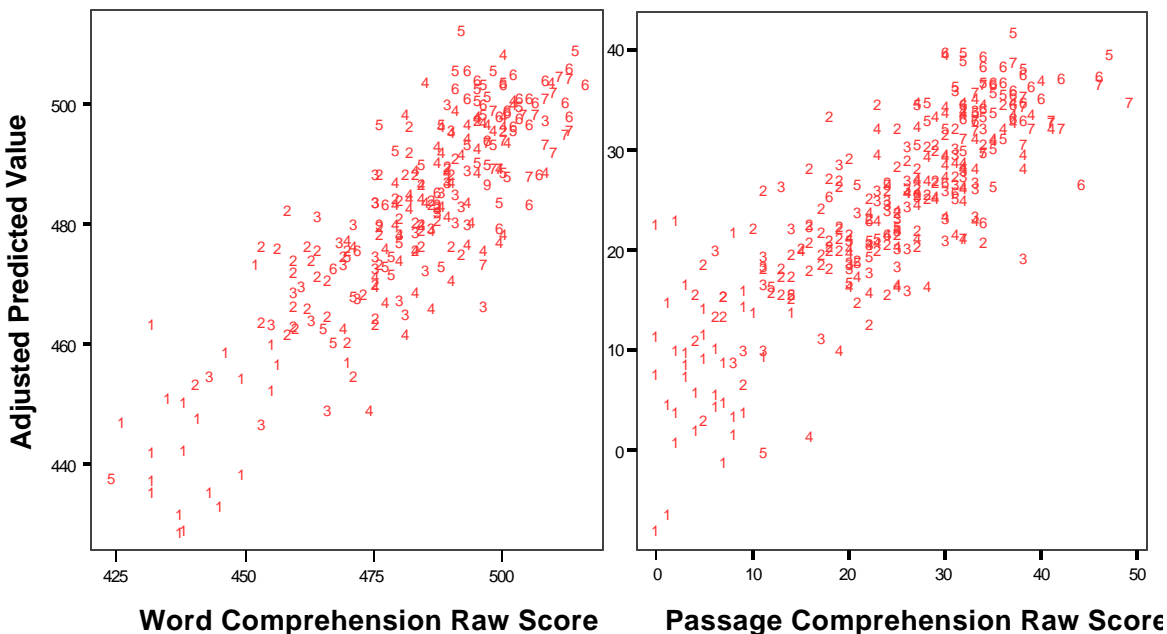
To improve on these results, we exploited additional information: question type, story level, and amount read. We encoded each student’s performance in terms of 54 predictor features: the proportion correct on each question type, and the number of correct and incorrect responses for each combination of question type and story level. The 4 question types and 8 story levels make 25 such combinations because defined words start at level C. We used backward regression in SPSS to build a separate model for each test by regressing test score against all 54 features, and then iteratively discarding insignificant predictors to optimize model fit. Applying the model to the feature values for a given student predicted that student’s test score. Scores correlate with grade. To avoid bias toward grade-level means, we did not use grade as a predictor, nor train a separate model for each grade, but we did use within-grade correlations to evaluate how much better the model predicted scores than grade alone.

We used two methods to estimate the performance of each model on unseen data from a similar distribution of students. The “leave-1-out” method, used in machine learning, adjusts the prediction for each student by training the same model without that student’s data, and evaluates these adjusted predictions against the actual scores. The adjusted R^2 measure, standard in statistics, penalizes model complexity based on the number of parameters.

Table 3: Predictive validity (Pearson Correlations) by grade and overall of models based on cloze test data

Grade:	1	2	3	4	5	6	7	9	All	Leave1	Adj. R^2
WRMT (total $N = 315$)	35	78	47	72	35	29	17	2			
Word Attack	0.25	0.67	0.55	0.59	0.70	0.54	0.51	1.00	0.72	0.66	0.50
Word Identification	0.21	0.65	0.64	0.62	0.85	0.57	0.40	1.00	0.84	0.82	0.69
Word Comprehension	0.44	0.73	0.65	0.71	0.85	0.71	0.55	1.00	0.86	0.84	0.73
Passage Comprehension	0.17	0.75	0.67	0.73	0.87	0.59	0.40	-1.00	0.85	0.83	0.71
GORT (total $N = 222$)	35	78	47	47	11	4	0	0			
Comprehension	0.49	0.55	0.53	0.45	0.62	0.74	.	.	0.72	0.66	0.53

Table 3 shows the predictive validity of the models, which did very well ($R=.84-.86$) on Word Identification,



Word Comprehension, and Passage Comprehension (all highly correlated), even using leave-1-out ($R=.82-.84$). They predicted scores better than grade alone, with $p<.01$ for all WRMT within-grade (2-6) correlations. Except for GORT, within-grade correlations were much lower in grade 1 (few 1st graders can read independently), and highest in grade 5. Figure 5 plots predicted scores (adjusted by the leave-1-out method) against actual scores.

Figure 5: Scatterplots of adjusted predicted scores by actual WRMT scores of 315 students in grades 1-9

4.7 Construct Validity

These predictions are based on the amount and level of material read, as well as the percentage of items correct. What skills do these items actually measure? The Reading Tutor reads the cloze questions aloud to the student, both the sentence prompt and the choices, unlike tests like the WRMT where the student must read the questions. So cloze items may measure listening comprehension, or at least comprehension of “assisted reading,” whereas WRMT measures independent reading skills. It might be interesting to see if presenting items silently improves their predictive validity, but we would want to restrict silent items to students who read well enough not to be frustrated by the lack of help – exactly the students we would expect to comprehend just about as well without it.

The multiple-choice format tests the ability to decide whether a given word fits in a particular context, whereas a fill-in format tests the ability to predict the word outright. The multiple-choice format is arguably more valid for

testing readers' metacognitive ability to judge whether they have identified a word correctly by seeing if it makes sense in context. Inserting questions throughout the reading of a story, rather than at the end, is arguably more valid for testing comprehension processes that occur while students are reading, but does not test retention. It might be interesting to see if the same items, administered at the end of the story, can help measure retention. Answering the items without guessing requires both knowing what the words mean, and judging which ones fit. Therefore questions about common words should tend to discriminate passage comprehension ability, while questions about rarer words should tend to discriminate students by vocabulary. However, these two skills are highly correlated, at least as measured by the WRMT, and are therefore hard to distinguish. Our test items vary randomly in the degree to which they require semantics or intersentential context to answer. Items with choices deliberately matched – or mismatched – by part of speech might help tease these skills apart.

5 Relation to Other Work

A literature search in the ERIC and INSPEC databases found numerous references to cloze questions. The most similar work investigated the use of word frequency and part of speech in automated generation of cloze tests (Coniam, 1997). Coniam compared three ways to pick which words to use as cloze items – every n^{th} word, a specified range of word frequency, or a specified word class such as nouns. He chose distractors from the 211 - million-word Bank of English tagged corpus, with the same word class and approximate frequency as the test word. He administered some of the resulting multiple-choice cloze tests (presumably on paper) to about 60 twelfth grade ESL students in Hong Kong. He evaluated the methods by percentage of acceptable test items generated, that is, with a facility index of 30%-80% and a discrimination index above 0.2.

The work presented here differs in several respects. Our data comes from over 300 students in 65 grade 1-9 classes at seven Pittsburgh-area schools, versus 60 students from two grade 12 classes in Hong Kong. Virtually all were native English speakers, not ESL students. Our cloze items were embedded in stories the students were reading on the Reading Tutor, not in a separate test checked by hand and administered to all students. We chose distractors from the same story, not from a large general corpus. We matched distractors by gross frequency range, but not by word class. The Reading Tutor presented cloze items aloud, not silently. Finally, we evaluated correlation to accepted measures of comprehension and vocabulary, not percentage of acceptable test items.

6 Conclusion

We have described and evaluated the automated generation of multiple-choice cloze questions to assess students' vocabulary and comprehension in Project LISTEN's Reading Tutor. We showed that performance on these items is highly reliable, and can be used to predict WRMT Word Identification, Word Comprehension, and Passage Comprehension scores with correlation better than .8, even for students the models were not tuned on.

What is this assessment useful for? We plan to report comprehension to teachers, who are eager to know it. We are especially interested in measuring student progress. Once we post-test students at the end of the 2002 school year, we will find out if their performance on cloze items over time predicted their gains from pre- to post-test.

References (see also www.cs.cmu.edu/~listen)

- Aist, G., & Mostow, J. (in press). Faster, better task choice in a reading tutor that listens. In P. DeCloque & M. Holland (Eds.), *Speech Technology for Language Learning*. The Netherlands: Swets & Zeitlinger Publishers.
- Beck, I. L., McKeown, M. G., & Kucan, L. (2002). *Bringing Words to Life: Robust Vocabulary Instruction*. NY: Guilford.
- Coniam, D. (1997). A preliminary inquiry into using corpus word frequency data in the automatic generation of English language cloze tests. *CALICO Journal*, 14(2-4), 15-33.
- Deno, S. L. (1985). Curriculum-Based Measurement: The emerging alternative. *Exceptional Children*, 52(3), 219-232.
- Drott, M. C. *The Cloze Test (free software)*. Retrieved April 3, 2002, from <http://drott.cis.drexel.edu/clozeproze.htm>
- Entin, E. B. (1984). Using the cloze procedure to assess program reading comprehension. *SIGCSE Bulletin*, 16(1), 448.
- Johns, J. L. (1977, December 1-3). *An Investigation Using the Cloze Procedure as a Teaching Technique*. Paper presented at the 27th Annual Meeting of the National Reading Conference, New Orleans, Louisiana.
- McKenna, M. C., & Layton, K. (1990). Concurrent Validity of Cloze as a Measure of Intersentential Comprehension. *Journal of Educational Psychology*, 82(2), 372-377.
- Mostow, J., & Aist, G. (2001). Evaluating tutors that listen: An overview of Project LISTEN. In K. Forbus & P. Feltovich (Eds.), *Smart Machines in Education*: MIT/AAAI Press.
- Spache, G. D. (1981). *Diagnostic Reading Scales*. Monterey, CA: McGraw-Hill.
- Wiederholt, J. L., & Bryant, B. R. (1992). *Gray Oral Reading Tests* (3rd ed.). Austin, TX: Pro-Ed.

Woodcock, R. W. (1998). *Woodcock Reading Mastery Tests - Revised (WRMT-R/NU)*. Circle Pines, Minnesota: American Guidance Service.